Third International Conference on Statistical Climatology

June 23-27, 1986,

Vienna, Austria

Third International Conference on Statistical Climatology

June 23-27, 1986,

Vienna, Austria

LIBRARY
ATMOSPHERIC ENVIRONMENT SERVICE
4905 DUFFERIN ST.
DOWNSVIEW, ONTARIO
CANADA M3H 5T4

OCT 22 1987

Medieninhaber: Österreichische Gesellschaft für Meteorologie
Hohe Warte 38, A-1190 Wien
Herausgeber: o. Univ.-Prof. Dr. Konrad Cehak
Redaktion: Univ.-Doz. Dr. Franz Nobilis
Herstellung: Verband der wissenschaftlichen Gesellschaften Österreichs, Lindengasse 37, A-1070 Wien

Editorial

When in 1980 a first attempt was made to gather statisticians and climatologists for a joint meeting, it was one of the first ventures of the newly introduced WCP. Only about 50 participants were present and about thirty papers were presented. Due to the great enthusiasm an agreement was made to continue with such meetings on a regularly basis. The success of the second meeting held 1983 in Sintra near Lisbon, Portugal, justified this opinion.

The editors are pleased to present this volume of the proceedings of the 3rd joint conference of climatologists and statisticians. This conference was conveaned by the Austrian and the American Society of Meteorology and by WMO. It was held on the premises of the International Atomic Energy Agency in Vienna. The Austrian Ministry of Science and Research has also donated funds for the conference. We thank all the sponsors for their help which enabled the conference to take place.

This third meeting was held in Vienna, Austria, from 23 to 27 June, 1986. There was a great attendence consisting of nearly a hundred participants from all over the world. They delivered 76 papers distributed over seven main topics and a number of minor ones.

Although we had hoped that all these topics would attract authors to approximately the same extent, some showed more and some less attractivity. The lack of interest in information theory and its application as well as the lack of interest in multi-dimensional statistics was disappointing for the editor, who was mainly responsible for the program of the conference. However, one can trace lines of collaboration between statisticians and climatologists which have been fruitful since the first conference in Tokyo and the present one and those which have been neglected. The main trend, at the moment, seems to be in the use of functional systems as for example the empirical orthogonal functions in representing climatological space distributions and also in looking for climatic change over large areas, e.g. a country.

The use of time dependent stochastic models was also much in the foreground, though in this area was very seldom any step beyond the use of ARMA processes, even an ARIMA is rarely to be found.

Taken as a whole, the conference was considered to have been a fruitful event by the participants and the demand was brought forward to have a further meeting in the Southern Hemisphere. The editor agrees with this wish.

The organizers of the conference would like to thank the IAEA and and the Lord Mayor of the City of Vienna for the help they extended to us in various ways. We especially thank Dr. F. Nobilis for his exceptionally hard work in the preparation of this report.

K. Cehak
Editor
and
Chairman of the Organization Committee
of the
3rd International Conference of Statistical Climatology

Shortly before the appearance of this publication, it is to our very deep regret that on the 18th of March 1987 Prof. Dr. Cehak died after a long severe illness, which he suffered in a remarkable way. Although strongly marked by this insidious disease he was trying to the very end to publish this report as quickly as possible, which now has become his last scientific work. The Austrian Society of Meteorology grieves over its president. His death is a great loss. In gratitude and in appreciation of his work in the field of meteorology in Austria the Austrian Society of Meteorology dedicates this publication to Prof. Dr. Cehak.

Innsbruck, 2nd April 1987

H. Pichler
President
of the
Austrian Society of Meteorology

T A B L E O F C O N T E N T S

I. Multivariate models of climatological data

PATTERNS OF SURFACE PRESSURE FIELD DURING JULY 1972-1983 IN SOUTHERN SOUTH AMERICA AND THE ANTARCTIC PENISULA
R. H. Compagnucci, W. M. Vargas 1

HOMOGENEOUS SUBREGIONS ACCORDING SURFACE PRESSURE VARIABILITY DURING JULY IN SOUTHERN SOUTH AMERICA AND THE ANTARCTIC PENISULA
W. M. Vargas, R. H. Compagnucci 15

COMPLETION OF AN INCOMPLETE DATA SET BY EOF-ANALYSIS: THE METHOD IS APPLICATED TO THE HISTORICAL SEA SURFACE TEMPERATURE DATA SET
R. Glowienka-Hense, A. Hense 22

AN ITERATIVE METHOD FOR THE PERIODICAL ANALYSIS OF LONG TIME SERIES
T. Szentimrey, O. Gulyás 28

THE USE OF CLUSTER ANALYSIS AND ROTATED EMPIRICAL ORTHOGONAL FUNCTION (EOF) ANALYSIS IN DESCRIBING THE MACROCIRCULATION PATTERN OF THE NORTHERN HEMISPHERE AND THE ATLANTIC-EUROPEAN REGION OF DIFFERENT HEIGHTS
J. Barthloy, A. G. Barnstone, R. E. Livezey 33

COMPARISON OF PRINCIPAL COMPONENTS AND FACTOR ANALYSIS METHODS FOR CLIMATOLOGICAL PARAMETERS
O. M. Essenwanger 40

THE USE OF THREE-MODE FACTOR ANALYSIS TO DETECT JOINT SPACE AND TIME VARIABILITY OF DROUGHT PATTERNS
P. A. Kay 46

ROTATION OF PRINCIPAL COMPONENTS (PARTS 1 and 2)
M. B. Richmann 52

II. Time-dependent models of climatological data

EXTREME DRY SPELLS - PERIODICITIES AND TRENDS IN ALPINE RIVER BASINS
F. Nobilis 66

INHOMOGENEOUS MARKOV MODELS FOR THE VARIATIONS IN CLOUD COVER
H. Madsen, P. Thyregod 71

TIME-DEPENDENT MODELS OF CLIMATOLOGICAL DATA
E. Suzuki, S. Hongo 77

PERSISTENCE OF THE WET SPELLS IN UGANDA
L. J. Ogalla, E. Bazira 85

MARKOV CHAIN MODEL FOR THE CLIMATOLOGICAL OBSERVATION SERIES ON WIND SPEED
N. Romanof, I. Elekes, R. Cotariu 99

THE BAYESIAN ASSESSMENT OF DIURNAL MEAN SURFACE TEMPERATURE DISTRIBUTIONS AT TWO FINNISH SITES WITH SHORT TEMPERATURE RECORDS
J. Helminen 107

MODELLING, ESTIMATION AND TESTING OF TIME-DEPENDENT CLIMATOLOGICAL PARAMETERS
G. Tunnicliffe Wilson, J. Belcher 112

III. Information theory in climatology, including design of networks

SOME PROBLEMS OF DESIGNING METEOROLOGICAL NETWORKS: APPLYING THE INFORMATION THEORY IN METEOROLOGY AND CLIMATOLOGY
F. Rákóczi 119

METEOROLOGICAL NETWORK DESIGN METHODS - A COMPARATIVE STUDY
T. Husain, M. A. Ukayli, H. U. Khan 129

OPTIMIZATION IN MESOSCALE METEOROLOGICAL NETWORK DESIGN
St. E. Jacobsen, I. Bruckner 139

IV. Extreme value statistics

EXTREME VALUE ANALYSIS OF CLIMATOLOGICAL DATA
T. A. Buishand 145

BI-VARIATE EXTREME VALUE DISTRIBUTIONS IN CLIMATOLOGY
K. Cehak 159

THE INTENSITY-DURATION-FREQUENCY CURVES OF THE ANNUAL MAXIMUM RAINFALL AT UCCLE (BRUSSELS)
G. Demarre, R. Sneyers 163

ESTIMATION, REGIONALIZATION, AND SPATIAL AVERAGING OF EXTREME RAINFALL AT SMALL TIME-STEPS
Ch. Obled, T. Lebel, M. Slimani 169

EXPONENTIALITY VERSUS GENERALIZED PARETO, QUICK TESTS
M. I. Gomes, M. A. J. van Montfort 185

NON-CLASSICAL EXTREME VALUE MODELS
M. I. Gomes, D. D. Pestana 196

THE STATISTICAL METHOD OF THE HEAVY RAINFALLS STUDY IN THE POLISH CARPATHIAN MTS
T. Niedźwiedź 201

ANALYSIS OF THE MAXIMUM PRECIPITATION DISTRIBUTION
V. Oancea, C. Boroneant 208

NEAREST NEIGHBOR SPACING DISTRIBUTIONS OF CUMULUS CLOUDS
R. F. Calahan 212

MISAPPLICATIONS OF EXTREME VALUE STATISTICS
A. Court 218

V. Statistical models for climatic variation research

STATISTICAL MODELS FOR CLIMATIC VARIATION RESEARCH
C.-D. Schönwiese 221

STATISTICAL STUDIES OF CENTRAL EUROPEAN PALAEOCLIMATE BASED ON BIOSPHERIC DATA OF THE RECENT MILLENIUM
W. Birrong 229

SOME STATISTICAL ASPECTS OF VOLCANISM-TEMPERATURE RELATIONSHIPS
C.-D. Schönwiese 235

A POTENTIAL PREDICTABILITY STUDY WITH A GENERAL CIRCULATION MODEL
F. W. Zwiers 241

CHARACTERISTICS OF SUMMER CLIMATE IN WESTERN EUROPE AND THE DETECTION OF THE CO_2-EFFECT
C. J. E. Schuurmans and A. J. Coops 251

USING NONPARAMETRIC RANDOMNESS TESTS FOR CLIMATE CHANGE ANALYSIS
V. T. Radiuhin, Ya. E. Mikhlin 256

ANOMALIES IN THE TEMPERATURE PATTERNS OF COLD PERIODS ON THE USSR AREA
A. M. Vysotsky, M. K. Zdonova, V. N. Razuzaev, N. N. Shevchenko 265

VI. Statistical methods for climate impact research

MONTE CARLO METHOD FOR RISK ANALYSES OF CLIMATIC DAMAGES AFFECTING THE YIELDS OF CROPS
Yu Xi-min, Pen Wei-bin 271

IMPROVING ESTIMATORS OF CROP YIELD DISTRIBUTIONS BY THE USE OF HISTORICAL RAINFALL DATA
T. A. B. Snijders 281

SAMPLING THEORY OPTIMALLY WEIGHTED DATA TO DETECT CLIMATIC SIGNALS
T. L. Bell, A. Abdullah 286

QUALITY/VALUE RELATIONSHIP FOR FORECASTS IN THE DYNAMIC COST-LOSS RATIO MODEL: INFINITE HORIZON, DISCOUNTED CASE
R. W. Katz, A. H. Murphy 289

THE IMPACT OF CLIMATE FLUCTUATIONS ON AIRLINE SCHEDULES
P. I. Robinson 295

VII. Statistical methods for the aims of applications of climatology

SOME STATISTICAL METHODS FOR THE APPLICATION OF CLIMATOLOGY
Jiuen Shi 301

OPTIMAL DESIGN OF THE REMOTE SOUNDING EXPERIMENTS WITH FINITE OBSERVATIONAL "WINDOWS"
V. Fedorov 311

THE USE OF CLIMATOLOGICAL STATISTICS IN THE PREDICTION OF EXTREME WIND SPEEDS IN TROPICAL CYCLONE REGIONS
P. N. Georgiou 322

AN ANALYSIS OF RAIN-PRODUCING SYSTEMS ON COASTAL GUYANA
A. B. Shaw 329

REGRESSION PREDICTION OF MONTHLY PRECIPITATION IN SUMMER
Hongxing Cao, Shanxian Yu 336

A NEW METHOD FOR INTERPOLATION OF CLIMATOLOGICAL DATA
Zhou Jiabin 341

AN INDIRECT MODEL FOR PROBABILITY DISTRIBUTION OF PRECIPITATION
Ding Juguo 346

VARIATION OF SPATIAL CORRELATION PATTERNS IN DENMARK
P. Allerup, H. Madsen 353

CARTOGRAPHY OF STATISTICAL PLUVIOMETRIC FIELDS WITH AN AUTOMATIC ALLOWANCE FOR REGIONAL TOPOGRAPHY
P. Benichou 360

EXTREME WIND SPEEDS DUE TO TROPICAL CYCLONES IN FRENCH DEPARTMENTS AND TERRITORIES OVERSEAS
D. Delaunay 367

COMPARISON OF METHODS TO ESTIMATE MISSING DATA AND TO BUILD LONG CLIMATOLOGICAL SERIES
J. Guiot 373

EXTRACTING INFORMATION FROM RAINFALL DATA
K. J. A. Revfeim 382

STATISTICS OF THE POLLUTANT CONCENTRATIONS DUE TO SINGLE SOURCES AND APPLICATIONS TO POLLUTANT DISPERSION NUMERICAL MODELLING
S. Tumanov 388

HYDROMETEOROLOGICAL DATA AND STATISTICAL ESTIMATION CHECKING UP BY MEANS OF STOCHASTIC MODELS
I. Zlate 394

THE EFFECT OF VIEWING ASPECT ON CLIMATOLOGICAL CLOUD DISTRIBUTION
D. D. Grantham, A. R. Boehm 400

PATTERN ANALYSIS OF GROWING SEASON PRECIPITATION IN SOUTH-CENTRAL CANADA
P. J. Lamb, M. B. Richman 408

SUBSTANTIATION OF THE CLIMATE MONITORING DATA SET COMPOSITION
G. V. Gruza 423

NON-PARAMETRIC METHOD FOR CONSTRUCTION OF TEST REFERENCE YEAR
L. Kajfež-Bogataj 430

OBJECTIVE PRECIPITATION FORECAST USING A SINGLE-STATION RADIOSONDE PROFILE
K. Pandzić 436

CLUSTER ANALYSIS OF WIND DIRECTION DISTRIBUTIONS
J. Urbančič 440

VIII. Other topics

SEASONAL VARIATION OF RAINFALL IN BELGIUM
R. Sneyers, Ch. Goossens 445

IMPLEMENTATION OF AN INFORMATION SYSTEM OF DATA SETS OF THE CLIMATOLOGICAL FIELD
B. Benesch 450

AN EVALUATION OF WIND DIRECTION PERSISTENCE AND PERIODICITY
St. E. Tuller 456

USING DIRECTIONAL DATA METHOD FOR ANALYSIS ON TYPHOON MOVEMENT AND ITS PROBABILITY DISTRIBUTION FEATURES
Jiuen Shi 461

A STATISTICAL MODEL FOR PREDICTING THE DURATION OF MEIYU RAIN BASED ON MINIMIZING MEAN SQUARE ERROR OF

PREDICTION
Shanxian Yu, Xiaoyun Chen 469

A SEQUENTIAL TEST METHOD FOR CORRELATION COEFFICIENTS AND ITS APPLICATION
Ding Yuguo 476

A PROBABLE AVENUE OF THE EFFEXT OF THE POLAR ICE ON THE NORTH PACIFIC SUBTROPICAL HIGH
Fang Zhifang 481

THE KUTZBACH EOF'S OF SUMMER MONSOON CIRCULATION AND RAINFALL FIELDS OVER INDIA
P. P. Kamte, S. V. Singh, R. H. Kripalani 489

RAINFALL AND WEATHER TYPES IN SOUTH WESTERN FRANCE
J. L. Champeaux, P. Peris 493

THE POTENTIAL PREDICTABILITY OF 70hPa HEIGHT AND RAINFALL FIELDS OVER THE INDIAN AND ADJOINING REGION DURING THE SUMMER MONSOON SEASON
R. H. Kripalani, P. P. Kamte, S. V. Singh 499

POLAR LOWS IN THE NORWEGIAN AND THE BARENTS SEA. A CLIMATOLOGICAL STUDY
I. Kanestrøm, T. Ese, K. Pedersen 502

MAXIMUM WINDS AND RAINFALL IN TAIWAN DURING A TYPHOON
H. P. Chu, H. Y. Tseng, Chin-Piao Pu 508

ERROR MEASURES FOR EVALUATING THE REPRESENTATIVENESS OF DIVISIONAL CLIMATIC DATA
A. J. Brazel, R. C. Balling Jr. 515

MESOSCALE CLIMATOLOGICAL STUDIES OF THE BLACK HILLS - AN ANALYSIS OF THE SPATIAL REPRESENTATION OF TEMPERATURES
A. A. Doneaud, J. R. Miller Jr. 521

THE SYNOPTIC CLIMATOLOGY OF MONTHLY PRECIPITATION AMOUNTS OVER THE UNITED STATES DURING WINTER IN RELATION TO THE SURROUNDING 700 MB HEIGHT FIELD
W. H. Klein, H. J. Bloom 527

THE SYNOPTIC CLIMATOLOGY OF MONTHLY MEAN SURFACE TEMPERATURE IN EUROPE IN RELATION TO THE 700 MB HEIGHT FIELD
W. H. Klein, Runhua Yang 535

THE FREQUENCY OF HAILDAYS IN CENTRAL SWITZERLAND. A COMPARISON HAILPAD-INSURANCE DATA
H. H. Schiesser 542

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

PATTERNS OF SURFACE PRESSURE FIELD DURING JULY 1972-1983 IN SOUTHERN SOUTH AMERICA AND THE ANTARCTIC PENINSULA

Rosa H. Compagnucci
Instituto Argentino de Nivología y Glaciología(IANIGLA)
C.C.330, 5500 Mendoza, Argentina

Walter Mario Vargas
Facultad de Ciencias Exactas y Naturales, Dpto. de Meteorología
Universidad de Buenos Aires
1428 Nuñez, Buenos Aires, Argentina

INTRODUCTION

The treatment of general circulation is usually required to establish climatic-synoptic description and forecasts of the variables and the meteorological phenomena. They allow a better understanding of the distribution of the variables and the reasons for their fluctuations.

One way to represent, although partialy, the general circulation is by means of surface pressure fields, which contain information about the weather behavior. Therefore, their study is a neccesary condition, but not always sufficient, for the analysis of the significant phenomena behavior, which help to define climatic anomalies.

The principal aim of this work is to study the surface pressure fields series and to render possible the analysis of the precipitation fluctuations of the Central Andes (Cordillera de Los Andes, South America between the parallels 30°s and 40°s).

Several authors have tried to connect the occurrence of precipitations in definite areas with types of surface pressure maps or variations of circulation indexes. Sawyer (1956) found a broad relationship between the distribution of rainfall and four depressions over the British Isles. Pittock (1975) studied spatial patterns of year-to-year variations of rainfall and their relation to the general circulation in Australia. He obtained significant correlation coefficients between the L-index of surface pressure of the subtropical Pacific anticyclone over the east coast of Australia and the precipitation patterns. Richman (1980) obtained sea level pressure patterns by principal components analysis, which were used to examine potential urban effect on storms (city induced rain).

Research applied to Southern South America was pionered by De Fina and Sabella (1958). They obtained a typification of the weather maps from an index built up with pressure surface values of Argentinian stations, suggesting the use of this classification for a semiobjective forecasts of precipitation areas.

There are also other researchers,who describe particular synoptic situations related with extreme phenomena. Nicolini and Norte (1980) used the type of synoptic surface maps as one indicator to classify the hail storms of northern Mendoza Province (Argentina). Schwarzkopf and Migliardo (1973) analized the synoptic situation that produced a strong tornado.

The rainfall regime of the Central Andes presents two principal characteristics: precipitations mainly occur in the winter months, being negligible the rest of the year, and, on the other hand, there is a great interanual variability with great snow accumulations in some years and null or scarce in others. Since the rivers of Cuyo have run-off related to snowfall, they pre-

sent great variability between years of great summer stream flow and other periods, some time long ones, with little flow. This affects the economy and development of the Argentine middle west area (Jornadas de Nivoglaciología, Mendoza, 1969). Considering this, the interest of this study is centered in the winter months and takes July's data as a first approach, which approximately represent the winter months, although more definite results will require taking account other months with a high likelihood of precipitation in the Central Andes. Nevertheless, July contains the majority of the possible winter surface pressure fields, because the sample (July 1972-1983) is not biased with respect to precipitation. Years with maximun and minimun rainfall were taken into account.

As Barry and Perry (1975) show, several more or less objective techniques to obtain types or patterns of pressure fields may be used. The sets of July's surface baric fields for the years 1972 to 1977 were employed by Compagnucci, Fornero and Vargas (1984) to obtain patterns of surface pressure. They used the multiple correlation methods of Lund (1963), linkage analysis of Mc Quitty (1957), and un-rotated principal components analysis with a T-mode input matrix. The results showed that the un-rotated principal components (PCs) was the technique which gave more appropiate results for the analysis of significant phenomena variability. Considering this, the patterns of the surface pressure fields are obtained, here, by the un-rotated (PCs).

DATA AND METHODOLOGY

Data from the 12z hour surface synoptic maps of the National Meteorological Service (Argentina) were used for this work.

From these data a sample containing the sea level pressure, reported by 81 stations (see fig.1) during 372 days of July from 1972 to 1983 was extracted. The main reason for this selection was that there exist complementary information (radiosonde observations, precipitation registers in the Central Andes, satellite imagery, etc.) and the large variability in the precipitations for the considered years.

When the pressure value in a station was non-existent, it was interpolated from the drawn basic field. Vargas and Compagnucci (1983) showed that this network gives redundant information and, therefore, substitution of interpolated data for those missing does not essentially modify the results. On the other hand, in the PCs analysis the small scale disturbance introduced in the fields, by errors in the interpolation, appears only in the higher order components. As the latter component account for the minimun percentage of the variance, they are not interesting for this study.

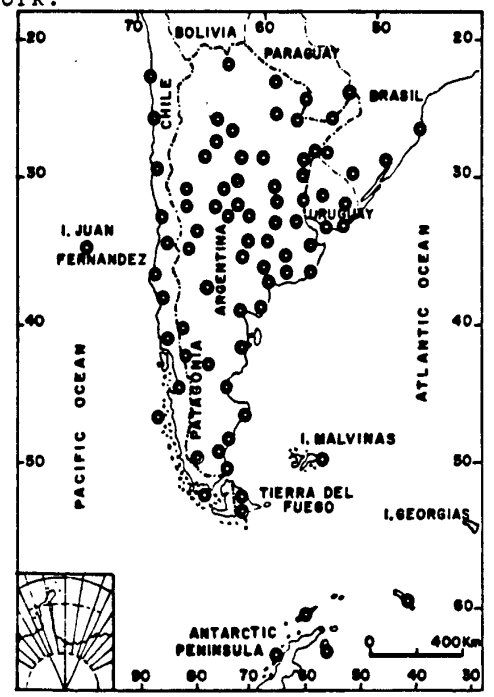


Fig. 1: GRID-POINT SELECTED FROM STATIONS REPORTING SEA LEVEL PRESSURE

A matrix $\mathbb{P}$ (81 rows by 372 columns) was column-standardized generating an input matrix of the T-mode as follows:

$$\mathbb{P} = \left\{ p_{ij} \right\}$$

p_{ij}: surface pressure for day-j and station-i

M: numbers of days

N: numbers of stations

j= 1, 2, ..., M

i= 1, 2, ..., N

$$\bar{p}_j = \frac{\sum_{i=1}^{N} p_{ij}}{N}$$

$\bar{p}_j$: spatial mean pressure for day-j

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{M} (p_{ij} - \bar{p}_j)^2}{N}}$$

σ_j : standard deviation relative to spatial mean pressure for day-j

$$\mathbb{T} = \left\{ t_{ij} \right\}$$

is a matrix of T-Mode with

$$t_{ij} = \frac{p_{ij} - \bar{p}_j}{\sigma_j}$$

t_{ij}: standardized surface pressure for station-i and day-j

The un-rotated PCs were calculated by means of the OFPRI and OFCOEF subroutines of the IMSL (1980) package. The factor-scores yield the spatial patterns. The factor-loadings time series allows to acertain which patterns constitute and orthonormal basis and it is possible to express each day as a lineal combination of them.

The period 1972 to 1983 was analyzed to obtain the basic theorical models. Climatic variability was studied dividing two sub-periods, 1972-1977 and 1978-1983. The relationship between the explained variance for each pattern and the monthly precipitation was analyzed calculating the PCs for each month.

RESULTS AND DISCUSSION

The variance extracted by each PC, given by the eigenvalue, is shown in table 1. For the whole period 1972 to 1983 the first ten PCs account for 95,4% of the variance, while in the sub-periods 1972 to 1977 and 1978 to 1983 only the first nine are needed in the first case, and the first ten must be included in the second to explain the same percentage.

PC_s	1972-1983			1972-1977			1978-1983		
	Eigenvalue	Per. Val.	Per.Cum.Val.	Eigenvalue	Per. Val.	Per.Cum.Val.	Eigenvalue	Per. Val.	Per.Cum.Val.
1	211.65	56.88	56.89	110.50	59.40	59.40	102.02	54.35	54.85
2	58.57	15.77	72.66	27.86	14.98	74.39	30.95	16.64	71.49
3	27.51	7.39	80.06	12.64	6.79	81.18	14.93	8.03	79.52
4	20.54	5.52	85.58	8.08	4.34	85.53	12.55	6.75	86.27
5	12.04	3.23	88.82	6.63	3.56	89.10	5.30	2.85	89.12
6	8.41	2.26	91.09	4.30	2.31	91.41	4.32	2.32	91.45
7	7.46	2.00	93.09	3.86	2.07	93.49	3.61	1.94	93.39
8	3.42	0.91	94.01	1.87	1.00	94.49	1.74	0.93	94.33
9	2.85	0.76	94.78	1.68	0.96	95.40	1.43	0.77	95.10
10	2.29	0.61	95.40	1.18	0.63	96.04	1.18	0.63	95.74
11	2.21	0.59	95.99	1.08	0.58	96.62	1.05	0.56	96.31
12	1.91	0.51	96.50	0.87	0.47	97.10	0.73	0.39	96.70
13	1.21	0.32	96.83						
14	1.18	0.31	97.15						
15	1.03	0.27	97.43						
16	0.89	0.24	97.67						

Table 1: EIGENVALUES, PERCENTAGE VARIANCE EXPLAINED AND CUMULATIVE PERCENTAGE EXPLAINED BY THE FIRST COMPONENTS FOR THE ANALYSIS OF DIFFERENT PERIODS.

A difference can also be observed in the percentages explained by each one of the first components between the two sub-periods. This indicates a difference in the distribution of the pressure field patterns for the 1972 to 1977 and 1978 to 1983. Therefore, ten or more years should be taken into account in the sample to obtain more stable climatic results and to assure a greater statistical significance. Consequently the results for the 1972-1983 period will be taken as reference in this report.

PC	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
1	67.15	49.49	72.21	63.55	75.13	39.82	44.64	61.71	48.23	59.98	58.73	68.40
2	12.49	21.92	11.66	16.44	12.41	20.09	26.38	17.03	14.44	14.37	16.29	15.61
3	5.69	7.75	4.86	8.70	4.74	13.31	9.96	6.23	11.11	7.02	7.07	6.92
4	3.64	5.83	4.07	3.08	2.62	8.06	6.22	5.01	9.51	5.34	5.61	2.44
5	3.32	4.13	2.09	2.50		5.45	3.30	2.83	4.96	3.32	3.12	1.63
6	2.05	3.01				3.88	2.30	2.03	4.08	2.78	2.07	
7		1.77				3.18	1.59		1.87	1.84	2.00	
8		1.19				1.34	1.17		1.14			
%T	94.83	95.12	94.90	94.30	94.93	95.17	95.59	94.87	95.38	94.69	94.92	95.01

Table 2: PERCENTAGE VARIANCE EXPLAINED BY THE FIRST COMPONENTS FOR MONTHLY ANALYSIS AND CUMULATIVE EXPLAINED PERCENTAGE.

There is a great interanual variability of the variance distribution explained for the first components as shown in table II.

In years such as 1976 nearly 95% of the variance is explained by the first four components, while in 1973 or 1977 it is neccesary to retain eight components to account for the same variance. This means a greater variability of the synoptic situation in July of 1973 and 1977.

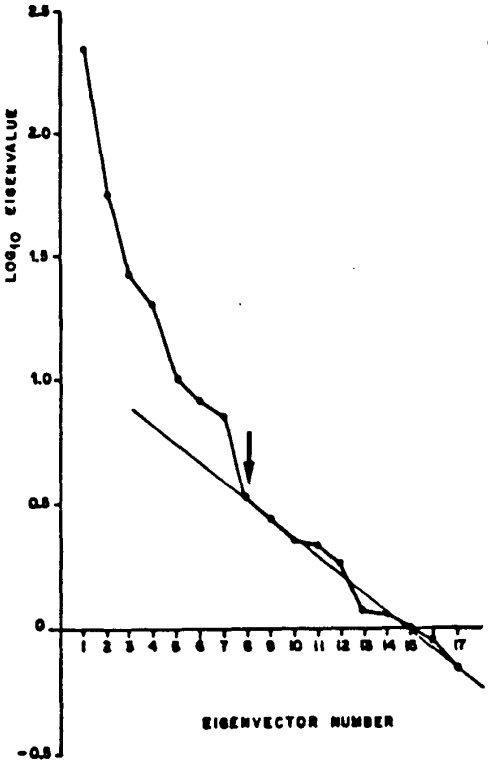


FIG. 2: THE LOG_{10} EIGENVALUE vs. EIGENVECTOR NUMBER (LEV) DIAGRAM FOR JULY 1972-1983 SAMPLE

Regarding the statistical significance of patterns, the Levś diagram (fig.2), according to Farmer (1971), suggest that only the seven first factors explaine significant variance (non-random data patterns).

Other criteria could be to retain the factors which account for a determinate percentage of the variance. This criterion is useful when the factor score orthonormal basis is used later on reconstructing fields. In the results showed in table I we can observe that the first seven components account for 93% of the variance and the eight first 94%, for the whole period 1972 to 1983.

Another criterion is to take into account the factor losdings time series amplitudes which reflect the weight of the corresponding pattern for each day. In figure 3, we see that the amplitudes for the ninth factor are small for all days, this means that this pattern does not include significant information for any of the days.

As the aim of this work is not to determine how many factors are to be retained or reflected, in order to avoid "noise" components, only the results for the first eight patterns are shown.

PATTERNS OF PRESSURE SURFACE FIELDS

Theorical patterns given by the PCs represent two possible fields: the flow and its inverse. This is due to the flip-flop property of the factors; however, the two possibilities are not always real.

The same structure of the patterns appeared in the whole period and in the sub-periods for low order components in the same order. The patterns of the 1972 to 1983 period are present here and included in the first pattern for the two sub-periods because of some small differences.

Pattern 1

The first pattern (see fig.4b) shows a zoned flow south to the 40°s parallel and to a relative flat field north of it.

One field corresponds to the western winds in the zone of meridional flow and pressures north to 40°s higher than the pressure south of 50°s parallel. There is a marked resemblance between this type and the mean sea level pressure for July 1930-1960 (see fig. 4a).

A small difference can be observed among the patterns for the period 1972-1983 (fig. 4b) and the sub-periods (fig. 4c and 4d). The sub-period 1978-1983 being more similar to the whole period pattern. This result shows that during 1972 to 1977 the influence of the zonal westerlies extends to lower latitudes and the subtropical anticyclones of the South Pacific and South Atlantic Oceans are not over the continent. On the other hand, in the first pattern for the 1978-1983 sub-period the subtropical anticyclones have influence over the continent between 20°s to 40°s.

The time series of the first factor loadings (fig. 3a) show that the majority of the days has negative loadings with absolute values larger than 0.8 .

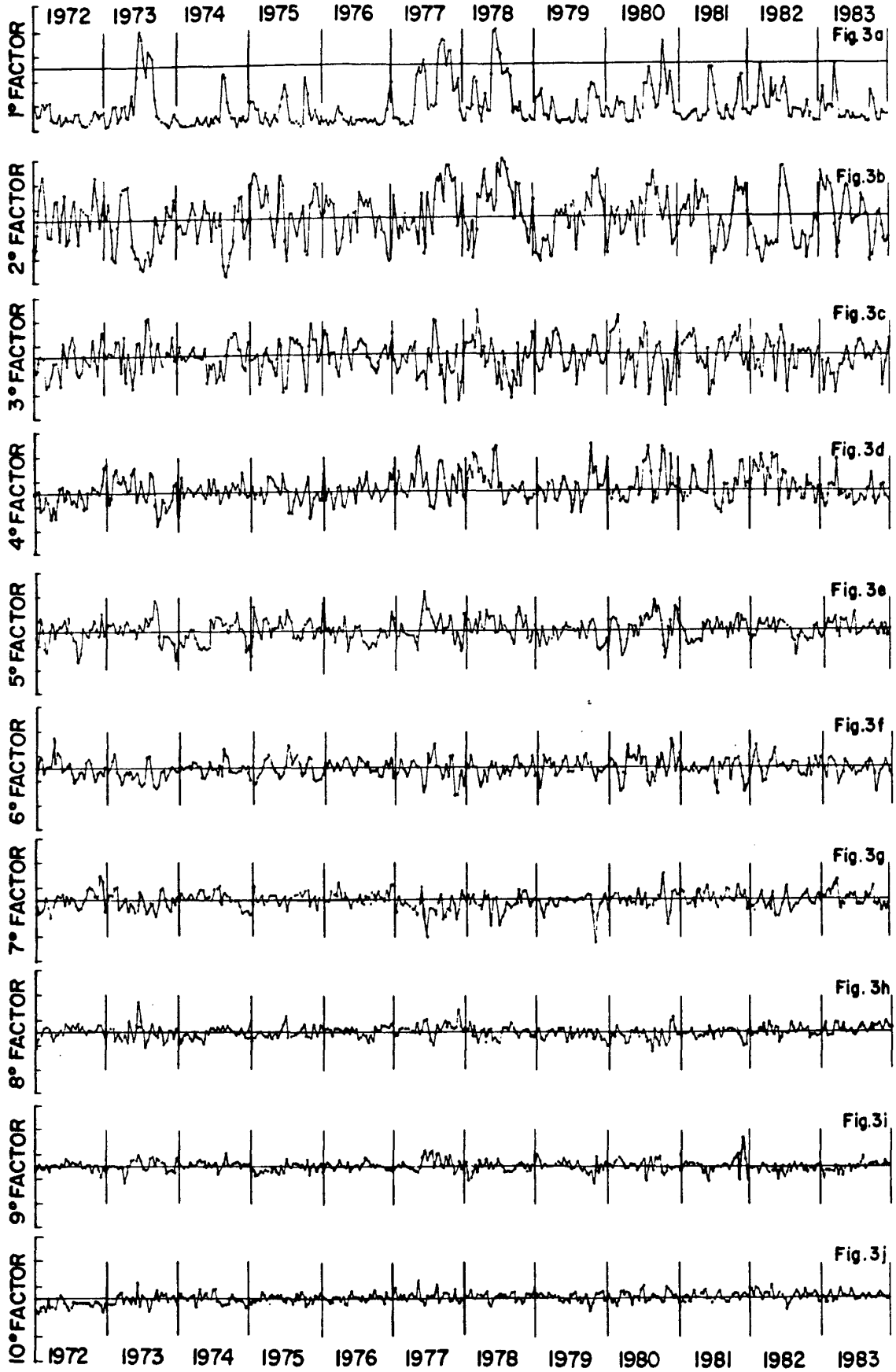


Fig. 3: FACTOR LOADINGS TIME SERIES FOR THE WHOLE PERIOD (JULY 1972-1983)

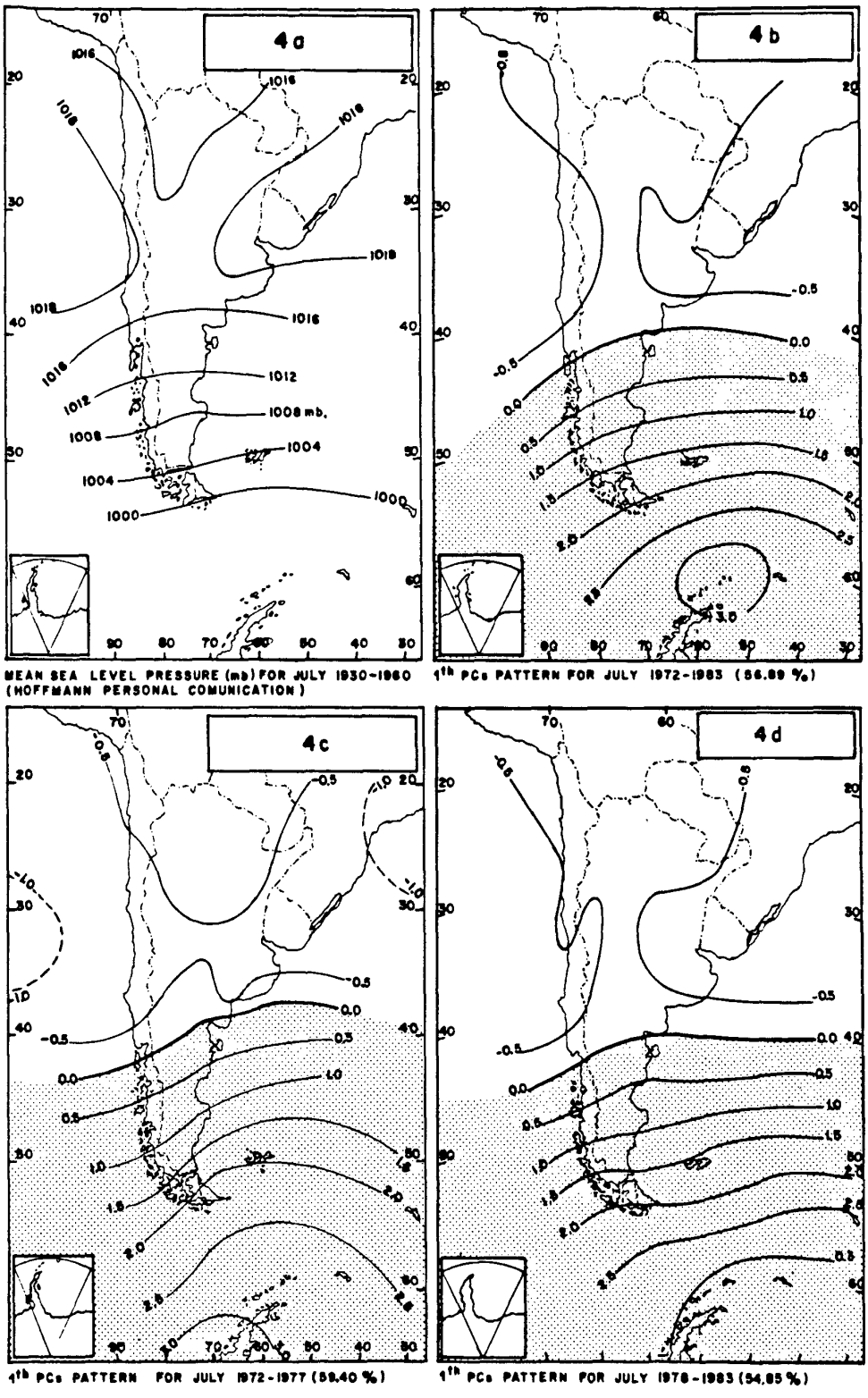
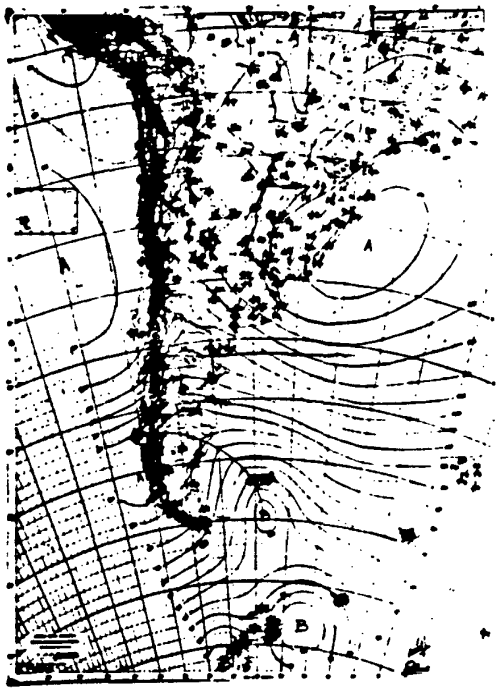


Fig. 4: MEAN SEA-LEVEL PRESSURE FIELD AND FIRST PCs PATTERNS OF JULY

This corresponds to the days with similar structure to the type described above, hence a great part of the daily real fields can be explained by this type. An example of a real case similar to this pattern is July 11th of 1979 (map 1 is a photo of the surface synoptic map obtained from the National Meteorological Service of Argentina).

The theorical inverse type presents low pressure north of 40°s with respect to the pressure south of 50°s, and zonal easterlies between 40°s and 60°s. This case is the inverse of the mean sea level pressure field. The days with positive factor loading have components of this type in their structure, but the time series of the factor loadings (fig.3a) show that only 17 of the 372 days have loadings between 0.6 and 0.2, and none over 0.6. We inferred that it is not likely to find, in July, an inverse flow to the mean surface level pressure and only few days have the same characteristics.



Map.1: 11th JULY OF 1979 (S.M.N. of Arg.)

Pattern 2

This second component pattern (fig.5) represents a closed center in the south of Patagonia and Chile, spreading its influence to the north of Argentina. It has a low index of circulation.

One of the types presents high pressure over the southern part of the continent.

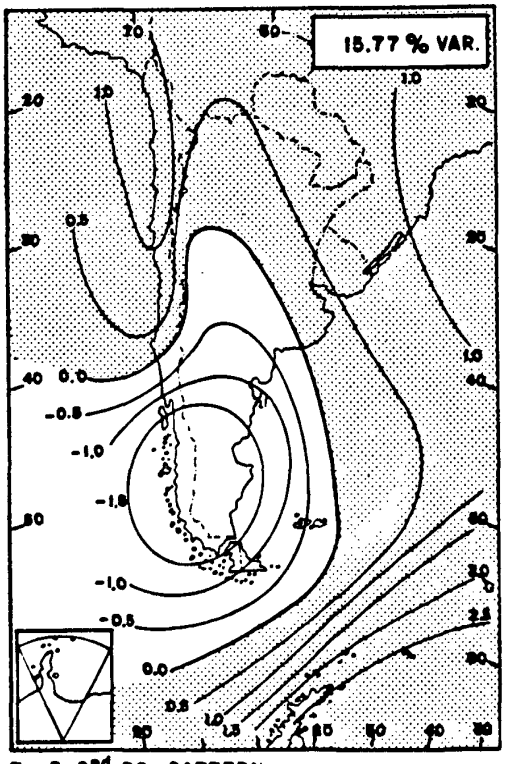


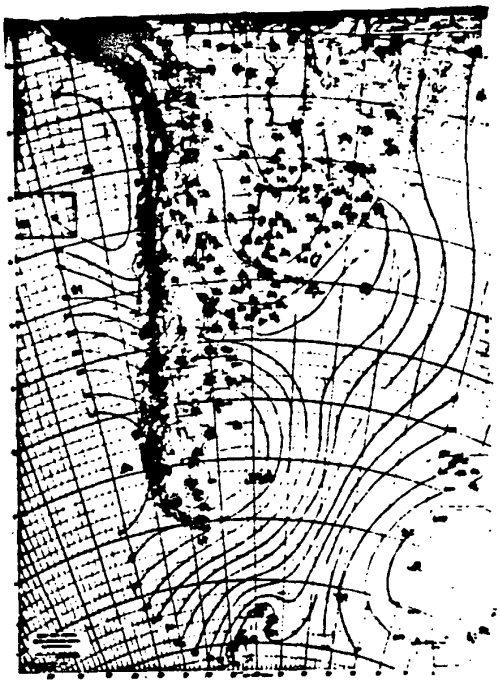
Fig. 5: 2nd PCs PATTERN



Map. 2 : 21st JULY OF 1974 (S.M.N. of Arg.)

The real days of similar structure have negative values of factor loadings (see fig.3b). Several daily pressure fields are explained almost totally by this model; an example is July 21st, 1974, shown in map 2.

The inverse model has a low pressure center in the southern continental part. Here, the days with similar structure have positive factor loadings (see fig. 3b) and also in this case various daily fields show similarity with the theorical model to a high degree; an example is 16th July of 1978, shown in map 3.



Map. 3 : 16th JULY OF 1978 (S.M.N. of Arg.)

Pattern 3

The third PCs pattern contains two centers between northeast Argentina and the Malvinas Isles (see in fig.6 the shadowy area) and a strong meridional gradient on the western part of the continent. This structure has a small circulation index.

One of the possible types is that with low pressure in northern Argentina and the Malvinas Isles, and a strong southern meridional flow over Chile and the western part of Argentina (Andean Cordillera between Argentina and Chile). The days that include this type in their structure flow have positive factor loadings (fig.3c). Some of these real fields present circulation nearly equal to this model, as July 7th, 1980, shown in map 4.

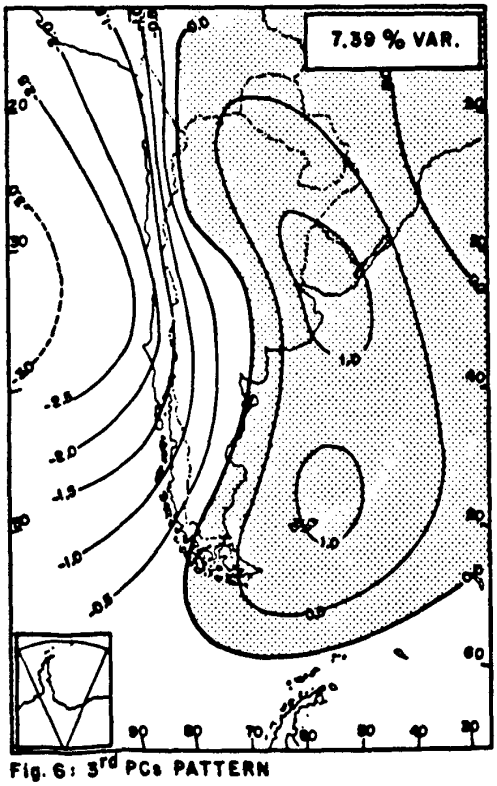
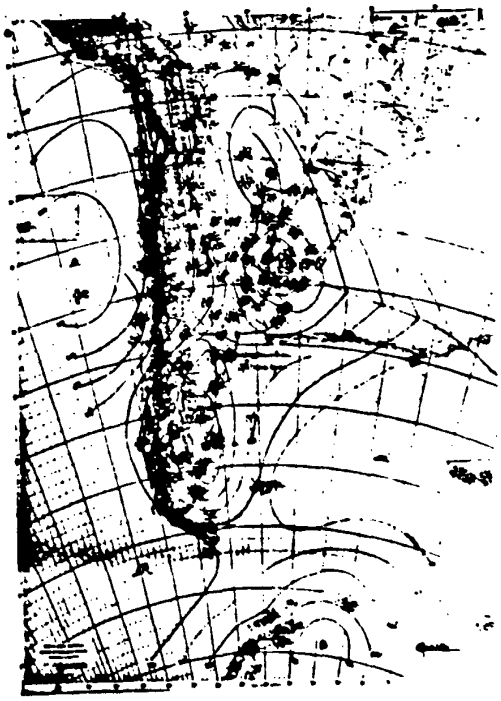
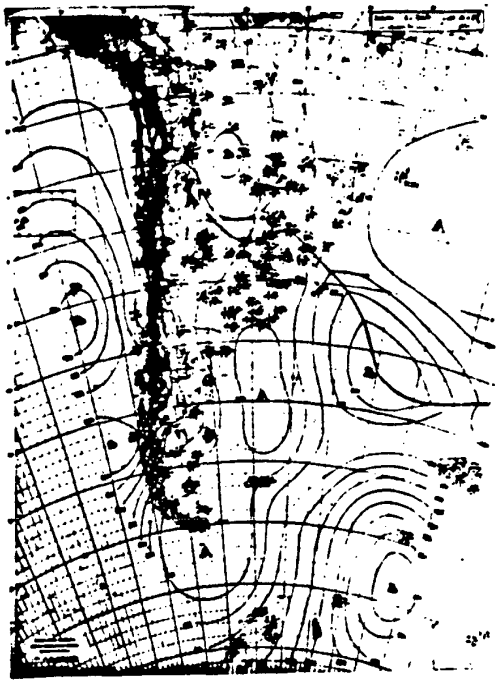


Fig. 6: 3rd PCs PATTERN



Map.4: 7th JULY OF 1980 (S.M.N. of Arg.)

The inverse type corresponds to a field with high pressure in northern Argentina and the Malvinas Isles, and a northern meridional flow over the western part of the area. The days for which the third factor loadings are negative (see fig.3c) have components of this model in their circulation fields; an example of a real flow that is similar to it is July 6th, 1978 (see map 5).



Map.5: 6th JULY OF 1978 (S.M.N. of Arg)

Pattern 4

The circulation of the fourth component shows: a center over Argentina that also extends to the central part of Chile (between 35°s and 50°s) and another center with same sign over the south-eastern part of the research area and with opposite sign, the north of Chile and an axis from north-east to south-west (shadowy area in fig.7).

One of the models shows high pressure in the shadowy area and low pressure over Argentina; these cases correspond to the fourth factor loadings positive (fig.3). Some days have similar fields as this model; as an example of a real case, July 25th of 1980, shown in map 6.

There are low pressures in the shadowy area and high pressures over Argentina in the inverse model. The days whose pressure fields have features of this type present a negative value for the fourth factor loadings (fig.3d). There are fewer cases here than when the shadowy areas have high pressure.

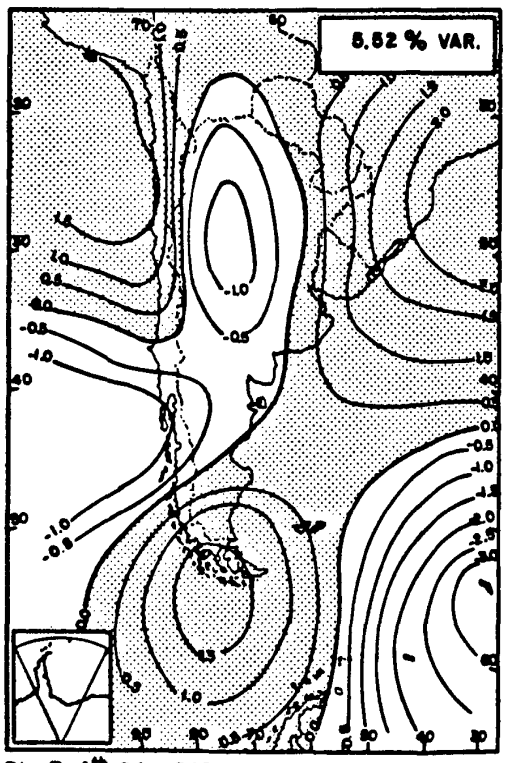
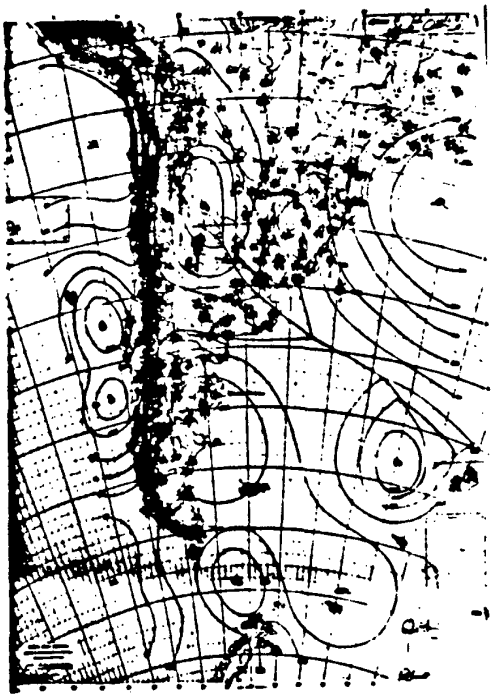


Fig. 7: 4th PCs PATTERN



Map.6: 25th JULY OF 1980 (S.M.N. of Arg.)

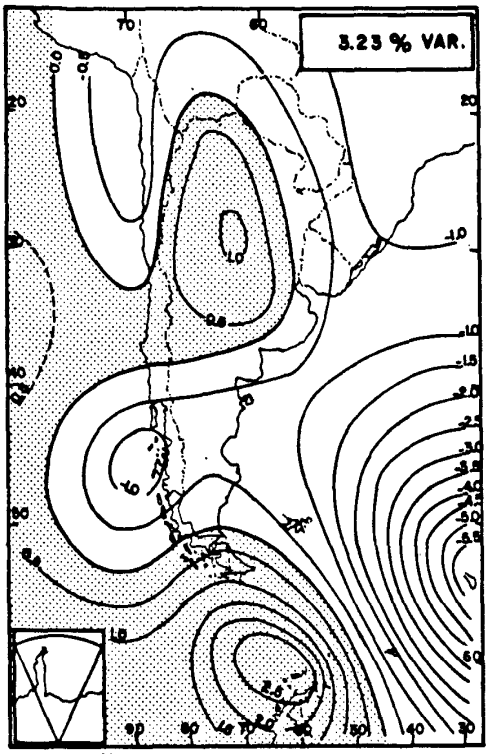


Fig. 8 : 5th PCs PATTERN

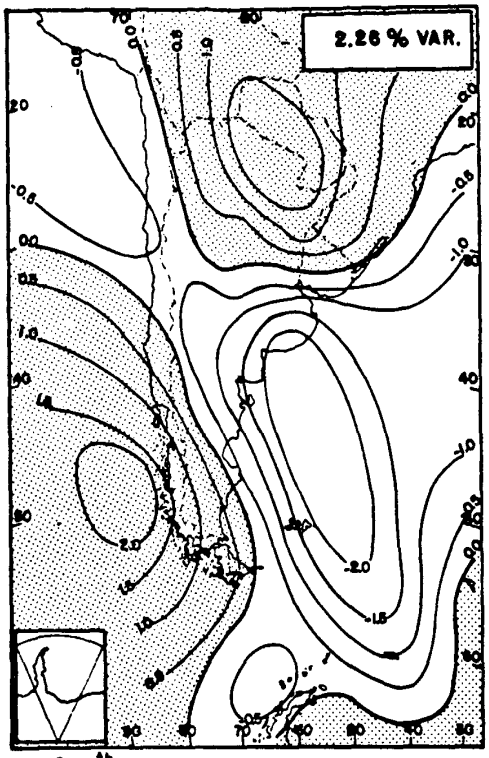


Fig. 9 : 6th PCs PATTERN

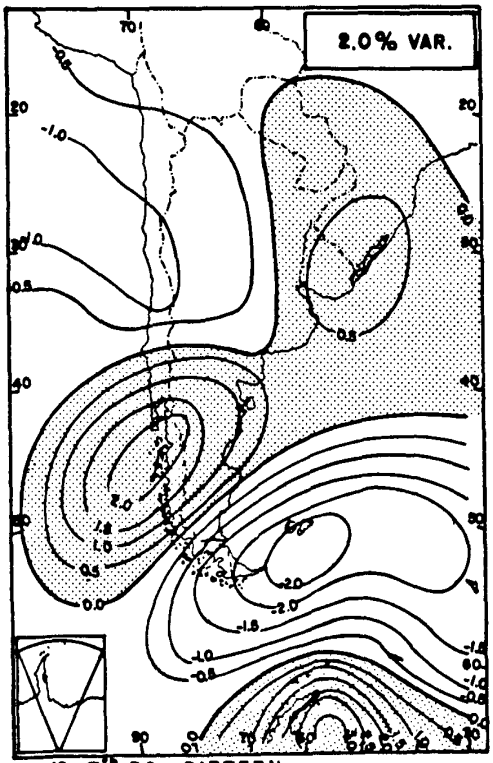


Fig.10 : 7th PCs PATTERN

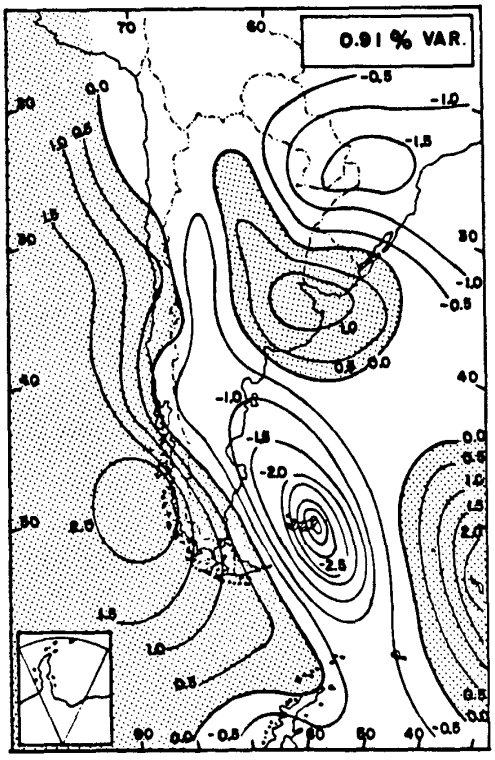


Fig. 11 : 8th PCs PATTERN

Patterns 5 to 8

None of the daily pressure fields can be adequately represented considering only one of the patterns given by the fifth to eighth PCs (see fig.8 to 11).

Nevertheless, some days have significant factor loadings for these fifth to eighth patterns (see fig.3e to 3h). They explain a considerable part of the variance in several real fields, and to reconstruct them in an approximate form, it is neccesary to consider these component patterns.

The remaining factors, higher order than eighth, have small factor loadings for all the days and confirm decreasing values of the factor loadings while increasing the order of the PCs.

PRESSURE PATTERNS AND PRECIPITATION IN SANTIAGO DE CHILE

It is desiderable, as mentioned in the introduction, to find a relationship between the surface pressure patterns fields and the precipitation in the Central Andes.

Vargas and Compagnucci (1985) demonstrated that Santiago de Chile's precipitation is representative of the winter snowfall in the Cordillera de Los Andes between the parallels 30°s and 40°s.

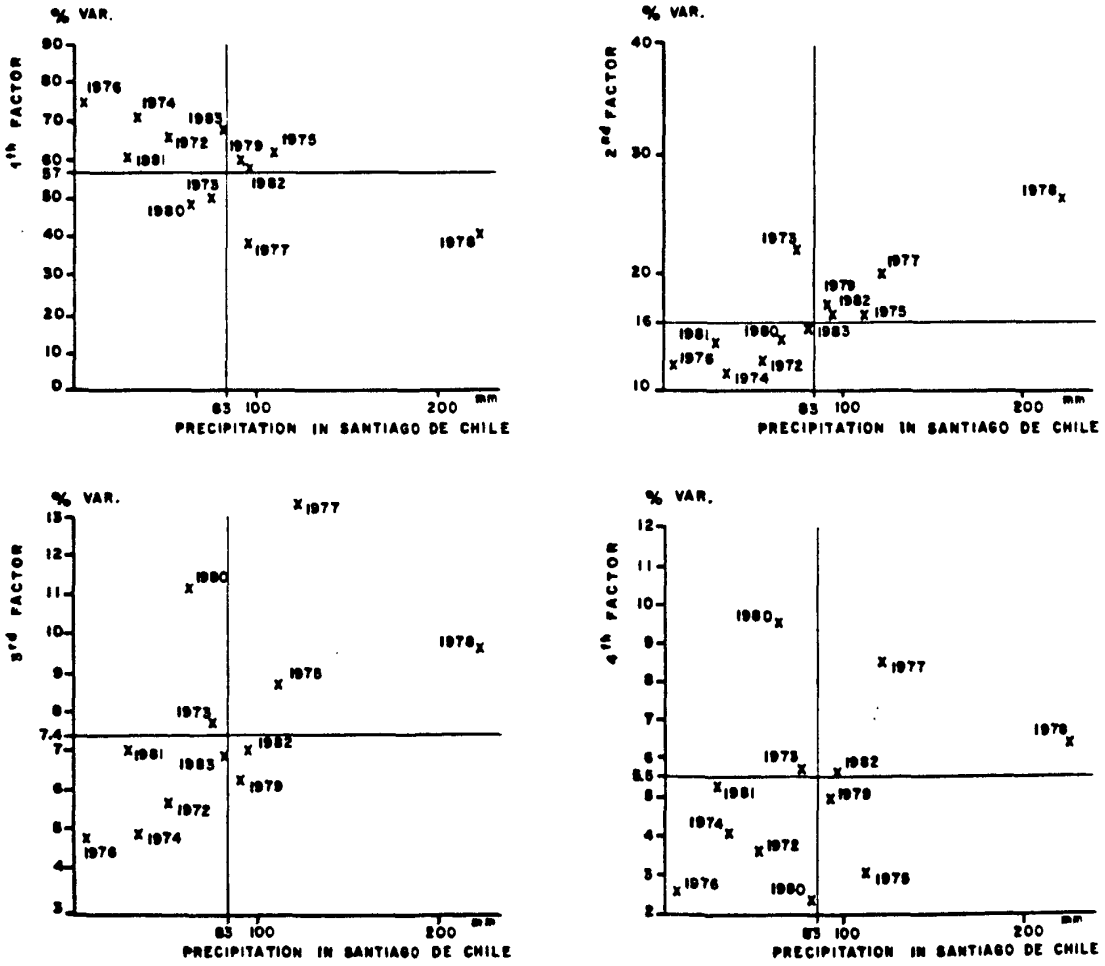


Fig.12: THE RELATION SHIP BETWEEN THE JULY'S PRECIPITATION IN SANTIAGO DE CHILE (mm) AND THE VARIANCE EXPLAINED BY THE FIRST FOUR PCs (%) THE HORIZONTAL AUXILIAR AXIS INDICATES THE VARIANCE EXPLAINED BY THE FACTORS IN THE WHOLE PERIOD 1972-1983 AND THE VERTICAL AUXILIAR AXIS SHOW THE MEAN OF JULY'S PRECIPITATION, TAKING INTO CONSIDERATION THE SAME PERIOD.

Therefore, the precipitation registered in Santiago de Chile for the Julies from 1972 to 1977 is presented versus the variance given by the first four low order PCs, show in fig. 12a to 12b.

The variance explained by the first component has an inverse relation with the precipitation: for the years in which the variance explained by this component is greater, the precipitation is generally lower as we cas see in fig. 12a, and,likewise, when the explained variance is smaller the precipitation is higher.

For the second and third factors, showed in fig.12b and 12c, the relation is direct: higher variance corresponds to higher precipitation and vice versa. The second factor, having the smallest dispersion, presents the clearest relationship. In the fourth factor there is a weak direct relation. In the other highest order component the dispersion increases and the relation is not clear.

The surface pressure fields with a strong content of the patterns given principally by the second factor but also by the third factor are associated with high precipitation whereas low precipitation may be expected when the first factor is dominant. We conclude, consequently, that a high percentage of the fluctuation in the precipitation time series of Santiago de Chile may be largely explained by variations in the explained variance by these first four surface pressure patterns.

CONCLUSIONS AND FUTURE WORK

The daily surface pressure fields of July can be expressed as a linear function of the first eight un-rotated PCs, losing only a low percentage of variance because these components account for 95% of it.

However, several days can be expressed considering only two or three factors, and it is also possible to find daily fields which are very similar to one of the first four theorical models.

The first pattern shows equal structure to the mean sea level pressure field, the inverse being impossible as a real case.

The second pattern is a high pressure center covering the major part of the continental area, centralized in Patagonia, and with equal probability, the inverse is a low pressure center.

Fluctuations in the variance explained by these patterns and the two following, reflect the interanual variability of the precipitation in Santiago de Chile.

These results suggest the importance of extending the analysis to the other winter months.

One potential use is to find the theorical probability of precipitation corresponding to the significant patterns given by low order PCs, and to use it later in objective forecasts of precipitation.

If a day of July can be expressed as a linear combination of these eight first patterns, we may expect that the probability of precipitation of that day will be a linear combination of the theorical precipitation probability of each model.

ACKNOWLEDGMENTS

We specially thank Mr. Luis Fornero for preparing the programs which made possible this work.

The linguistic assistance of Prof. Ester Car de Dubanced (SECEDOC-CRICYT) is much appreciated.

We also thank the personnel of Comunicación Visual and SECECOM (CRICYT), specially to Dr. Richard Branham, director of SECECOM,for their help.
(CRICYT : Centro Regional de Investigaciones Cientificas y Tecnológicas)

BIBLIOGRAPHY

- Altinger de Schwarzkopf, M.L. and Migliardo, R.C., 1973: El tornado de San Justo, Provincia de Santa Fe, ocurrido el día 10 de Enero de 1973. Meteorológica, Vol.4, 65-87 pp.
- Barry, R.G. and Perry, A.H., 1973: Synoptic Climatology Methuen Co Ltd. London, 555 pp.
- Compagnucci, R.H., Fornero, L. and Vargas, W.M., 1984: Algunos Métodos estadísticos para tipificación de situaciones sinópticas: discusión metodológica. 13 Reunión Científica Argentina de Geofísica y Geodesta, Bariloche. (Aviable from CRICYT-ME. CONICET Mendoza, Argentina). To publishe in Geoacta.
- De Fina and Sabella, 1958: La predicción de las áreas de lluvia por el nuevo método casi objetivo RDF. Publicación N°61 Institutos de suelos y Agrotecnia, Buenos Aires, ó, Revista de Investigaciones Agrícolas, Argentina, Tomo 12.
- Farmer, S.A., 1971: An investigation into the results of principal components analysis of data derived from random numbers, The Statistician, Vol.20, N°4, 63-72 pp.
- IMSL, 1980: International Mathematical and Statistical Library, Fortran Subrutines for Mathematics and Statistics Problem Solving Software System. IMSL. Coustomer Relations, Texas, USA.
- Jornadas de Nivo-Glaciología, 1969: Mendoza, Argentina (Aviable from CRICYT-ME., CONICET, Mendoza).
- Lund, I.A., 1963: Map pattern classification by statistical methods. Journ. of Applied Met., 2, 56-65 pp.
- Mc Quitty, L., 1957: Elementary linkage analysis for orthogonal and oblique types and typal relevancies. Educ. Psychol. Measurement, 17, 207-229 pp.
- Nicolini, M. and Norte, F., 1980: Characteristics of Mendoza Hailstorms and their environment. Proceedings of " III third W.M.O. Sci. Conf.on Weather Mod. ", Clermont-Ferrand, 641-148 pp.
- Pittock, A.B., 1975: Climatic change and the patterns of variation in Australian rainfall Search, 6, 498-504 pp.
- Richman, M.B., 1980: Mar typing patterns associated with potential urban enhanced precipitation. M.S. thesis, University of Illinois, Urbana, (Aviable from Illinois State Water Survey, Champaing), 123 pp.
- Sawyer, J.S., 1956: Rainfall of depressions which pass eastward over near the British Isles. Mett. Off. Porf. Notes N°118.
- Vargas, W.M. and Compagnucci, R.H., 1983: Methodological aspects of principal component analysis in meteorological fields. Proceding of " II International Meeting on Statistical Climatology " Lisboa, Portugal Instituto Nacional de Meteorología y Geofísica, 5.3.1-5.3.9 pp.
- Vargas, W.M. and Compagnucci, R.H., 1985: Relaciones del régimen de precipitación entre Santiago de Chile y las series de la región cordillerana (To publishe in GEOACTA, Aviable from CRICYT-ME, Mendoza, Argentina).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

HOMOGENEOUS SUBREGIONS ACCORDING SURFACE PRESSURE VARIABILITY DURING JULY IN SOUTHERN SOUTH AMERICA AND THE ANTARCTIC PENINSULA

Walter Mario Vargas
Facultad de Ciencias Exactas y Naturales, Dpto. de Meteorología
Universidad de Buenos Aires
1428 Nuñez, Buenos Aires, Argentina

Rosa H. Compagnucci
Instituto Argentino de Nivología y Glaciología(IANIGLA)
C.C.330, 5500 Mendoza, Argentina

INTRODUCTION

The purpose of this paper is to describe the spatial patterns of surface pressure variability over the southern part of South-America and the Antarctic Peninsula during July.

The results of this analysis will allow the region to be divided into subregions of high spatial coherence, which can be used to separate permanent effects due to orographic conditions and the geographical location of the area with regard to the general circulation.

The southern part of South America is a narrow continental cone extending meridionally to the south of parallel 20°S between 45°W and 80°W. The Cordillera de los Andes, located in its occidental area, is a mountain range of approximately 5000 mts. situated from north to south as an immense wall perpendicular to the westerlies. To the south of parallel 35°S it is lower and narrower, offering less resistance to the westerlies circulation. The north part of the western and eastern coasts is under the influence of the semipermanent subtropical anticyclones of the Pacific and Atlantic Oceans. The subpolar front is located between Tierra del Fuego and the Antarctic Peninsula.

We seek here to define the subregions of surface pressure variability in an objective way and to look for possible changes in the characteristic spatial patterns. These results complement the ones obtained by the July surface pressure fields in the period 1972 to 1983 given by Compagnucci and Vargas (1986).

DATA AND METHODOLOGY

The surface pressure registered by eighty one stations of the southern part

of South America and the Antarctic Peninsula (fig.1) during 372 days of the Julies from 1972 to 1983 are taken into account for this analysis.

The data belong to the hour 12Z and were obtained from the synoptic maps of the National Meteorological Service of Argentina. This information matrix was also used in the study "Patterns of surface pressure fields during July 1972-1983 in Southern South America and the Antarctic Peninsula" (Compagnucci and Vargas, 1986).

As the aim of this paper is to analize the relationship among the stations, the network points (81) are considered here as variables and the days (372) as observations.

The data matrix $\mathbb{P}$ (81 by 372) was standardized by rows, which means to standardized the time series of each station, giving an S-Mode input matrix as following:

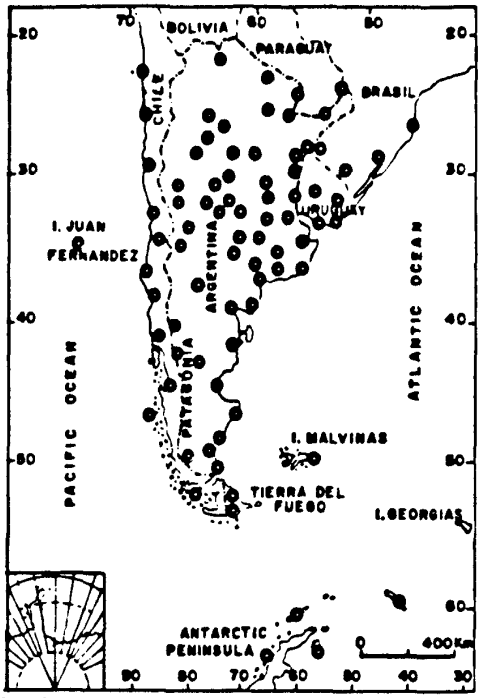


Fig. 1: GRID-POINT SELECTED FROM STATIONS REPORTING SEA LEVEL PRESSURE

$$\mathbb{P} = \left\{ p_{ij} \right\}$$

$\mathbb{P}$: data matrix (N x M)

N : number of stations and M : number of days

pij: pressure for the station-i and day-j

$$\bar{p}_i = \frac{1}{M} \sum_{j=1}^{M} p_{ij}$$ $\bar{p}_i$: mean pressure of the station-i

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{M} \left(p_{ij} - \bar{p}_i \right)^2}{M}}$$ σ_i : standard deviation for the station-i

$$\mathbb{S} = \left\{ s_{ij} \right\} \quad \text{with} \quad s_{ij} = \frac{p_{ij} - \bar{p}_i}{\sigma_i}$$ is the S-Mode input matrix

The spatial variability patterns are given by the factor loadings in the principal component analysis for the S-Mode input matrix. The principal components (PCs) were calculated using the OFPRI and OFCOEF subroutines of the IMSL (1980) package.

We used Lund's method (1963), which applied simple linear correlation to define accurately the subregions in which the area is divided. The procedure is as follows: the station in the network which has the largest number of stations correlated with it, using an arbitrary threshold of r= 0.7, is selected as first type. After abstraction of these, the station with the next highest number of related stations is designed as second type, and so on.

RESULTS

We examined the temporal stability of the PCs patterns by dividing the data into two periods: 1972-1977 and 1978-1983.

The PCs analysis for the whole period 1972-1983, and the subperiods show that the variances explained by low order components in each period were similar (see Table 1).

PC_s	1972-1983			1972-1977			1978-1983		
	Eigenvalue	Per.Val.	Per.Cum.Val.	Eigenvalue	Per.Val.	Per.Cum.Val.	Eigenvalue	Per.Val.	Per.Cum.Val.
1	47.22	58.30	58.30	45.82	56.57	56.57	49.03	60.53	60.53
2	11.84	14.61	72.92	12.61	15.57	72.15	11.07	13.67	74.20
3	6.97	8.60	81.53	7.10	8.77	80.92	6.77	8.36	82.57
4	3.12	3.87	85.40	3.33	4.11	85.04	3.17	3.91	86.49
5	2.94	3.69	89.10	2.95	3.64	88.68	2.92	3.61	90.10
6	1.67	2.06	91.17	2.03	2.51	91.19	1.49	1.84	91.94
7	1.24	1.53	92.70	1.36	1.68	92.88	1.03	1.28	93.22
8	0.99	1.22	93.93	0.96	1.19	94.04	0.96	1.18	94.41
9	0.77	0.96	94.89	0.82	1.02	95.09	0.65	0.80	95.20

Table 1: EIGENVALUES, PERCENTAGE VARIANCE EXPLAINED AND CUMULATIVE PERCENTAGE EXPLAINED BY THE FIRST COMPONENTS FOR THE ANALYSIS OF DIFFERENT PERIODS

Considering that Kaiser's test (1959) indicates as significant those PCs whose eigenvalues are greater than one, the variability of the surface pressure for July can be explained taking into account the models given by the seven first PCs, which account for 93% of the total variance.

Independently, Lund's method gives, for the three analized periods, a division of the studied area into seven subregions (see fig.4a to 4c), agreeing with the Kaiser's test results.

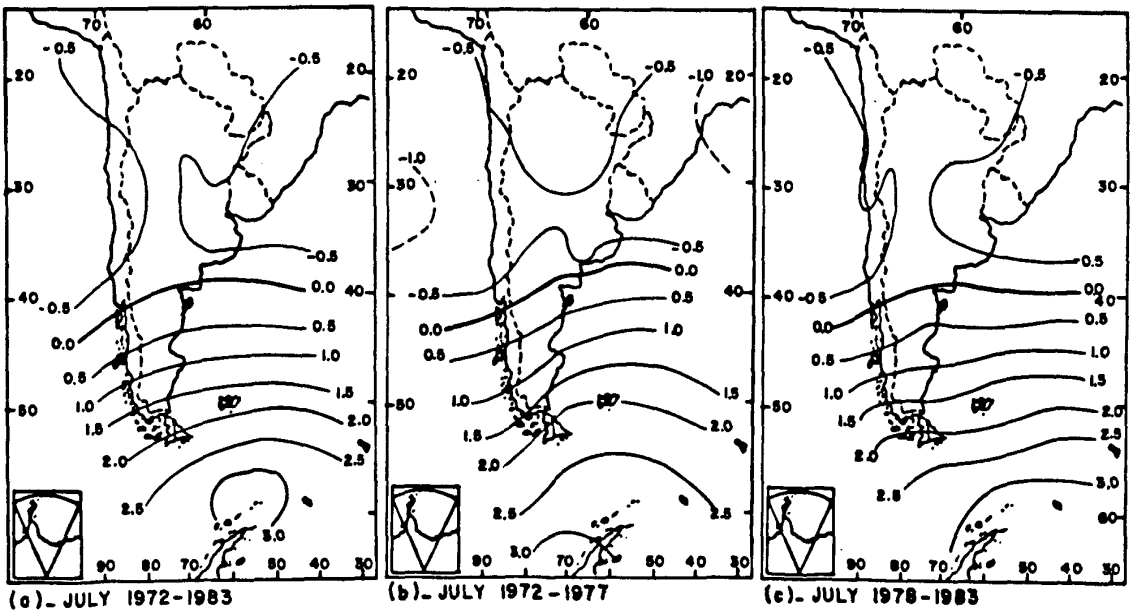


Fig.2 : THE FIRST PATTERNS OF DAILY SURFACE PRESSURE FIELDS FOR JULY OBTEINED BY PRINCIPAL COMPONENT ANALYSIS WHITH T-MODE INPUT MATRIX. (COMPAGNUCCI AND VARGAS, 1986)

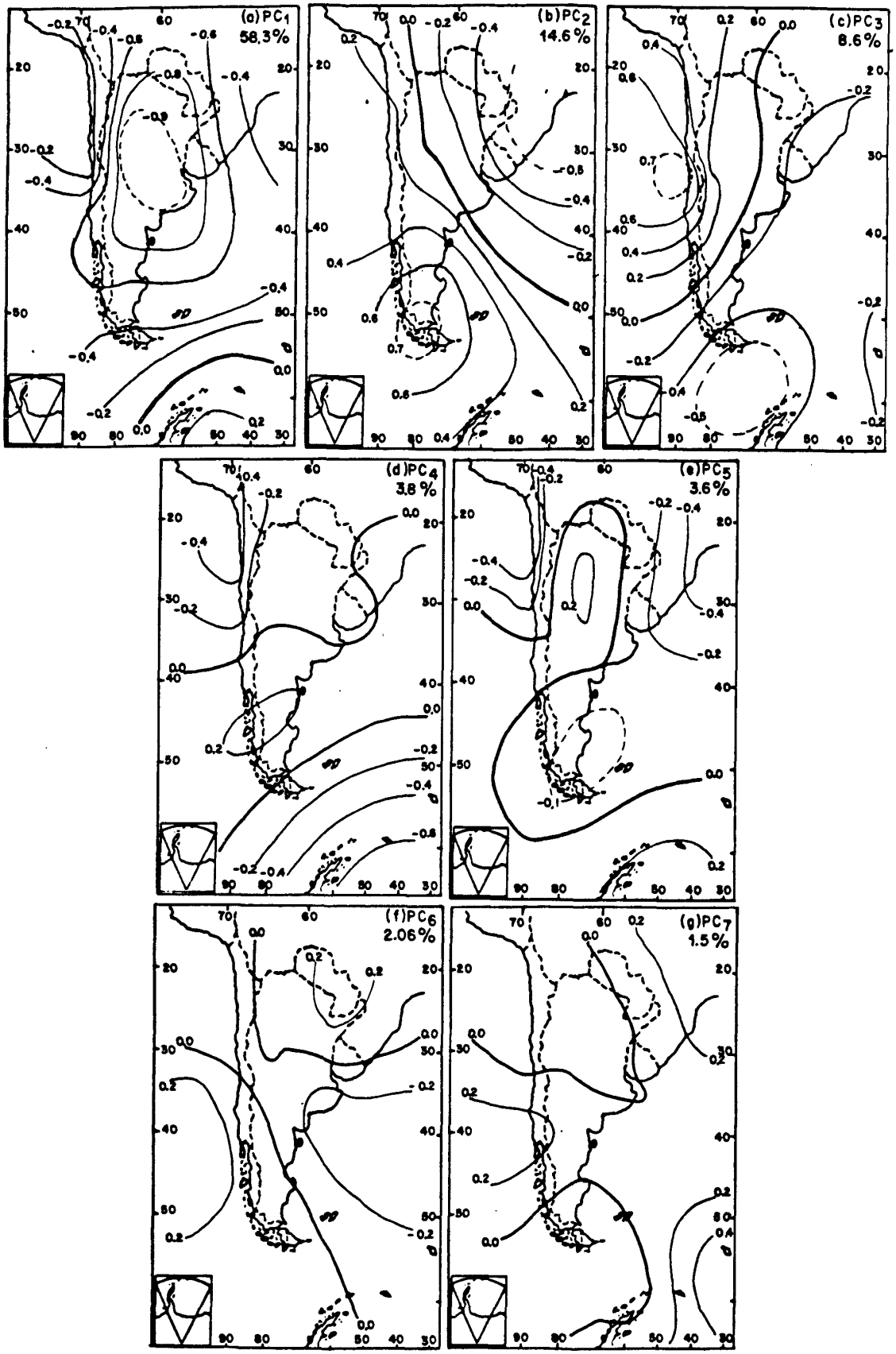


Fig. 3: PRINCIPAL COMPONENT PATTERNS FOR JULY DURING 1972-1983

Variations in the results are observed in differences of extension from the subregions corresponding to the 1972-1977 and 1978-1983 subperiods (see figs.4b and 4c), the regionalization obtained for 1972 to 1983 being similar to that of 1978 to 1983. These variations are produced by differences in the atmospheric circulation for the subperiods.

Compagnucci and Vargas (1986) obtained the types of the surface pressure fields. The results show that the subperiods present some discrepancies in the distribution of the circulation patterns.

Also, there were some differences among the first pattern structure (see fig.2a to 2b) that resulted from the analysis for the two subperiods and that for the whole period.

For 1972-1983 and 1978-1983 the results show that the semipermanent anticyclones of the Atlantic and Pacific Oceans are more pronounced on the continent, and the westerlies affect higher latitudes than in the 1972-1977 period (see fig.2a to 2c).

The patterns that result from the remaining significant PCs were equal for the three periods, so they can not be the reason for the difference in the subregions extension.

PATTERNS OF SURFACE PRESSURE VARIABILITY

The patterns given by the seven first components for the whole period 1972-1983 are shown in fig.3a to 3g. The same basic patterns occur in the two subperiods.

The first pattern (fig.3a) shows variations affecting the region as a whole, with maximun loadings over central Argentina corresponding to the first coherent region identified by Lund's method (fig.4a). The strong gradient between this area and the north part of Chile is an indication of the marked difference between leeward and windward produced by tha Cordillera de los Andes to north of 35°S because of its morphological characteristics.

The second PCs pattern (fig.3b) shows variations of opposite signs northeast and south-west of a zero line, extending from the north to Chile towards the south-east, with anomalies in the south of Brasil and the north-eastern part of the area, and in the opposite sense in Chile and the south-west of Argentina. The south-east to north-west division of surface pressure anomalies is evident throughout the year because the frontal system has in general a south-west to north-east movement for the research area. This characteristic is accentuated in the winter months.

The third pattern (fig.3c) shows anomalies of opposite sign occuring in the west and east of the southern cone of South America, with the largest anomalies in the central part of Chile.

The fourth component pattern (fig.3d) indicates a north western and south eastern portion of the area with the largest anomalies of the same sign and variations of opposite sign extending from north-east to south-west.

For the fifth component the pattern (fig.3e) shows the largest anomaly in northern Chile, and with the opposite sign the variations in central and northern Argentina.

The sixth model (fig.3f) distinguishes the north-eastern part of the research area with the largest anomaly and, in the same sense, the southwest.

The seventh model (fig.3g) shows the largest variation in the south-eastern part of the area covering the Orcadas and the Georgias Islands with the same sign as the variations in the north of Chile.

For higher order PCs the variations accounted for are small, and since our aim is to identify relatively important patterns such detail is not relevant here.

SUBREGIONS OF COHERENT SURFACE PRESSURE VARIATIONS

The first PCs patterns suggest seven main regions. According to Lund's method (with r≥0.7) we can identify seven subregions of high spatial coherence. Figures 4a to 4c show the regionalization for the whole period 1972-1983 and for the subperiods.

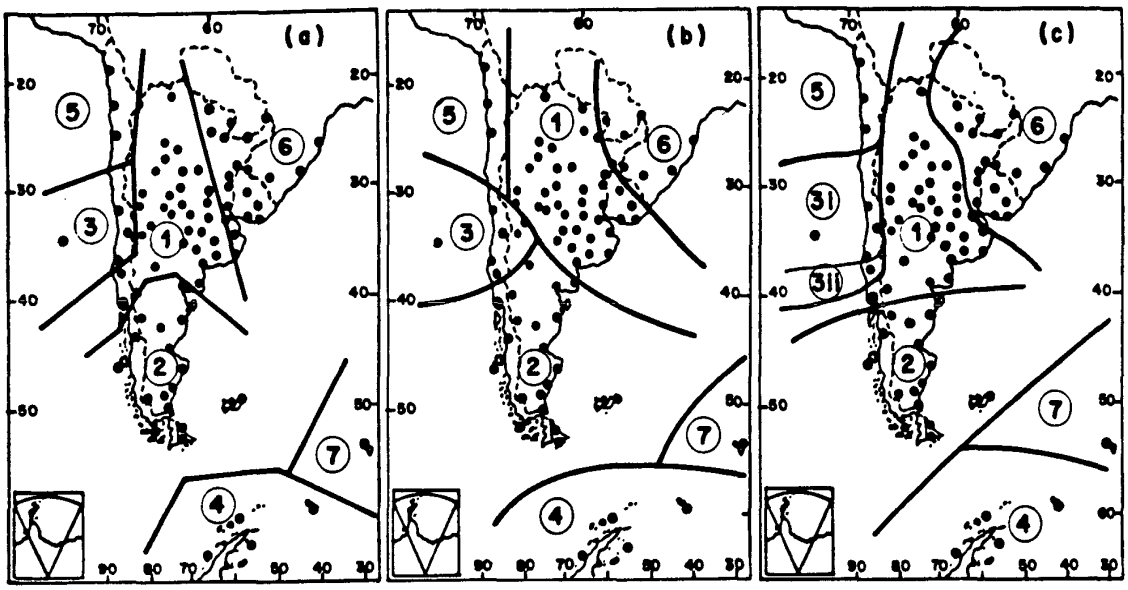


Fig.4: THE SUBREGIONS DEFINED BY THE LUND'S METHOD WITH AND ARBITRARY THRESHOLD r=0.7, FOR THE WHOLE PERIOD 1972-1983 (fig. 2-a), THE SUBPERIODS 1972-1977 (fig.2-b) AND 1978-1983 (fig.2-c)

The first subarea corresponds to the central part of Argentina; it is the most continental part of the studied area and where the effect of the semipermanent anticyclones of the Atlantic and Pacific Oceans is less.

The area of the second subregion covers Patagonia, the Malvinas Isles, Tierra del Fuego, and the south of Chile. This is the zone of westerlies influence. There is not a difference between leeward and windward of the Cordillera de los Andes because, as it is lower and narrower, it does not disturb the flow such as occurs north of parallel 35°S.

The third subregion corresponds to the area on the continent where the Pacific anticyclon lower limit oscilates. Minetti (1982) called it a "climatic sensitive area" because parameters take place in it, showing more marked fluctuations than the ones which occur in the rest of Chile. The results of the 1978-1983 subperiod (fig.4c) show this subregion divided into two parts (3i) and (3ii). But analyzing the correlation values between them, the values are 0.6 or higher, so they can be considered as only one subregion.

The Antarctic Peninsula and Orcadas Isles correspond to the fourth subregion.

The fifth subarea comprises the northern part of Chile which is under the permanent influence of the Pacific anticyclon.

The sixth subregion covers, for 1972-1977, the south of Brasil and the northeast of Argentina (fig.4b) spreading out to the west for 1978-1983 and 1972-1983. This subregion is affected by the influence of the Atlantic anticyclon.

The seventh subregion is defined by the meteorological station in the Georgias Islands. Its record has correlation coefficient of 0.8 with respect to the record of the Orcadas Isles but correlations of less than 0.3 with respect to the other stations of the network. Perhaps this difference is due to

the disturbance that the continent provokes in the synoptic waves not having an evident influence in the Georgia Station, and this resulted in a high percentage of error. Nevertheless, this last hypothesis would be contradicting the high correlation with the Orcadas station

CONCLUSION AND FUTURE WORK

The surface pressure variability in July for the southern part of South America can be explained by the seven first PCs patterns, which account for 93% of the total variance.

The variances explained by the first PCs are similar in both subperiods, 1972-1977 and 1978-1983, and to the whole period 1972-1983. This allows us to insure the stability in the regionalization not withstanding some years being singular, as are the "El Niño" years.

An application of Lund's method (with $r \geq 0.7$) to the three periods studied shows the research area divided into seven subregions that only vary slightly in extension without altering their number and location.

The differences in the extension of the subregions for the two subperiods show that the influence of the Pacific and more notably the Atlantic anticyclons is greater in 1978-1983 than in 1972-1977. And also we can see the westerlies in lower latitudes for 1972-1977 than in 1978-1983. This confirms the results obtained by the typification of the daily surface pressure fields (Compagnucci and Vargas, 1986).

Considering these results we think that it will be worthwhile to study the stability of the regionalization for the other months. An analysis of typical summer months could indicate the maximum possible deviation of this regionalization obtained for July. Likewise, this procedure will be applied to anomalous years such as pressumably the years of "El Niño". We hope in this way to determine the variations of surface pressure deviations for the research area, the southern cone of South America.

ACKNOWLEDGMENTS

We specially thank Mr.Luis Fornero for preparing the programs which make possible this work.

The linguistic assistance of Prof.Ester Car de Dubanced (SECEDOC-CRICYT) is much appreciated.

We also thank the personnel of MAGRAF (CRICYT) and SECECOM (CRICYT),specially to Dr.Richard Branham, director of SECECOM for their help.

BIBLIOGRAPHY

- Compagnucci, R.H. and Vargas, W.M. (1986): Patterns of surface pressure fields during July 1972-1983 in Southern South America and Antarctic Peninsula, Third International Conference on Statistical Climatology, June 23-27, 1986, Vienna, Austria.
- IMSL (1980): International Mathematical and Statistical Library Fortran Subroutines for Mathematics and Statistics Problem Solving Software Systems. IMSL, Customer Relations, Texas, USA.
- Kaiser (1959): The application of electronic computers to factor analysis; Symp on the application of computer to psychological problem. Am. Psy. Ass.
- Lund, I.A. (1963): Map-pattern classification by statistical methods; Journ. of Applied Met. 2, 56-65.
- Minetti and others (1982): La actividad anticiclónica y las precipitaciones en Chile y en la zona Cordillerana Central Andina, Instituto Panamericano de Geografía e Historia, Geofísica 16, 145-157.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

COMPLETION OF AN INCOMPLETE DATA SET BY EOF-ANALYSIS: THE METHOD IS APPLICATED TO THE HISTORICAL SEA SURFACE TEMPERATURE DATA SET

Rita Glowienka-Hense, Andreas Hense
Meteorologisches Institut, Bonn
Hans von Storch
Max Planck Institut für Meteorologie, Hamburg

Estimates of principal vectors the socalled empirical orthogonal functions (EOF) from meteorological data are, due to their convergence properties, very valuable for statistical testing and data compression. This has been discussed in detail eg. by Storch and Kruse (1985). Here another property of EOF analyses will be demonstrated. Up to now the analysis has only been performed for data sets that were thought to be reliable in the sense of being homogeneous and complete. It was generally accepted that such sophisticated mathematical methods can only be useful for good data. In the following it will be shown that because of their special convergence properties EOF can, under certain constraints, also be a helpful tool for noisy and even incomplete data sets. The information still inherent in these data is optimally filtered out through the dominant eigenvectors (EV), which are then available for eg. statistical testing. Additionally the EOF-expansion provides a method to interpolate spatially missing field values as long as the data coverage is not too sparse. Historical sea surface temperatures (HSST) for the Atlantic have been used as an example to test the above mentioned thesis.

DATA

The data set used is the HSST data set of the German Seewetteramt for the Atlantic covering the period 1850 to 1960, where ship observations of SST have been averaged on an irregular grid (Höflich 1974).

The data set is a classical example for a data set containing the aforementioned insufficiencies. The available data are greatly reduced during the first thirty to forty years and during the two world wars. Variing ship routes led to variing measurement locations within the grids which is especially precarious in coastal areas or other regions of sharp temperature gradients. (See figure 1 for data coverage)

METHODS

In case of a complete finite data set Y(t,i) the EOF-analysis provides a series of orthonormal functions which fully describe the time series under discussion. The requirement of an optimal representation through these functions leads to the following eigenvalue problem:

$$R_{ij} \; X_j = l_j X_j$$

$$R_{ij} = \frac{1}{N-1} \sum_t Y(t,i)Y(t,j)$$

where R is the correlation matrix and l,X are its eigenvalues and EV respectively. The question tested in the following is whether the correlation matrix for a noisy and incomplete data set like the HSST can be estimated with sufficient accuracy to provide reasonable values of the eigenvalues and EV. The formula for the estimated matrix r_{ij} from the incomplete data set replacing R_{ij} is:

$$r_{ij} = \frac{1}{t(i,j)-1} \sum_{t(i,j)} Y(t,i)Y(t,j)$$

where t(i,j) are the time steps where measurements at grid point i and j are available. A rule of thumb for the differences in the eigenvalues and EV arising from the incomplete sampling can be derived following North et al (1982).

Estimates of the amplitude time series of the derived EV were computed according to the following formula:

$$C_i(t) = \frac{M}{g(t)} \sum_j Y(t,j)X_i(j)$$

where C_i is the time amplitude function associated with EV X_i, g(t) the number of available data at time t and M is the number of grid points. To care for bad estimates of the time series in case of missing data, the time amplitudes were replaced by fitted spline values, which were weighted according to the available data, if there were less than forty grid elements with data at a certain time.

The derived EV and amplitudes are by definition global values, which means they are defined at every grid point and at every time. Thus putting

$$Y^*(t,j) = \sum_{i=1}^{N} C_i(t) \; X_i(j)$$

where N is small compared tp M provides a filtered and spatially interpolated field Y^* of the original data as long as the amplitudes $C_i(t)$ are reliable.

The procedure of our analysis was the following. First there were set certain thresholds subjectively fixed in view of the data available and the demands of statistics. Individual monthly means were accepted if they were computed from at least five measurements. The EOF-analysis was then done - for each calendar month separately - out of those grids with individual monthly means for at least fifty years. These criteria might seem very low but otherwise the analysis would have been based on too few grid points which in turn could

diminish the stability of the EV. The included number of grids was between 113 and 125 for the different months.

Figure 2 shows the first two EV of the HSST for January and their respective time series. There are two hints which seem to support their validity. First they reveal gross structures of coherent patterns and secondly the EV of the following months look very similar. In order to perform a statistical test on the reality of the EV, we profit from the fact that the SST have a memory of more than a month. Thus cross spectra were computed between the time series of EV of succeeding months. Examples of the results with base months January and July are shown in figure 3. In view of the manipulations of the time series through the fitted data the cross spectra have of course to be interpreted with care. But for the high frequency part of the spectrum the smoothing out of data gaps should even diminish the real coherency. The spectra show a memory of the EV of two to three months except for the autumn months of EV 2. So it can be concluded from this self-consistency test of the analysis that the HSST variance represented through these EV describes a true signal of the SST.

The estimate of the eigenvalue was tested by determining the its ratio to the mean square time amplitude. It turned out that the first eigenvalue is nearly twice as large as the mean square of the time amplitude, whereas the second eigenvalue is about 20 % larger than the amplitude variance. Nevertheless the first two EV explain in the mean about 20 % of the variance of the HSST which is considering the large section of the Atlantic covered by the analysis surely a notable signal in the climate system.

CONCLUSIONS

The HSST data set of the German Seewetteramt was used to demonstrate that even for an incomplete data set the method of EOF-analysis can be successfully used to filter physical information out of noisy data.

ACKNOWLEDGEMENT

The research was sponsered by the Climate Research Program of the Rheinisch-Westfälische Akademie der Wissenschaften, Düsseldorf.

LITERATURE

Höflich O.,1974, Part I of GATE Report No. 2, WMO, Geneva, Switzerland.

North G., et al. 1982, Mon. Wea. Rev., 110, 699-706.

Storch H. von, Kruse H.A. 1985, TELLUS A,37A,361-377.

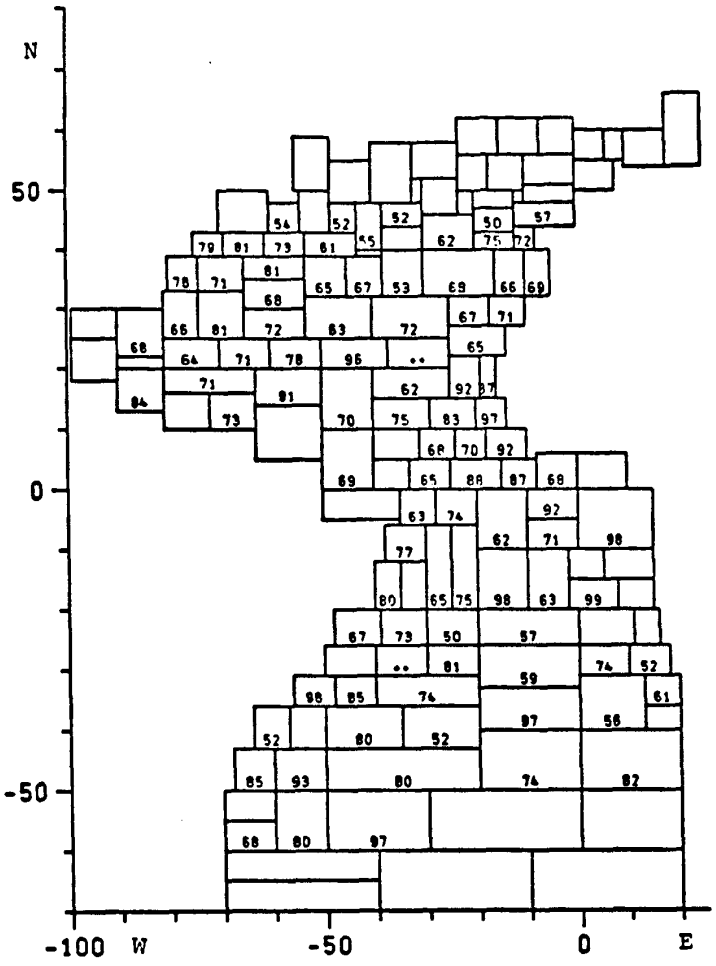


Fig.1: Number of monthly means for January at the single grid points.

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
eigenvalue/C_1^2	1.5	1.8	1.8	1.8	1.8	1.8	1.9	1.7	1.8	1.8	1.5	1.6
eigenvalue/C_2^2	1.3	1.1	1.4	1.4	1.2	1.4	1.2	1.0	1.0	1.0	1.4	1.0

Table 1: Ratio between the first and second eigenvalue and the respective mean square time amplitudes.

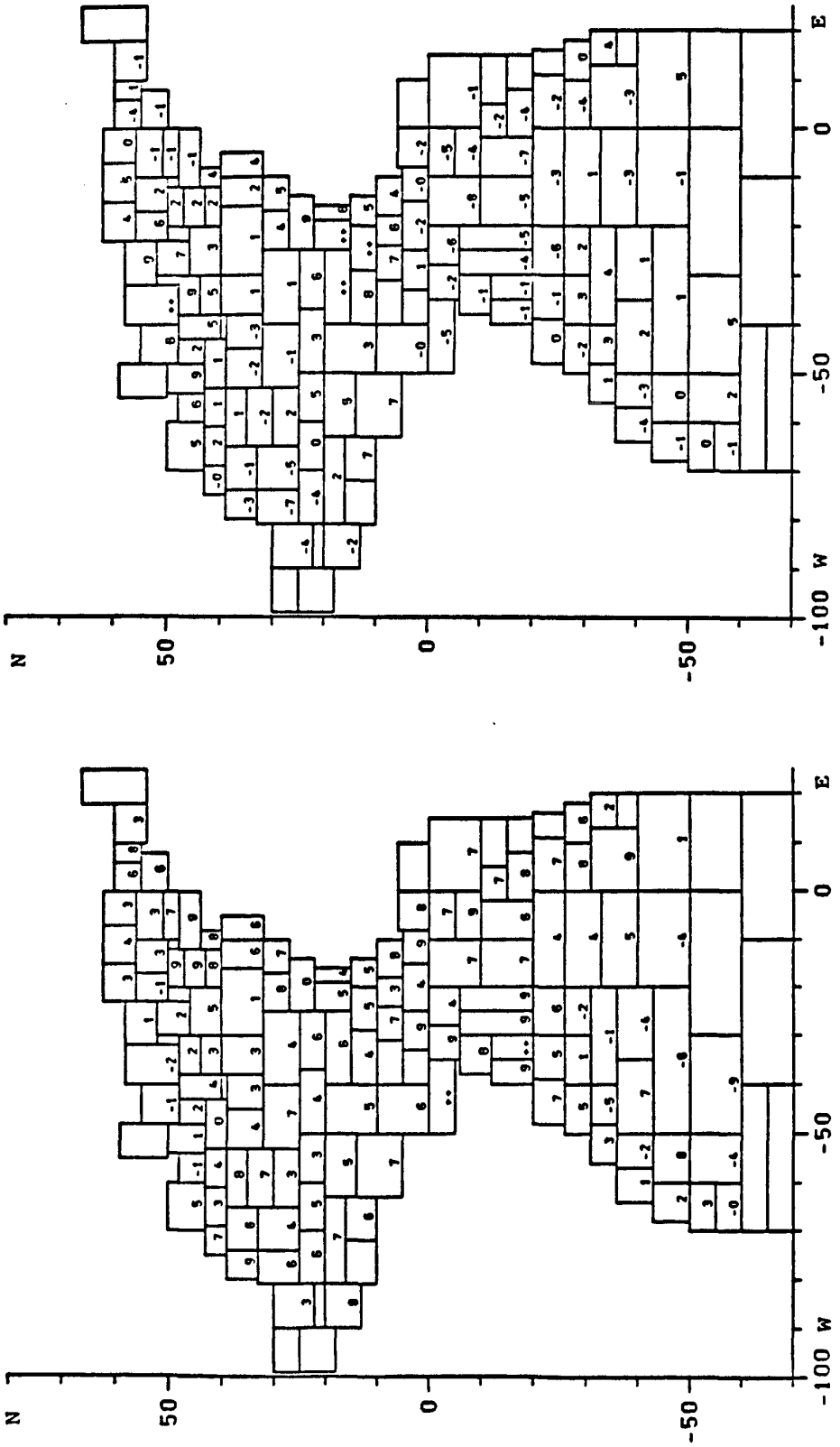


Fig.2a: First and second EV of HSST for January.

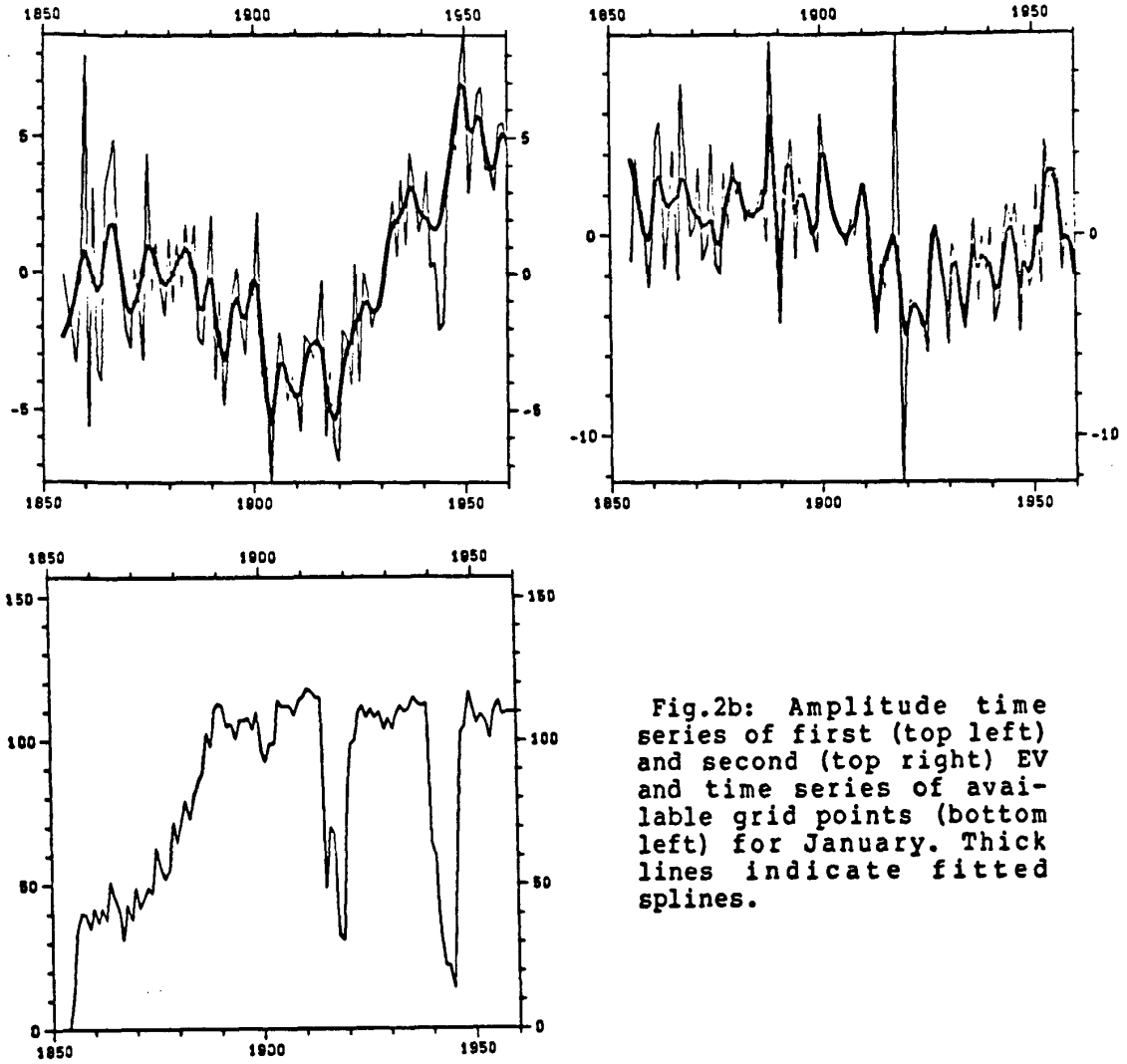


Fig.2b: Amplitude time series of first (top left) and second (top right) EV and time series of available grid points (bottom left) for January. Thick lines indicate fitted splines.

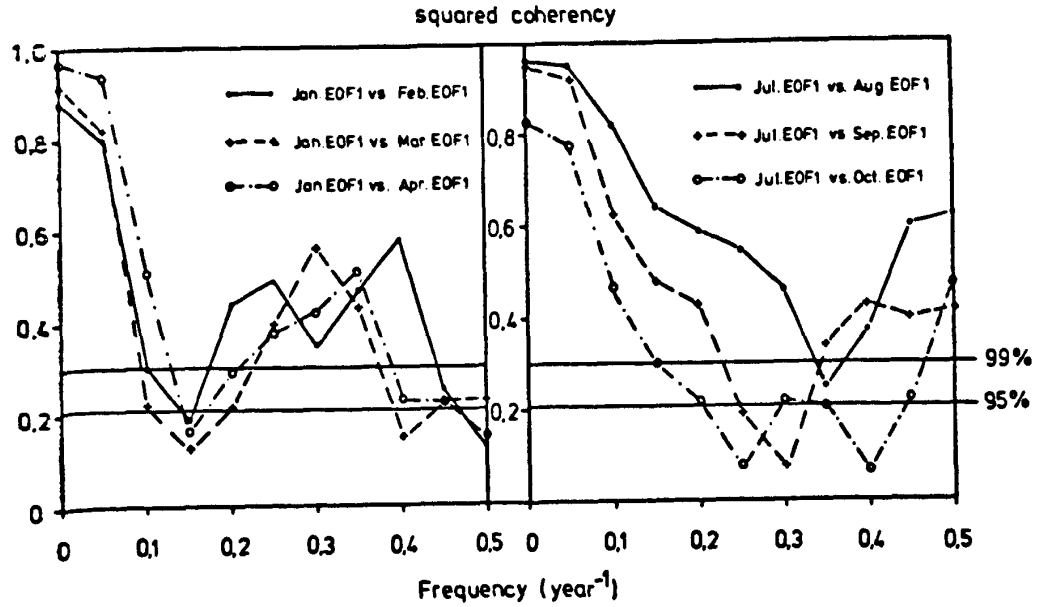


Fig. 3: Squared coherencies of cross spectra between time amplitudes of EV1 for three succeeding months with base months January and July.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

AN ITERATIVE METHOD FOR THE PERIODICAL ANALYSIS OF LONG TIME SERIES

Szentimrey, T. - Gulyás, O.
/Central Institute for Meteorology, Budapest/

One of the basic climatological problems is the detection of hidden periodical components of the meteorological time series. These series are usually modelled like statistical time series, and the discrete Fourier series expansion is mostly applied on purpose to detect the periodical components. It is the so called Schuster's method. But there is a serious problem in connection with this method, namely that it makes the analysis possible only for orthogonal components. This fact is a strict restriction in respect of the period lengths, consequently the allowed periodical components cannot be "physically" interpreted in most of the cases. The more general method than this one, i.e. the method of least squares does not have similar restriction in connection with the period lengths, but in this case the joint investigation of numerous periodical components is not a simple problem, /see e.g., Anderson, 1971, Priestley, 1981/.

Namely, let us suppose that

$$x(t) \qquad t=1,2,\ldots,N$$

is a statistical time series, and

$$\left\{ \varphi_i(t) \right\}_{i=1}^{L} \qquad t=1,2,\ldots,N$$

is a system of linearly independent functions. In our case the elements of this system are trigonometric functions, more exactly

$$\varphi_{2i-1}(t) = \cos \omega_i t \qquad i=1,2,\ldots,K$$

$$\varphi_{2i}(t) = \sin \omega_i t \qquad L=2K$$

and the problem of the method of least squares is the following minimization:

$$\min_{\underset{\sim}{c}} \sum_{t=1}^{N} \left(x(t) - \underset{\sim}{c}^T \underset{\sim}{\varphi}(t) \right)^2$$

where

$$\underset{\sim}{\varphi}(t) = \left[\varphi_1(t), \varphi_2(t), \ldots, \varphi_L(t) \right]^T .$$

In this paper we present an iterative method to calculate the solution of this minimum problem. This iteration is based on a potential type "learning algorithm" that is applied in connection with the recognition of statistical formations /Ajzerman et.al, 1970/. Let us see the suggested iterative algorithm and its properties in a little more general form.

Theorem

Suppose that the domain of the functions $x(t), \varphi_1(t), \ldots, \varphi_L(t)$ is $\{1,2,\ldots,N\}$, and the functions $\varphi_i(t)\ (i=1,2,\ldots,L)$ are linearly independent on this set. Let vector $\underline{c}$ be the solution of the former minimum problem, i.e.

$$\sum_{t=1}^{N}\left(x(t)-\underline{c}^T\underline{\varphi}(t)\right)^2=\min_{\underline{\tilde{c}}}\sum_{t=1}^{N}\left(x(t)-\underline{\tilde{c}}^T\underline{\varphi}(t)\right)^2$$

Furthermore let us assume that the variates $t_1, t_2, \ldots, t_n, \ldots$ are totally independent and identically uniform distributed on the set $\{1,2,\ldots,N\}$ i.e. $P(t_i=j)=\frac{1}{N}\ (j=1,2,\ldots,N)$; the white noise $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n, \ldots$ with expectation 0 and variance σ_i^2 is independent of the variates $t_1, t_2, \ldots, t_n, \ldots$ and $y_i = x(t_i)+\varepsilon_i\ (i=1,2,\ldots,n,\ldots)$.

The iterative algorithm:

$$\underline{c}_1=\tau_1\underline{\varphi}_1\ ,\quad \underline{c}_n=\underline{c}_{n-1}+\tau_n\underline{\varphi}_n$$

where

$$\tau_1=\gamma_1 y_1\ ,\quad \tau_n=\gamma_n\left(y_n-\underline{c}_{n-1}^T\underline{\varphi}_n\right)$$

$$\underline{\varphi}_n=\left[\varphi_1(t_n),\varphi_2(t_n),\ldots,\varphi_L(t_n)\right]^T$$

and $\{\gamma_n\}_{n=1}^{\infty}$ is a sequence of positive numbers. This procedure might be seen as a very simple one.

1. Statement

If $\sum_{n=1}^{\infty}\gamma_n=\infty$ and $\sum_{n=1}^{\infty}\gamma_n^2<\infty$ then

i, $\underline{c}_n$ is an asymptotically unbiased estimator of coefficient $\underline{c}$, and

ii, $\underline{c}_n$ is a consistent estimator of $\underline{c}$.

It is fulfilled e.g. in case of $\gamma_n=\frac{1}{n}$.

In our case, when the functions $\varphi_i(t)\ (i=1,2,\ldots,L)$ are trigonometric functions, i.e.

$\varphi_{2i-1}(t) = \cos \omega_i t \quad , \quad \varphi_{2i}(t) = \sin \omega_i t \qquad (i = 1,2,\ldots,k \; ; \; L = 2k)$

the following statements are true as well, being connected with the rate of convergence.

2. Statement

$$E\left(\|\underline{c}_n - \underline{c}\|^2\right) \le \|\underline{c}\|^2 \prod_{i=1}^{n} \Lambda_i + K(\sigma_1^2 + \sigma_2^2)\left(\sum_{i=1}^{n-1} \gamma_i^2 \prod_{j=i+1}^{n} \Lambda_j + \gamma_n^2\right)$$

$$E\left(\|\underline{c}_n - \underline{c}\|^2\right) \ge \|\underline{c}\|^2 \prod_{i=1}^{n} \lambda_i + K(\sigma_1^2 + \sigma_2^2)\left(\sum_{i=1}^{n-1} \gamma_i^2 \prod_{j=i+1}^{n} \lambda_j + \gamma_n^2\right)$$

where

$E\left(\|\underline{c}_n - \underline{c}\|^2\right)$ is the expectation of the deviation $\|\underline{c}_n - \underline{c}\|^2$;

$\sigma_1^2 = \frac{1}{N} \sum_{t=1}^{N} \left(x(t) - \underline{c}^T \underline{\varphi}(t)\right)^2$ i.e. the mean-square deviation of the approximation;

Λ_i is the largest eigenvalue and λ_i is the least one of the matrix $(I - \gamma_i(2 - \gamma_i K)\Phi)$ where

$$\Phi = \frac{1}{N} \sum_{t=1}^{N} \underline{\varphi}(t)\, \underline{\varphi}^T(t) \; .$$

It is worthy to mention that if the functions $\varphi_i(t)$ $(i = 1,2,\ldots,L)$ constitute an orthogonal system, then there are equalities instead of unequalities.

3. Statement

$\lambda_i < 1 \; , \; \Lambda_i < 1$ if and only if $0 < \gamma_i < \frac{2}{K}$

Consequently if $\gamma_i > \frac{2}{K}$ $(i = 1,2,\ldots,n)$ and the vectors $\underline{c}_i$ $(i = 1,2,\ldots,n)$ are the corresponding estimators, furthermore $0 < \gamma^* < \frac{2}{K}$ and $\underline{c}^* = \gamma^* \gamma_1 \underline{\varphi}_1$ then the estimator $\underline{c}^*$ is more efficient than the estimators $\underline{c}_i$ $(i = 1,2,\ldots,n)$ because

$$E\left(\|\underline{c}_i - \underline{c}\|^2\right) > E\left(\|\underline{c}^* - \underline{c}\|^2\right) \qquad i = 1,2,\ldots,n$$

With respect to the above statements of theorem, the sequence of $\gamma_n = \frac{1}{K/2 + n}$ was applied by us in the practice.

Last but not least a meteorological example is presented.

The analysed meteorological time series was the decade averages of 80-year temperature data series of Budapest, during the period 1901-1980, i.e. the number of elements was 2880.

The investigation consisted of two parts.

1. The comparison of the iterative method with the direct calculation method in case of 500 orthogonal periodical components.

After 15000 iteration steps the coefficients were found almost one and the same, and only the one-year periodical components differed significantly from 0.

Direct method:

$$x(t) = 11.21 - 10.45 \cos\frac{2\pi}{36}t - 3.38 \sin\frac{2\pi}{36}t + \varepsilon_1(t) \qquad (t=1,2,\dots,2880)$$

Iterative method:

$$x(t) = 11.13 - 9.52 \cos\frac{2\pi}{36}t - 3.56 \sin\frac{2\pi}{36}t + \varepsilon_2(t) \qquad (t=1,2,\dots,2880)$$

where $\varepsilon_1(t), \varepsilon_2(t)$ are some series of noise, and their standard deviations are $D(\varepsilon_1) = 2.63$, $D(\varepsilon_2) = 2.74$.

2. The computation of the coefficients of 0.5 - 1 - 1.5 .... 24,5 - year periodical components, i.e. the solution of the following minimum problem by the iterative method:

$$\min_{A_0, A_i, B_i} \sum_{t=1}^{2880} \left(x(t) - A_0 - \sum_{i=1}^{50} (A_i \cos\omega_i t + B_i \sin\omega_i t) \right)^2$$

where $\omega_i = \frac{2\pi}{18 \cdot i}$ $\qquad i = 1, 2, \dots, 50$.

In this case the periodical components, except the one-year ones, do not significantly differ from 0 too, and so

$$x(t) = 11.14 - 10.05 \cos\frac{2\pi}{36}t - 3.76 \sin\frac{2\pi}{36}t + \varepsilon_3(t) \qquad (t=1,2,\dots,2880)$$

where the standard deviation of $\varepsilon_3(t)$ is 2.67 , being in accordance with the former results.

References

Ajzerman, M.A. - Braverman, E.M. - Rozonoer, L.I. /1970/: Metod potencialnüh funkcij v teorii obucsennija masin. Izd. Nauka. Moszkva.

Anderson, T.W. /1971/: The statistical analysis of time series. John Wiley Sons. New York - London - Sydney - Toronto.

Priestley, M.B. /1981/: Spectral analysis and time series. Vol. 1, 2. Academic Press. London - New York - Paris.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE USE OF CLUSTER ANALYSIS AND ROTATED EMPIRICAL ORTHOGONAL FUNCTION /EOF/ ANALYSIS IN DESCRIBING THE MACROCIRCULATION PATTERN OF THE NORTHERN HEMISPHERE AND THE ATLANTIC-EUROPEAN REGION ON DIFFERENT HEIGHTS

Judit Bartholy, Central Met. Institute, Budapest, Hungary
Anthony G. Barnston, Climate Analysis Center, NMC/NWS/NOAA, Washington,D.C. USA
Robert E. Livezey, Goddard Lab. for Atm., NASA, Greenbelt, Md., USA

Introduction

As early as in the 1950s and 1960s climatological and macrosynoptical classifications were made by several researchers for example Dzerdzejevskij /1946/, Vangengejm /Bolotinskaja - 1964/, Hess and Brezowsky /1969/, Lamb /1972/ etc., with different aims and for different geographical regions. In each case different criteria were taken as bases of classification /the geographical position of cyclones and anticyclones, the direction of ridges, etc./. A common feature of these early classifications was that both creating the classes and arranging the phenomena were subjective processes and were carried out with the help of the human eye and synoptical practice.

With the widespread use of high-capacity computers more objective methods have become possible. A new branch of mathematics, cluster analysis deals with the problems of classification algorithms. Some good summaries on this topic are given by Anderberg /1973/, Späth /1980/, Juhász /1980/, Romesburg /1984/. Several attempts have been made in the recent past for meteorological application of clustering procedures: Sulochana /1980/ made a precipitation classification for the region of India, Hartman /1984/ classified the tropical cloud configurations, Panagiotis /1984/ gave a classification of weather situations in Greece, and Ronberg /1985/ in Israel. Kruizinga /1979/ and Marvon /1985/ made typisation on the basis of 500 mb height fields.

In the last years we have also applied several kinds of clustering techniques in our researches and created objective macrocirculation systems for large regions. The aim of this work was to eliminate the subjectively coded Hess-Brezowsky macrosynoptic system from the analogue forecasting model used by the Central Forecasting Institute of the Meteorological Service of Hungary. In our first work /Ambrózy-Bartholy-Gulyás /1983// we made a system for the Atlantic-European region. With the increased validity period of the forecasts however, it became necessary to make a hemispheric scale system. So, later on, in our experiments we made attempts of hemispherical clustering of the 500 and 700 mb height levels by means of different classification algorithms. Probably due to the high number of dimensions, these classes were not separated from each other sufficiently. Therefore, prior to the use of the classification algorithms a procedure for features extraction and for data reduction was carried out on the entire data base: the rotated empirical orthogonal function analysis. This part of the work was done at the National Meteorological Center /NMC/ Climate Analysis Center /CAC/ in Washington D.C., USA, jointly with Anthony G. Barnston and Robert E. Livezey /NASA, Goddard Laboratory for Atmospheres/.

Attempts for clustering

Since the meteorological application of clustering is not yet very widely used, I'm going to give a brief summary of the principles of the procedures we used and their place among the methods of clustering. Cluster analysis, as a rule, has a dual aim: first, to explore the structure of the set of objects; second, to select the separate objects in such a way that the similar ones get into the same group, while the differing ones are placed in different classes.

The cluster analysis methods can be devided into two main types: the hierarchical and the non-hierarchical methods. The methods included in the first group constitute a hierarchical system of clusters in such a way that any two clusters are either disjoint or one of them implies the other. The hierarchal methods belong either to the agglomerative or the divisive type, depending on whether they reach the hierarchical system of clusters by unification or by division. Non-hierarchical methods can also be divided into two parts, namely, the overlapping and the disjoint classifications, depending on whether the grouping allows overlappings between the cluster or not.

Another important component of cluster analysis, the connection function, is also of two types: similarity function or non-uniformity function, depending on whether they take their maximum values in case of similarity or non-uniformity of the objects. In the past years in our researches hierarchal agglomerative and non-hierarchal disjoint methods were used, and in every case the classification was performed with the use of a connection function of the similarity type.

Atlantic-European region

In our first clustering attempt based on the daily 500 mb height fields, we constructed a seasonal system for the Atlantic-European region. Here the separate types represented the macrocirculational situations most characteristic of the season. For determining the classes the dynamical "k-means" method worked out by McQueen /1967/ was used which belongs to the non-hierarchical disjoint methods. In the algorithm the computation was carried out with similarity connection functions according to Euclidean metrics. In the iteratively approximating version of the method, the stability of the system during the separate iterations was measured by the number of objects regrouped in a new type /non-identical with that of the previous iteration/. Figure 1 illustrates the pressure maps of AT_{500}, $RT_{500/1000}$ and of surface level of the 9^{th} spring type. The first one was obtained as a result of clustering and the latter two were computed on the basis of the archive of the full time-series. The seasonal cluster system contains 19 spring, 8 summer, 15 autumn and 17 winter types. To measure the effectiveness of the types we considered the external and internal distances of the systems which, by definition, represent the distances between the class centers and the internal radius of the various classes respectively. Internal radius: the average distance of all the fields included in the given type, from the cluster center.

With this classification we succeeded in reducing the internal distances by 42 per cent, and in increasing the external distances in average by 60-70 per cent in comparison with the Hess-Brezowsky macrosynoptic system. This means a better filling and spanning of the physically given 80-dimensional space /the Atlantic-European region is represented by 80 grid point values/. The good results of this systematization were confirmed by the comparative verification results obtained after its insertion in to the long-range forecasting model.

The Northern Hemisphere

In our further experiments the investigated geographic region was expanded to the northern hemisphere; for technical reasons we switched over to the 700 mb level /on 500 mb there was a significant lack of data for this region in the data set at our disposal/; and we chose a larger time scale: the fields of decade averages instead of daily fields. Like Folland /1985/ and Maryon /1985/ in their classifications for the Atlantic-European region, we also made our hemispherical cluster system for 6 natural seasons /Jan.-Febr., March.-Apr. etc./. The low sample size does not cause a problem: as running means of decade fields were used in our researches, there was no reduction of the data set compared to the daily field set; only a smoothing was applied. The analysis was carried out on the data series of a 35 year period /1950-1984/ of the NMC, Washington D.C.,

Two attempts were made for clustering the running decade fields. In one case the hierarchical agglomerative, in the other case the non-hierarchical disjoint, non-dynamical "k-means" method was used, similarly to the investigation of the Atlantic-European region. In both cases 10 types were determined for each season. In the investigations each hemispheric running means of 700 mb decade field was characterized by 358 grid-point-values, each representing approximately equal territories. The results satisfied us moderately: The types of the class systems did not separate well, presumably due to the high dimension numbers.

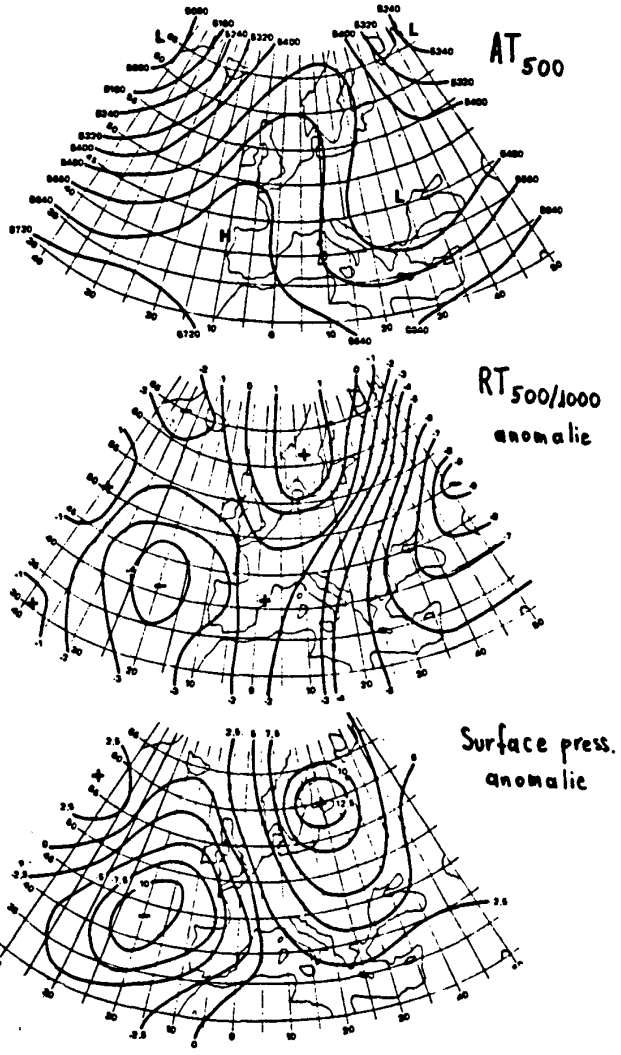


Fig.1: 9th class for spring season of the cluster system for the Atlantic-European region

We were confident that carrying out an empirical orthogonal function analysis /EOF/ prior to the implementation of the clustering procedures, the reduction in dimension number allowed by concentration of the information will give better results. A comparative evaluation of the cluster systems will be discussed later.

Joint application of the rotated EOF analysis and the clustering

For reasons described above we applied a combined method, in which clustering was done after having carried out the EOF analysis of the fields and varimax rotation of the amplitudes. The EOF analysis was first used by Lorenz /1956/ and Gilman /1957/. In Craddock's investigations /1969/ the hemispherical 500 mb fields are represented with the expansion coefficients of the fields expanded by eigen vectors of the correlation matrix and, the coefficients of the field are used as the data base of an analog forecasting procedure. So the procedure is applied as a feature extraction and data reduction method. The EOF analysis - besides its several advantages described by Wallace /1972/. Barnston-Livezey /1985/ etc. - was used by us mainly in the Craddockean sense. In the procedure the eigenvalue equation of the correlation matrix of the standardized data set is solved. The obtained eigenvectors are orthogonal, but their physical interpretation is difficult. Carrying out the varimax rotation analysed in detail among others by Horel /1981/ and Harman /1960/, we obtain results that can be interpreted better physically - at a price of a little deterioration of orthogonality. Inside the separate modes rotation maximalizes variance and eliminates smoothing caused by averaging and so it represents another feature extraction procedure.

After the EOF analysis and the rotation every field of every season is characterized by 10 coefficients and, after carrying out our iterative clustering procedure on these data as data base, 10 types are obtained for each season. In the clustering procedure several attempts were made for selecting the initial cluster centers in order to accelerate the convergency of the method. Naturally, in order to obtain the final class centers, the inverse transformation of the rotation and the restoration of the 700 mb cluster centers have to be accomplished on the basis of the EOF coefficients and the eigenvectors. As an example, Figure 2 illustrates the 700 mb height and anomaly fields of the type No. 7 of the winter season /January-February/.

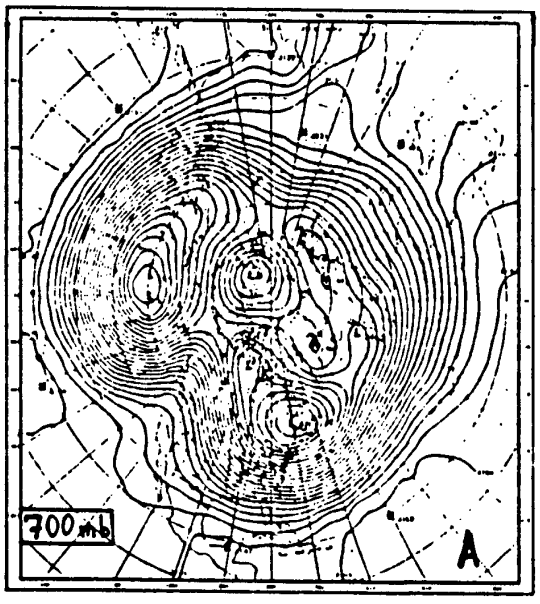


Fig. 2B and explanation
see next page

Comparative evaluation

During the 3 clustering experiments for the northern hemisphere the stability of classifications was checked separately for each method. The mathematical background of the methods ensures the convergence of the algorithms in the classification procedures used. The invariance of the procedures on the randomly chosen initial cluster centers was checked.

In these methods each iteration sweeps through the whole data set once and every fields placed in a class. So, counting the fields placed in a new class /differing from the previous inclusion/ a good index can be obtained for the stability of the cluster system. In table I. the final stability indices are demonstrated /for each method uniformly 12 iterations were made/. It can be seen that, while in the case of the clustering after the rotated EOF analysis, in the last iteration only 0.3 per cent of the fields were regrouped into a new class /altogether 7 fields/, in the case of the hierarchical method their number was almost 10 times higher.

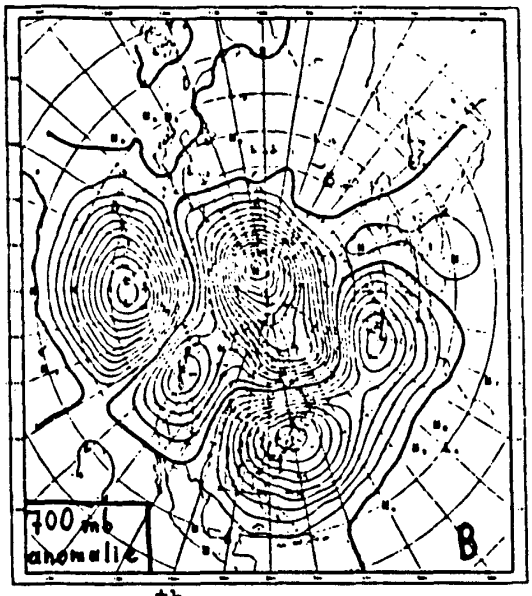


Fig.2: 7th class for winter season of the northern hemipherical cluster system /the method used rotated EOF analysis and dynamical cluster technics/

The external and internal distances of the three cluster systems of the hemispherical region were compared. These results are shown in table II. It can be seen that the procedure applying the EOF analysis compared to the hierarchical method; 1/ ensures a separation between the clusters by about 30 per cent better /increase of the external distances/; 2/ the different clusters are concentrated around the type centers by 22 per cent better /reduction of the internal radius of the classes/. These results are mean values, after averaging for 6 seasons.

The relatively little mean internal and great mean external distances can be evaluated positively if it is not accompanied by extreme frequency

	Hierarchical	Dynamical	EOF + Dynam.
	m e t h o d s		
Instability indices	2.5 %	1.3 %	0.3 %

I table - Instability indices for the 3 methods /the recoded field numbers by the last iteration given in percentage/

distribution for each type. But this condition is also fulfilled, because the amount of data included in the separate types is not less than 6 and not more than 17 per cent of the whole data set. The life-time of the separate types is in average two weeks, but, naturally, it greatly varies depending on the seasons and types.

It was demonstrated that out of the 3 classification methods, the system with clustering after a rotated EOF analysis is the best one in all aspects, therefore we are going to use this in our long-range forecasting analog method. Insertion into the model and the verification work is being done.

The analysing work has unfortunately not been accomplished yet for technical reasons. Continuation of the analysis is intended, jointly with R. Livezey and A. Barnston.

We are planning to make a comparison of the described 3 cluster systems and the rotated EOF analysis of Livezey and Barnston /B. Livezey is giving an account on this topic in another session/. All the 4 investigations were carried out on the same data base and for the same region. It would be also desirable to make a cross-checking with previous teleconnection-analyses of other authors.

	Hierarch. - Dynam.	Dynam. - EOF+dynam.	Hierarch. - EOF+dyn
	m e t h o d s		
Increasing of the external distances	11 %	18 %	29 %
Decreasing of the internal distances	8 %	14 %	22 %

II table - Separation indices between the classes and concentration indices inside the classes /changes of external and internal distances in perentage/, calculated for each pars of the 3 methods

REFERENCES

Ambrózy,P.-Bartholy,J.-Gulyás,O./1984/: A system of seasonal macrocirculation pattern for the Atlantic-European region, Időjárás, Vol. 88. No. 3. pp. 121-133.

Anderberg,M.R./1973/: Clusteranalisis for applications, Academic Press, New York

Barnston,A.G.-Livezey,R.E./1985/: High resolution rotated EOF analysis of northern hem. 700mb heights for predictive purposes, 9. Conf. on Prob. and Stat in Atm. Sci. Virg. Beach, Am.Met.Soc. pp. 290-298.

Bartholy,J.-Kaba,M./1985/: Further development of seasonal climate forecasts for the territory of Hungary, Időjárás, Vol. 89. No. 4. p. 185.

Craddock,J.M.-Flood,C.R./1969/: Eigenvectors for representing the 500mb geop. surface over the Northern Hem. Quart.J.R. Met.Soc. No. 405. p. 576.

Dzerdzeevskij,B.L./1946/: Typification of circulation mechanism in Northern Hem. and characteristics of synoptic seasons, Moscow, Gidrometeoizdat

Esbensen,S.K./1984/: A comparison of intermonthly and interannual telecon. in the 700 mb geopot. height field during the Northern Hem. winter, Mon.W.Rev.Vol.112.p.2016.

Folland,C.K.-Colman,A./1985/: A multivariate tech.for use in long-rane forecasting, Proc. 1 WMO Conf.Diag.Pred. Monthly-Seasonal Atm. Variations over Globe, Maryland

Gilman,D.L./1957/:EOF-s applied to thirty-day forecasting, Sci.Rep. No.1. Contract AF19 /604/-1283, Dept. of Met., M.I.T. Cambridga, Mass.

Harman,H.H./1960/: Modern factor analysis, The University of Chicago Press, London

Hartman,D.L./1984/: Some implication of the mesoscale circ.in trop. cloud clusters for large-scale dyn. and clim. J. of the Atm.Sci., Vol.41. No.1. pp. 113-121.

Hess,P.-Brezowsky,H./1969/: Katalog der Grosswetterlagen Europas, Ber. Wetterd. Offenbach, No. 113.

Horel,J.D./1981/: A rot. principal comp. anal. of the interannual variability of the Northern Hem. 500mb height field, Mon.W.Rev. Vol.109. pp. 2080-2092.

Juhász,F.-Mályusz,K./1980/: Problems of cluster anal. from the viewpoint of numerical anal.,Colloquia Math.Soc.János Bólyai 22. Numerical Methods, Keszthely, p. 405.

Kruizinga,S./1979/: Objective classification of daily 500mb patterns, Bull. of six Conf. on Prob. and Stat. in Atmospheric Science of the Amer. Met. Soc.

Lamb,H.H./1972/: British Isles weather types and a register of the daily sequence of circulation patterns, Geophys. Mem., London 16. No. 116.

Lorenz,E.N./1956/: EOF-s and statistical weather prediction, Sci.Rep. No.1. Stat. Forecasting Project, Dept. of Met., M.I.T. Cambridge, Mass.

MacQueen,J./1967/: Some methods for classification and analysis of multivariate observations, Proc. of Fifth Berkeley Symp. on Prob. and Stat.

Maryon,R.H.-Storey,A.M./1985/: A multivariate statistical model for forecasting anomalies on half-monthly mean surf. pressure, Jour.of Clim. Vol.5. pp.561-578.

Panagiotis,M./1984/: Weather-type classification by factor anal. Jour.of Clim.Vol.4.p437

Romesburg,H.C./1984/: Cluster anal. for researchers, Li.Le.Publ., Belmount, California

Ronberg,B./1985/: An objectiv weather typing system for Israel, Ninth Conf. on Prob. and Stat. in Atm. Sci., Virginia Beach

Späth,H./1980/: Cluster analysis algorithms for data reduction and classification of objects, John Wiley and Sons, New York

Sulochana,G./1980/: Cluster analysis of rainfall stations of the Indian peninsula, Quart.J.R.Met.Soc. 106. pp. 873-886.

Wallace,J.M./1972/: Empirical Orthogonal representation of time series in the frequency domain, Jour.Appl.Met. Vol. 11. No. 6. pp. 887-900.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

COMPARISON OF PRINCIPAL COMPONENTS AND FACTOR ANALYSIS METHODS FOR CLIMATOLOGICAL PARAMETERS

DR. OSKAR M. ESSENWANGER
RD&E CENTER, ARMY MISSILE COMMAND

ABSTRACT:

Four different methods for estimating the communalities in factor analysis of climatological elements were studied: principal components, least squares and maximum likelihood. Although these methods lead to different estimators, the factors exhibit similarity after rotation and provide proper climatological characteristics. Details and peculiarities of these methods are discussed.

INTRODUCTION

Rapid data processing on modern computers has aided the climatologist in the utilization of the most sophisticated statistical techniques. This trend continues especially by the availability of "canned" programs. Although these "canned packages" have simplified the technical task, a thorough understanding of the methodology is still necessary; take e.g. the routine on factor analysis. Here the user may have to make a choice between four different methods, all based on specific and sound statistical principles. Since the primary concern of the climatologist should be the disclosure of the climatological characteristics the choice between individual methods may be puzzling at first. Thus an attempt was made to delineate some of the statistical background and study the outcome of the factor analysis in terms of the climatological characteristics. It can be demonstrated that four different statistical approaches lead to differences in the unrotated factor loads. However, after rotation the differences in the climatological characteristics of the factors were small in the orthogonal rotation.

STATISTICAL BACKGROUND

Factor analysis was originally developed in behavioral science (e.g. Spearman, 1904, 1927; Cattell, 1952, 1965) and has not been utilized very often in climatology. Only in the last two decades some articles can be found (e.g. Christensen and Bryson, 1966, Kutzbach, 1967 or Buell, 1971).

The basic model of factor analysis can be formulated as follows:

$$M_X = M_A M_F + M_E \qquad (1)$$

where M_X is a data matrix (the only known matrix in eqn 1), M_A is a coefficient matrix of factors, M_F the factor matrix, and M_E an error matrix. M_A is also called the factor loading matrix or factor pattern. It is postulated that neither the factors nor the factors and errors are correlated. Then eqn 1 has the following mathematical solution:

$$M_X = M_A \Phi M^T_A + (\Psi) \qquad (2)$$

where $\Phi = M^T_F M_F$ is the factor covariance matrix and Ψ is a diagonal matrix related to errors $\Psi = M_D M_E$.

In its standard form M_X is a correlation matrix with unity in its diagonal. This system is called "closed" or "principal components analysis." The true factor analysis is based on the assumption that not all factors are known a priori. Mathematically this fact is accounted for by reducing the diagonal of the matrix M_X. These diagonal elements are called "communalities." More details can be found in the literature such as Essenwanger (1976).

Guttman (1956) has suggested the substitution of the diagonal communalities by the multiple correlation coefficient and Cattell (1952) has recommended other correlation criteria (see also Essenwanger, 1976). In recent times statistical principles have been proposed. In the unweighted least squares (ULSQ) method U is a minimum for

$$U = (1/2)\ \mathrm{tr}\ (M_S - M_X)^2 \qquad (3)$$

where M_S is the correlation matrix with estimators in the diagonal (communalities) and tr means the trace. The generalized least square (GLSQ) method requires a minimum for

$$G = (1/2)\ \mathrm{tr}\ (I_n - M_S^{-1} M_X)^2 \qquad (4)$$

where I_n denotes a diagonal matrix of unity, M_S and M_X see previously defined. The maximum likelihood principle (MXLI) minimizes:

$$M = \mathrm{tr}\,(M_X^{-1} M_S) - \ln \left| M_X^{-1} M_S \right| - n \qquad (5)$$

see Jöreskog (1967), where n is the number of variables, | | the determinant.

	1	2	3	4	5	6	7	8	9
1 CEIL	1.0	.07	-.18	-.16	-.32	-.37	-.20	.26	-.80
2 Ln VIS		1.0	.43	.47	.33	.15	-.48	-.20	-.19
3 WD			1.0	.51	.43	.37	-.13	-.18	.11
4 WS				1.0	.38	.27	-.27	-.22	.06
5 TEMP					1.0	.93	-.05	-.18	.20
6 DEWP						1.0	.31	-.15	.28
7 REHU							1.0	.07	.26
8 PRES								1.0	-.17
9 CL. AMT									1.0

CEIL = Ceiling, Ln VIS = Ln Visibility, WD = Wind Direction, WS = Windspeed
TEMP = Dry Bulb Temperature, DEWP = Dewpoint Temperature
REHU = Relative Humidity, PRES = Station Pressure, CL.AMT = Cloud Amount (Skycover)

TABLE 1. CORRELATION MATRIX (JANUARY, STUTTGART)

	UNROTATED				ORTHOGONAL ROT.			
	1	2	3	4	1	2	3	4
CEIL	-.58	-.60	.41	.18	-.02	-.20	-.91	-.14
Ln VIS	.46	-.70	-.10	-.01	.79	.16	-.20	.16
WD	.67	-.30	.00	-.19	.62	.41	.16	-.01
WS	.63	-.42	-.16	-.19	.73	.25	.15	.06
TEMP	.85	-.02	.42	.07	.29	.89	.11	.07
DEWP	.80	.24	.52	.13	.02	.97	.18	.06
REHU	-.03	.73	.36	.18	-.73	.35	.21	-.04
PRES	-.40	.03	.49	-.76	-.13	-.08	-.14	-.97
CL. AMT	.44	.68	-.42	-.36	-.08	-.10	.94	.04
VAR	3.09	2.17	1.20	.81	2.18	2.17	1.91	1.00

TABLE 2. FACTOR LOADS (JANUARY, STUTTGART)

UNROTATED AND ROTATED FACTORS

Table 1 provides an example of a correlation matrix with nine climatological elements for January at Stuttgart (Fed Rep. Germany) for the period 1946-1953. It is apparent that a simple interprepation of the interrelationship between elements is difficult based on the correlation matrix alone. (The correlation with the wind direction has been computed as described in Essenwanger, 1964).

Table 2 shows the results of a factor analysis with the principal components method (PR.CO) using the entry M_X from Table 1. The first four factors comprising 81% of the variance are included (truncated principal components analysis). It is apparent that the first factor is a mixture of several elements although the temperature and dewpoint dominate. This unrotated solution is a valid set of factors but the factors are related to a mixture of elements.

The right side of Table 2 displays a different set of factors after orthogonal rotation. Now the relationship to the elements is simplified. Although the first factor exhibits still a mixture related to wind, visibility, and relative humidity, the second factor provides the temperature effect of the system. The third factor is associated with clouds (ceiling and cloud amount) and the last factor renders the effects of pressure. Earlier Buell (1971) had thought that rotation may not be necessary for climatological elements. This study supports, however, that in a complex system of elements a rotation of the original factors is beneficial.

Table 3 serves two goals. It demonstrates the differences between four methods of estimating communalities as well as orthogonal and oblique rotation. In order to enhance the perception of the factors, the signs and correlations of less than 0.4 have been omitted from the table except where necessary for continuity or maximum in a row. The top part shows the factor load after an orthogonal rotation. The four methods differ only in minor details if the sequence of factors is altered to match similar factors. We learn that ULSQ and PR.CO lead to the same factors differing only slightly in variance (VAR) and numerical values. ULSQ and GLSQ diverge in the last factor. The relationship to pressure plays an insignificant role in GLSQ. The same fact is apparent in the MXLI method.

The oblique rotation differs little from the orthogonal rotation for the first three methods. One peculiar fact emerges for the MXLI method. It can only be understood, however, if both structure matrix and factor pattern (Table 4) are inspected. The factor pattern contains the regression coefficients of the system in the oblique rotation. The product variance (PROD. VAR) is $\Sigma s_{ij}.p_{ij}$ where s_{ij} is the i-th row of the j- th column in the stucture matrix and p_{ij} the same in the factor pattern. The sum can be positive or negative. A close perusal of this PROD. VAR sum in Table 3 reveals that results are comparable to the VAR of the orthogonal rotation for the first three methods but differ for the MXLI method. In fact, the sum of the first three factors exceeds the sum of 6.31 of the orthogonal rotation case. Thus with the last factor the surplus is subtracted and the PROD. VAR is reduced to 6.29 which is in line with the top part. This is accomplished, however, by virtually duplicating the second factor. Since this singularity is not present for all months (not included here) it may be concluded that it is not a programming error. However, it could be a peculiarity how the oblique rotation is calculated for the maximum likelihood method.

Table 5 lists the communalities for the first method. Apparently the low value for the pressure element in the GLSQ and MXLI method may explain why the pressure is not significantly related with the factors in these two methods (see Table 3). It should be noted that the PR.CO method leads to the highest percentage approximation of the total variance.

Table 6 lists the individual elements in sequence of weight within the system of 9 elements. This weight is $x^2/9$, x being the element in the rotated factors (e.g. Table 3). The maximum can be 11.11%. No adjustment with reference to the total sum (Table 6, last line) has been made. A larger gap appears in the % figures between the fifth and sixth element (GSLQ 6th and 7th). Under these considerations ceiling, cloud amount, dewpoint and temperature appear in the group of five (six) elements while pressure is replaced by the relative humidity for GSLQ and MXLI. Thus the climatological features appear relatively robust. For more details see Essenwanger, 1986.

JANUARY, STUTTGART (UNIT 0.01)

ORTHOGONAL ROTATION

	PR.CO				ULSQ				GLSQ				MXLI			
CEIL		91				85					79				91	
Ln VIS	79				76					44		40	32			65
WD	62		41		52		33		16			98		23		65
WS	73				62							46				68
TEMP			89				86		94					93		
DEWP			97				100		94					93		
REHU	73				58					98		13	95			26
PRES				98				98			17					28
CL.AMT		94				91					99				86	
VAR	218	191	217	100	171	172	204	100	193	132	174	149	111	187	172	161
$\sum$VAR				726				647				648				631

OBLIQUE ROTATION

CEIL		93				87					80				90	
Ln VIS	81				78					47		43	50			
WD	66		44		55				28			99		39		40
WS	76				64							49				
TEMP			91				88		96					96		99
DEWP			99				100		97					99		94
REHU	69				56					99		15	99			
PRES				98				99			18				26	
CL.AMT		95				92					99				83	
PROD. VAR	223	190	214	100	173	171	202	99	198	133	171	146	349	937	157	-814
$\sum$				727				645				648				629

TABLE 3. COMPARISON OF ORTHOGONAL AND OBLIQUE ROTATION

		STRUCTURE MATRIX				FACTOR PATTERN			
1	CEIL	-.18	-.36	-.90	-.31	-.06	-.57	-.88	.45
2	Ln VIS	-.50	.19	-.03	.30	-2.39	7.21	-.13	-6.87
3	WD	-.16	.39	.18	.40	-2.26	8.09	.06	-7.64
4	WS	-.29	.30	.19	.34	-2.41	8.10	.08	-7.73
5	TEMP	-.13	.96	.26	.99	-.25	.86	.00	.14
6	DEWP	.23	.99	.24	.94	.08	.98	.00	.00
7	REHU	.99	.23	-.01	-.03	.98	.08	.00	.00
8	PRES	.08	-.16	-.26	-.16	.95	-3.35	.23	3.22
9	CL.AMT	.23	.26	.83	.20	.46	-.74	.83	.75
$\sum s_{ij}p_{ij}$						3.49	9.37	1.57	-8.14

TABLE 4. STRUCTURE MATRIX AND FACTOR PATTERN FOR MXLI, OBLIQUE ROTATION

VARIATE	PR.CO	ULSQ	GLSQ	MXLI
1 CEIL	.89	.77	.67	.87
2 Ln VIS	.71	.63	.42	.56
3 WD	.58	.40	1.00	.48
4 WS	.62	.46	.33	.50
5 TEMP	.89	.85	1.00	1.00
6 DEWP	.98	1.00	1.00	1.00
7 REHU	.70	.43	1.00	1.00
8 PRES	.99	1.00	.07	.13
9 CL.AMT	.90	.86	1.00	.76

TABLE 5. COMMUNALITIES (JANUARY, STUTTGART)

	PR. CO		ULSQ		GLSQ		MXLT	
1	DEWP	10.54	DEWP	11.11	CL. AMT	10.98	CEIL	10.05
2	PRES	10.49	CL. AMT	10.95	WD	10.67	REHU	9.97
3	CL. AMT	9.82	PRES	10.13	REHU	10.60	TEMP	9.59
4	CEIL	9.21	TEMP	8.15	DEWP	10.40	DEWP	9.52
5	TEMP	8.79	CEIL	7.96	TEMP	9.64	CL. AMT	8.21
6	Ln VIS	6.87	Ln VIS	6.41	CEIL	8.74	WS	5.06
7	WS	6.00	WS	4.38	WS	2.36	WD	4.73
8	REHU	5.86	REHU	3.78	Ln VIS	2.18	Ln VIS	4.67
9	WD	4.31	WD	3.00	PRES	0.34	PRES	0.90
Σ		71.89		65.87		65.91		62.70

TABLE 6. MAXIM. WEIGHT IN FACTORS (ORTHO ROT, UNIT %)

SEASON TREND

A quick look at the variation of the four factors during the year may be of interest. This fluctuation is depicted in Table 7. For this purpose a 10 year period (1946-1956) at Stuttgart was chosen. Furthermore, the cloud amount was omitted because in the 9 elements analysis for 7 years the cloud amount was always combined with the ceiling. Since the PR.CO method provided the highest approximation of the variance the result for this method is chosen here. Again, for better perception correlations less than 0.4 were omitted.

Examination of all months renders that the pressure always forms an individual factor. The wind components and temperature elements fall into one individual factor each. Spring and summer however, add the temperature to another factor of mixtures, so is the windspeed in October. The similarity of April and July is noteworthy. As expected, the weights of the individual element fluctuate during the season.

	JAN				APRIL				JULY				OCT			
CEIL			88		43	61			44	58					77	
Ln VIS	78				78				71					54	65	
WD	68					80				80				82		
WS	80					79				80			80	84		
TEMP		92			64		75		63		75		93			
DEWP		97					92				95		98			
REHU	63		47		92				91				84			
PRES				99				96				99				93
VAR	221	205	125	100	212	170	151	100	204	165	150	101	259	183	110	110

TABLE 7. PR.CO FACTOR LOAD, ORTHOG. ROT., STUTTGART (1946-1956), UNIT 0.01

CONCLUSION

Four different methods leading to estimators for the "communalities" in factor analysis were studied. These four methods are based on statistical principles such as least squares, maximum likelihood and principal components analysis.

It was expected that the communalities for the four individual methods differed. Consequently one finds dissimilarity in the original "factor loads." However, after rotation the four factors disclosed a meaningful climatological pattern, although the weights of these factors differed. All four methods had in common that the two temperature elements appeared together in one factor and the wind components in another one. Pressure was only a significant factor in the PR.CO and ULSQ method. The elements cloud ceiling, visibility and relative humidity fluctuated in assignment for the individual methods and during the seasons. Furthermore, the weight of the factors within the system also displayed some variations.

For the data sample on which this analysis was based the orthogonal and oblique rotation was almost alike except for the maximum likelihood method. The (truncated) principal components analysis disclosed the highest percentage approximation of the total variance. If the goal is a maximum of representation this method and an orthogonal rotation is certainly the simplest mathematical scheme. From the viewpoint of climatological characteristic the individual methods differ little but the oblique rotation in the "canned program" for the maximum likelihood method showed some peculiarities with two of the oblique factors highly correlated with each other. This may not be the most desirable result from the climatological point of view.

REFERENCES

Buell, C. E., 1971. Integral Equations Representation for Factor Analysis. J. Atmosph. Sci., 28, 1502-1505.

Cattell, R. B., 1952. Factor Analysis. Harper, New York, pp. 462.

Cattell, R. B., 1965. The Configurative Method for Surer Identification of Personality Dimensions, Notably in Child Study Psych. Rep. 16, 269-270.

Christensen, W. I. Jr. and R. A. Bryson, 1966. An Investigation of Component Analysis and Weather Classification. Mon. Weather Rev, 94, 697-709.

Essenwanger, O. M., 1964. The Cumulative Distribution for Wind Direction Frequencies. Meteorol. Rundsch. 17, 131-134.

Essenwanger, O. M., 1976. Applied Statistics in Atmospheric Science Elsevier, Amsterdam, pp. 412.

Essenwanger, O. M., 1986. A Comparison of Methods for Factor Analysis of Visibility. ARO Report 86-X, Proceedings of the Thirty-first Conference on the Design of Experiments in Army Research, Development and Testing.

Guttman, L., 1956. "Best Possible" Systematic Estimates of Communalities Psychrometrika, 21, 273-285.

Jöreskog, K. G., 1967. Some Contributions to Maximum Likelihood Factor Analysis Psychometrika, 32, 443-482.

Kutzbach, J. E., 1967. Empirical Eigenvectors of Sea-Level Pressure Surface Temperature and Precipitation Complexes Over North America, J. Appl. Meteorol., 6, 791-802.

Spearman, C., 1904. General Intelligence Objectively Determined and Measured. Am. J. Psychol, 15, 201-293.

Spearman, C., 1927. The Abilities of Man. Mac Millan, London, pp. 87.

ACKNOWLEDGEMENT

The author's thanks go to Dr. Dorathy A. Stewart for her critical review of the manuscript. Mr. Roger Betts deserves the credit for the preparation of the computer programs to obtain the data. Last, not least, Mrs. Gloria McCrary must be thanked for her patience and diligence during the process of establishing the manuscript from the first draft to the final version.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE USE OF THREE-MODE FACTOR ANALYSIS TO DETECT JOINT SPACE AND TIME VARIABILITY OF DROUGHT PATTERNS

Paul A. Kay, Department of Geography, University of Utah
Salt Lake City, Utah 84112, U.S.A.

INTRODUCTION

In climatological studies the data cube has axes or modes of space (the station network), field (the variables of interest), and time (the length of record). In the study of spatial distributions of fields, forms of eigenvector analysis are often used to elucidate dominant patterns of variability. The patterns thus identified are time-invariant. It may be asked, however, if this temporal stability of spatial pattern is really the case. It is also possible to slice the data cube in such a way as to define temporal patterns in space-invariant mode. Standard models enforce assumption of temporal or spatial stability that in fact may not be the case. The purpose of this paper is to introduce to climatology a factor analytical technique that exposes the interaction of temporal and spatial variability. In the way of illustration, the technique is applied to a question of drought patterns in a portion of the western United States.

TWO-MODE FACTOR MODELS

There exists a large literature of applications of eigenvector techniques to meterological and climatological questions. For example, Richman (1983) reports more than 60 in the first third of the 1980s. In the following discussion, the general term "factor" is used *sensu lato* to refer to both components and factors.

Common to all applications has been the partitioning of the data cube into two-mode matrices, holding the third axis constant; this yields six possible analytical strategies, four of which are available to the study of spatial pattern. In Q-mode, correlations of fields varying across space are analyzed for a given time, and the factor scores are mapped to depict regions defined by common variability of the fields. R-mode is similar, except it is the correlations of the stations that are analyzed and the loadings of the stations onto the factors that are mapped. In S-mode, correlations of temporal series for the stations are analyzed for a given field, and the factor scores for the stations are mapped to depict regions of common variability. T-mode is similar, expect it is the correlations of the stations that are analyzed and the loadings of the stations onto the factores that are mapped. In R- and T-modes, the factor scores are temporal series of correlations of observations with factor patterns. Those patterns themselves are invariant; it is the contribution of the fixed patterns to an observation that is measured by the scores. Thus, the researcher is limited to investigation of only one field at a time or of several fields at only one time. The techniques of two-mode analyses are incapable of examining temporal evolution of patterns defined by spatial interaction of multiple fields, without having an imponderable number of maps.

Recent expansions of the classical techniques have sought to allow the analysis of variability jointly in space and time. Weare and Nasstrom (1982) used lagged covariance matrices to introduce time slices. This techinque is practically limited in the number of time steps that can be efficiently sampled. Horel (1984) used complex principal components analysis to identify propagating features as well as standing waves. The interpretation of the relationships of phase and amplitude is not easy, and the technique would seem to have limited applicability to fields whose temporal change are not clearly amenable to cyclical representation.

THREE-MODE ANALYSIS

Three-mode factor analysis was developed in psychology for problems of measuring change, for example, interpretation of the effectiveness of learning strategies from a matrix of scores on a number of tests administered to a number of students at a number of times (Tucker, 1966). A few applications in economic and urban geography to study the temporal evolution of spatial patterns (Cant 1971; Langlois 1983) suggest that the technique may have much to offer in climatological research where questions of a similar nature occur.

For the data matrix X_{ijk}, where i is the dimension of the space mode, j is the dimension of the field mode, and k is the dimension of the time mode, the fundamental model is

$$x_{ijk} = \Sigma_m \Sigma_p \Sigma_q a_{im} b_{jp} c_{kq} g_{mpq}$$

where x_{ijk} is an element of the data matrix, a_{im}, b_{jp}, and c_{kq} are elements of the factor matrices for the three modes, and g_{mpq} is an element of the core matrix. This equation can be rewritten into a set of equations, one for each mode, using combined-mode matrices

$$_iX_{(jk)} = {}_iA_m G_{(pq)} ({}_pB_j \text{@} {}_qC_k)$$

and similarly for modes j and k using combined modes (ik) and (ij), respectively; the symbol @ signifies the Kroenecker matrix product. Eigenvalues and the eigenvectors in the matrix ${}_iA_m$ may be determined in the usual way (such as principal components analysis), retaining only a few of the principal roots (m<i) by some usual selectrion criteria; and similarly for ${}_jB_p$ and ${}_kC_q$. The number of factors extracted in each mode need not be identical. The core matrix is calculated from an equation of the form

$$_mG_{(pq)} = {}_mA_i X_{(jk)} ({}_jB_p \text{@} {}_kC_q)$$

or similarly for combined modes (mp) or (mq).

The factor matrices A, B, and C may be interpreted in ways similar to that in conventional two mode analyses (Figure 1). In A are the factors grouping the stations according to their similar variability in time and with respect to the fields in the analysis; the factors represent idealized regions. In B are the factors grouping the variables according to their similar variability through time and across space; the factors represent idealized "climate types". In C are the factors grouping the times according to their similar variability across space with respect to the variables; the factors represent idealized episodic behavior. The core matrix G contains scores of each factor from one mode on each factor of a second mode for each factor in the third mode. That is, it displays the interrelationships between the factors on each of the three modes. If the scores for a spatial mode factor with respect to the temporal mode factors are identical, it may be interpreted that that spatial pattern is invariant across different time scales.

APPLICATION

Drought has been a recurrent, approximately periodic, phenomenon (Mitchell *et al.*, 1979) in the United States, but each occurence has differed in intensity, duration and location (Karl and Koscielny 1982; Karl 1983). The present study was part of a project investigating climate variability in Utah, and so is only a small subset of the national pattern. However, the area is of some interest as an example of variability in space as the boundaries between three regions of drought experience lie in the study area (Karl and Koscielny, 1982). The 1930s drought, for example, was very pronounced in the West North Central region, containing northern Utah, but was absent in the South West, including southern Utah. Karl and Koscielny (1982) also indicated that durations were different between those two regions and the West, but they reported no significant periodicities in any region.

In this study, the data are Palmer Drought Severity Indices, a measure of meteorological drought reported for climatologic divisions (see Karl, 1983, for discussion). The spatial mode of the data cube comprises the sixteen climatic divisions in and adjoining Utah. The field mode comprises the twelve months of the year treated individually. The temporal mode comprises the fifty-one years 1931-1981. The results of the analysis include regions, seasonal patterns of drought intensity occurence, and episodic behavior of drought occurence. Only the first few factors in each mode are reported here; they are analogous to unrotated principal components.

The first two regions show general, but not absolute, spatial coherence (Figure 2). Pattern I emphasizes a contrast between the northwest and the central parts of the region. Pattern II, and others not shown here, exhibit a north-south differentiation. The patterns correspond approximately to the above-mentionned regionalization. The patterns no doubt would be clarified if rotation to simple structure were applied.

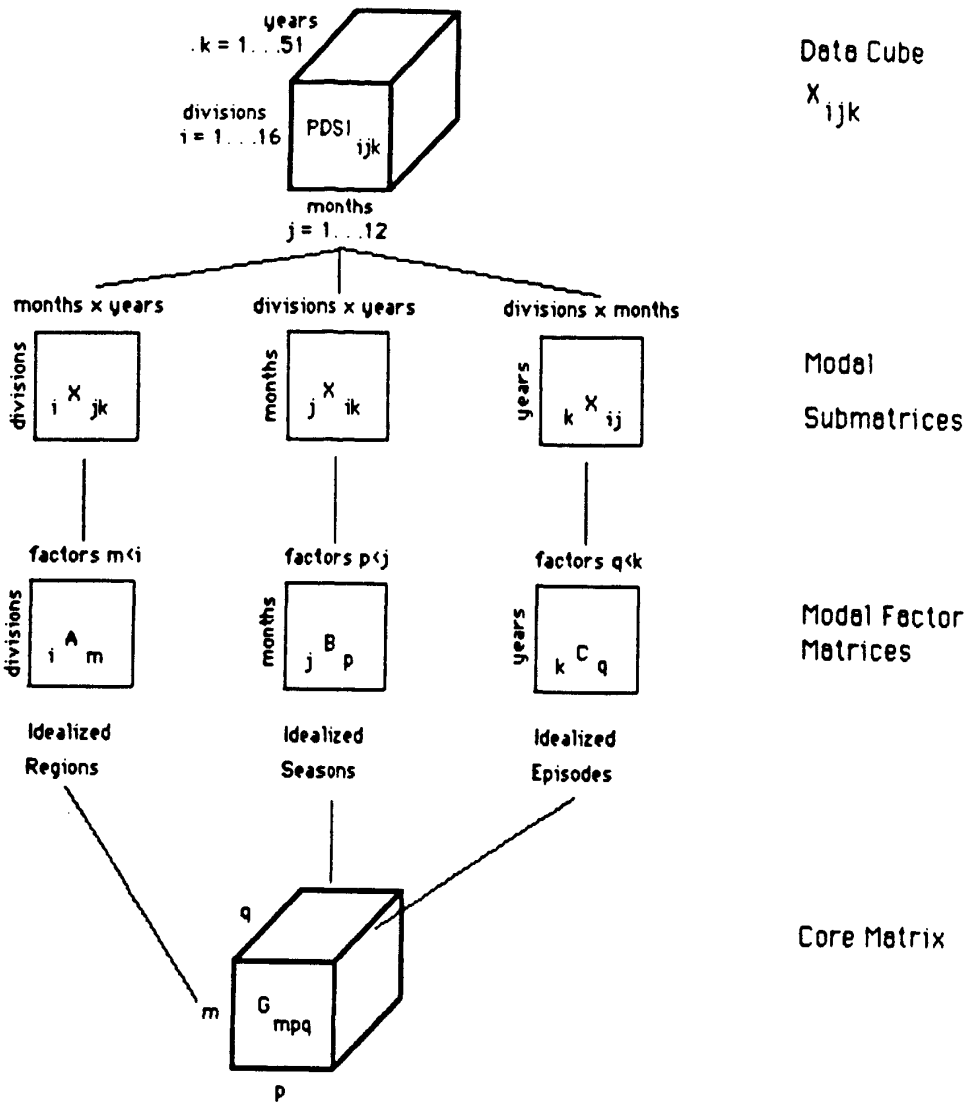


FIGURE 1. Strategy of three-mode factor analysis.

The first three seasonal regimes are shown in Figure 3. Regime I represents rather uniform drought intensity all year, with a slight increase in summer. Regime II exhibits seasonal variability, with contrasting signs between winter months and spring and autumn months, and also contrasting signs within summer. Regime III shows a semi-annual contrast, suggesting that drought is set or broken more in winter than summer, which in any event is dry and therefore a time of moisture stress in the study area.

The first three episodic patterns are shown in Figure 4. Pattern I emphasizes a quasi-periodicity of 20 to 30 years for drought. Pattern II emphasizes higher frequency variation on the order of one to two years. Pattern III displays very long-term persistence.

The largest value in the core matrix (Table 1) is for the first factor from each of the three modes. This suggests that the first spatial pattern is associated with year-long drought pattern tending to occur quasi-periodically, with occurences in the 1930s, 1950s-mid 1960s, and briefly in the late 1970s. The inverse pattern appears when long-term persistence is highlighted, but does not necessarily appear when other seasonal regime or episodic patterns are considered. Most of the scores in the core matrix are low, however, suggesting that the system is dominated by the concurrence of the first factors.

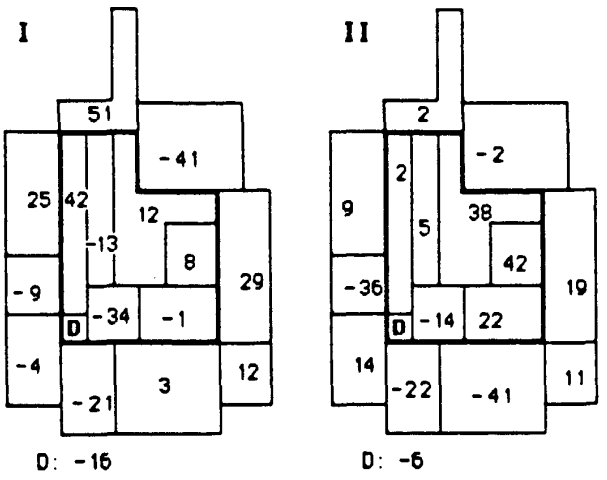


Figure 2. Schematic representation of the first two regional patterns.

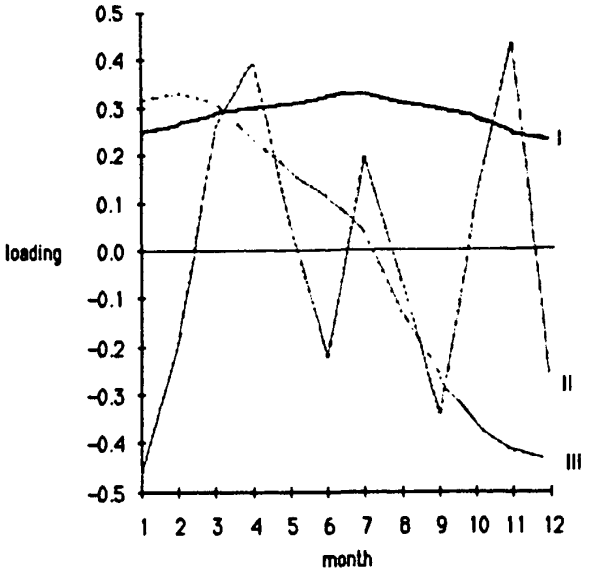


Figure 3. The first three seasonal regime patterns.

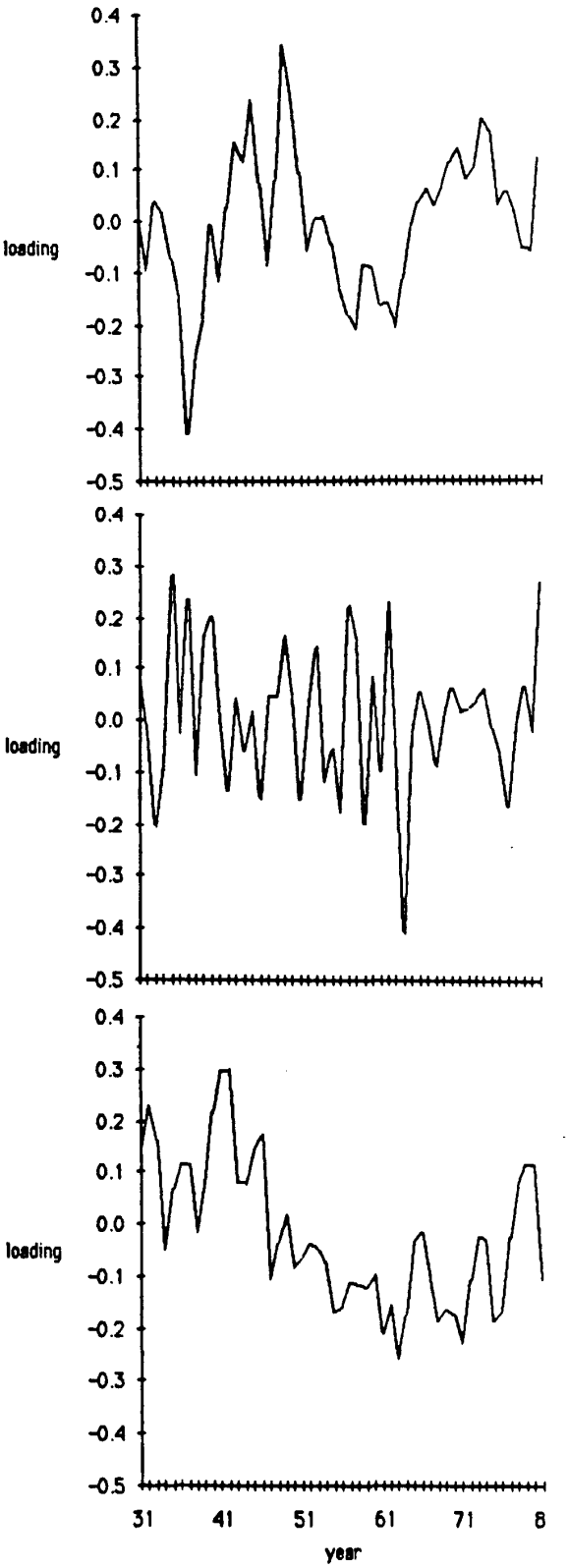


Figure 4. The first three episodic patterns.

		episode I	episode II	episode III
region I	regime I	19.14	-0.73	-3.16
	regime II	1.64	0.12	0.43
	regime III	-3.60	1.17	0.11
region II	regime I	-1.66	2.26	3.06
	regime II	-0.06	0.33	0.09
	regime III	0.93	-0.62	-0.65

TABLE 1. Core matrix, G, for the first two regional patterns, the first three seasonal regimes, and the first three episodic patterns.

CONCLUSIONS

The technique presented in this paper is subject to all the usual concerns of eigenvector techniques, including choice of factoring strategies, number of factors to retain, and choice of rotation to simple structure. A useful discussion of these matters may be found in Richman (1983). In the solution presented above, as in unrotated principal components, the first factors contain most of the variance but may lie between natural patterns. Some rotational strategy would probably clarify the patterns. The example is also constrained by the choice of the PDSI variables, which by definition have strong intermonthly persistence, and by the choice of climatic divisions, which may not really be homogenous areas but certainly are of different sizes containing different numbers of stations.

Despite these cautions, the three-mode factor analysis does appear to have considerable potential for application in climatological studies. Its main contribution is the ability to consider both spatial and temporal variability jointly, rather than artificially holding one or the other constant. In the example presented here, there is evidence that spatial patterns are not consistent across all time scales of variability, but rather that some patterns are more characteristic of certain time frames than others. This suggests that the causitive mechanisms, such as atmospheric circulation, may make differential contributions to spatial patterns dependant on the time scale of variability. Clearly this possibility requires further study.

There is a considerable literature, almost entirely in the social sciences, on extensions and variants of three-mode analysis; the model presented here is the most fundamental (Law *et al.* 1984). Differences among the available models reduce to their concept of latent factors. In the fundamental model, factors are thought to be intrinsic to each mode. Here, it is the interactions of patterns in the spatial mode with those of the temporal mode that are of interest. In other three-mode models, latent factors are conceived to be beyond the modes, to govern the entire situation; that is, the spatial and temporal factors identified are merely expressions of more fundamental behavior that is common to the system as a whole. Do such transcendant factors exist in climatology? Is it possible that the factors defined on the temporal or spatial modes, as in this paper, are really only partial expressions of some larger behavior, say "zonality/meridionality"? The answers to these issues will come more from a qualitative understanding of the reality we are trying to model than from any additional mathematical sophistication with regard to the general linear model of which eigenvector techniques are a part (cf Gould 1984). Three-mode analyses may be useful for illuminating these questions, by permitting the study of spatial and temporal interactions that are not addressed in two-mode analyses.

REFERENCES

Cant, R.G., 1971: Changes in the location of manufacturing in New Zealand 1957-68: an application of three-mode factor analysis. New Zealand Geographer 27, 38-55.

Gould, P., 1984: Destructive retrieval in the realm of three-mode thinking. In, H.G. Law, *et al.*, *q.v.*, 536-554.

Karl, T.R., 1983: Some spatial characteristics of drought duration in the United States. Journal of Climate and Applied Meteorology 22, 1356-1366.

Karl, T.R., and A.J. Koscielny, 1982: Drought in the United States: 1895-1981. Journal of Climatology 2, 313-329.

Langlois, A., 1983: Les transformations de l'espace social de la ville: une application de l'analyse factorielle à trois entrées. The Canadian Geographer 27, 67-73.

Law, H.G., C.W. Snyder, Jr., J.A. Hattie, and R.P. McDonald, eds., 1984: Research Methods for Multimode Data Analysis. Praeger Publishers, 686 pp.

Mitchell, J.M., Jr., C.W. Stockton, and D.M. Meko, 1979: Evidence of a 22-year rhythm of drought in the western United States related to the Hale Solar Cycle since the 17th century. Solar Terrestrial Influences on Weather and Climate, B.M. McCormack and T.A. Seliga, eds., Reidel, 125-144.

Richman, M.B., 1983: Rotation of principal components in climatological research. Part I: Theoretical considerations, suitable applications and advantages over unrotated solutions. Proceedings, 8th Conference on Probability and Statistics in Atmospheric Sciences, American Meteorological Society, 59-68.

Tucker, L., 1966: Some mathematical notes on three-mode factor analysis. Psychometrika 31, 279-311.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

ROTATION OF PRINCIPAL COMPONENTS (PARTS 1 and 2)

Michael B. Richman
Climate and Meteorology Section
Illinois State Water Survey
Champaign, Illinois 61820-7407, U.S.A.

1. INTRODUCTION

The eigentechniques have been used with increasing frequency since 1980; no fewer than eighty applications of empirical orthogonal function (EOFs), principal components (PCs) and common factor analysis (CFA) were published in the meteorological literature during this period. A characteristic of this proliferation has been the varied types of ways in which analyses are specified. For instance, there are different dispersion matrices used to relate data in these analyses (e.g., correlation, covariance, cross-products), three basic eigenmodels (EOFs, PCs, CFA), numerous procedures to determine the optimum number of significant eigenmodes, a number of alternative solutions for interpreting the output (e.g., orthogonal rotations, oblique rotations, Procrustes target rotations), and six different ways to specify a data set prior to the analysis, within these alternative solutions, each of which brings out different relationships among the data. Consequently, an analysis of a single data set can yield a number of physically plausible interpretations base on the options outlined, and all of the options need to be carefully considered in order for the analysis to yield physically meaningful results.

The question concerning the applicability of an alternative solution or linear transformation (known as a 'rotation') has not ben addressed fully in the meteorological literature. All three eigenmodels (EOF, PCA, CFA) allow for an infinite number of alternative final solutions which satisfy the models' equations, a fact which will serve as motivation for this paper, which will present the theory behind simple structure rotation and outline the meteorological benefits of rotating various data sets. One eigenmodel (PCA) is then developed and shown to be valid for any non-signular linear transformation, a characteristic which is capitalized on in the detailed discussion of simple structure and its graphical presentation. The analytic rotations are individually explored and then compared in a Monte Carlo analysis and the resulting findings are useful to meteorologists in that they provide a more finely calibrated tool for the optimal choice of a solution (e.g., unrotated, orthogonally rotated, obliquely rotated) in order to most accurately determine the modes of variation in an analysis.

2. WHY BOTHER WITH ROTATION?

A question which immediately arises is: why not simply interpret the unrotated EOFs/PCs? Unrotated solutions offer a number of potential advantages, such as their economy, ability to extract the maximal variance from a data set, their spatial and temporal orthogonality and their pattern insensitivity to the number of PCs retained, whereas the rotated solutions have a set of jargon associated with them (e.g., hyperplanes, intercomponent corelations, oblique angles, etc.) which can be discouragingly foreign to uninitiated meteorologists. Additionally, these characteristics make unrotated solutions good candidates for situations where pure data reduction is sought, in cases where the PCs will be used for regresion without individually interpreting each mode and even in certain exploratory applications. The reason why meteoorlogists should at least consider the possibility of applying a rotation to a data set is that unrotated solutions exhibit four characteristics which hamper their utility to isolate individual modes of variation. These four characteristcs are domain shape dependence, subdomain instability, sampling problems and inaccurate portrayal of the physical relationships embedded within the input matrix. These are individually explored in Richman (1986, J. Climatology, 6, 293-335).

3. PRINCIPAL COMPONENT MODEL

Basic equations

The principal component model will be developed for application to the rotated solutions. Most of the results are directly applicable to EOF and FA also. The PC model can be defined from an N x n data matrix $\mathbf{Z} = \{z_{ij}: i=1, \ldots, N; j=1, \ldots, n\}$ where i indexes individuals or cases and j indexes variables. For every, i, j, z_{ij} denotes the variable value of some physical field or some derived field. The PC model representation of z_{ij} is

$$z_{ij} = \sum_{m=1}^{r} f_{im} a_{mj}^{T} \qquad \begin{matrix} i = 1, \ldots, N; j = 1, \ldots, n; \\ m = 1, \ldots, r; r \leq n \end{matrix} \qquad (1)$$

where f_{im} is the rth principal component score for the ith individual and a_{jm} is the rth principal component loading on the jth variable.

In matrix form (1) becomes

$$\mathbf{Z} = \mathbf{F}\mathbf{A}^{T} \qquad (2)$$

The principal component scores (sometimes called 'amplitudes' in the meteorological literature) and loadings can be developed with the following definitions (assumed to be in standardized form and divided by the number of cases although not a necessary condition):

$$R \equiv Z^T Z \ (n \times n) \text{ (data correlation matrix)} \quad (3)$$

$$\Phi \equiv F^T F \ (r \times r) \text{ (principal component score correlation matrix)} \quad (4)$$

$$S \equiv Z^T F \ (n \times r) \text{ (principal component primary structure matrix)} \quad (5)$$

Since $Z^T = AF^T$ can be substituted in (5),

$$Z^T F = AF^T F \quad (6)$$

$$S = A\Phi \quad (7)$$

which is an alternative solution of the PC primary structure matrix (5) which represents the correlations between the PCs and the variables. The equations (3), (4) and (6) are used to define the PC model in terms of the correlation matrix (R) representation: $R = Z^T Z = AF^T FA^T$

$$R = A\Phi A^T \quad (8)$$

The relationship (8) is the fundamental PC equation in terms of the correlation matrix and is a combination of A (the PC primary pattern matrix) and $A\Phi$ (the PC primary structure matrix). In the initial solution (and under orthogonal rotation) $\Phi = I_r$, therefore $A\Phi = A$ and both are referred to as PC loadings.

Transformational indeterminacy

In the introduction it was stated that an infinite number of alternative solutions satisfied the PC model. Applying the relationships which define the PC model in (2) and (8), it will be shown that an inherent indeterminacy exists, since a unique solution is not possible without additional constraints. This arises because PCs may be linearly transformed to an infinite number of other sets of PCs having the same properties by using arbitrary non-singular linear transformations (T) such as

$$F^* = FT^T \quad (9)$$

$$A^* = AT^{-1} \quad (10)$$

The data matrix representation in (2) then becomes

$$Z = FA^T = FT^T(T^T)^{-1}A^T = F^*A^{*T} \quad (11)$$

For the correlation matrix representation in (8), Φ^* can be defined as $\Phi^* = T\Phi T^T$ and application of (10) yields

$$R = A\Phi A^T = AT^{-1}T\Phi T^T(T^T)^{-1}A^T = A^*\Phi^*A^{*T} \quad (12)$$

From (11) it can be seen that F and F^* have the same properties and are indistinguishable, accounting for the same amount of total variance. From (12) it can be seen that A can be linearly transformed to A^* to yield a new set of PC loadings. To determine a unique set of PCs, a set of r basis vectors must be defined so that the PCs can be referred to these vectors. The basis vectors can be chosen to be orthogonal to each other, so that the PCs are uncorrelated in an orthogonal rotation, or can be chosen to be obliquely related, so that the PCs are correlated in an oblique rotation. In both orthogonal and oblique solutions, the basis vectors are typically chosen such that the PC loadings (A^*) [orthogonal case] or the PC primary pattern matrix (A^*) [oblique case] have specific variance properties. The matrix A^* is typically transformed with the constraint that it contains the properties of a concept known as 'simple structure'.

4. SIMPLE STRUCTURE

4.1. General concept

Thurstone (1947) developed a set of five requirements or criteria for a transformed solution, which he called simple structure, to help ensure that the matrix A^* is easily interpretable. The definition of these five requirements will be presented in Section 4.3. Some background information will now be offered to make the exact definition more meaningful and logical. The goal of simple structure is to aid in 'the scientific problem of discovering an underlying order in the domain'. This implies that the concept is useful in exploratory work. Thurstone reasoned that the variable correlations (R) defined a geometrical configuration of variable vectors without a reference frame. Since the PCs could be represented by the axes of the frame, it was necessary to locate a reference frame in the variable configuration where it could be used for meaningful scientific interpretation of the configuration.

Thurstone noted that, in the majority of analyses, variable vectors would cluster or group together. Furthermore, the angular separations between the variable vectors within a group were relatively small, whereas the separations between groups of clustered vectors were markedly larger. When these subgroups or clusters were examined, the specific variable vectors within any one were frequently so highly related that one causal or summary influence could be assigned to each of the variables within a cluster. Rotating the PCs into such a position where they explained these

clusters (were co-linear with them) would facilitate easier interpretation since each PC would represent one summary influence (or a subset of influences) and would not be related to other summary influences in other data clusters.

4.2 Technical aspects of simple structure

The fundamental rationale behind simple structre is the compact scientific explanation of a variable with the smallest number of PCs. For example, if there were n variables and r PCs retained in an analysis, the solution would fulfill the principles of simple structures when each variable required fewer than r PCs to account for most of the variable's variance. Moreover, the individual PCs would have a compact interpretation if each PC was associated with only a subgroup of the original variables. Since a simple structure solution requires that much of the variables' variance must be accounted for by a linear combination of fewer than r PCs, this directly implies that to achieve simple structure, the variables' variance must exist in a subspace having no more than r - 1 dimensions in r-dimensional PC space. By definition, a subspace of r - 1 dimensions is known as a hyperplane or co-ordinate hyperplane. Under simple structure rotation, the problem is to find those subspaces which completely contain the significant variance of various subsets of the original variables. The device which is used to locate these subspaces is known as a reference axis. A reference axis (as opposed to a primary axis) is a hypothetical random variable of unit length in PC space which is orthogonal to a hyperplane in the r-dimensional PC space. Since a reference axis is orthogonal to every vector contained in the hyperplane, a vector in PC space which is orthogonal to an observed variable is potentially a reference axis and identifies a corresponding co-ordinate hyperplane to which it is orthogonal (Mulaik, 1972). Given variables for which simple structure applies, then, since there are n observed variables and, at most, r linearly independent reference axes in r-dimensional PC space, some of the observed variables must be orthogonal to the same reference axes, and the significant part of these variables' variances must share the same subspaces of the PC space. Mulaik (1972) suggests a way to find a simple structure solution using the n x r reference structure matrix (V)

$$\mathbf{V} = \mathbf{R}_{ZY} = \mathbf{Z}^T\mathbf{Y} \qquad (13)$$

where $\mathbf{Z}^T$ is a matrix of n row vectors z_1^T, ..., z_n^T which represent the normalized variables Z_1, ..., Z_n and Y is a matrix consisting of r row vectors y_1^T, ..., y_r^T representing the r reference axes. The entries in V are correlation coefficients between the vectors z_1^T, ..., z_n^T, y_1^T, ..., y_r^T.

4.3 Explicit definition of simple structure

From the definition of V in (13), a simple structure can be determined which satisfies Thurstone's (1947, p. 335) five criteria. Thurstone notes that variables of complexity r should be eliminated prior to a final rotation and interpretation since these variables do not contribute to the location of the reference frame. This point is frequently overlooked. All that these criteria require of the r reference axes is that they be a set of linearly independent vectors in PC space.

1. Each row of the [oblique] reference structure matrix V should have at least one zero. This criterion satisfies the basic concept (explained in Section 4.1).
2. For each column m of the matrix V there should be a distinct set of at least r linearly independent observed variables whose correlations (as found in the mth column of V) with the mth reference axis variable are zero.
3. For every pair of columns in V there should be several variables, whose entries are zero in one column but not in the other.
4. With four or more PCs, each pair of columns of V should have a large proportion of corresponding zero entries.
5. For every pair of columns in V there should preferably be only a small number of entries in both columns which do not vanish.

Thurstone notes that when these conditions are met a graphical plot of each pair of columns of V shows (i) a large concentration of points in two radial streaks, (ii) a large number of points at or near the origin, and (iii) only a small number of points off the two radial streaks. For a PC analysis of r dimensions, there would be 1/2r (r-1) graphical plot diagrams. Thurstone (1947, p. 335) goes so far as to note that 'In the last analysis it is the appearance of the diagrams that determines, more than any other criterion, which of the hyperplanes of the simple structure are convincing and whether the whole configuration is to be accepted as stable and ready for interpretation.'

A common point of confusion is that the reference vectors (Y) mentioned above are not the PCs (F). The reference axes are treated as if they were PCs in a simple structure solution since they form a basis in PC space. The PCs are detemined after the hyperplanes have been determined by the

reference axes. The reference axes form a biorthogonal system of vectors with the PC vectors. Richman (1986) formally relates the two matrices showing that the zeros in A^* correspond to the zeros in V and the primary pattern matrix has coefficients which are proportional by columns to the matrix of reference structure coefficients. Therefore, in a definition of simple structure configurations, either the reference structure matrix or the primary pattern matrix can be used. This will allow identification of simple structure in the pairwise graphical plots directly from the PCs.

4.4 Objective analytic rotation

The simple structure criteria listed in section 4.3 have been used as a springboard to create analytic computer programs to approximate at least a portion of the concept. There is a difficulty in defining the five criteria objectively since they are topological concepts originally designed for hand-rotating the axes. The advantages gained with analytic rotation programs arise from the objectivity gained by automating an algebraic formula which replaces human judgement in placing the reference axes and the time saved by computers. This last feature has been such a strong motivating force that no fewer than nineteen orthogonal and oblique rotations exist today; they are listed in Table 1.

Table I. Analytic simple structure rotations

Rotation name	Type	Reference(s)
Quartimax	Orthogonal	Carroll (1953); Neuhaus and Wrigley (1954)
Varimax	Orthogonal	Kaiser (1958)
Transvarimax	Orthogonal	Saunders (1962)
Parsimax	Orthogonal	Crawford (1967)
Equamax	Orthogonal	Kaiser (1974)
Quartimin	Oblique	Carroll (1953)
Biquartimin	Oblique	Carroll (1957)
Covarimin	Oblique	Kaiser (1958)
Binormamin	Oblique	Kaiser and Dickman (1959, 1977)
Maxplane	Oblique	Cattell and Muerle (1960)
Oblimax	Oblique	Saunders (1961)
Harris-Kaiser Class II	Oblique	Harris and Kaiser (1964)
Harris-Kaiser Class III	Oblique	Harris and Kaiser (1964)
Promax	Oblique	Hendrickson and White (1964)
Direct oblimin	Oblique	Jennrich and Sampson (1966)
Functionplane	Oblique	Katz and Rohlf (1974)
H-A (Types A and B)	Oblique	Hakstian and Abell (1974)
Casey's method	Oblique	Kaiser and Cerny (1978)
DAPPFR	Oblique	Tucker and Finkbeiner (1982)

Most of the analytic rotations are simple algebraic expressions which attempt to approximate simple structure through the application of specially designed mathematical algorithms which distribute the PC loadings such that the dispersion of the loadings is maximized by maximizing the number of large and small coefficients (e.g., Varimax). This, in turn, has led to the common misconception that simple structure involves maximizing the number of near-zero and large loadings. This is not what simple structure attempts to do since simple structure makes no assumptions about how large the salient coefficients are; maximization of the near-zero coefficients is the most important aspect. Nevertheless, the advantages gained by such analytic programs has resulted in their widespread use and the newer rotations (i.e., mid-1970's to present) more faithfully reproduce all of Thurstone's rules.. The important point to remember is that all of the analytic rotation algorithms listed in Table I were created as attempts to fulfill Thurstone's simple structure principles, as simple structure itself was considered to be a highly desirable concept by the various authors.

5. GRAPHICAL PRESENTATION OF SIMPLE STRUCTURE

5.1 Construction of pairwise plots

How can it be determined whether an analytic rotation is yielding results close to an ideal simple structure solution? Thurstone provides a clue when he states that the inspection of graphical pairwise plots of PCs is of utmost importance in the interpretation of an analysis (see section 4.3). Such a plot involves setting two PCs as the x- and y-axes. The rotated PC loadings (or pattern coefficients) are plotted up with respect to the two PCs as a reference frame. An example is provided in Figure 1 which plots up the PC primary pattern coefficients in Table II for the three sets

Table II. Hypothetical PC loadings for graphical plot example shown in Figure 1 (Harman, 1976, p. 359)

Variable	PC1	PC2	PC3
1	0·997	0·000	−0·002
2	0·008	0·994	−0·006
3	−0·001	0·013	0·985
4	0·471	0·759	0·007
5	0·406	−0·008	0·863
6	0·008	0·411	0·800
7	0·728	0·522	0·001
8	0·863	0·014	0·411
9	−0·013	0·773	0·474
10	0·615	0·643	0·009
11	0·678	−0·001	0·672
12	−0·007	0·624	0·645
13	0·990	0·030	−0·011
14	−0·047	0·993	0·050
15	0·019	−0·046	0·996
16	0·293	0·378	0·697
17	0·645	0·472	0·301
18	0·970	−0·036	0·009
19	0·055	0·962	−0·055
20	−0·009	0·038	0·950

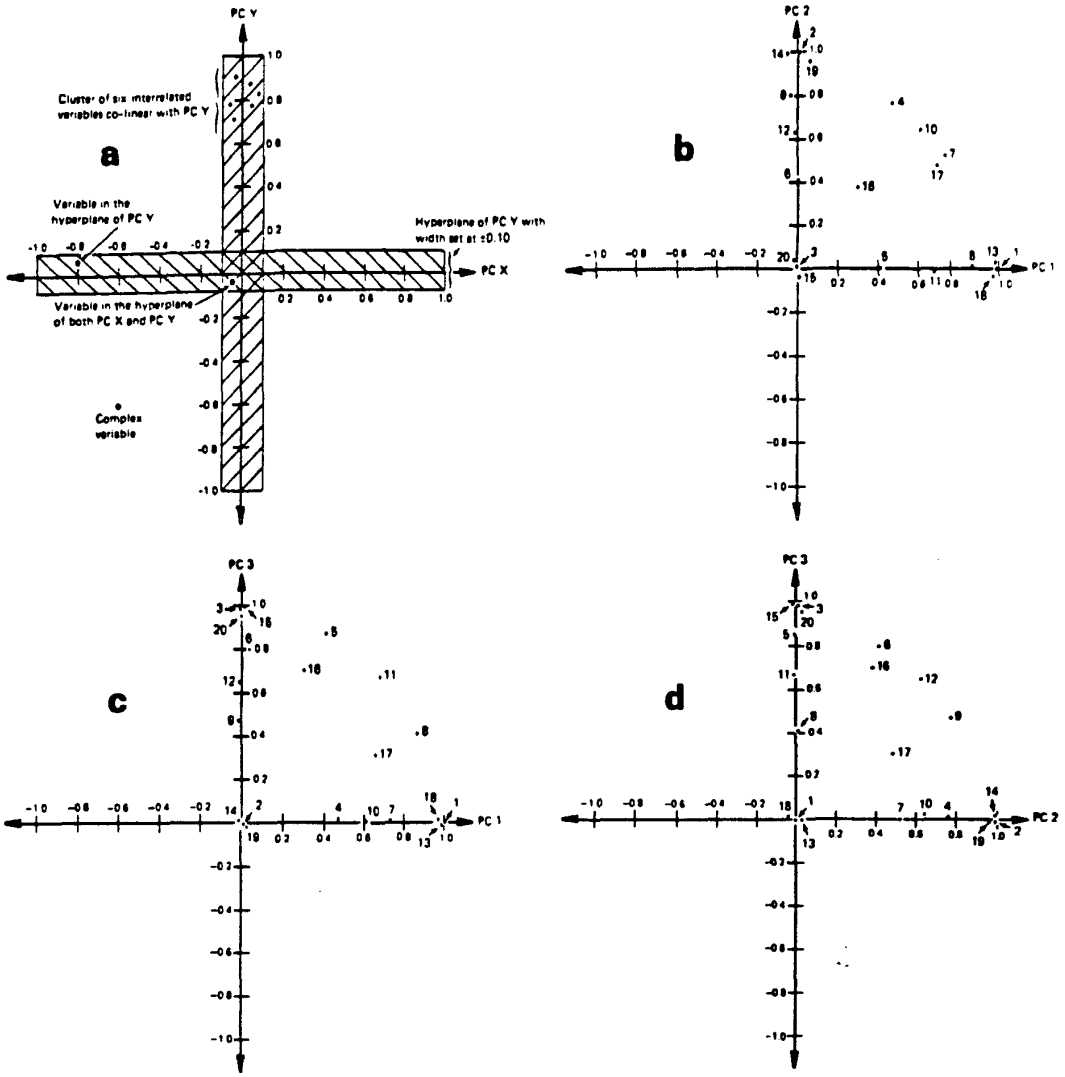


Figure 1. Graphical pairwise plots of the obliquely rotated primary pattern coefficients for the three PCs listed in Table II and an illustrative guideline for interpretation (panel (a)). The plot of PC1 vs. PC2 is panel (b), PC1 vs. PC3 is panel (c) and PC2 vs. PC3 is panel (d). Note the clustering of subsets of variables along the PC axes. The numbering of the points in panels (b), (c), (d) corresponds to variable numbers in Table II.

of pairwise plots. In addition to this, Figure 1(a) has been constructed as a guideline to illustrate the hyperplanes which are orthogonal to the PCs. Variables which fall close to a hyperplane are termed 'in the hyperplane of that PC' (see below for definition of hyperplane width); variables located near the origin are termined 'in the hyperplanes of both PCs', and variables which hae large coefficients on both PCs are termed 'complex variables' (not to be confused with $\sqrt{(-1)}$). Examining Figure 1(b), which plots PC1 vs. PC2, indicates that variables 2, 3, 6, 9, 12, 14, 15, 19, and 20 are nearly zero with respect to PC1. Strictly speaking, the hyperplane is infinitely narrow and can only be formed by an exact zero loading. The question arises of how the problem of sampling errors and other noise affects the magnitude of the loadings as it is unlikely that an exact zero loading will occur in any analysis. Since a subset of variables are 'close' to zero, Tucker (1955) suggested that variables considered to be in the hyperplane should have coefficients falling within a certain narrow band on either side of zero. Typically, values between ± 0.05 and ± 0.20 are used with the hyperplane width being directly proportional to the number of input observations and the amount of noise in the analysis. For this example a value of ± 0.10 will be used, as shown in Figure 1(a). Returning to Figure 1(b), variables 2, 3, 6, 9, 12, 14, 15, 19 and 20 all define the hyperplane of PC1. Similarly, variales 1, 3, 5, 8, 11, 13, 15, 18 and 20 define the hyperplane of PC2. PC1 clusters variables 1, 5, 8, 11, 13, and 18 whereas PC2 clusters variables 2, 6, 9, 12, 14 and

19 (Figure 1(b)). Similar interpretations can be made by examining Figure 1(c) and (d).

5.2 Amount of simple structure present in an analysis

Using the graphical plots presented in Figrue 1 as a guide, the amount of simple structure present in an analysis will now be schematically illustrated in Figure 2. This is vital to the development of the Monte Carlo technique in sections 9 and 10. Strong simple structure is present in Figure 2(a), as there are many variables in the hyperplane with distinct clusters of variables radiating out along the axes and few complex variables. Figure 2(b) illustrates moderate simple structure in which there are more complex variables than the previous case, and slightly fewer variables in the hyperplane. Figure 2(c) depicts weak simple structure in which there are a good number of complex variables, poorly defined hyperplanes and indistinct clusters. Finally, Figure 2(d) depicts a random configuration with many complex variables and no clusters, in which case any PC position would be equally valid to describe the variables. Fortunately, meteorological data are rarely in a random configuration; those that are should not be subject to an eigenanalysis (particularly EOF/PC), as it is possible for the coefficients to appear as signal when they are artifacts of the techniques (see Richman, 1985).

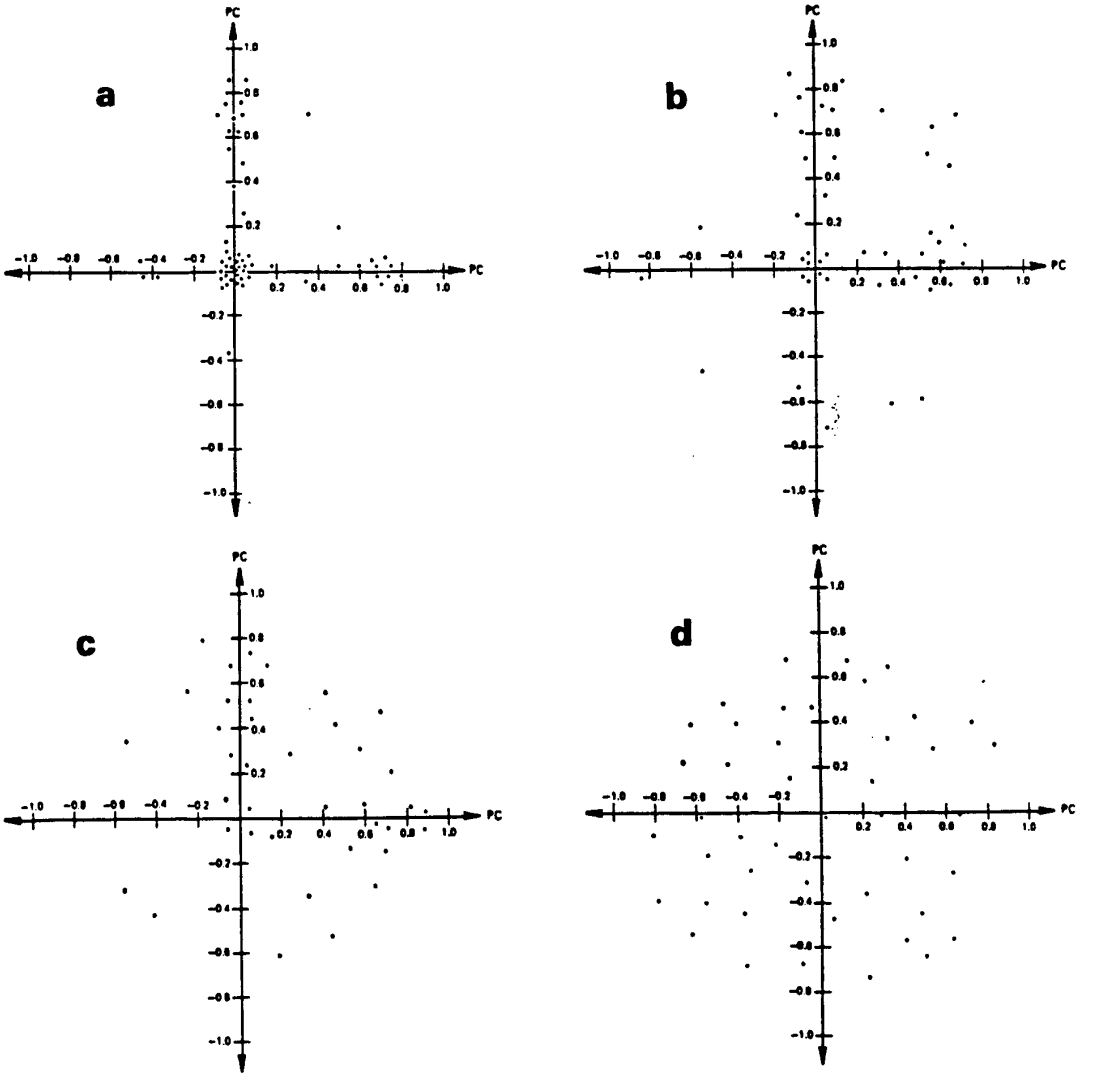


Figure 2. Hypothetical schematic examples of various degrees of simple structure using the graphical pairwise plots of rotated PC loadings as a guide. Panel (a) corresponds to a strong simple structure. Panel (b) corresponds to a moderate simple structure. Panel (c) corresponds to a weak simple structure. Panel (d) corresponds to a random configuration of input correlations. These plots can be compared to actual pairwise graphical plots of meteorological data to assess the degree of simple structure in any rotated data set

5.3 Relating real data to the amount of simple structure present

Using real examples of meteorological data, the amount of simple structure present in pairwise plots of PCs can be determined. The first example, Figure 3(a), is taken from the Lamb and Richman (1983a) precipitation analysis, where PC loadings refer to climatological stations, exhibiting a large number of near-zero loadings in the hyperplane and clusters of salient loadings aligned with the PCs indicating a strong simple structure. The second example, Figure 3(b), is taken from the Walsh and Richman (1981) sea-surface temmperature data set which exhibits moderate simple structure, as there are a number of complex variables between the PCs and clusters of variables which are not as well defined as in Figure 3(a). The third example, Figure 3(c), is taken from a study by Richman (1981) of 500 mb geopotential data (not shown in that article) in which there are more ambiguous clusters of variables with fewer variables in the hyperplanes corresponding to a weak simple structure. One point should be made clearly: the results in Figure 3 do not imply that all precipitation data exhibit strong simple structure, all sea-surface temperature data exhibit moderate simple structure or all geopotential data exhibit weak simple structure. In any given analysis some pairwise plots may exhibit strong simple structure, whereas others may exhibit moderate simple structure; therefore, all possible 1/2r (r-1) pairs of plots should be examined in order to assess the overall degree of simple structure in the data (as illustrated by Richman and Lamb, 1985).

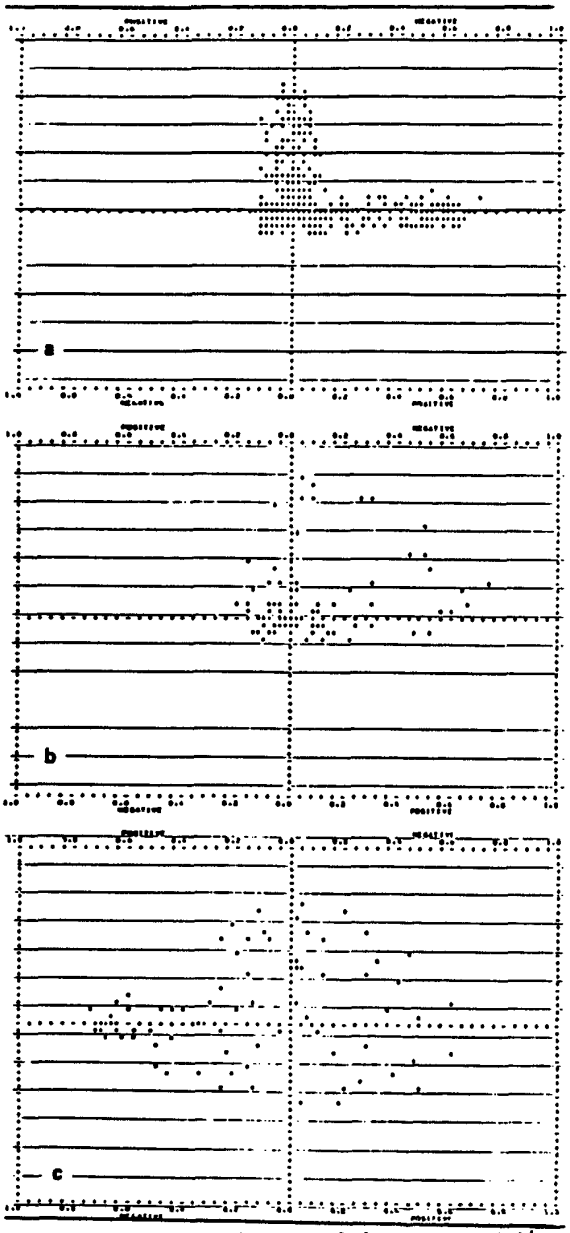


Figure 3. Computerized output of three representative graphical plots of meteorological/climatological data sets. Panel (a) illustrates a strong degree of simple structure with a Varimax rotation using the information in Figure 2 as a guideline. Panel (b) illustrates moderate simple structure with a Varimax rotation. Panel (c) illustrates a very weak simple structure with a Varimax rotation.

6. ANALYTIC SIMPLE STRUCTURE ROTATION

In section 4, the concept of simple structure was introduced along with a method of assessing the amount of simple structure present in a database (section 5). The focus of this portion of this work involves comparing the analytic rotation techniques available on the statistical packages. Thurstone's (1947) five simple structure criteria have been partially incorporated into algebraic expressions which attempt to approximate a simple structure solution of which 19 were identified in Table 1. A survey of the major mathematical/statistical packages (BMDP, 1981; IMSL, 1979; SAS, 1982; SPSS, 1979) revealed that, of the 19, only a subset of eight orthogonal and oblique rotations are readily obtainable (Table III) because some are outdated whereas others are so new that they are not yet available. Of these eight, all five oblique rotations and the most accurate orthogonal rotation will be examined and compared for accuracy over a

Table III Analytical simple structure rotations available on the four major mathematical statistical packages. Asterisk denotes Procrustes which is generally used as a confirmatory transformation

	Packages			
BMDP (1981)	IMSL (1979)	SAS (1982)	SPSS (1979)
Orthogonal rotations			
Equimax Quartimax Varimax	Equimax Quartimax Varimax	Equimax Quartimax Varimax	Equimax Quartimax Varimax
Oblique rotations			
Direct oblimin Harris-Kaiser II IC Harris-Kaiser II B'B	Harris-Kaiser II IC Harris-Kaiser II B'B Procrustes* Promax	Harris-Kaiser II IC Harris-Kaiser II B'B Procrustes* Promax	Direct oblimin

variety of different data sets. Algebraically, the orthogonal rotations must satisfy equation (10), $B = A^* = AT^{-1}$, where B is an n x r rotated PC loading matrix corresponding to A^* in (10), A is an n x r initial unrotated loading matrix and T is an r x r orthonormal transformation matrix which attempts to impact the properties of simple structure on B. (Some of the packages define T^{-1} as T; therefore, in these situations, $B = AT$). Since T is an orthonormal matrix, $T^{T}T = TT^{T} = I_r$. The oblique rotations also use (10) but relax the condition that $T^{T}T = TT^{T} = I_r$ so that Φ may have off-diagonal values. Each criterion is mathematically explored in Richman (1986).

7. MONTE CARLO SIMULATION

7.1 Rationale

In the past, the choice of which analytic rotation(s) to apply has been a 'blind' decision; that is, there have been few studies comparing how accurately (see Section 7.4 for specifics) the rotations can uncover the underlying structure for different types of data in various analyses. An initial analysis would typically be followed by a readily accessible orthogonal or oblique rotation. This was the motivating force behind this Monte Carlo study: to examine how well various factor solutions (i.e., unrotated, orthogonally rotated, obliquely rotated and random spin) can uncover input factor structures which were randomly generated. The basic Monte Carlo procedure and format were developed by Tucker (1983) to test a new rotational technique and the theoretical development is given by Tucker et al. (1969) and Tucker and Finkbeiner (1982). The technique uses factor analysis as opposed to PC. Although the factor model may have a slight advantage in accurately uncovering the underlying modes of variation, the results found will be valid for either technique. A brief discussion of some of the parameters will ensue in section 7.3.

7.2. How Thurstone's simple structure concept was incorporated into the simulation

Thurstone's simple structure concept can be seen by examining parwise graphical PC plots as shown in section 5.2. The relationships shown in Figure 2 will form the basis for objectively stating the amount of simple structure in the variable configurations. Recalling that the simple structure plots have many near-zero loadings and radial streaks or clusters aids in objectifying the concept. The values for the plots in the Monte Carlo procedure are randomly generated numbers between 0 and ±1. These input random numbers are used to simulate realistic looking simple structure fields by treating them as if they are factor pattern coefficients. In order to create these 'mock' fields, all values which are less than the specific pre-set value shown in Table IV are changed to zero to ensure well-defined hyperplanes. To ensure that the hyperplane values are separate from the 'clusters', a minimum number of high loadings on one factor which are low loadings on the second must be present, and a specific number of low loadings must be present. All three of these values vary with the degree of simple structure as determined by Tucker (1983) and shown in Table IV.

Table IV. Relationships for various degrees of simple structure in first-order factor loadings for 25 and 50 variable simulations (Tucker, 1983)

	Simple Structure		
25 variables/5 factors	Strong	Moderate	Weak
Critical loading to be set to zero	≤0·70	≤0·55	≤0·40
Minimum number of zero loadings	15	12	10
Minimum number of loadings high on factor *X* and low on factor *Y*	5	4	3
50 Variables/10 factors			
Critical loading to be set to zero	≤0·70	≤0·55	≤0·40
Minimum number of zero loadings	30	25	20
Minimum number of loadings high on factor *X* and low on factor *Y*	5	4	3

7.3 Outline of simulation

From the preceding, it can be seen that the basic configuration contains the properties of simple structure as defined by the random numbers and the amount of simple structure required. The purpose of the simulation is to test various procedures (i.e., unrotated, rotated solutions) in their ability to retrieve the mock simple structure fields generated according to the program. To make the results closer to actual data (which contain signal and noise) a random noise component is added to the mock simple structure fields. In order to ensure control over the obliquity of the primary factors, the matrix Φ is defined initially. Since is a correlation matrix of factor scores, it can be defined as $\Phi = A_2{}^{T}A_2$ where A_2 are the randomly generated loadings for the second order (see Richman (1983a) for discussion on higher-order factoring). The program uses a form of the correlation equation (8) to generate unrotated common factors from the mock simple structure pattern coefficients:

$$B\Phi B^{T} = BVV^{T}B^{T} \quad (14)$$

$$\Phi = VV^{T}$$

$$A = BV \quad (15)$$

where B is an n x r matrix of simple structure factor pattern coefficients, Φ is the r x r factor score intercorrelation matrix, V is the r x r transformation matrix and A is the n x r unrotated factor loading matrix. At this point the white noise men-

tioned above is added to the matrix A, as specified by the user. The matrix A is then transposed and multiplied by itself to yield a product matrix, C.

$$C = A^T A \tag{16}$$

This product matrix is then decomposed into a real eigensolution via a principal axis factor solution and the eigenvalue series is checked to ensure full rank. The unrotated output and the outputs of the various rotations with noise are all compared to the original mock simple structure input (without noise). A random rotation is also used, which spins the axes by randomly generated values prior to the comparison, for a baseline. The results will be computed for analyses with the significant loadings being of one sign only (one-sided solution) as well as analyses with both positive and negative significant loadings (two-sided solution) with the remainder of the parameters set at the program's constants. Strong, moderate and weak simple structure conditions will be examined. It is important that both one-sided and two-sided solutions be computed as different meteorological data sets fall into each category as illustrated under the discussion of the DAPPFR rotation, in Richman (1986, p. 321-322).

7.4. Matching factors or components

In the previous section it was shown how the unrotated, rotated and randomly spun solutions would be compared to the initial structure. In order to quantitatively assess the goodness-of-match, the coefficient of congruence will be used. The coefficient of congruence (g) between any factors [components] A and B is given by (Harman, 1976)

$$g_{AB} = \frac{\sum_{j=1}^{n} b_{jA} b_{jB}}{\left[\left(\sum_{j=1}^{n} b_{jA}^2\right)\left(\sum_{j=1}^{n} b_{jB}^2\right)\right]^{\frac{1}{2}}} \tag{17}$$

where b are the component/factor loadings or pattern coefficients. The form of the coefficient is similar to a correlation coefficient, as the possible values range from + 1.0 for total agreement through 0 for no relationship to -1.0 for total inverse agreement, and it corresponds to the cosine of the angular separation between the pairs of loadings. However, the coefficient of congruence uses loadings (b's) which are not deviations from their respective means. Specifically, the magnitude of the loadings is an important aspect of component identification. Korth and Tucker (1975) used a Monte Carlo approach to obtain normative data on the distriubtion of the congruence coefficient, with a view to establishing a baseline for evaluating its significance. This led Tucker (1983) to attach the following goodness-of-match names to specific ranges of absolute congruence coefficients: 0.98 to 1.00 (exellent match); 0.92 to 0.98 (good match); 0.82 to 0.92 (borderline match); 0.68 to 0.82 (poor match); less than 0.68 (terrible match). This was done since the congruence coefficient is biased towards a higher value than the corresponding correlation coefficient. Therefore, the meaning of a congruence coefficient between two sets of PC loadings of 0.8 is not nearly as strong as would be intuitively expected with a correlation coefficient of 0.8. Any coefficient of less than approximately |0.70| is not any better than would be expected by randomly spinning the PC axes prior to matching the sets of PC loadings.

8. RESULTS OF THE MONTE CARLO SIMULATION

The rotations which were compared were largely available on the four mathematical/statistical packages in Table III. Varimax was the sole orthogonal rotation available on the packages used, for the reason mentioned in section 6.2 (which was confirmed in an earlier Monte Carlo study). However, the orthogonal Procrustes experimental transformation is also included to attempt to assess whether it is more accurate than Varimax. The oblique rotations tested are also available on the packages (with the exception of the one- and two-sided DAPPFR solutions) and represent the second-generation (i.e., mid-sixties) types of linear transformations. Some of the original oblique rotations were also tested (e.g., Binormamin and Oblimax) [not shown herein] an the results were clearly inferior to those rotations tested. Both the Harris-Kaiser Case II IC and Harris-Kaiser Case II B^1B rotations were tested along with Direct oblimin with $\delta = 0$ (default on packages) and both Promax k = 2 and Promax k = 4. The unrotated solution is also included along with a random rotation in which the factor axes are spun by a randomly generated angle. The standard deviation of the random noise was set at 0.05, which represented a middle value of those tested (0.025 to 0.075). This value would represent a moderate amount of noise in an actual analysis as 95 per cent of the loadings would be expected to be perturbed by values $\leq$0.10 from their ideal values. There were 200 replications of each rotation type within each cell to obtain a representative distribution of results. The results will be subdivided into four categories: (i) one-sided twenty-five variables/five factors retained, (ii) one-

Table V. Results of 200 replications of the Monte Carlo simulation for 25 variables, 5 factors and 1-sided correlations. Degree of simple structure varies from strong (a), to moderate (b) and to weak (c)

		Root mean square congruence coefficient distribution									
Subjective match	Congruence coefficient	Unrotated	Normal Varimax	Direct oblimin	Harris-Kaiser II IC	Harris-Kaiser II B^TB	Promax $k=4$	Promax $k=2$	DAPPFR	Orthogonal Procrustes	Random transformation
(a) Strong simple structure											
Excellent	$g \geq 0{\cdot}98$		6	194	195	195	194	195	195	96	
Good	$g \geq 0{\cdot}92$		186	6	5	5	6	5	5	101	
Borderline	$g \geq 0{\cdot}82$	5	8							3	
Poor	$g \geq 0{\cdot}68$	192									90
Terrible	$g \leq 0{\cdot}67$	3									110
(b) *Moderate simple structure*											
Excellent	$g \geq 0{\cdot}98$		1	157	27	145	56	158	185	29	
Good	$g \geq 0{\cdot}92$		180	33	156	51	132	31	15	157	
Borderline	$g \geq 0{\cdot}82$	1	19	10	17	4	12	11		14	
Poor	$g \geq 0{\cdot}68$	198									141
Terrible	$g \leq 0{\cdot}67$	1									59
(c) Weak simple structure											
Excellent	$g \geq 0{\cdot}98$			30	1	15	2	53	99	2	
Good	$g \geq 0{\cdot}92$		131	100	56	93	120	101	75	138	
Borderline	$g \geq 0{\cdot}82$	1	61	67	109	85	67	46	26	59	10
Poor	$g \geq 0{\cdot}68$	194	8	3	25	7	10			1	158
Terrible	$g \leq 0{\cdot}67$	5			9		1				32

sided fifty variables/ten factors retained, (iii) two-sided twenty-five variables/five factors retained, and (iv) two-sided fifty variables/ten factors retained. The five and ten factor runs are in the range of the majority of meteorological studies; therefore the results should be applicable to most research designs.

The first result is for the one-sided 25 variable/5 factor design with strong simple structure (Table V(a)). The analysis in this cell reveals that the oblique rotations tested could all reproduce the input structure almost perfectly. Varimax did not do quite as well, with 93 per cent of its replications having a matching congruence coefficient between 0.92 and 0.98 (good range) and the orthogonal Procrustes was slightly more accurate. It is noteworthy that the unrotated results did poorly in recovering the input structure, with random rotation being the worst. With the simple structure at the moderate level and all other parameters, except those listed in Table IV held constant, the analysis (Table V(b)) still indicates mostly good to excellent matches for the various rotations. DAPPFR (one-sided solution) is now the most accurate rotation, and Promax k = 2, Direct oblimin, and Harris-Kaiser Case II B^TB also do well. Any of these three would generally yield good results. Promax k = 4 and Harris-Kaiser Case II IC did not do as well, and Varimax was also below these, which is probably due to the orthogonality constraint, as the orthogonal Procrustes was only slightly better. The unrotated and random rotation were the worst, as before. The simple structure was next dropped to the weak level, indicating very little clustering and fewer variables in the hyperplane, and all other parameters except those listed in Table IV were held constant (Table V(c)). The results point out that DAPPFR (one-sided solution) does best with 87 per cent of the replications in the good to excellent range, Promax k = 2 was the best package rotation (77 per cent good to excellent), followed by direct oblimin (65 per cent good to excellent). The remaining rotations were a little less accurate, although they fared better than either the unrotated solutions or a random rotation. These results point to the importance of choosing one of the more accurate rotations for these types of condition, as there are a wide range of matches.

Next, the number of variables was increased to fifty and the number of factors retained to ten. This is a much more difficult situation to achieve good matches, since, even if one factor out of ten is poorly matched, it will drag the resulting RMS congruence coefficients down considerably. (Some runs were also performed at fifty variables and five factors retained, yet the matches were almost as good as twenty-five variables and five factors, indicating that the number of factors retained was the most critical aspect in degrading the matches). The simple structure was set as one-sided strong (Table VI(a)) with the best rotation being the one-sided DAPPFR with 100 per cent replications having $g \geq 0.92$. Of the available package rotations, Promax k = 2 was best with almost 97 per cent of the replications in the good to excellent range with the other oblique solutions being slightly less accurate. Orthogonal Procrustes was more accurate than Varimax, which did not do

Table VI. Results of 200 replications of the Monte Carlo simulation for 50 variables, 10 factors and 1-sided correlations. Degree of simple structure varies from strong (a), to moderate (b) and to weak (c)

		Root mean square congruence coefficient distribution									
Subjective match	Congruence coefficient	Unrotated	Normal Varimax	Direct oblimin	Harris-Kaiser II IC	Harris-Kaiser II B^TB	Promax $k=4$	Promax $k=2$	DAPPFR	Orthogonal Procrustes	Random transformation
(a) Strong simple structure											
Excellent	$g \geq 0{\cdot}98$		1	33	1	38	2	83	148	17	
Good	$g \geq 0{\cdot}92$		100	142	127	141	180	110	52	155	
Borderline	$g \geq 0{\cdot}82$	1	99	25	72	21	18	7		28	
Poor	$g \geq 0{\cdot}68$	182									6
Terrible	$g \leq 0{\cdot}67$	17									194
(b) *Moderate simple structure*											
Excellent	$g \geq 0{\cdot}98$								42		
Good	$g \geq 0{\cdot}92$		4	6	1	3	3	18	72	5	
Borderline	$g \geq 0{\cdot}82$		132	153	25	102	104	145	85	143	
Poor	$g \geq 0{\cdot}68$	198	64	41	165	95	93	37	1	52	26
Terrible	$g \leq 0{\cdot}67$	2			9						174
(c) Weak simple structure											
Excellent	$g \geq 0{\cdot}98$										
Good	$g \geq 0{\cdot}92$										
Borderline	$g \geq 0{\cdot}82$		38			1	6	29	35	34	
Poor	$g \geq 0{\cdot}68$	186	156	185	50	191	170	169	163	162	43
Terrible	$g \leq 0{\cdot}67$	14	6	15	150	8	24	2	2	4	157

well, with almost half of its replications in the borderline range (0.82 to 0.92). The unrotated results were poorer than any of the rotations, with the random spin coming in last. With simple structure at the moderate level, and all other parameters except those listed in Table IV held constant (Table VI(b)), the results indicate a large range in the ability of the solutions to recover the input data. The one-sided DAPPFR rotation was clearly best, with 57 per cent of the replications in the good to excellent range, and there was a large drop-off to the next best solution, Promax k = 2, with Direct oblimin slightly below this along with the two orthogonal rotations. An important point to note is that although Promax k = 2 and Direct oblimin had most replications in the borderline range, they are still better than the unrotated results. The simple structure was dropped to the weak level (one-sided) with all other parameters held constant (Table VI(c)). The results are not particularly encouraging, as even the best rotations are only marginally better than the unrotated results and are in the poor range. Varimax was best (19 per cent of the replications in the borderline DAPPFR (one-sided) was only slightly behind (18 per cent) followed by orthogonal Procrustes (17 per cent) and Promax k = 2 (15 per cent) with the remainder of the solutions being in the same range as the unrotated results (poor). This points to a problem in dealing with many factors/PCs and weak simple structure for all solutions; hence caution should be applied prior to physically interpreting such analyses without additional corroborating evidence.

The second phase of the simulation incorporated two-sided rotations into the analysis, as both significant positive and negative loadings were allowed. The results for the 25 variable/5 factor solution (Table VII(a)) are similar to the corresponding one-sided case, as all of the oblique rotations do very well with over 90 per cent of the replications in the excellent range. As before, the Varimax rotation has 65 per cent of its cases in the good range and 35 per cent in the excellent, which is lower in accuracy than any of the oblique methods, but still may be sufficient for most research purposes, whereas orthogonal Procrustes fares slightly better (42 per cent excellent). The unrotated solution does poorly in recovering the input structure and the random spin does worst. When the amount of simple structure was reduced to the moderate level for two-sided data (Table VII(b)), the results indicated a wide range in the ability of the various solutions to recover the input structures. The two-sided DAPPFR rotation is clearly superior to the rest with 83 per cent of the replications having congruence coefficients $\geq 0{\cdot}92$. Promax k = 2, Harris-Kaiser Case II B^TB and Direct oblimin are well behind DAPPFR, with orthogonal Procrustes and Varimax falling slightly behind these and the unrotated solution doing poorest. As the amount of simple structure was again dropped to the weak level for the two-sided solution, the results (Table VII(c)) indicate a further degradation of the quality of match as most of the rotations are in the poor to borderline range. The two-sided DAPPFR rotation was the best with 10

Table VII. Results of 200 replications of the Monte Carlo simulation for 25 variables, 5 factors and 2-sided correlations. Degree of simple structure varies from strong (a), to moderate (b) and to weak (c)

		Root mean square congruence coefficient distribution									
Subjective match	Congruence coefficient	Unrotated	Normal Varimax	Direct oblimin	Harris-Kaiser II IC	Harris-Kaiser II $B^{T}B$	Promax $k=4$	Promax $k=2$	DAPPFR	Orthogonal Procrustes	Random transformation
(a) Strong simple structure											
Excellent	$g \geq 0{\cdot}98$		70	188	181	189	187	188	191	83	
Good	$g \geq 0{\cdot}92$		128	12	18	11	12	12	9	116	
Borderline	$g \geq 0{\cdot}82$	1	2		1		1			1	
Poor	$g \geq 0{\cdot}68$	199									93
Terrible	$g \leq 0{\cdot}67$										107
(b) *Moderate simple structure*											
Excellent	$g \geq 0{\cdot}98$		14	50	30	51	42	59	134	26	
Good	$g \geq 0{\cdot}92$		142	115	80	115	115	103	32	136	
Borderline	$g \geq 0{\cdot}82$	3	34	33	86	31	40	33	30	33	
Poor	$g \geq 0{\cdot}68$	197	10	2	4	3	3	5	4	5	86
Terrible	$g \leq 0{\cdot}67$										114
(c) Weak simple structure											
Excellent	$g \geq 0{\cdot}98$										
Good	$g \geq 0{\cdot}92$		9	10	9	12	11	12	20	10	
Borderline	$g \geq 0{\cdot}82$	5	88	109	80	102	98	98	93	95	
Poor	$g \geq 0{\cdot}68$	192	101	81	106	85	90	89	87	94	90
Terrible	$g \leq 0{\cdot}67$	3	2		5	1	1	1		1	110

per cent of the replications in the good range, the Harris-Kaiser Case II $B^{T}B$ was the best package rotation (6 per cent in the good range), and the unrotated solution was clearly lower than any of the rotations, yet the distance between the two was closer than in previous two-sided runs.

Next, the number of variables was increased to fifty and the number of factors was increased to ten for two-sided data with simple structure set as strong, and the results (Table VIII(a)) indicate that DAPPFR (two-sided) had the most accuracy by a wide margin with 92 per cent of the replications having congruence coefficients ≥ 0.92. Promax k = 2 and Harris-Kaiser Case II $B^{T}B$ were the most accurate package rotations with between 68 and 70 per cent of the cases in the good to excellent range. Of the orthogonal rotations, Procrustes has 54 per cent of the replications in the good to excellent range with Varimax having 45 per cent in the good range; the unrotated was poorest. When the simple structure was reduced to the moderate level for two-sided solutions, the resulting accuracy for all rotated solutions dropped considerably (Table VIII(b)) as all rotated solutions, with the exception of the two-sided DAPPFR, were very close to the unrotated accuracy. This is a

Table VIII. Results of 200 replications of the Monte Carlo simulation for 50 variables, 10 factors and 2-sided correlations. Degree of simple structure varies from strong (a), to moderate (b) and to weak (c)

		Root mean square congruence coefficient distribution									
Subjective match	Congruence coefficient	Unrotated	Normal Varimax	Direct oblimin	Harris-Kaiser II IC	Harris-Kaiser II $B^{T}B$	Promax $k=4$	Promax $k=2$	DAPPFR	Orthogonal Procrustes	Random transformation
(a) Strong simple structure											
Excellent	$g \geq 0{\cdot}98$			1		4	1	13	114	10	
Good	$g \geq 0{\cdot}92$		90	80	69	135	118	123	70	97	
Borderline	$g \geq 0{\cdot}82$		101	109	122	57	72	58	16	89	
Poor	$g \geq 0{\cdot}68$	196	9	10	9	4	9	6		4	1
Terrible	$g \leq 0{\cdot}67$	4									199
(b) *Moderate simple structure*											
Excellent	$g \geq 0{\cdot}98$										
Good	$g \geq 0{\cdot}92$								10		
Borderline	$g \geq 0{\cdot}82$	1	5	2	3	17	8	17	41	15	
Poor	$g \geq 0{\cdot}68$	183	166	156	135	158	161	167	139	163	2
Terrible	$g \leq 0{\cdot}67$	16	29	42	62	25	31	16	10	22	198
(c) Weak simple structure											
Excellent	$g \geq 0{\cdot}98$										
Good	$g \geq 0{\cdot}92$										
Borderline	$g \geq 0{\cdot}82$										
Poor	$g \geq 0{\cdot}68$	189	7	13	9	8	8	9	9	7	3
Terrible	$g \leq 0{\cdot}67$	11	193	187	191	192	192	191	191	193	197

decline from the one-sided results in Table VI(b) and points to the difficulty in accurately reproducing modes with large positive and negative features. DAPPFR does somewhat better with 5 per cent of the cases in the good range and over 25 per cent having congruence coefficients ≥ 0.82 (borderline or better). The only package rotations slightly better than the unrotated solution are Promax k = 2 and , possibly, Harris-Kaiser Case II $B^T B$. Finally, when the simple structure was reduced to the weak level, indicating very little clustering and few variables in the hyperplane, the results (Table VIII(c)) indicate that this is the most difficult situation for any solution to accurately capture, and further indicate that simple structure rotation does not work well when the two-sided data exhibit little simple structure and many factors are retained. The most obvious results is that the unrotated solution does best in this one case, although the best replications are in the poor range, indicating a questionable correspondence to the input data. Every simple structure rotation is even worse than the unrotated solution, being near or at the random spin accuracy level. Data in this configuration might benefit from a Procrustes transformation if there is sufficient prior knowledge concerning the nature of the modes of variation. If this is not the case, another type of analysis technique may be warranted.

9. SUMMARY AND DISCUSSION

This study has outlined some pitfalls of the conventional use of unrotated principal components, examined the theory behind simple structure and compared the accuracy of both unrotated and rotated solutions.

Four basic disadvantages of unrotated solutions have been shown to be reduced by using the rotated solutions in Richman (1986). Domain shape dependence does not appear to be a problem, subdomains have similar patterns to the full domain, sampling errors are greatly reduced even if the neighbouring eigenvalues are almost equal, as long as there is structure within the input data (e.g., see Tables V(a), VI(a), VII(a), and VIII(a)) and, finally, the rotated solutions frequently do a better job of agreeing with the physical structure of the data as embedded within the correlation or covariance input matrix. These four advantages of simple structure rotations make them good candidates to accurately identify modes of variation in climatological data sets. A number of studies have recently applied rotation to the unrotated CFs/PCs/EOFs with successful results, including those of Horel (1981, 1984), Richman (1981), Walsh and Richman (1981) Balling and Lawson (1982), Karl and Koscielny (1982), Karl et al. (1982), Cohen (1983), Hayden (1983), Ashbaugh et al. (1984, 1985), Ronberg and Wang (1985), Barnston and Livezey (1985), Englehart and Douglas (1985) and Thurston and Spengler (1985). These studies involved meteorological fields such as temperature, rainfall, sea-surface temperature, cyclone frequency, chemical ions and geopotential height, yet there have been many other unrotated EOF/PC studies involving parameters (e.g., vertical velocity (Le Drew, 1980); IR radiation (Weickmann, 1983); vertical temeprature profiles (Uddstrom and Wark, 1985) etc.) which have never been rotated. Application of rotation to such fields would be interesting to assess the amount of simple structure inherent in such data.

A suggested analysis 'plan of attack' for a data set would first include choosing the most physically meaningful mode of decomposition (e.g., p, s, t, etc.) from Richman (1986, p. 314-316), relating the data with one of the dispersion matrices, choosing the most appropriate eigenmodel and deciding on the number of EOFs/PCs/factors to retain, followed by applicaiton of one of the more accurate simple structure rotations under a wide variety of conditions. This work points to two: DAPPFR or Promax k = 2 if a mathematical/statistical package is available. By using one of these two, the chances of arriving at the optimal simple structure solution are maximized prior to interpreting the results. The construction of the graphical pairwise plots of the rotated PC loadings illustrated in section 5 is a very important step in an analysis. By inspecing the 1/2r (r - 1) plots and comparing these to Figure 2, the overall degree of simple structure within the data set can be assessed. The next step involves scanning the input dispersion matrix for the magnitudes and signs of the relationships and rotated loadings (primary pattern matrix in oblique cases) to detect whether salient or significant loadings should be all of one sign (one-sided) or contain both large positive and negative values (two-sided). At this point, information on the number of variables, number of rotated PCs retained, signs of the loadings and overall degree of simple structure can be assembled and compared to the results of the Monte Carlo simulation in Table V-VIII to find what solution (unrotated, rotated) should be the most accurate for the data under investigation. If the initial guess was not one of the better solutions, the principal components can be re-rotated (i.e., to another criterion) if necessary and the pairwise plots of PC loadings inspected to ensure a reasonable simple structure. Analysis can then be

made directly from the primary pattern matrix (under oblique rotation) or from the PC loading matrix (orthogonal rotation or unrotated solutions).

Acknowledgements

This work was partly supported by NSF Grant ATM 83-12653 and has been published in the June 1986 (Volume 6, pp. 293-335) issue of Journal of Climatology (Royal Meteorological Society). The author would like to thank Peter Lamb, John Walsh, John Horel, Bob Livezey, Rudy Preisendorfer, John Bistline and two anonymous reviewers for their beneficial suggestions and comments during various stages of this work. Additional thanks go to Rebecca Runge for typing the manuscript and to John Brother for drafting the Figures. The author extends particular gratitude to Ledyard Tucker for sharing his keen insight into the nature of simple structure and for supplying a source program of the simulation package for modification without which this work would not have been possible.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

EXTREME DRY SPELLS - PERIODICITIES AND TRENDS IN ALPINE RIVER BASINS

by

Franz Nobilis

Hydrographisches Zentralbüro, Wien
Marxergasse 2, A-1030 Wien

1. INTRODUCTION

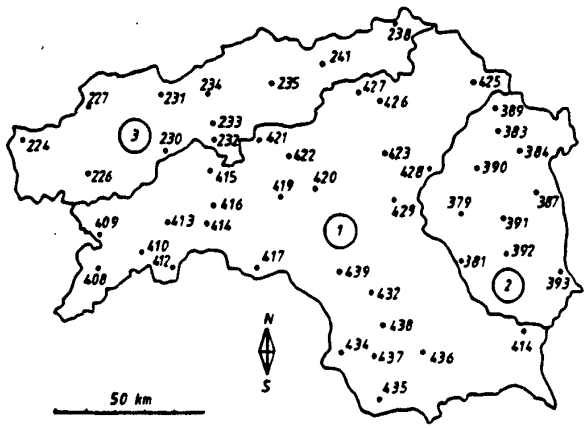
With regard to droughts and dry spells as recurrent phenomena, the question of existing periodicities arises. For example, Mitchell (1968) examined the "Palmer Drought Index" by using spectral analysis (summer months in New York, 1826 to 1966) and determined some peaks in the spectrum, only one of which was significant (3 years). Other peaks at 8, 13 and 40 years were also considerable but did not exceed the 95% significance level.

The task of this paper is to investigate if extreme dry spells in Austria (threshold value 0.2 mm precipitation according to the national definition) show periodicities and trends. Some considerations (including a number of extensions) of the author, who also published a comprehensive book on dry spells, written in German (Nobilis, 1985), are presented in this paper.

2. BASIC MATERIAL AND METHODS OF INVESTIGATION

2.1 Basic material

The observations at 48 stations of river basins of the rivers Mur, Raab and Enns in Styria, a federal province of Austria (Fig.1), were used (for details see Nobilis, 1985).



The observed extreme dry spells during the period 1901-1978 were divided in samples for the year and for the vegetation period (April to September). The samples differ significantly at the 5% level. The results of individual stations are discussed as examples, e.g.those of the station Weiz (No.379; 15°38'E, 47°04'N; 480 m altitude). In this way relevant data regarding the river basins, which cover an area of 15 619 km^2, were calculated.

Figure 1 - River basins of the rivers Mur (1), Raab (2) and Enns (3) in Styria, Austria, including all observation stations

2.2 Spectral analysis

In Austria, spectral analysis was used to examine periodicities for hydrological parameters (e.g. Cehak, 1973; Cehak and Kessler, 1976; Cehak, 1975a and 1976) and meteorological parameters (e.g. Koncek and Cehak, 1968; Cehak, 1969; Sabo, 1972). These investigations resulted in statements on the statistical significance of variations of climate in Austria (Cehak, 1975b and 1977).

The statistical spectral analysis was developed to investigate non-randomness of time series. Tukey (Blackmann and Tukey, 1958) prepared the spectral analysis, which goes back to the generalized Fourier-analysis (Wiener, 1930), for practical use. In this paper some variance spectra of dry spells will be indicated. A "red-noise" spectrum with 95% confidence limits (WMO, 1966) was calculated to prove the significance of the periodicities.. The test by Anderson (1958) was used to investigate if the stochastic process is a Markov-process of the first order. The null hypothesis was not rejected at the selected significance level.

2.3 Results of the spectral analysis

For the year and the vegetation period the following results were derived for the Styrian river basins of the rivers Mur (1), Raab (2) and Enns (3):

River basin No.	No.of obs. stations	Periodicities (years) for the annual series
1	27	generally 3-4 years (partially significant)
2	10	5-6 years occur more frequently than 3-4 years, latter significant only at Fürstenfeld, station No.393
3	11	generally 3-4 years (partially significant)
		Periodicities (years) for the vegetation period
1	27	generally 2-3 years (partially significant)
2	10	some stations 2-3 years (signif.), some 5-6 years
3	11	generally 2-3 years and 3-4 years not significant (signif. at 8 years only at Hohentauern, No.232, and Trieben, No. 233)

Examples for the different spectra in the river basins for the annual series and the vegetation period are given in Fig.2.

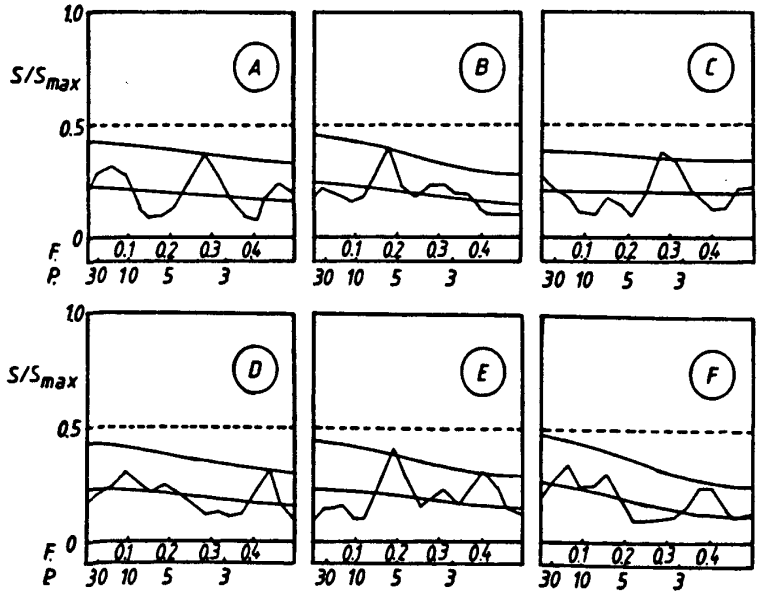


Figure 2 - Variance spectra for the annual series of extreme dry spells (A:St.Johann a.T. No.415; B:Birkfeld, No. 390; C:Schladming, No. 224) and the series for the vegetation period (D:Bruck a.d.M., No.423; E:Vorau, No.384; F:Wildalpen, No.241)

2.4 Trend analysis

The spectral analysis includes the investigation with regard to persistence. In order to examine the question if another type of non-randomness has occured in the data a test developed by Sneyers (1963), based on considerations by Mann (e.g. Cehak, 1975a),was used.

For the year and the vegetation period, the following results were obtained with regard to the Styrian river Basins of the rivers Mur (1), Raab (2) and Enns (3):

River basin No.	No.of obs. stations	Trends (years) in the annual series (p = period)
1	27	until 1910 shorter p., 1910-1920/30 longer p.; 1950 and thereafter few fluctuations (partially significant)
2	10	1901-1920 longer p., 1920-1940 shorter p., most of these trends being significant
3	11	1901-1910 shorter p., 1910-1930 longer p. and after 1930 shorter p. (signif. only at No.224, Schladming (1940-60) and No.230, Oppenberg (1970-75))
		Trends (years) in the series of the vegetation period
1	27	until 1920 longer p., 1920-30 shorter p., 1950 and after: shorter p. (partially significant)
2	10	until 1920 longer p., later on shorter p., but no marked trend
3	11	until 1915 shorter p., later on longer p. (partially significant)

3. CONCLUSIONS CONCERNING PERIODICITIES AND TRENDS

The results regarding periodicities for the annual series may be summarized as follows: Partially significant periodicities between 3 and 4 years occur, and in the series concerning the vegetation periods there are mainly periodicities of 2 to 3 years, most of which are significant.

Summarizing the results of the trend analysis for the annual series, a tendency to shorter periods until 1910 should be mentioned. They are followed from 1910 to 1920/30 by a tendency to longer periods and from 1920/30-1940 partially to shorter periods. Since that time, the situation has been fluctuating. With regard to the vegetation period, the trend to shorter periods, which began in the fifties, is significant in some cases.

4. CASE STUDY OF AN EXTREME DRY SPELL

In October/November 1983 an extensive extreme dry spell, which varied in different areas according to orographic influences, occured in Austria. A practical method to estimate the intensity of dry periods is shown for the above-mentioned Styrian river basins. In a preliminary investigation, the samples for the extreme spells (annual series and series for the vegetation period covering the years from 1901-1978) were fitted to Fisher-Tippett distributions, as developed by Jenkinson (1955). Confidence limits of 68% according to the method of Borgmann (1961) were used for the extreme values of the samples. To illustrate this method, Fig. 3 shows the results for Weiz (No.379).

The extreme dry spell lasted from 18 October to 25/26 November 1983, which constituted a maximum of 40 days (national definition). The absolute number of dry days is shown in Fig. 4 above. The estimate of the return periods was carried out by recalculating them on the basis of the results obtained for the annual series (Fig. 4 below). The figure shows the orographically influenced differentiation where the absolute number of dry days not necessarily means the same return period. This method appears to be very helpful for estimating extreme dry spells, in addition to the usual investigations such as these concerning percentage of precipitation with regard to average or deciles of precipitation (Gibbs, 1975).

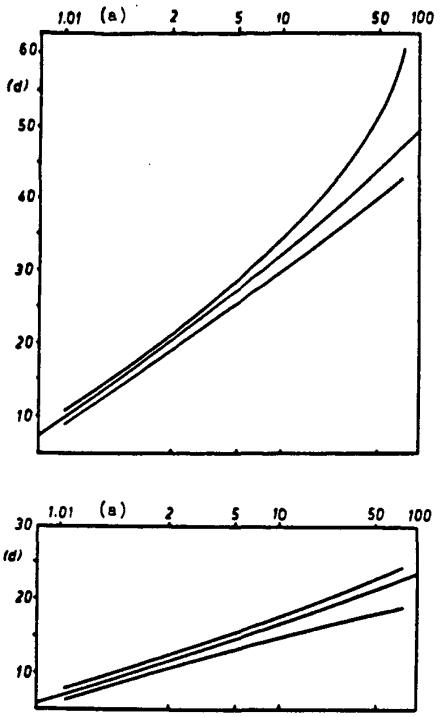


Figure 3 - Extreme value distributions for the extreme dry spells at Weiz, No.379 (data for the year are given above, for the vegetation period below)

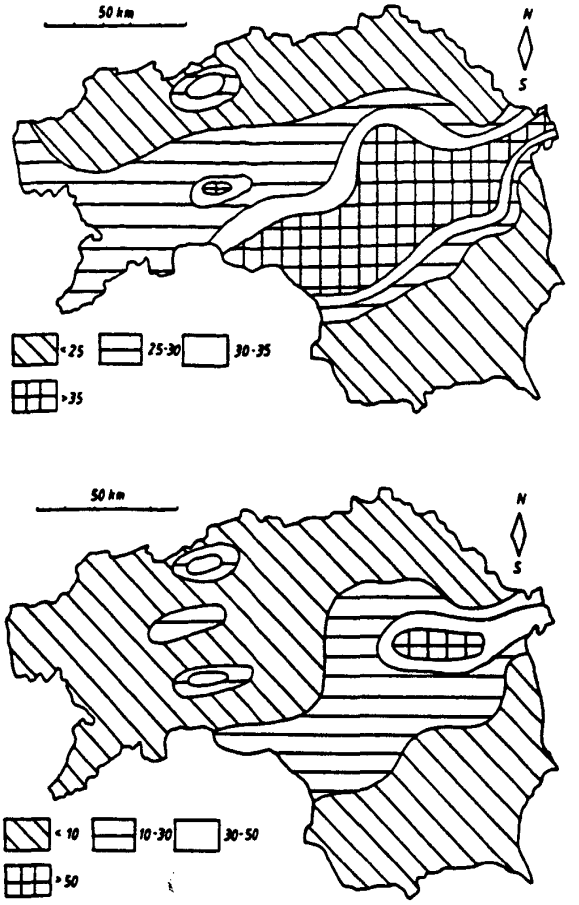


Figure 4 - Absolute number of dry days during the extreme dry spell of October/ November 1983 (above) and the calculated return periods (below)

REFERENCES

Anderson, T.W., 1958: Introduction to Multivariate Statistical Analysis. Wiley & Sons, New York, 374 pp.

Borgmann, L.E., 1961: The Frequency Distribution of Near Extremes. J.Geoph. Res., 66, 3295-3307.

Blackman, R.B. and Tukey, J.W., 1958: The Measurement of Power Spectra. Dover Publications, New York, 190 pp.

Cehak, K., 1973: On flood probabilities of east alpine rivers. J.Hydrol., 20, 65-82.

Cehak, K., 1975a: Spiegelschwankungen eines Alpensees in den letzten 50 Jahren. Wetter und Leben, 27, 71-76.

Cehak, K., 1975b: Climatic fluctuations in the north-eastern pre-alpine region. Physical and Dynamic climatology, Proc.Symp. on phys. and dyn. climat., WMO No.347, 393-400.

Cehak, K., 1977: Über die statistische Sicherung der jüngsten Schwankungen des Klimas in Österreich. Arch.Met.Geoph.Biokl., Ser.B, 25, 209-219.

Cehak, K. and Kessler, A., 1976: Varianzspektrumanalyse der Seespiegelschwankungen des Titicaca-Sees (Südamerika). Arch.Met.Geoph.Biokl., Ser.B, 24, 201-208.

Gibbs, W.J., 1975: Drought - its definition, delineation and effects, WMO Special environmental report No.5, 11-39.

Jenkinson, A.F., 1955: The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. Quart. J. Royal Met.Soc., 87, 158-171.

Koncek, N. and Cehak, K., 1968: Säkuläre Temperaturschwankungen in Mitteleuropa während der letzten 190 Jahre. Arch.Met.Geoph.Biokl., Ser.B, 16, 1-17.

Nobilis, F., 1985: Trockenperioden in Österreich. Verlag des Verbandes der wissenschaftlichen Gesellschaften Österreichs, Wien, 293 pp.

Sabo, P., 1972: Untersuchungen über Periodizitäten der Temperaturen und der Höhen von Druckflächen über Wien. Arch.Met.Geoph.Biokl., Ser.B, 21, 35-44.

Sneyers, R., 1963: Sur la détermination de la stabilité des séries climatologiques. Arid zone Research, CHANGES of Climate, Proceedings of the Rome Symposium, organized by UNESCO and WMO, Paris, 37-44.

Wiener, N., 1930: Generalized Harmonic Analysis. Acta Math. 55, 117-258.

WMO, 1966: Statistical Analysis and Prognosis in Meteorology. Techn.Note No.71, WMO-No.178, TP 88, 191 pp.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

INHOMOGENEOUS MARKOV MODELS FOR THE VARIATIONS IN CLOUD COVER

Henrik Madsen and Poul Thyregod

The Institute of Mathematical Statistics and Operations Research
The Technical University of Denmark
DK-2800 Lyngby, Denmark

1. INTRODUCTION

Markov models are widely used to describe climatological phenomena. For example, Coe and Stern (1982) have used first- and second-order Markov chain models for describing rainfall occurrence and Roldan-Cañas et al. (1982) have used first-order Markov models to describe the variation of the horizontal direction of daily wind. Generally, only Markov models in discrete time are considered.

In a previous paper (Madsen et al., 1985) first-order *homogeneous* Markov chain models were used to describe hourly observations of cloud cover at a site near Copenhagen. Models in discrete as well as in continuous time were considered. In order to describe the annual variation a homogeneous model was set up for every month in the year; the diurnal variation of cloud cover, however, was not described by the homogeneous model.

The purpose of the present paper is to show that the diurnal variation of cloud cover can be described by considering *inhomogeneous* Markov models in which the parameters depend on the time of the diurnal cycle. Both discrete and continuous time models are discussed. In analogy with the homogeneous case the matrix of transition rates belonging to the embedded Markov process in continuous time is shown to have a very simple structure, due to physical restrictions for transitions of cloud cover.

2. THE DATA

The investigation is based on 15 years of hourly cloud cover observations, taken at the synoptic observation station at Værløse Airport about 15 km northwest of Copenhagen, Denmark. The observations are total cloud cover in octas in accordance with the WMO code (WMO, 1972, 1974). Thus, the cloud cover is measured as an integer between 0 and 9, where 0 corresponds to completely clear sky, 8 to completely overcast sky, and the observation 9 indicates that the cloud cover is unobservable.

3. THE ANNUAL AND DIURNAL VARIATION

Previously, in Madsen et al. (1985) a very clear annual variation was found. The diurnal variation of the cloud cover mean is illustated in Fig. 1a. The mean belonging to a specific hour in the day and night has been calculated as the average for all twelve months and the 15 years.

The diurnal variation of the cloud cover mean in Fig. 1 shows that it is more cloudy during the day than at night. The diurnal variation is, however, less than the annual variation. An analysis of the diurnal variation of the distribution of cloud cover has shown a higher frequency of clear sky catego-

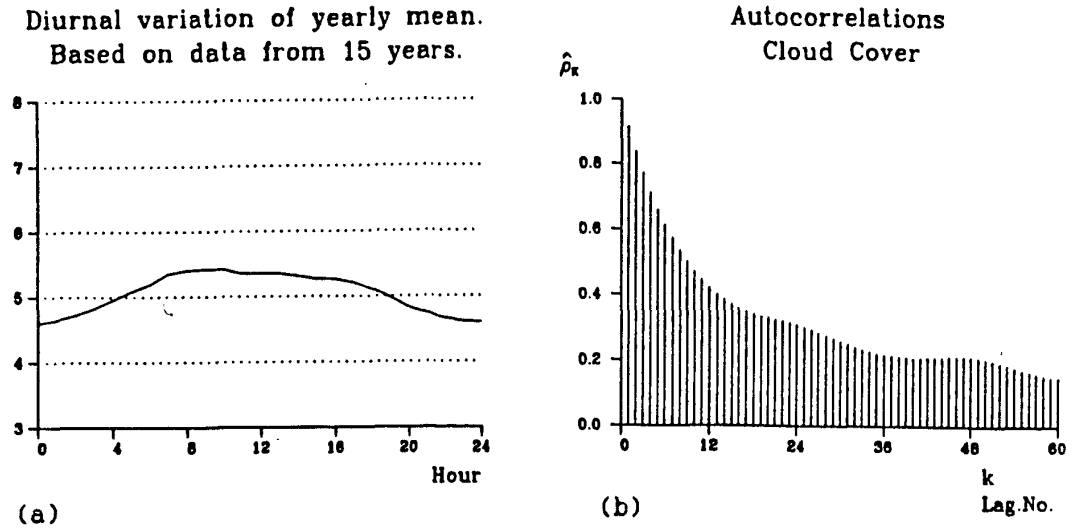


Fig. 1. (a) The overall diurnal variation of the cloud cover mean.
(b) The estimated autocorrelation function.

ries (0-1) at night than during the day. Conversely the frequency of middle categories is smaller at night than during the day. The frequency of overcast sky (7-8) is nearly constant during day and night. Finally the observation (9) which mostly indicates foggy weather is much more frequent at night than at daytime.

4. THE HOURLY CORRELATIONS

Due to inertia of variations according to the dynamics of the physical system the cloud cover observations show correlation in time - the autocorrelation. The estimated autocorrelation function, based on all observations, is shown in Fig. 1b. It is seen that the autocorrelation dies out almost exponentially, but disturbed by a weak 24-hour oscillation. Due to the exponential decaying of the autocorrelation function it is concluded that a major part of the statistical information about the development of cloud cover is contained simply in the present observation. Thus it is natural to consider first-order Markov models for describing the variations of cloud cover observations.

Stochastic processes with parameters that do not vary with time are called homogeneous, otherwise they are called inhomogeneous. In Madsen et al. (1985) the annual variation is described by setting up a homogeneous model for each month in the year. Thus, the set of homogeneous models describe the annual variation and the dominating exponential decay of the autocorrelation function. But, due to the observed diurnal variation of cloud cover and the assumption that the weak 24-hours oscillation in the autocorrelation function is caused by the diurnal variation, an improved first order model must involve parameters varying with the time of the diurnal cycle, i.e. inhomogeneous first-order Markov models have to be considered.

5. A DISCRETE TIME INHOMOGENEOUS MARKOV CHAIN MODEL

In the following the observations will be denoted $X_1, X_2, \ldots, X_N$, where $X_t \in \{0,1,\ldots,9\}$ denotes the cloud cover category at time t. Since cloud cover is measured as an integer it is reasonable to use a model with discrete states, i.e. a Markov chain model.

A Markov chain for describing hourly observations is characterized by the one-hour transition probabilities p_{jk}. In the inhomogeneous case the process is characterized by timevarying one-hour transition probabilities which are defined as (Cox and Miller, 1980)

$$p_{jk}(t) = \text{prob}\{X_{t+1} = k \mid X_t = j\} \tag{1}$$

Considering the variations of cloud cover it is reasonable to expect that the probability for transition from state j to state k at a given hour of the day remain the same for nearby days. As long as observations of cloud cover within a limited part of the annual cycle are considered, it is natural to assume that

$$p_{jk}(t+24) = p_{jk}(t) \tag{2}$$

That is, the transition probabilities, defined by (1), are constrained to be a function of the time s of the diurnal cycle. The matrix of one-hour transition probabilities then becomes

$$P(s) = \{p_{jk}(s)\} \tag{3}$$

Since the observations are by the hour, we have that $s \in \{1,2,\ldots,24\}$. Thus, the conditional likelihood function becomes

$$L(P(1),P(2),\ldots,P(24)) = \prod_s \prod_j \prod_k p_{jk}(s)^{n_{jk}(s)} \tag{4}$$

where $n_{jk}(s)$ is the number of jumps from j at time s to k at time s+1 in the series of observations.

The ML-estimate of $p_{jk}(s)$ is

$$\hat{p}_{jk}(s) = n_{jk}(s) \,/\, \sum_k n_{jk}(s) \tag{5}$$

In discrete time a model can simply be based on the estimates of P(1), P(2),..., P(24). One disadvantage of the model in discrete time is, however, the enormous number of parameters, namely (90 * 24) for each month.

6. CONTINUOUS TIME INHOMOGENEOUS MARKOV CHAIN MODELS

Model and assumptions

In analogy with the homogeneous case described in Madsen et al. (1985), the required number of parameters for a continuous time model is considerably reduced as compared with the discrete time model. The reduction of the number of parameters is due to the fact that in continuous time only changes to neighbouring cloud cover categories are allowed - the cloud cover does not change momentarily from, e.g. state 2 to state 5.

There are, however, exceptions to this principle. A further analysis in Madsen et al. (1985) has shown that transitions between clear sky (0) and foggy weather (9) must be allowed.

Thus, introducing ($t < u$)

$$p_{i,j}(t,u) = \text{prob}\{X(u) = j \mid X(t) = i\} \tag{6}$$

the model is based on the following assumptions for $\Delta u \to 0$:

$$
\begin{aligned}
p_{i,j}(u,u+\Delta u) &= o(\Delta u) \ , \ |i-j| > 2 \\
p_{i,i}(u,u+\Delta u) &= 1 - q_i(u)\Delta u + o(\Delta u) \\
p_{i,i-1}(u,u+\Delta u) &= w_i(u)q_i(u)\Delta u + o(\Delta u) \\
p_{i,i+1}(u,u+\Delta u) &= (1 - w_i(u))q_i(u)\Delta u + o(\Delta u)
\end{aligned} \tag{7}
$$

where $i \in \{0,1,\ldots,9\}$, and $q_i(u)$ is the transition rate and $w_i(u)$ the frequency of transitions to a lower cloud cover category for state i at time u. The notation includes transitions between cloud cover categories 0 and 9, if we define state (-1) = state (9) and state (10) = state (0).

The assumptions (7) leads directly to Kolmogorov's differential equation for inhomogeneous Markov processes (forward equation)

$$
\frac{\partial P(t,u)}{\partial u} = P(t,u)Q(u) \tag{8}
$$

where $P(t,u) = \{p_{i,j}(t,u)\}$. The matrix of transition rates Q(u) has the same structure as in the homogeneous case in Madsen et al. (1985), but the elements are time varying. From (7) the matrix of transition rates Q(u) becomes

$$
Q(u) = \begin{bmatrix}
-q_0(u) & (1-w_0(u))q_0(u) & 0 & \cdots & 0 & w_0(u)q_0(u) \\
w_1(u)q_1(u) & -q_1(u) & (1-w_1(u))q_1(u) & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & w_i(u)q_i(u) & -q_i(u) & (1-w_i(u))q_i(u) & 0 \\
\vdots & & & \ddots & \ddots & \vdots \\
(1-w_9(u))q_9(u) & 0 & \cdots & 0 & w_9(u)q_9(u) & -q_9(u)
\end{bmatrix} \tag{9}
$$

A simple solution to Kolmogorov's differential equation (8) is obtained by assuming that Q(t) is constant in the period [t, t+T[, thus

$$
P(t,t+T) = e^{Q(t)T}P(t,t) = e^{Q(t)T} \tag{10}
$$

For T = 1 hour we obtain the one-hour transition probabilities

$$
P(t,t+1) = P(t) = e^{Q(t)} \tag{11}
$$

By using the assumption (2) it follows that the parameters of Q(t) are restricted to be a function of the time s in the diurnal cycle. The number of parameters for the continuous time model compared to the discrete time model is reduced from 90 * 24 to 20 * 24.

An additional reduction of the number of required parameters is obtained by modelling the diurnal variation of $q_i(s)$ and $w_i(s)$ by simple functions. A consideration of some simple estimates (Madsen, 1985) has shown that it is reasonable to use the first two terms of a Fourier expansion as an approximation to the diurnal variation, i.e.

$$
\begin{aligned}
q_i(s) = q_{i0} &+ q_{i1}\sin 2\pi\tfrac{s}{24} + q_{i2}\cos 2\pi\tfrac{s}{24} + \\
&q_{i3}\sin 4\pi\tfrac{s}{24} + q_{i4}\cos 4\pi\tfrac{s}{24}
\end{aligned} \tag{12}
$$

and analogously for $w_i(s)$. This series expansion of the diurnal variation causes a further reduction in the number of parameters to 20 * 5 for every months. In the next section estimates with and without the Fourier approximation are compared.

A further reduction may be obtained if the diurnal variation is negligible for some cloud cover categories, i.e. $q_i(s)$ or $w_i(s)$ constant for some i.

ML-estimates

From (4) we obtain the logarithm of the conditional likelihood function

$$\log L(P(1),P(2),\ldots,P(24)) = \sum_s \sum_i \sum_j n_{ij}(s) \log p_{ij}(s) \qquad (13)$$

The ML-estimates of $q_i(s)$ and $w_i(s)$ are those which maximize (13). The maximum is found numerically by methods described in Madsen (1985). Due to numerical problems Q(s) is restricted to be constant in periods of 3 hours. The problems are associated with the number of observations which, due to the large number of parameters in the inhomogeneous case, becomes too small in some situations.

An indication of the possible estimation error under the model is obtained by using the second-order derivative of the logarithm of the likelihood function as an estimate of the inverse of the variance matrix for the parameters of Q(s).

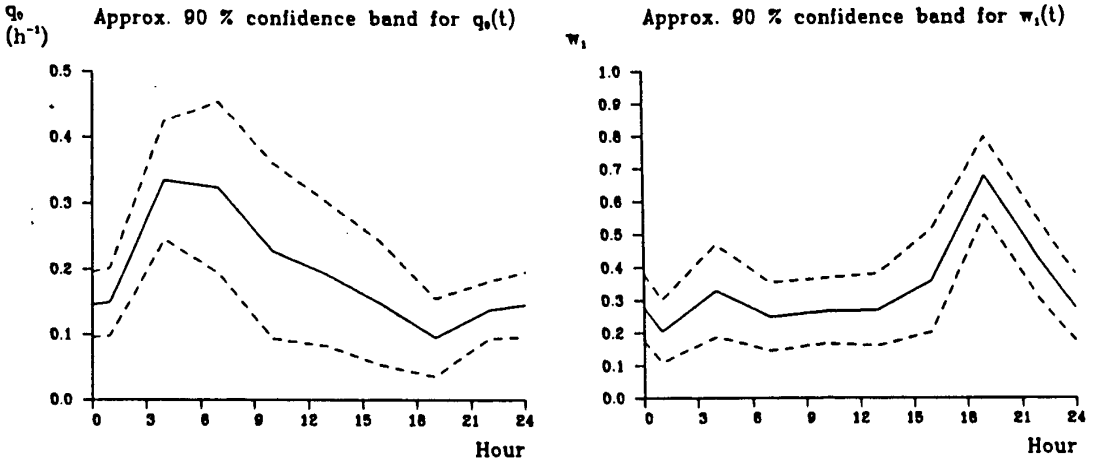


Fig. 2. ML-estimate of $q_0(s)$ and $w_1(s)$ for March. Corresponding confidence bands of approximately 90% are also shown.

The ML-estimate of $q_0(s)$ and $w_1(s)$ are shown for March in Fig. 2, together with confidence bands of approximately 90%. The uncertainty of $\hat{q}_0(s)$ is larger at daytime than at night. This is most likely due to the fact that clear sky (0) is more rarely present during daytime than during nighttime.

The ML-estimates of the variation of $q_0(s)$ and $w_1(s)$, modelled by the first two terms in the Fourier expansion (12), are shown in Fig. 3. In order to illustrate how the ML-estimate of the first two terms of the Fourier expansion forms the diurnal variation, the ML-estimates, described previously, are also depicted.

Considering the uncertainty of the estimates illustrated in Fig. 2, the Fourier expansion of the diurnal variation seems to be a useful way to reduce the number of required parameters.

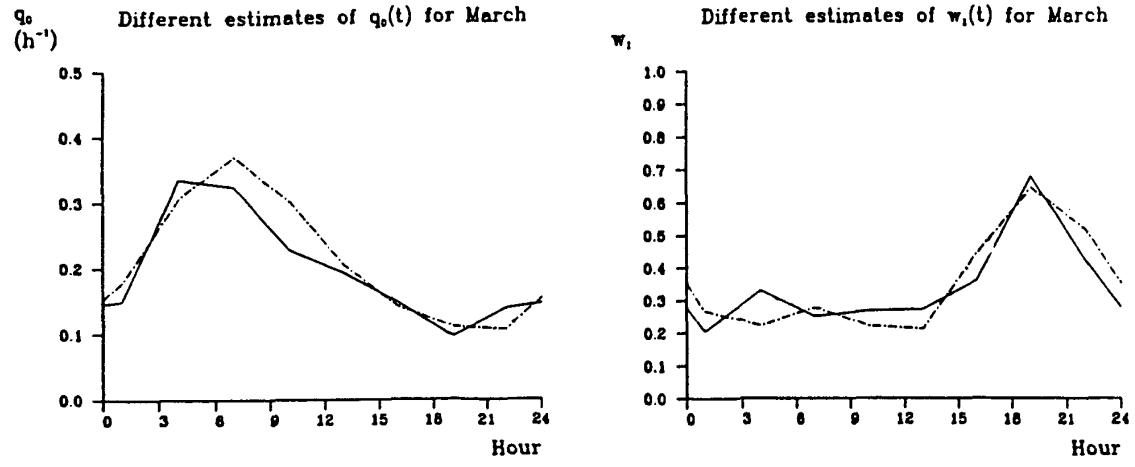


Fig. 3. Estimates of $q_0(s)$ and $w_1(s)$. ML-estimates (——) and ML-estimate of the Fourier expansion (12) (-·-·-).

7. SUMMARY

In a previous work (Madsen et al., 1985) homogeneous Markov chain models are considered, and the discussion is then focusing on the advantage of considering Markov models in continuous time. Using the homogeneous approach the annual variation is described by setting up a model for every month in the year, whereas the diurnal variation is neglected. In the present paper it is shown that the diurnal variation can be described by considering inhomogeneous models where the parameters are restricted to be functions of the time of the diurnal cycle. Both models in discrete and continuous time are considered. The model in continuous time is shown to have a very simple structure due to physical restrictions for transitions of cloud cover, and the number of parameters is considerably reduced by using the continuous time model compared to the discrete time model. An additional reduction of the number of parameters is obtained by modelling the diurnal variation of the parameters in the continuous time model by some terms of the Fourier expansion.

REFERENCES

Coe, R. and R. D. Stern, 1982: Fitting Models to Daily Rainfall Data. *J. Appl. Meteor.*, 21, 1024-1031.

Cox, D. R. and H. D. Miller, 1980: *The Theory of Stochastic Processes* Chapman and Hall Ltd, 398 pp.

Madsen, H., 1985: *Statistically Determined Dynamical Models for Climate Processes.* Thesis. IMSOR, The Technical University of Denmark, DK-2800 Lyngby. 428 pp.

Madsen, H., H. Spliid, and P. Thyregod, 1985: Markov Models in Discrete and Continuous Time for Hourly Observations of Cloud Cover. *Journal of Climate and Applied Meteorology*, Vol. 24, pp. 629-639.

Paltridge, C. W. and C. M. R. Platt, 1976: *Radiative Processes in Meteorology and Climatology.* Elsevier, 318 pp.

Roldan-Cañas, J., A. Garcia-Guzman and A. Losada-Villasante, 1982: A Stochastic Model for Wind Occurrence. *J. Appl. Meteor.*, 21, 740-744.

World Meteorological Organization, 1972, 1974: *Manual on Codes. Vol. 1 and 2.* WMO, Post Box No. 5, CH-1211, Geneva 20, Switzerland.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

Time dependent models of climatological data

Eiichi Suzuki, Masanori Miyata and Shigeru Hongo
Information Science Research Center
Aoyama Gakuin University
Shibuya, Tokyo 150, Japan

Introduction

As the main time dependent models of climatological change process and prediction, both physico-statistical equations and stochastic-statistical models have been proposed by many authors. However, the former were summarized in detail by the WMO Report "Physical Basis for Climate Prediction(1982)" already, so the latter would be more desirable to be reviewed in the abridged forms.

Fundamental parts of the former equations are taken up rather simply at first.

1. Fundamental physical equations

(a) Climatic time-change of temperature deviation T(t) from the quasi-steady state will be written as

$$\frac{dT(t)}{dt} + \frac{T(t)}{\tau^*} = F(t) \qquad (1.1)$$

in the earth-scale as has been shown by North, G.R. and R.F. Cahalan (1981) already, where τ^* is a climatic characterized factor and F(t) is a stochastic irregular external force constrained by the following two conditions:

$$\langle F(t) \rangle = 0, \quad \langle F(t)F(t') \rangle = f^2\delta(t-t')$$

where $\langle\rangle$ is ensemble average and δ is Diracs' delta function. Then we have the next non-stationary standard deviation of T.

$$\sigma(t) = \langle (T-\langle T \rangle)^{1/2} = f(\tau^*/2)^{1/2}(1-\exp(-2t/\tau^*)^{1/2} \qquad (1.2)$$

(b) Climatic time-change based on energy balance can be written as

$$C\frac{\partial T(\hat{\ell},t)}{\partial t} - R^2 D\nabla^2 T(\hat{\ell},t) + BT(\hat{\ell},t) + A = QS(\hat{\ell})a(\hat{\ell}) \qquad (1.3)$$

by North, G.R. (1975), where C is heat capacity over the unit earth-surface, $\hat{\ell}$ is unit vector towards the objective earth-surface from the center of the earth, $T(\hat{\ell},t)$ is climatological temperature of such the surface at time t, R is radius of the earth, $BT(\hat{\ell},t)+A$ is vertical heat diffusion by Radiation-Law obtained by Budyko, Q is solar constant/4, S is annual mean radiation and a is surface albedo respectively. Under the replacement $F(\hat{\ell},t)=C^{-1}QS(\hat{\ell})a(\hat{\ell})$, $F(\hat{\ell},t)$ can be usually expressed

by the expansion of spherical harmonics, and the several physico-stastical equations are contrived by the several theoretical climatologists. (All of the main results are omitted here.)

2. Special cases of ARiMA model

As is well known, ARiMA (integrated Auto-Regressive and Moving Average) model is proposed by Box, G.E.P. and G.M. Jenkins(1976), and it is an effective model applicable for non-stationary time series such as climatic change. However, no direct application of this model is seen in statistical climatology at present, and its special cases (ARMA, AR etc.) have been applied under the assumption of no trend as follows:

(a) Brown,B.G., Katz,R.W. and A.H.Murphy (1984) proposed a general model for simulate and forecast windspeed and windpower. These climatological factors have the diurnal non-stationary character and non Gaussian distribution generally, but as the actual results of suitable order estimation, they adopted a AR(p) process having Yule-Walker coefficients $\hat{\phi}_1$, ... ,$\hat{\phi}_p$ approximately.
After all, point forecast is possible by using of Yule-Walker's recurrsion procedure.

(b) Finzi,G., Bonelli,P. and G.Bacci (1984) studied a stochastic treatment of surface wind speed by applying a simple ARMA model and gave its application example to serve for air quality-control purpose.

(c) K.E.Trenberth (1985) investigated widely persistence of daily geopotential heights at 1000mb and 500mb over the Southern Hemisphere, for both Summer and Winter. And he obtained the next main results:

(1) low-order Auto-Regressive (AR) models fit to the data, generally.
(2) a red noise AR(1) model provides the best fit especially in midlatitudes.
(3) AR(2) and AR(3) processes provide the best fit for local cases. The next well-known appendix is commonly referred by him.

appendix : second-order discrete and continuous AR processes can be shown as follows:

$$x_t-\mu = \alpha_1(x_{t-1}-\mu)+\alpha_2(x_{t-2}-\mu)+\varepsilon_t \qquad (2.1)$$

$$a_2\frac{d^2x}{dt^2}+a_1\frac{dx}{dt}+a_0\{x(t)-\mu\} = \varepsilon(t). \qquad (2.2)$$

where ε_t is purely random process.
If centered finite differences are used, then we have

$$a_2 = (1-\alpha_2)/2,\ a_1 = (1+\alpha_2)/2,\ a_0 = 1-\alpha_2-\alpha_1. \qquad (2.3)$$

Moreover if we set

$$\beta^2 = 2\alpha_1(1-\alpha_2)-1+6\alpha_2-\alpha_2^2$$

then complementary functions for (2.2) are as follows:

(i) if $\beta^2>0$,

$$x(t) = A\cdot\exp\{-(1+\alpha_2+\beta)/(1-\alpha_2)t\} + B\cdot\exp\{-(1+\alpha_2-\beta)/(1-\alpha_2)t\} \quad (2.4)$$

(ii) if $\beta^2<0$ and $\gamma^2=-\beta^2$

$$x(t) = \exp\{-(1+\alpha_2)/(1-\alpha_2)t\}\{A\cdot\cos\frac{2\pi t}{P} + B\cdot\sin\frac{2\pi t}{P}\} \quad (2.5)$$

where $P = 2\pi(1-\alpha_2)/\gamma$ is quasi-periodicity.

(d) Pao-shin Chu and R.W.Katz (1985) tried the application of ARMA process to Southern Oscillation (SO) fluctuations, in order to verify the index consisting of difference of normalized sea-level pressure departures between Tahiti and Darwin.
SOI (Southern Oscillation Index) are contributed by many climatologists.
Time series modelling and forecasting are based on ARMA(p,q) as follows :

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t - \theta_q a_{t-1} - \cdots - \theta_q a_{t-q} \quad (2.6)$$

a_t : white noise process $\sim N(0,\sigma_a^2)$.

(2.6) is simply shown by

$$\Phi(B)X_t = \Theta(B)a_t$$

where B is backward shift operator.

Baysian Information Criterion is written by

$$BIC(p,q) = N\cdot\ln\hat{\sigma}_a^2(p,q)+S\cdot\ln N \quad (2.7)$$

where N:total sample size,

$\hat{\sigma}_a^2(p,q)$:estimated variance of error term by fitting ARMA(p,q)

S:total number of parameters.

Forecasting of this model is formulated by

$$\hat{X}_t(\ell) = \phi_1 X_{t+\ell-1} + \cdots + \phi_p X_{t+\ell-p} + a_{t+\ell} - \theta_1 a_{t+\ell-1} - \cdots - \theta_q a_{t+\ell-q} \quad (2.8)$$

Variance of forecasted value $V(\ell)$ is given by

$$V(1)=\sigma_a^2, \quad V_\ell=(1+\sum_{j=1}^{\ell-1}\psi_j^2)\sigma_a^2, \quad \ell=2,3,\cdots \quad (2.9)$$

(refer Box. Jenkins (1976) p137)

As actual cases, the next ARMA(1,7;1)

$$\hat{X}_t = \phi_1 X_{t-1} + \phi_7 X_{t-7} + a_t - \theta_1 a_{t-1}$$

is applied for SOI.

3. Markov model and persistence model

(a) H.Madson et al (1984) applied Markov models in discrete and continuous time series for hourly observations of cloud cover (hour to hour correlation $\hat{\rho}=0.92$).

As the discrete model, let us observe $X_1, X_2, \cdots, X_n$, then put

$$p_{jk} = P_r(X_{t+1}=k \mid X_t=j).$$

We have $L(p)=\prod_j\prod_k p_{jk}^{n_{jk}}$ as likelifood function and M.L. estimator $\hat{p}_{jk}=n_{jk}/n_j$. respectively. Moreover, they propose the following well-known continuous models simply:

$$\begin{aligned} &P_{ij}(\Delta t)=o(\Delta t) && \text{for } |i-j| \geq 2 \\ &p_{ii}(\Delta t)=1-q_i\Delta t+o(\Delta t) \end{aligned} \qquad (3.1)$$

and in general, for cloud cover matrix

$$\frac{dP(t)}{dt} = P(t)\cdot Q, \qquad P(o)=I \qquad (3.2)$$

is considered by this paper.

(b) R.M. Dole (1986) tried space analysis of EOF by using the data of Extratropical Northern Hemisphere circulation in winter at first, and investigated persistence and variability in time averages, in order to clarify persistent anomalies of the upper data secondly.

(c) Perrone,T.J. and R.G. Miller (1985) has shown the various relations between GEM (Generalized Exponential Markov) model and MOS (Model Output Statistics), to use for a comparative verification of statistical forecasting.

(d) K.E. Trenberth (1984) studied persistence of daily geopotential heights over the Southern Hemisphere, and verified some effects of finite sample size and persistence on meteorological statistics with some examples.

4. Primitive decomposition models

Generally, the following two decompositions are well known:

$$\begin{aligned} X(t) &= T(t)\cdot C(t)\cdot S(t)\cdot I(t) \\ X(t) &= T(t)+C(t)+S(t)+I(t) \end{aligned} \qquad (4.1)$$

where, X(t) is obtained time series, T(t) is deterministic trend, C(t) is cyclic movement, S(t) is seasonal variation and I(t) is irregular disturbance respectively.

The latter case is mainly taken up by climatologists and functional expressions of T(t) are shown by many authors:

Now, the several examples are introduced as follows:

(a) Kingtse C. Mo and H.V. Loon (1985) assumed the two guadratic trends for sea-surface pressure and 500mb height of main world stations in winter season

$$Z(t) = \bar{Z}+a+bt+ct^2 \qquad (4.2)$$

where $\bar{Z}$ is the 31 years mean of period 1951-1981, a, b and c are computed by the ordinary least squares.

Moreover, the same expressions are applied for geopotential height field, surface air temperature and upper air temperature field etc,.

(b) Swanson,G.S. and K.E. Trenberth (1981) verified the similar deterministic trends in the southern hemisphere troposphere circulations.

(c) Vining,K.C. and J.F. Griffiths tried trend analysis based on regression line and performed special analysis of the ranked monthly maximum to investigate climatic variability at ten stations across the US having the data 1900 before.

(d) T.A. Buishand (1985) considered the effect of seasonal variation S(t) and serial correlation $\bar{\theta}\tau$ on the extreme value distribution of rainfall data with their numerical examples.

5. Spectrum and cross (auto) correlation model

(a) Landin,M.G. and L.F. Bosart (1985) have proposed application of harmonic analysis and have shown the numerical results for diurnal variability of precipitation in the Northern US. They called Rainfall synopis such numerous Tables and Figures obtained in winter and spring.

(b) K.E. Trenberth (1984) listed up the two methods A and B of auto correlated time series as follows:

In general, sample auto-cov.: $C_L = \frac{1}{N}\sum_{i=L+1}^{N}(x_{i-L}-\bar{x})(x_i-\bar{x})$,

Variance: $C_0 = \frac{1}{N}\sum_{j=1}^{N}(x_j-\bar{x})^2$

then, they have $\hat{\theta}_L = C_L/C_0$.

Method A: $x_{i,j}$ is the i-th value of the j-th sample, method A defines

$$C_{Lj} = \frac{1}{N}\sum_{i=L+1}^{N}(x_{i-L,j}-\bar{x}_j)(x_{i,j}-\bar{x}_j)$$

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N}x_{i,j} \qquad \text{(sample mean)}$$

sample auto-correlation at lag L is

$$\hat{\gamma}_{L,j} = C_{Lj}/C_{0j}, \quad \hat{\gamma}_L = \frac{1}{J}\sum_{j=1}^{J}\hat{\gamma}_{L,j}$$

In contrast, method B defines

$$C_{Lj} = \frac{1}{N}\sum_{i=L+1}^{N}(x_{i-L,j}-\bar{x})(x_{i,j}-\bar{x})$$

$$\bar{x} = \frac{1}{NJ}\sum_{j=1}^{J}\sum_{i=1}^{N}x_{i,j}$$

$$C_L = \frac{1}{J}\sum_{j=1}^{L}C_{Lj}, \qquad \hat{\gamma}_L = C_L/C_0$$

For analysis of red noise time series

$$x_t = \alpha x_{t-1}+\varepsilon_t \quad (\alpha=0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9) \tag{5.1}$$

The next result is obtained by method A:
Approximate expected value of C_{Lj} is

$$\langle C_L\rangle \approx \frac{N-L}{L}\left(R_L-\frac{T_0}{N}\right)\sigma^2$$

$$\approx \frac{N-L}{L}\left[\left\langle\frac{1}{N-L}\sum_{i=L+1}^{N}x_{i-L,j}x_{ij}-\langle\bar{x}_j^2\rangle\right\rangle\right] \tag{5.2}$$

where R_L is population correlation of lag L.

(c) Rinne,J and S. Järvernoja (1985) proposed a method based on error spectrum. Let the analysis of the 500mb geopotential height be given in term of empirical orthogonal functions (EOF)

$$Z(i,t) = Zm(i)+\sum_{\nu=1}^{175}C\nu(t)f_\nu(i) \tag{5.3}$$

where Z(i,t) represents height, Zm(i) is a long term mean of Z and t, i refer to time and grid point.
Then, we have the next coefficients:

$$C_\nu(t) = \sum_i [Z(i,j)-Zm(i)]f_\nu(i)dA_i \tag{5.4}$$

where dA_i is areal element.
Here, if random error ε_ν, systematic error ε_ν, syst. are taken into account, then analysis error is generally included in C_ν coefficients as follows:

$$C_{\nu,anal}(t) = s_\nu\cdot C_{\nu,true}(t)+S_\nu\cdot\varepsilon_\nu(t)+S_\nu\cdot\varepsilon_{\nu,syst} \tag{5.5}$$

6. Climatic data analysis and prediction

T. Schlatter (1975) tried some experiments with a multi-variate objective analysis and R.A. Cristensen (1985) proposed Information Theoretical Models (ITM), in order to do seasonal precipitation forecasting with a 6~7 month lead time in the PN (Pacific Northwest) districts.
P.D. Thompson (1985) indicated the two stages of predictions, i.e., prediction of probable errors of climatological predictions.

Concluding remarks and future problems

At last we would like to point out the following items and future problems:

(1) Special cases of ARiMA are applied for climatic change, so direct and modified ARiMA will be expected to be further applied, since trend of ARiMA is simply eliminated by sequential difference.

(2) Two suggestive trend models are not yet used at present, the one of them is normal regression theory $X_t \sim N(\sum_{i=1}^{k} a_i \psi_i(t), \sigma^2)$ wherer $\{\psi_i(t)\}$ have the wide generality (Polynorminal, Fourier series and the other orthogonal function expansions are possible to be use as special cases), the other of them is stochastic trend defined by M.H. Quenouille (1968) as follows:

$$\mathbf{X}_t = \mathbf{Y}_t + \mathbf{Z}_t, \quad \mathbf{Y}_t = U_1 \mathbf{Y}_{t-1} + A_0^{-1} \varepsilon_t$$

$$\mathbf{Z}_t = \mathbf{Z}_{t-1} + B_0^{-1} \eta_t \qquad (\rho(Z_{i\,t}, Z_{j\,t+s}) = \pm 1)$$

where $\mathbf{X}_t$ is non-stationary observation, $\mathbf{Z}_t$ is trend, $\mathbf{Y}_t$ is random variation and both ε_t and η_t are irregular fluctuations respectively.

(3) Kalman filter and smoother models would be possible to use for the case of non-stationary processes, but their applications are not seen in the latest years.

(4) Analysis of persistence, spectral analysis and Markov chain modelling are well studied by many authors as before, but synthetic multi-variate time series analysis is not always applied sufficiently for climatological data, even if EOF is frequently taken up. For example, cross covariance (or correlation) matrix with time lag $\tau(>0)$ would be powerful statistic in future.

After all, as has been pointed out by P. Newbold (1984), the matrix statistic

$$\Gamma_\tau = E(\mathbf{X}_t \cdot \mathbf{X}'_{t-\tau}), \quad E(\mathbf{X}_t) = 0$$

obey the next difference equation

$$\Gamma_\tau = \phi_1\Gamma_{\tau-1} + \cdots + \phi_p\Gamma_{\tau-p} \qquad (\tau=q+1, \cdots)$$

and decision of p and rational estimations of $\phi_1,\cdots,\phi_p$ are important problems in future, I think.

References

In order to prepare this report, an enormous collection of copied papers related to our title during the recent five years, by using the following main research magazines:

(1) Monthly Weather Review
(2) Journal of Climate and Applied Meteorology
(3) Journal of Applied Meteorology
(4) Journal of Atmospheric Sciences

Thus, listing of each reference is suspended, since all of suggested papers introduced by this report could be find in either one of these magazines and also for saving pages.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

PERSISTENCE OF THE WET SPELLS IN UGANDA

L.J. Ogallo and E. Bazira

Department of Meteorology
University of Nairobi, Kenya

INTRODUCTION

Daily characteristics of the wet and dry spells have significant influence on all rain-dependent activities. In Uganda where the economy depends entirely on the exports of rain-dependent agricultural products, the seasonal characteristics of the wet/dry spells generally determine the quality and quantity of the agricultural exports.

A part from determining the availability of water for agricultural use,the spatial and temporal patterns of the spells also influence day to day agricultural activities like ploughing, planting, weeding, fertilizer applications and many others. They also have significant influence on water and wind erosions, together with the occurrence and spread of many diseases and pests.

In Uganda the seasonal distribution of the wet/dry spells also determines the amount of water available for domestic and industrial use, and for hydroelectric power generation.

The frequency and probability of wet spells over parts of East Africa have been examined by Alusa and Gwage (1978), Ogallo and Chillambo (1982), Oteng'i and Ogallo (1984). In this study the following components of the wet spells in Uganda were investigated (i) Probability of a wet spell lasting for at least k days.

(ii) The persistence of the wet spells (weather the occurrences of wet spells depends on rainfall events in the previous days). (iii) The use of Markov Chain models in describing the occurrences of wet spells.

Uganda is enclosed within latitudes 5°N - 1°S and longitudes 29°E - 36°E. The typical patterns of seasonal rainfall in Uganda are given in figures 1a - 1d. Most of the county receives substantial rainfall throughout the year due to the influence of topography, and the existence of many large lakes.

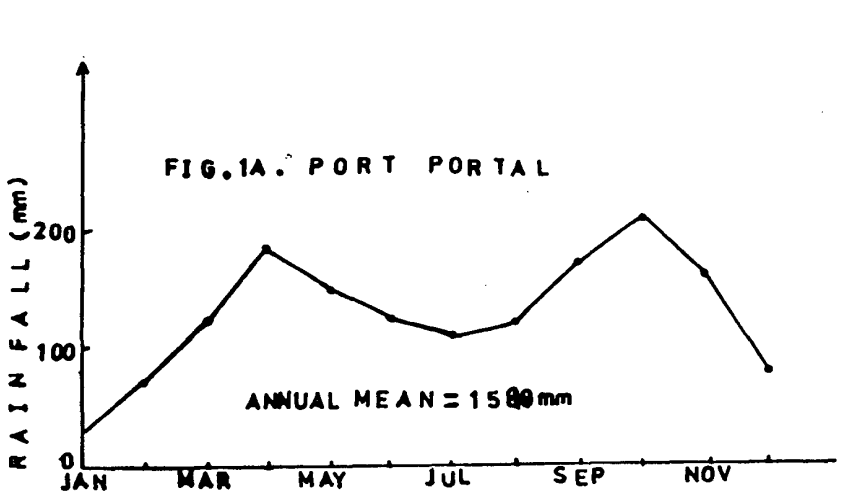
The rainfall peaks centred around March-may and September-November are associated with the seasonal movements of the Inter Tropical Convergence Zone (ITCZ). The July/August peak is however related to the westerly influx of moisture from Atlantic and the moist Congo/Zaire basin.

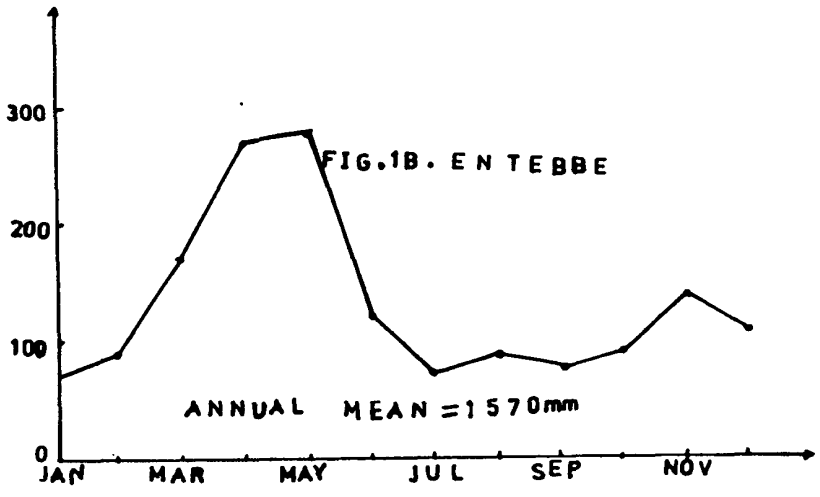
Figure 2 gives the network of the rainfall stations used. The distribution of the thirteen stations were based on the homogeneous regional grouping by Ogallo (1980). The daily rainfall records were for the period (1948 - 83). A thresh hold value for a rainy day (wet spell) was taken as 1 mm.

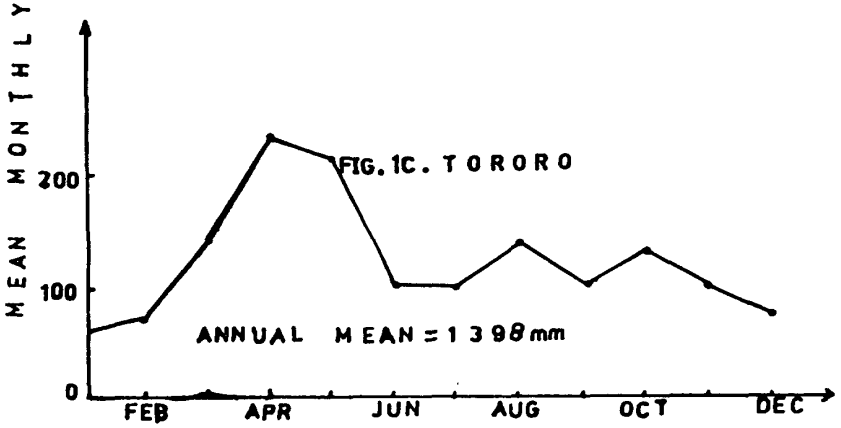
PROBABILITY OF OCCURRENCE

The daily rainfall records were used to derive the monthly frequencies of the wet spells. The frequency (F_j) of all wet spells in any given month j may be expressed as:

$$F_j = \sum_{i=1}^{n} \sum_{k=1}^{m} F_{ijk} \qquad (1)$$







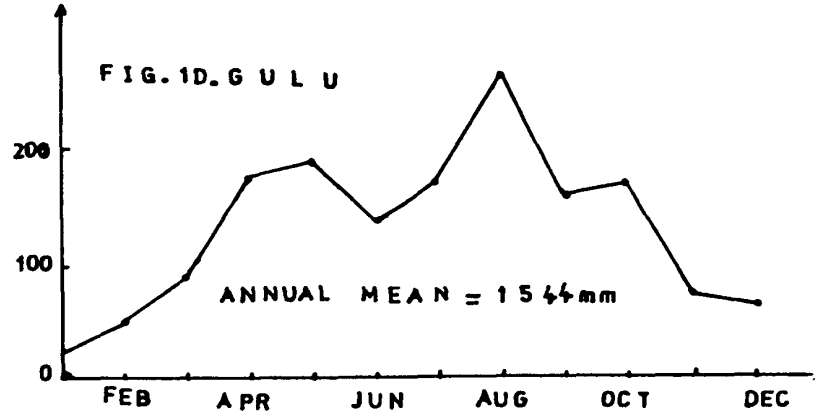


FIG. 1A–1C. MEAN SEASONAL PATTERNS OF RAINFALL

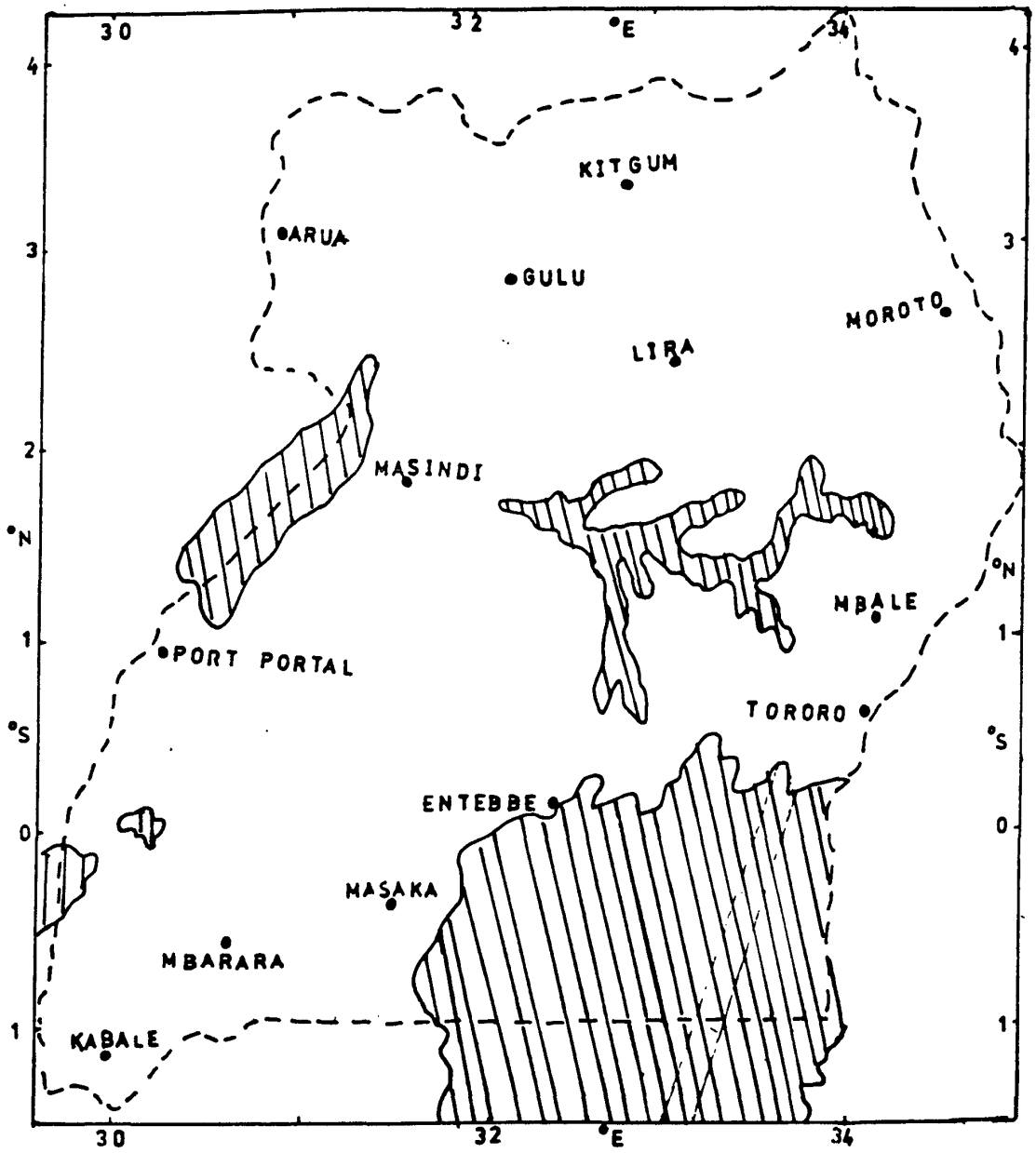


FIG. 2. NETWORK OE THE STATIONS

Where m is the number of years used, and F_{ijk} the frequency of i runs of wet spells during month j and year k.

F_{ijk}, F_j and the other related frequencies can be used to obtain probability limits for the various occurrence of various wet spells. The frequency (F_{ij}) of wet spells lasting at least i days during month

j can for example be obtained usisng cumulative values of F_{ijk};

$$F_{ij} - F_j - \sum^{I-1} \sum_{k=1}^{m} F_{ijk} \qquad (2)$$

Corresponding probabilities are derived from the frquency of the occurrences.

MARKOV CHAIN MODELS

Parameters for Markov Chain models are derived from conditional properties. The occurrences of any rain is assumed to depend on the occurrence or absence of rains during the previous days and independent of the rainfall events in the preceding days.

A Markov Chain of order r is a probability model based only on the rainfall events during the previous r days.

The conditional probabilities for the Markov Chains can be used to derive various properties of rainfall reccurrence. Detailed account of the conditional probabilities and the Markov Chains can be obtained from Parzen (1960), Gabriel and and Neuman (1962), Weiss (1964), Feyerhem and Bark (1965), Todorovic and Woolhiser (1965), Katz (1977) and many others.

The two conditional probabilities P_w and P_D for the first order Markov Chain model may be expressed as:

$$P_w = P(W/W) = F(WW)/F(W) \qquad (3)$$

$$P_D = P(W/D) = F(WD)/F(W) \qquad (4)$$

When P(W/W) and F(WW) are respectively the

probability and frequency of a wet day given that the previous day was wet. P(W/D) and F(WD) are the corresponding probability and frequency when the previous day was dry.

Equations (3) and (4) can be extended to obtain conditional probability of higher order Markov Chains. Gabriel and Neuman (1962), found the distribution of the spells to be Geometric with the probability P(w = k) of a wet spell of length k given as:

$$P(w = k) = (1 - P_w)P_w^{k-1} \qquad (5)$$

In this study conditional probabilities of upto the fifth order of the chain - dependent models were computed. X^2 - test was then used to determine the most adequate model.

PERSISTENCE RATIO

Many methods have been used to determine the degree of persistence of the weather. A simple persistence ratio as given by Brooks and Carruthers (1953) has been used here to further describe seasonal characteristics of rainfall in Uganda. The persistence ratio R may be expressed as:

$$R = (1 - P)/(1 - P(w/w)) - 1 \qquad (6)$$

Where P is the probability of a wet day and P(w/w) retains the previous meaning.

If P = P(w/w), the wet spells would be independent and the value of R = 0. If P(w/w) = 1, the degree of persistence would be high and R = ∞. For P(w/w)<P R will be negative, and this represents a tendency for alternations of the occurrences and non-occurrence (Yao 1983). The ratio R was computed for each month at each station.

RESULSTS AND DISCUSSIONS

Typical patterns exhibited by the frequency of the wet spells are shown in figures 3a - 3d. In general, most of the rainstorms lasted for one day. The frequency of the wet spells then fell off exponentially from one day to the longest duration. The longest spells were observed during the wettest seasons and years. The maximum length of a wet spell recorded in Uganda between 1948 - 83 was 45 days. This was observed at Fort Portal from 21st September to 4th November, 1954.

Seasonal expectations for a rainy day are given in table 1. The observed patterns reflect the seasonal influence of the ITCZ during March - May and September - October, together with the contributions from the westerly moisture influx from Atlantic and Zaire/Congo basin which is at maximum during the months of July/August.

When Markov Chain models were fitted to the sequences of the wet spells first order model was the best fit over many areas, especially during the wet months. Monthly values of conditional probabilities for the first order Markov Chains are given in tables 2 and 3 for the verious locations and months.

The seasonal characteristics of rainfall over the various regions are clearly reflected in these patterns.

The computed values of the persistence ratio were positive throughout the year in many areas. The peak values were centred around the peak rainy months (Figure 4).

	MONTH											
STATION	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEPT	OCT	NOV	DEC
GULU	.08	.15	.36	.57	.60	.49	.53	.66	.62	.65	.43	.20
MOROTO	.04	.09	.16	.33	.34	.22	.36	.33	.20	.18	.12	.08
ENTEBBE	.35	.39	.59	.72	.62	.43	.31	.30	.31	.46	.56	.48
KITGUM	.07	.15	.30	.53	.57	.46	.57	.56	.48	.59	.33	.12
MBALE	.15	.18	.33	.53	.59	.53	.64	.51	.51	.51	.36	.22
ARUA	.11	.22	.33	.48	.46	.46	.56	.64	.62	.63	.44	.16
LIRA	.14	.18	.30	.49	.53	.41	.44	.55	.55	.56	.31	.18
MBARARA	.26	.30	.46	.57	.37	.12	.11	.29	.49	.53	.62	.38
MASAKA	.32	.31	.44	.62	.52	.20	.15	.23	.30	.42	.43	.38
TORORO	.26	.33	.42	.71	.74	.57	.42	.51	.52	.62	.52	.33
FORTPORTAL	.22	.33	.64	.83	.59	.38	.33	.46	.79	.81	.78	.38
MASINDI	.20	.25	.46	.60	.53	.41	.42	.55	.67	.62	.59	.27
KABALE	.38	.51	.60	.75	.46	.18	.11	.26	.42	.55	.74	.55

Table 1. Probability of a rainy day.

	MONTH											
STATION	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
GULU	33	18	39	53	57	44	45	63	50	60	46	45
MOROTO	30	29	42	52	58	35	43	44	43	43	40	50
ENTEBBE	47	43	56	64	60	46	37	26	29	43	61	54
KITGUM	20	27	36	49	45	41	43	43	43	52	46	33
MBALE	27	14	36	51	47	36	51	37	41	43	43	35
ARUA	38	32	42	43	44	38	40	57	46	53	43	33
LIRA	30	29	36	47	46	40	38	45	50	49	45	39
MBARARA	26	32	41	43	73	33	31	38	35	39	46	36
MASAKA	35	26	38	55	47	27	27	24	27	39	47	43
TORORO	32	40	49	60	57	45	36	38	49	49	53	44
FORT PORTAL	38	33	57	62	58	46	40	59	60	66	65	46
MASINDI	40	29	44	52	44	37	42	40	45	47	56	40
KABALE	36	49	52	64	53	46	50	42	49	48	61	45

Table 2: Conditional Probability of a wet spell with previous day wet (X100)

	MONTH											
STATION	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
GULU	05	09	12	21	20	21	23	24	24	20	17	08
MOROTO	02	06	08	12	11	12	16	14	08	08	05	04
ENTEBBE	20	12	21	21	19	18	15	17	17	19	16	17
KITGUM	04	08	14	20	23	20	24	23	20	20	13	06
MBALE	08	12	16	20	24	28	23	24	22	21	16	11
ARUA	06	11	14	19	20	21	23	24	24	22	18	08
LIRA	07	10	15	20	22	18	20	22	20	21	13	09
MBARARA	15	15	20	24	18	06	06	13	23	24	24	18
MASAKA	11	17	21	26	20	11	08	12	17	19	18	17
TORORO	12	15	17	21	23	23	20	23	22	22	18	14
FORT PORTAL	10	15	20	24	20	15	15	14	23	20	20	15
MASINDI	11	13	19	21	22	19	18	24	24	24	19	12
KABALE	18	29	22	20	16	08	05	11	17	21	21	22

Table: 3 Conditional probability of a wet spell with the previous day dry (X100)

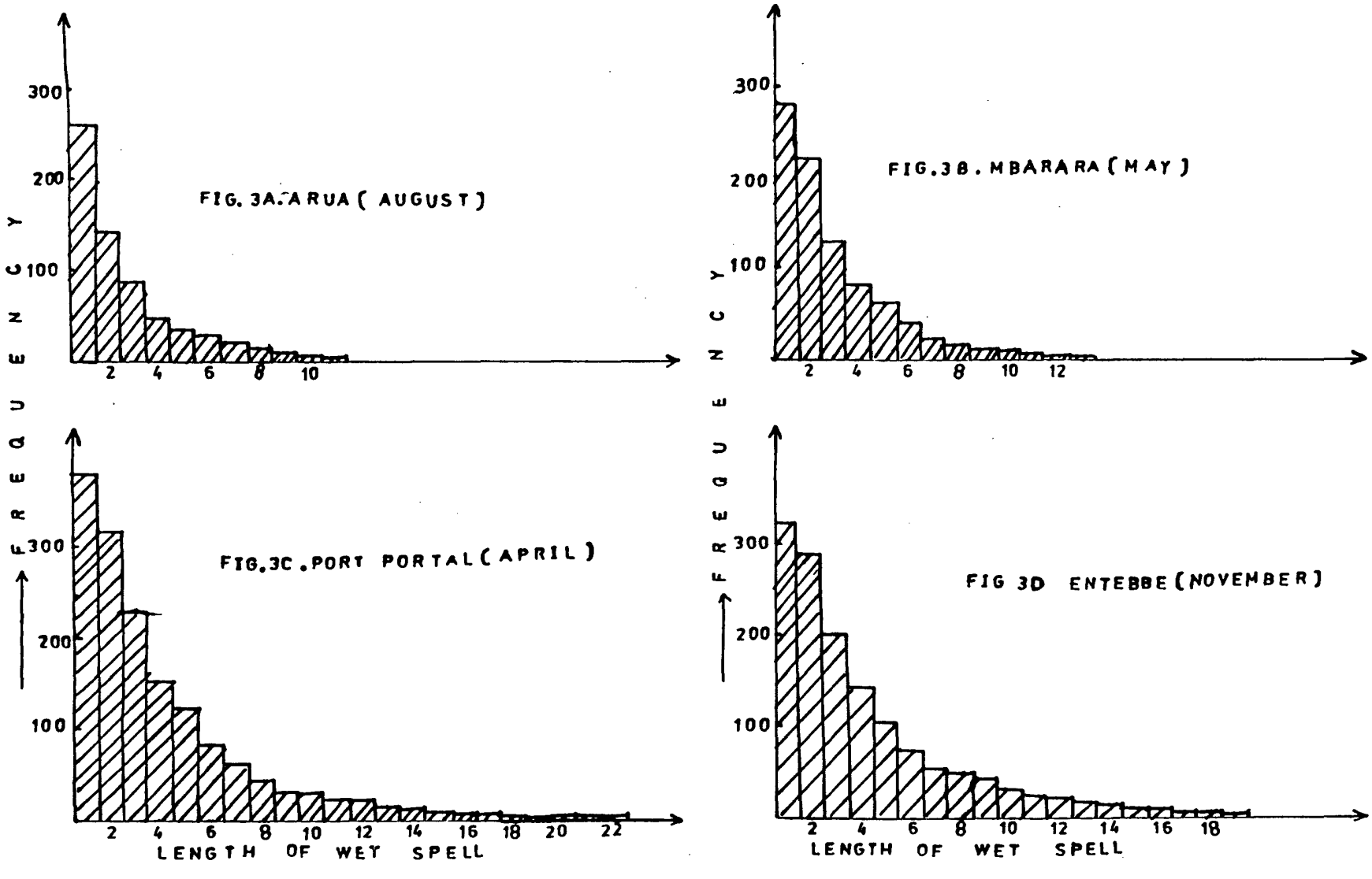
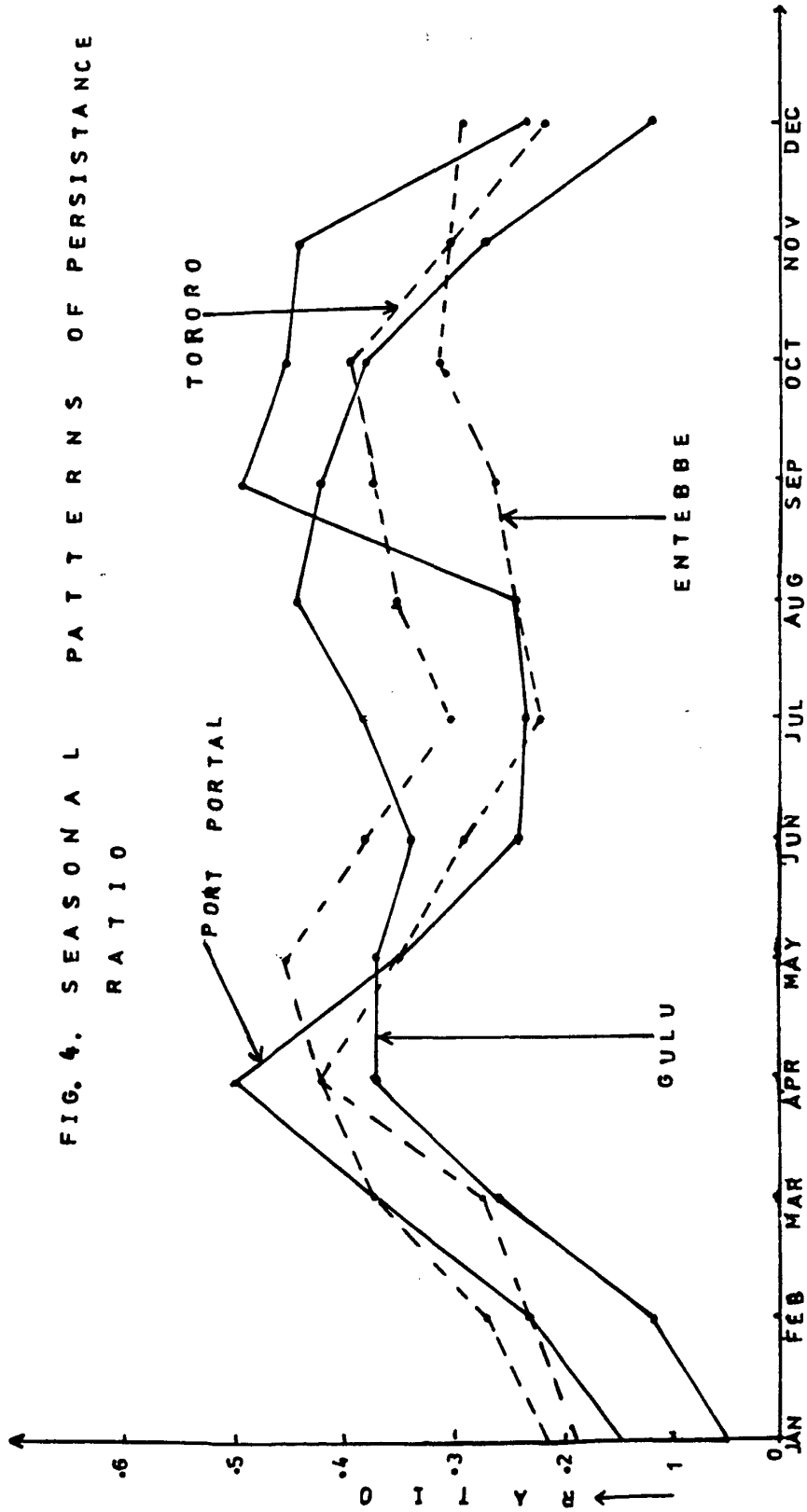


FIG. 3A-3D: OBSERVED FREQUENCIES OF WET SPELLS


FIG. 4. SEASONAL PATTERNS OF PERSISTANCE RATIO
TORORO
PORT PORTAL
ENTEBBE
GULU
RATIO
.6
.5
.4
.3
.2
1
0
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
MONTHS

Large positive values were however obtained throughout the year over some areas which are wet throughout the year. The values were close to zero during the dry months in some areas, indicating the high degree of independence of the wet spells during the dry seasons.

CONCLUSIONS

The results from the study indicate that although many rainstorms lasted for one day, there was a tendency for a wet spell to follow a wet spell, especially during the wet seasons. The first order Markov Chains fitted the sequences of the wet spells better than higher order chain-dependent models.

The prevailing meteorological conditions in Uganda during the rainy seasons may be used to explain the good fit of the chain-dependent models during these months. Rainfall during the peak rainy months of March - May and September - November are associated to the ITCZ. The July/ August rainfall peaks are connected to the influx of moisture from Atlantic and the moist Congo/Zaire basins by the westerly winds. These systems together with the local factors like topography and the large lake generate stable rainfall conditions which often persist for more than one day. They however respond to the day to day fluctuations in the dominant general circulation parameters.

REFERENCES

Alusa, A.L. and P.M. Gwage, 1978. The occurrence of dry spells during the East African long rains. East Afr. Inst. Met. Report 2/78 Kenya.

Brooks, C.P.E. and N. Carruthers, 1953. Handbook of statistical methods in meteorology. HMSO 412 pp.

Feyerhem, A.M. and L.O. Bark. 1965. Statistical methods for persisence precipitation patterns. J. App. Meteor., 4, 320 - 328.

Gabriel, K.R. and J. Neuman 1962. A Markov Chain model for daily rainfall occurrence. Quart J. Roy. Meteor. Soc. 88, 90 - 95.

Katz, R.W. 1977. Precipitation as a chain-dependent process. J. App. Meteor., 16, 671 - 676.

Parzen E. 1960. Modern probability and its applications. Wiley & Sons, New York.

Ogallo L.J. 1980. Regional classification of East African rainfall records into homogenious groups Stat. Climate Developments Afm. Sc., Elsevier, 13, 253 - 261.

Ogallo L.J. and W. A. Chillambo 1982. The characteristics of wet spells in Tanzania. E. Afr. Agric. For J. 47(4), 87 - 95.

Oteng'i S.B. and L.J. Ogallo 1984. Persitence of daily rainfall over parts of Kenya. WMO Regional Scientific conf. on GATE, EAMEX and Tropical Met. in Africa, 10 - 14 Dec. 1984 Dakar Senegal.

Todorovic P. and D.A. Woolhiser 1975. A stochastic model of n-day precipitation. J. App. Meteor. 14, 17 - 24.

Weiss L.L. 1964 Sequences of wet or dry days described by a Markov chain probability model. Mon. Wea. Rev. 92, 169 - 176

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

MARKOV CHAIN MODEL FOR THE CLIMATOLOGICAL OBSERVATION SERIES ON WIND SPEED

Nicolae Romanof, Istvan Elekes, Roland Cotariu
Institute of Meteorology and Hydrology
Bucharest - Romania

INTRODUCTION

The series of values supplied by observations on meteorological elements are framed from a statistical viewpoint as random variable chains generally dependent statistically. The series of values make up samples on account of which the parameters characteristic to some theoretical distribution functions are estimated, the empirical distributions are studied, etc.

The statistical time dependence is studied in terms of the estimates for the autocorrelation function and spectral density or through other methods in the theory of dynamic series. Similarly, the Markov chain model is used with a finite number of states defined according to a classification of the observations(Gabriel and Neuman,1972; McAllister,1969; Gringorten,1972; Romanof and Elekeş,1975; Romanof,1982).

The paper is an application of the Markov model to the study of the statistical interdependence of the observation series consisting of wind speed data recorded at 8o Romanian meteorological stations. For reasons connected with the use of such method in the study of atmospheric pollution, the states are defined as:

state 1: $u \leq 2$ m/s

state 2: $u > 2$ m/s

with u denoting the wind speed.

The conditional probabilities and the probabilities of occurrence for certain state sequences are estimated. The order of the Markov chain is established by means of the χ^2 test using chains of different orders up to the fourth.

A measure for the state 1 persistence is obtained in terms of Besson's persistence coefficients.

MATHEMATICAL MODEL

Let, i_1, $i_2 \ldots i_n$ be the states observed at moments $t_1, t_2 \ldots t_n$; the $i_1, i_2 \ldots i_n$ variables can take a finite number of values,S. Noting down by:
$P(i_n, t_n / i_{n-1} t_{n-1} \ldots i_1, t_1)$ the conditional probability that the in state should be achieved at moment t_n if at moments $t_1 \ldots t_{n-1}$, $i_1 \ldots i_{n-1}$ states were achieved.

If the conditional probability depends on m previous states only, an m-order chain is obtained and consequently:
$P(i_n t_n / i_{n-1}\ t_{n-1} \ldots i_1 t_1) = P(i_n\ t_n / i_{n-1}\ t_{n-1} \ldots i_{n-m}\ t_{n-m})$

Chains of random variables will be considered, derived from observations carried out at equal time intervals and the $t_1 \ldots t_n$ moments will not be noted down in the relations below. In the case of a Markov chain of the m-order the probability for a sequence of n states $i_1 \ldots i_n$ to be achieved is:

$$P(i_1 \ldots i_n) = P(i_1) P(i_2/i_1) P(i_3/i_1, i_3) \ldots P(i_n / i_{n-1} \ldots i_{n-m})$$

The estimation functions for conditional probabilities in the case of a homogeneous (stationary) chain are given below (Anderson and Goodman, 1957; Mihoc and Craiu, 1972).

$$\hat{P}_{i_1,i_2\ldots i_{k-1}i_k} = \hat{P}(i_k/i_{k-1},\ldots i_1) = \frac{n_{i_1\ldots i_n}}{n_{i_1\ldots i_{n-1}}}$$

where, k=2,3,...m, and
$n_{i1\ldots ik}$ is the number of sequences of states $i_1,i_2,\ldots i_k$ estimated on account of a sample of n lengh (much larger than m). When n tends to the infinite the value $Pi_1\text{-}i_k$ converges towards the conditional probability $Pi_1\text{-}i_k$.

The variable:

$$\chi^2 = \sum_{i_1,\ldots i_{m+1}=1}^{s} n_{i_1\ldots i_m} \frac{\left(\hat{P}_{i_1\ldots i_{m+1}} - \hat{P}_{i_2\ldots i_{n+1}}\right)^2}{\hat{P}_{i_2\ldots i_{m+1}}}$$

is asymptotically distributed χ^2 with $S^{m-1}(S-1)^2$ degrees of freedom.

Using the χ^2 test and the χ^2 estimation the "m-1 order chain" hypothesis is tested as against its alternative hypothesis the "m order chain". The measure for an i state persistence is given by Besson's persistence coefficient (Brooks and Carruthers, 1953).

$$r_B^i = 1 - \left(\frac{1-P_{ii}}{1-P_i}\right)^2$$

where, $r_B^i = 0$ corresponds to the lack of persistence,
and $r_{B-1}^i = 1$ indicates a maximum persistence.

RESULTS

In order to represent graphically the statistical parameters evaluated by means of the computation model presented above, a rectangular zone was defined and covered by a square grid of 24×24 points. In each grid point a value was estimated from data supplied by one or more meteorological stations. A meteorological station was considered to possibly contribute to maximum 9 grid points around it, the value in that point being computed by means of the formula below:

$$V = \frac{\sum_{i=1}^{n} V_i\, e^{-\lambda d_i^2}}{\sum_{i=1}^{n} e^{-\lambda d_i^2}}$$

where, V is the value in the grid point, n is the number of meteorological stations possibly contributing to the grid point being considered, V_i are the values at these stations and di the distances between the grid point and those stations that contribute to the grid point.

Parameter λ was determined experimentally from similar graphical representations was assigned the value of /8/.

In order to obtain a graphical representation, a CDC 9335 type of graphical linear print was achieved capable to cover about 95o points along a line. For the representation of the results in our experiment 54o lines were used having 672 points on each(six vertical points were included on each line). Thus the values got denser in each grid square, that is 784 intermediate values could be computed for each grid square. A two-linear function of the form below was used to get the values denser /8/.

$$v = ax + by + cxy + d$$

where, v is the value in an intermediate point while the a,b,c and d coefficients are determined as follows: a rectangular system of axes is considered for each grid point with values v1,v2,v3 and v4 as corner values (v1 being in the origin of the coordinate axes, while the others are opposing the clockwise sense) and the restrictions below are forewarded:

$$\begin{array}{llll} x = o & \text{and} & y = o & v = v1 \\ x = 1 & \text{and} & y = o & v = v2 \\ x = o & \text{and} & y = 1 & v = v4 \\ x = 1 & \text{and} & y = 1 & v = v3 \end{array}$$

This way the coefficients will take the form:

$$\begin{aligned} a &= v2 - v1 \\ b &= v4 - v1 \\ c &= (v3 + v1) - (v2 + v4) \\ d &= v1 \end{aligned}$$

In order to draw the isolines the values of the parameters being considered were classified into seven intervals obtained in terms of the ratio $\frac{\log_{max} - \log_{min}}{7}$ where $\log_{max}$ and $\log_{min}$ are the logarithms of the extreme values of the parameters.

A particular problem was raised by the points situated at a distance two times longer than the diagonal of the square grid when no climatological station contributes data to it (i.e. points outside the country's territory. In drawing the isolines in the neighbourhood of these points all the grid points were initialized at the value of o.ooool. After the values in the grid points were assessed, the initial value of o,ooool in the points lacking contributions from climatological stations and situated within a distance equal to the diagonal of a grid square was replaced by a value determined through the formula:

$$V = - \frac{\sum_{i=1}^{n} V_i \, e^{-\lambda d_i^2}}{\sum_{i=1}^{n} e^{-\lambda d_i^2}}$$

where, V_i are the values in the points situated at a maximum distance of a diagonal from the point being considered and the d_i are the distances from this point.

The graphical representation was achieved by means of a FELIX C-1024 type of computer along the following stages: (i) according to a programme written in FORTRAN language the statistical parameters were estimated and the results entered a data bank, (ii) a programme written in FORTRAN language was used to compute the values in the grid points as well as the ones obtained in squares with denser values,(iii)the graphical representations were achieved by means of a CDC 9335 connected on-line to the data bank.

The model input data consisted of series of values on wind speed measured four times a day at o1.oo;o7.oo;13.oo;19.oo o'clock, local time, recorded in the time interval 1961-1985 at 8o stations.

The estimations performed put forth the fact that the series being analysed behave as Markov chains of at least the fourth order (significance level 5%). Provided the hypothesis of the independence of events $P(i_1,...i_n)$ was computed and it was noticed that the value differs greatly from the value estimated directly from the series of data $\hat{P}(i_1,...i_n)$, $n \leq 5$.

The same computation was performed considering a first order Markov chain and led to similar values for p and $\hat{P}$ although the χ^2 test led to chains of at least the fourth order. Over the Romanian territory the variability of the estimated statistical parameters was emphasized: P_1 ; P_{11}, P_{22}, r_B', etc. Some of these parameters are directly represented on the isolines automatically drawn for the entire surface of Romania.

References

Anderson, T.W., Goodman, L., 1957: Statistical inference about Markov chains, Ann Math. Statist. 28, 89-110.

Brooks, C.E.P., Carruthers, N., 1953: Handbook of Statistical Methods in Meteorology, Her Majesty's Stationary Office, London.

Drăghici, I., Pepasian, I., 1974: Contribuţii la rezolvarea problemei trasării automate a izoliniilor diverselor cîmpuri meteorologice (Contributions to the solution of the problem of automated isoline drawing for meteorological fields). Culegere de lucrări IMH, 1974, 47-58.

Gabriel, K.R., Neman, I.A., 1962: A Markov chain model for daily rainfall occurrence at Tel Aviv, J.R.Met.Soc. 88, 90-95.

Mc.Allister, C.R., 1969: Cloud-cover recurrence and diurnal variation, J. App. Met. 8, 769-775.

Mihoc, G., Crain, M., 1972: Inferenţa Statistică pentru Variabile Dependente (Statistical inference for dependent variables). Ed.Academiei R.S.România, Bucureşti.

Romanof, N., Elekeş, I., 1975: Sur la dependence du type chaîne Markov des observations sur la vitesse du vent. Meteorology and Hydrology, 1 , 47-51.

Romanof, N., 1982: A Markov chain model for the mean daily SO_2 concentrations, Atm.Env. 16, 1895-1897.

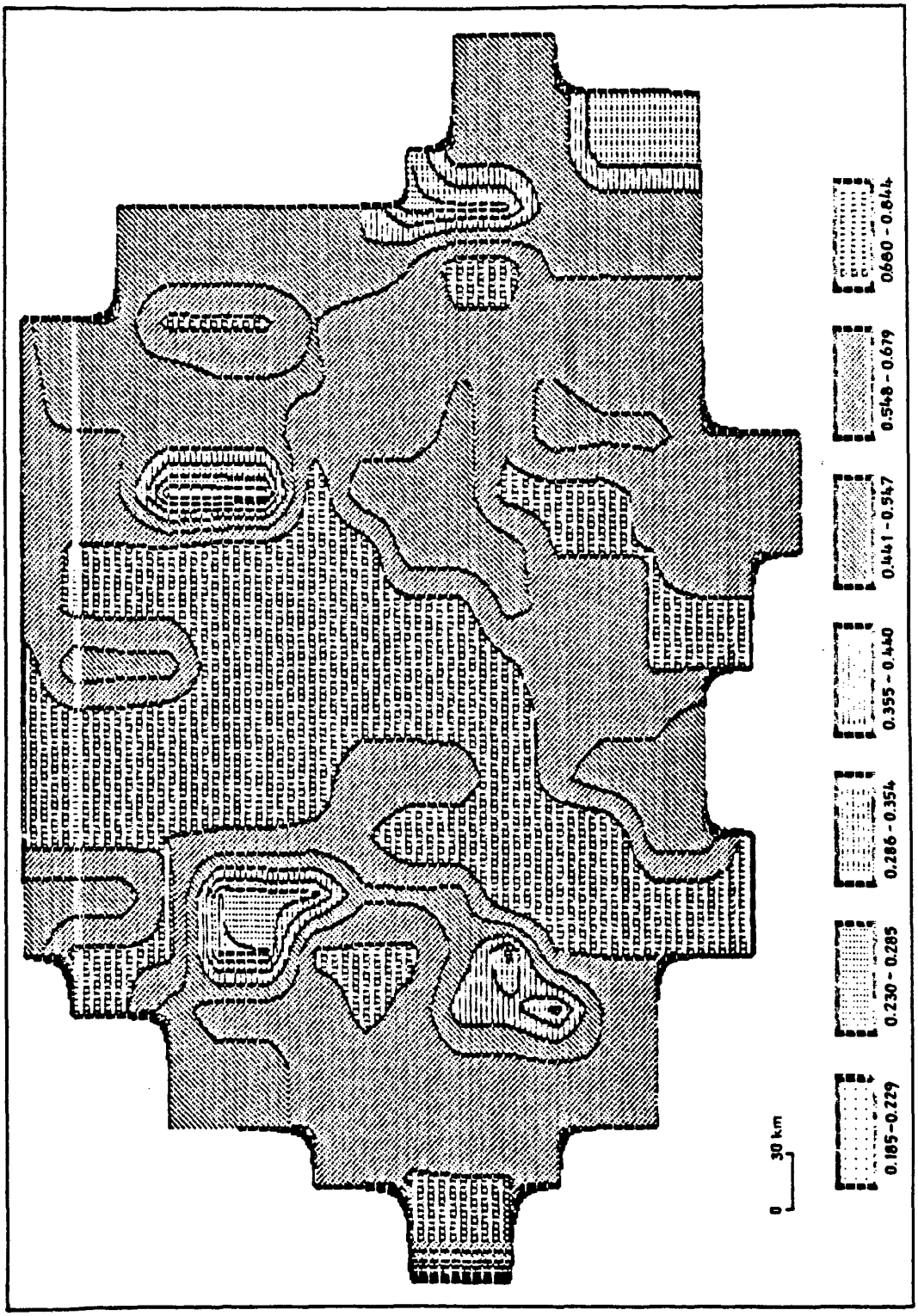


FIG.1 Probability of occurrence for a certain state 1(u≤2 m/s)

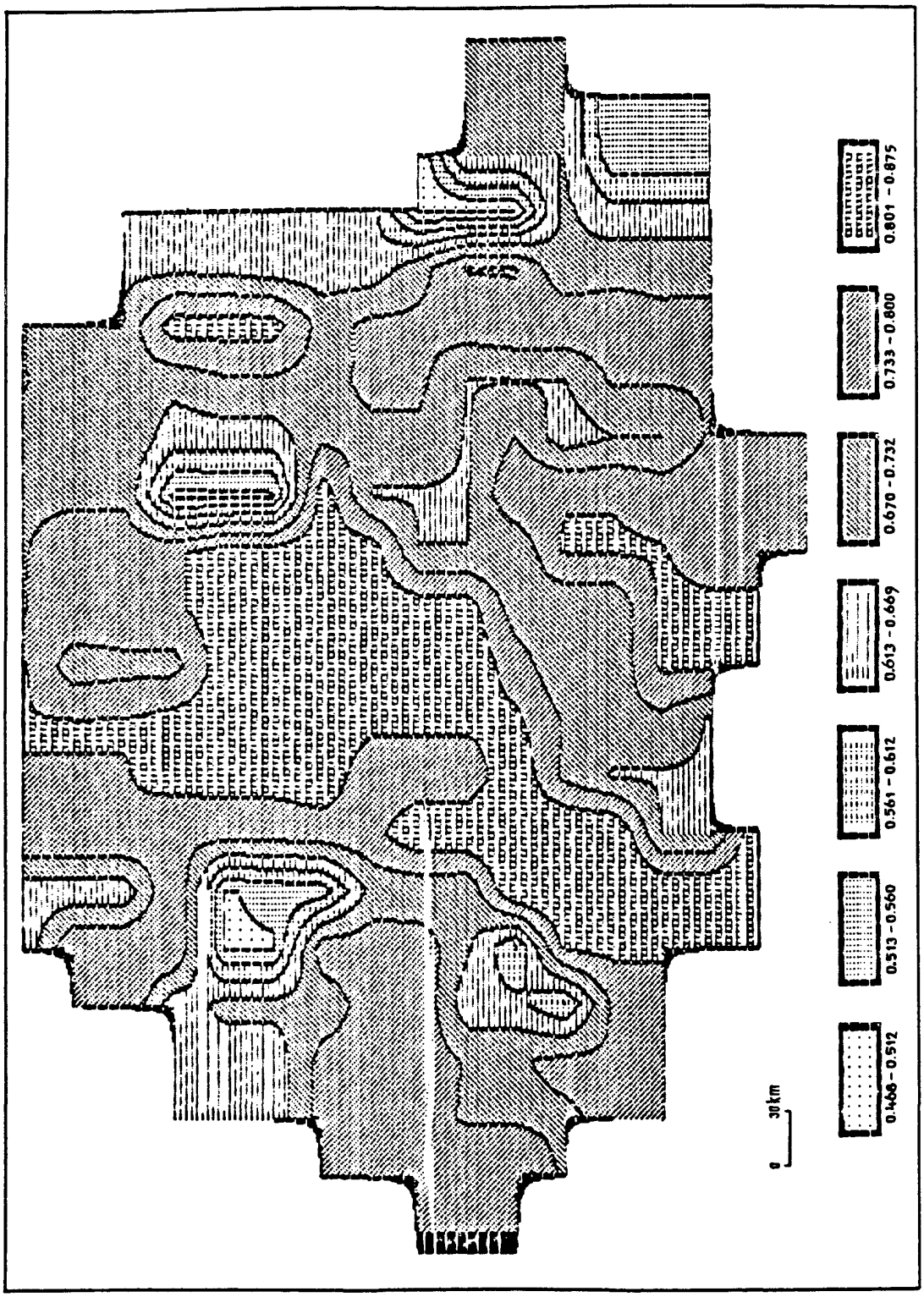


FIG.2 The conditional probability P_{11}

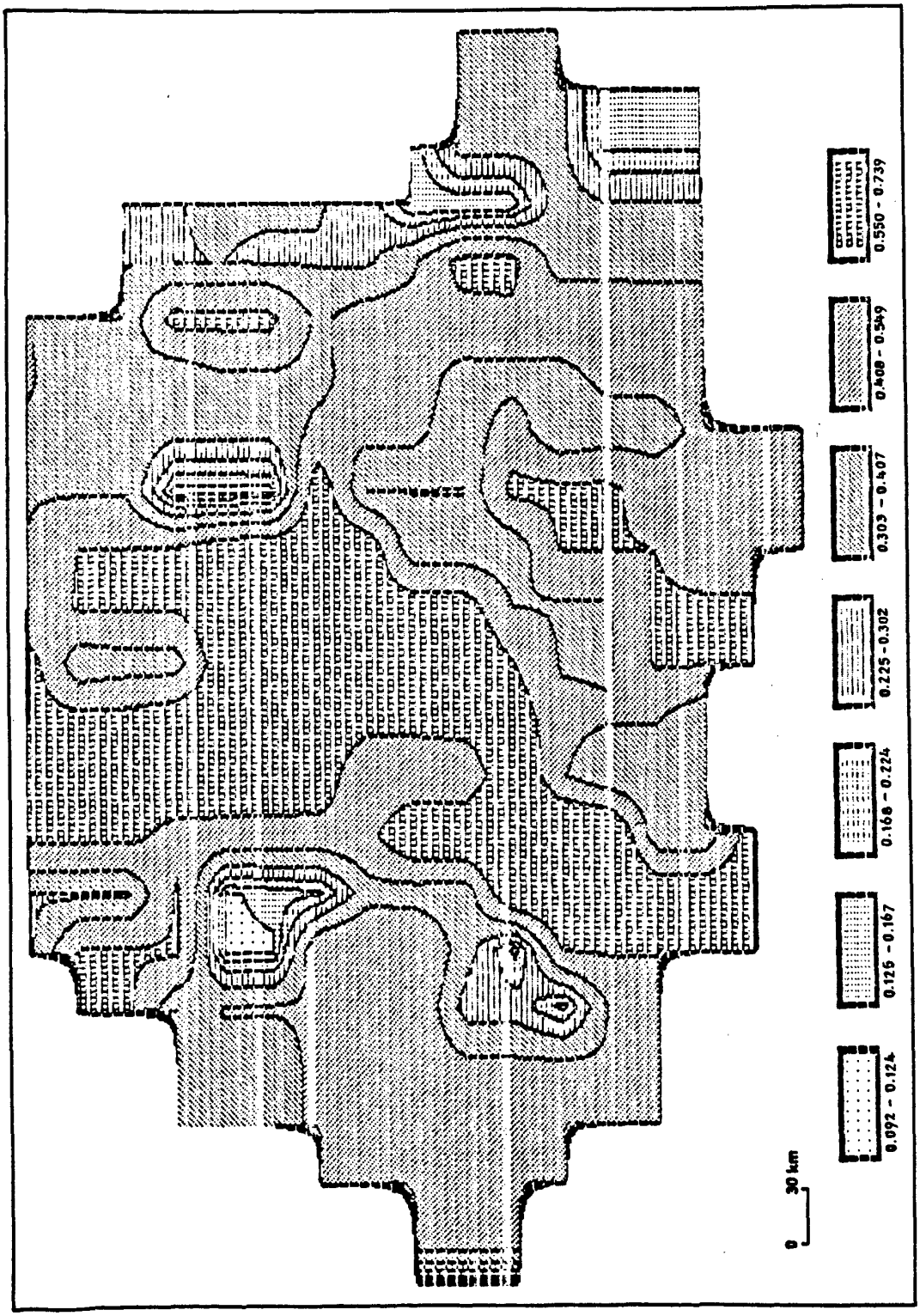


FIG.3 Probability of successive occurrence for the states 11, P(11)

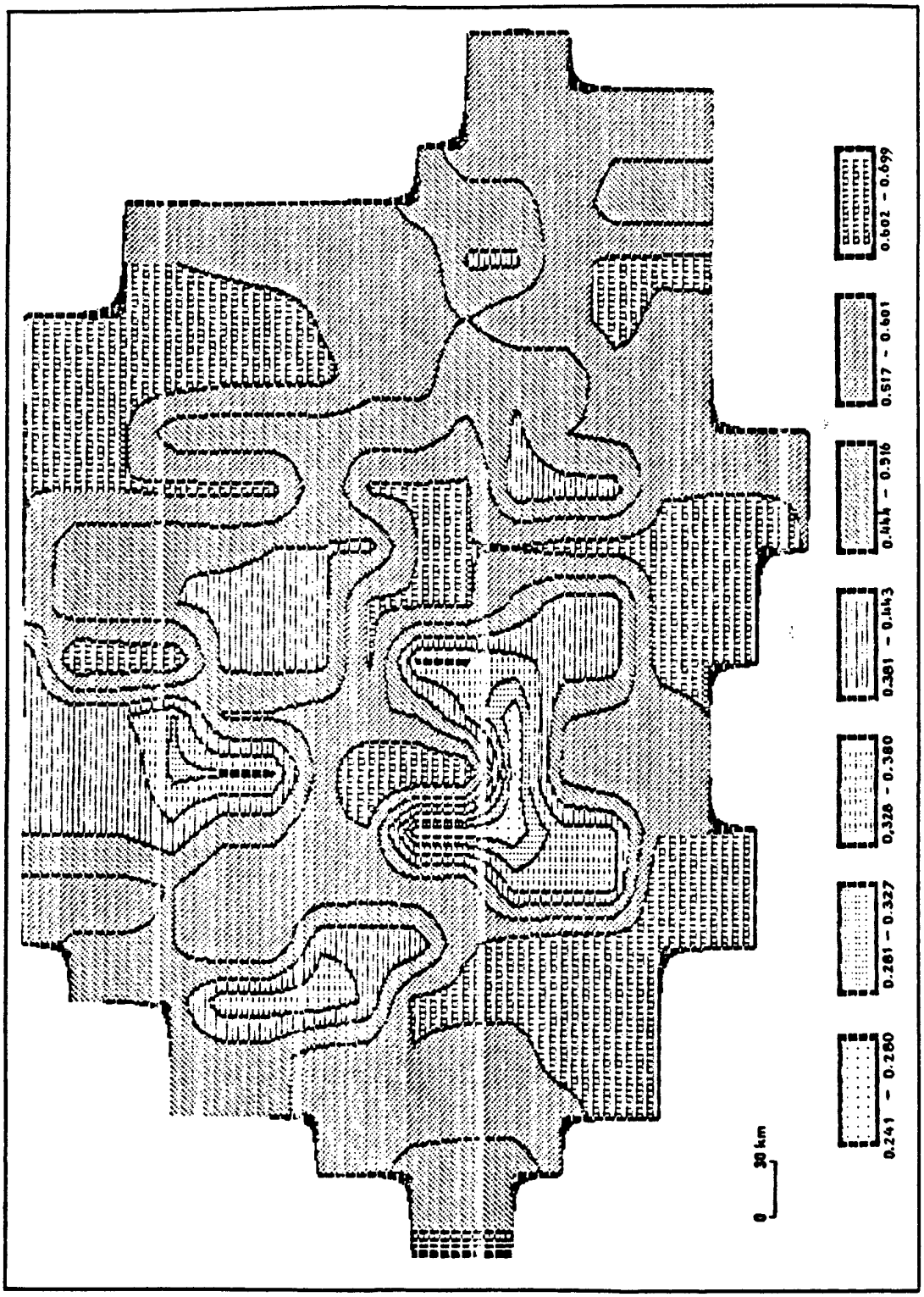


FIG.4 Besson's persistence coefficient r'_B

Third International Conference on Statistical Climatology
June 23-27, 1986, Veinna, Austria

THE BAYESIAN ASSESSMENT OF DIURNAL MEAN SURFACE TEMPERATURE DISTRIBUTIONS AT TWO FINNISH SITES WITH SHORT TEMPERATURE RECORDS

Jaakko Helminen

Finnish Meteorological Institute
P.O.Box 503, SF-00101 Helsinki, Finland

1. INTRODUCTION

The climatologist faces frequently situations where the record of climatological variables at a given site is too short to make any statistically meaningful estimates of their distributions. Then to use as a first guess distributions based on experience and distributions of some neighbouring sites the Bayesian approach offers a systematic method to be considered. In the following the Bayesian approach is applied to the diurnal mean surface temperature distributions at two Finnish sites, Kumlinge in the southwestern archipelago area and Kilpisjärvi in northwestern Lapland.

The role of a strong a priori distribution was considered in this context where the variable was quite continuous spatially.

2. THE BAYESIAN APPROACH

The Bayesian approach in the climatological context can be based on the experience of the climatologist, appropriate theoretical results and climatological records of neighbouring sites. From these elements and the cahracteristics of the variable in question the climatologist forms a subjective view of the appropriate data generating process and the distributions of its parameters, like the mean and the variance in case of a normal data-generating process. This view in terms of the parameter distribution is called the a priori distribution. With new observations at the site under consideration the view of the distribution usually changes and one obtains the a posteriori distribution. Even if the general form of the a priori and the a posteriori distributions might be the same, the values of the moments might differ considerably.

To consider the relationships between the a priori and the a posteriori distributions one can use Bayes' theorem according to which

$$f_1(\pi|\omega) = \frac{f_{dgp}(\omega|\pi)\, f_0(\pi)}{\int_{-\infty}^{\infty} f_{dgp}(\omega|\pi)\, f_0(\pi)\, d\pi}. \quad (1)$$

Here the values of observations are denoted by ω and the small set of parameters governing the data generating process by π. Then $f_0(\pi)$ and $f_1(\pi|\omega)$ represent the a priori and the a posteriori probability distribution functions of the parameters, respectively. The function $f_{dgp}(\omega|\pi)$ is viewed here as the likelihood function of π.

Based on the Central Limit Theorem we have assumed in this work that the diurnal mean surface temperature follows the normal data-generating process. Furthermore we considered its precision $h = 1/\sigma^2$ to be known with the value h_*. Then the only parameter of the process was μ. Given its a priori distribution in the form

$$f_0(\mu) = f_N(\mu \mid m', n'h_*) \quad (2)$$

(m' and n' parameters) it can be shown (Epstein, 1985) that

$$f_1(\mu|h=h_*, m', n', \bar{x}, n) = f_N(\mu|m'', n''h_*) \quad (3)$$

where

$$n'' = n+n' \text{ and} \quad (4)$$

$$m'' = (n\bar{x} + n'm')/(n+n'). \quad (5)$$

One can also form the posterior predictive distribution function of $\bar{x}$ by taking into account the given precision h_* and the posterior distribution function of μ (cf. (3)). According to Epstein (1985) one gets

$$f(\bar{x} \mid n,m'',n'') = f_N(\bar{x} \mid m'', n_u h_*), \quad (6)$$

where

$$n_u = (nn'')/(n + n''). \quad (7)$$

3. DATA AND THE A PRIORI DISTRIBUTIONS

The observations covered diurnal mean surface temperatures of February and July at two Finnish sites, Kumlinge 60°18'N, 20°45'E; southwestern archipelago area) for years 1984-85 and Kilpisjärvi (69°03'N, 20°47'E; norhtwestern Lapland) for the years 1979-85 (cf. fig. 1). In order to ensure a sample of independent cases only diurnal mean surface temperatures of every sixth day were included. This sampling procedure was considered to be adequate for typical Finnish conditions.

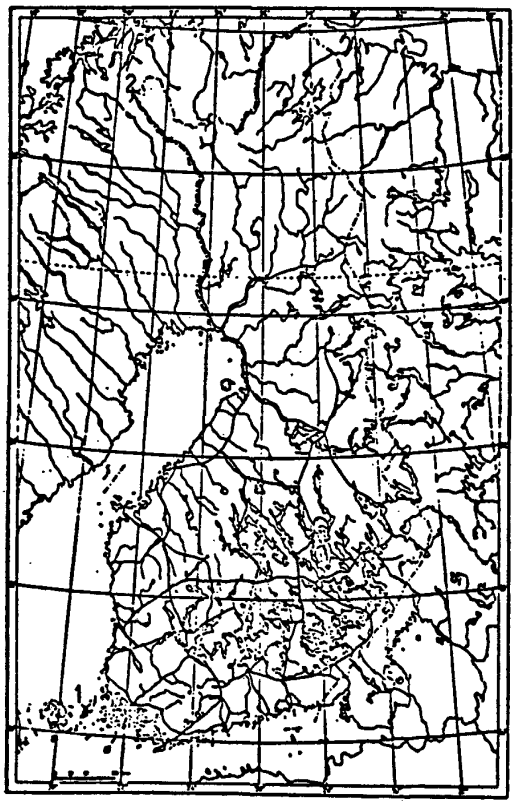


Fig. 1. The two Finnish sites under consideration (x). 1. Kumlinge, 2. Kilpisjärvi. The neighbouring sites for Kumlinge are marked with small squares.

The values of n, $\bar{T}$, and σ_T based on the observations are given in Table 1.

Table 1. The values of n, $\bar{T}$, and σ_T for the sites under consideration ($[\bar{T}] = [\sigma_T] = {}^{o}C$)

Site	February n	$\bar{T}$	σ_T	July n	$\bar{T}$	σ_T
Kumlinge	10	-6.5	5.5	12	15.6	1.5
Kilpis-järvi	35	-14.6	10.5	42	10.8	2.8

In order to determine the a priori distributions we used in case of Kumlinge the distributions of the following neighbourhood stations: Mariehamn (60°07'N, 19°54'E), Utö (59°47'N, 21°23'E), and Nyhamn (59°58'N, 19°58'E) with temperature records covering the period 1959-85. The values of n_1,m', and σ_T were as given in Table 2.

Table 2. The values of n_1, m', and σ_T for Kumlinge

Neighbourhood station	February n_1	m'	σ_T	July n_1	m'	σ_T
Mariehamn	135	-4.9	5.7	162	15.6	2.2
Utö	135	-3.9	5.1	156	15.9	2.3
Nyhamn	135	-3.4	4.7	162	15.5	2.2

The representative value of m' was taken as the arithmetic mean $\tilde{m}'$ over the three stations. Also h_* was determined as $h_* = 1/\sigma_T^2$. On the other hand $\sigma_{\bar{T}} = (1/\sqrt{n_1})\,\sigma_T$. Then

$$\sigma_{\tilde{m}'} = \left[\widetilde{\sigma_{\bar{T}}^2} + (\tilde{m}' - \bar{T})^2\right]^{\frac{1}{2}}. \quad (8)$$

Finally we get

$$n'h_* = 1/\sigma_{\tilde{m}'}^2. \quad (9)$$

For Kumlinge we obtained the values of h_*,$\tilde{m}'$,$\sigma_{\tilde{m}'}$, and n' as written in Table 3.

Table 3. The values of h_*,$\tilde{m}'$, $\sigma_{\tilde{m}'}$, and n' for Kumlinge

February h_*	$\tilde{m}'$	$\sigma_{\tilde{m}'}$	n'	July h_*	$\tilde{m}'$	$\sigma_{\tilde{m}'}$	n'
0.037	-4.1	0.89	34.4	0.207	15.7	0.27	64.3

For Kilpisjärvi the only neighbourhood station used was the old Kilpisjärvi weather

station ($69^{o}03'N$, $20^{o}48'E$) for which the values of n_1,m', and σ_T were as given in Table 4.

Table 4. The values of n_1,m', and σ_T for Kilpisjärvi

Neighbourhood station	February			July		
	n_1	m'	σ_T	n_1	m'	σ_T
	100	-13.1	7.8	120	10.6	3.5

Then the values of h_*, $\widetilde{m}'$, $\sigma_{\widetilde{m}'}$, and n' were as written in Table 5.

Table 5. The values of h_*, $\widetilde{m}'$, $\sigma_{\widetilde{m}'}$, and n' for Kilpisjärvi

February				July			
h_*	$\widetilde{m}'$	$\sigma_{\widetilde{m}'}$	n'	h_*	$\widetilde{m}'$	$\sigma_{\widetilde{m}'}$	n'
0.016	-13.1	1.69	21.3	0.082	10.6	0.60	34.1

4. RESULTS

Using the values of Tables 1, 3, and 5 the following parameter values of the a posteriori and the a posteriori predictive distributions were obtained (cf. Tables 6 and 7).

Table 6. The parameter values m",$n''h_*$, σ_{pt}, $n_u h_*$, and σ_{ptpr} for Kumlinge. ($\sigma_{pt} = \sigma_{posterior}$, $\sigma_{ptpr} = \sigma_{posterior\ predictive}$) ($[m''] = [\sigma_{pt}] = [\sigma_{ptpr}] = {}^{o}C$)

February				
m"	$n''h_*$	σ_{pt}	$n_u h_*$	σ_{ptpr}
-4.6	1.64	0.78	0.30	1.82
July				
m"	$n''h_*$	pt	$n_u h_*$	ptpr
15.7	15.8	0.25	2.15	0.68

Table 7. The parameter values m", $n''h_*$, σ_{pt}, $n_u h_*$, and σ_{ptpr} for Kilpisjärvi. ($\sigma_{pt} = \sigma_{posterior}$, $\sigma_{ptpr} = \sigma_{posterior\ predictive}$) ($[m''] = [\sigma_{pt}] = [\sigma_{ptpr}] = {}^{o}C$)

February				
m"	$n''h_*$	σ_{pt}	$n_u h_*$	σ_{ptpr}
-14.0	0.93	1.04	0.35	1.68
July				
m"	$n''h_*$	σ_{pt}	$n_u h_*$	σ_{ptpr}
10.7	6.21	0.40	2.21	0.67

5. DISCUSSION

A comparison of the values of m" with the monthly mean surface temperatures for the period 1931-1960 (Kolkki, 1966) shows a reasonable correspondence (cf. Figs. 2 and 3).

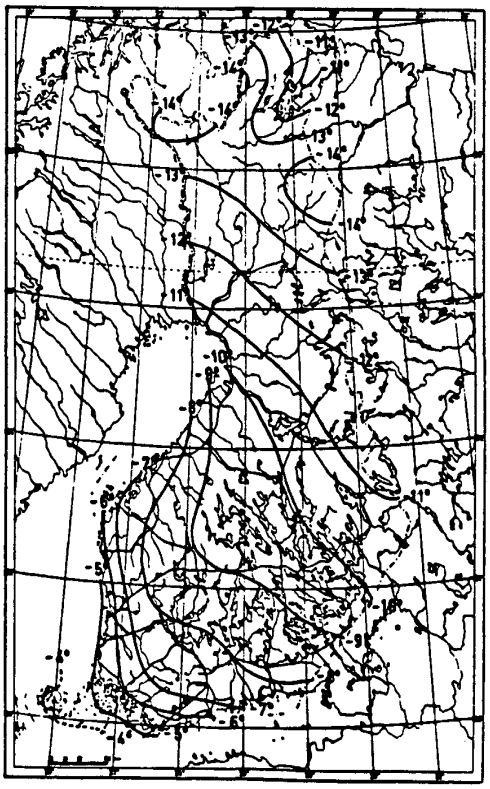


Fig. 2. The mean surface temperature over Finland in February (1931-1960) (in centigrade) (after Kolkki, 1966)

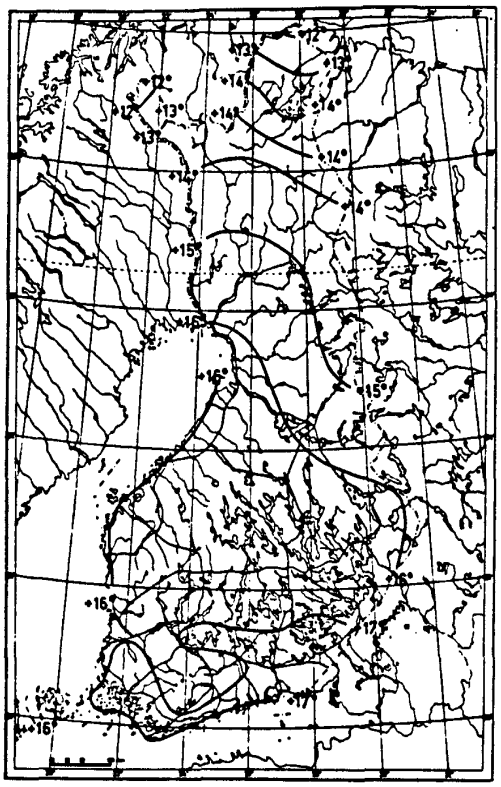


Fig. 3. The mean surface temperature over Finland in July (1931-1960) (in centigrade) (after Kolkki, 1966)

By comparing the values of m" with the values of $\bar{T}$ one can see that the influence of the relatively strong prior has been considerable in February, especially in Kumlinge. The large differences of February led to further climatological considerations.

The climate of the Finnish winter 1984-85, including February 1985, was quite exceptional. In fact February 1985 was among the coldest Februaries of the century and moreover in a unique manner persistently cold. Then one would suspect that the sample covering February 1985 does not consist of independent cases as presumed by the Bayesian approach. To avoid the possible bias caused by February 1985 the results were recalculated by excluding February 1985. Then we obtained results as given in Tables 8 to 11.

Table 8. The values of $n, \bar{T}, h_*, \widetilde{m'}, \sigma_{\widetilde{m'}}$, and n' of February for Kumlinge. (The year 1985 excluded)

n	$\bar{T}$	h_*	$\widetilde{m'}$	$\sigma_{\widetilde{m'}}$	n'
5	-2.3	0.038	-3.9	0.89	33.1

Table 9. The values of m", $n"h_*$, σ_{pt}, $n_u h_*$, and σ_{ptpr} of February for Kumlinge (The year 1985 excluded)

m"	$n"h_*$	σ_{pt}	$n_u h_*$	σ_{ptpr}
-3.7	1.45	0.83	0.17	2.44

Table 10. The values of $n, \bar{T}, h_*, \widetilde{m'}, \sigma_{\widetilde{m'}}$, and n' of February for Kilpisjärvi. (The year 1985 excluded)

n	$\bar{T}$	h_*	$\widetilde{m'}$	$\sigma_{\widetilde{m'}}$	n'
30	-13.5	0.016	-13.1	1.69	21.3

Table 11. The values of m", $n"h_*$, σ_{pt}, $n_u h_*$, and σ_{ptpr} of February for Kilpisjärvi. (The year 1985 excluded)

m"	$n"h_*$	σ_{pt}	$n_u h_*$	σ_{ptpr}
-13.3	0.84	1.09	0.31	1.79

By comparing the values of m" in Tables 6 and 9 one can see that the value for Kumlinge changed considerably. However, the small sample size should be taken into account before further inferences are made. Nevertheless the difference $\widetilde{m'}$ - m" became smaller and the posterior based on the data of 1984 only would likely be a better choice to proceed with from 1986 onwards.

In Kilpisjärvi the exclusion of February 1985 revealed that the difference between the $\bar{T}$ values of the old and the new station might not be as large as indicated in Tables 5 and 7.

6. SUMMARY

In this work we have demonstrated the use of a relatively strong prior, when dealing with a rather continuous, long-term mean variable. The results obtained were quite reasonable and the values of m" differed from the values of m'. It should also be emphasized that in climatology we have also variables which might need relatively weak priors (e.g. precipitation).

Also one should keep always in mind that the data sample should consist of independent cases, although this requirement is met well in quite a few climatological considerations.

Acknowledgements: The author acknowledges his gratitude to the Academy of Finland for the financial support of this work.

7. REFERENCES

Epstein, E. S., 1985: Statistical Inference and Prediction in Climatology: a Bayesian Approach. Meteor. Monogr., 20, No. 42, 199 p.

Kolkki, O., 1966: Tables and Maps of Temperature in Finland during 1931-1960. Supplement to the Meteorological Yearbook of Finland, 65, Part 1a - 1965. The Finnish Meteorological Office, Helsinki, Finland.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

MODELLING, ESTIMATION AND TESTING OF TIME-DEPENDENT CLIMATOLOGICAL PARAMETERS

G. TUNNICLIFFE WILSON and J. BELCHER

Mathematics Department, University of Lancaster, U.K.

1. THE STATISTICAL CONTEXT OF PARAMETER VARIATION

If we are looking for a change in any system, then it is important to know (a) what we take as the normal state of the system, and (b) what kind of change we are looking for. When the normal state of the system *is* one of change, we need to consider more carefully what questions we are asking.

As an example, consider the measurements of the Southern Oscillation Index (SOI) from 1964 to 1978 shown in Figure 1. Suppose we ask whether this Index is changing. Obviously successive measurements are different, but we take that as due to short term random fluctuations and measurement error. So is the underlying level changing ? The assumption of no change, corresponds here to taking the observations to be of the form $z_t = m + e_t$, where m is a fixed mean level, and e_t is random error. Testing for change can be done by taking two separate sections of the series, estimating their means and the error variance, then carrying out a standard test for the difference in these means. More commonly in time series contexts, a *running mean* using the Exponential Smoothing filter is employed:

$$\hat{m}_t = (1-\theta)(z_t + \theta z_{t-1} + \theta^2 z_{t-2} + \dots)$$

implemented by the recursion $\hat{m}_t = (1-\theta)z_t + \theta\hat{m}_{t-1}$.

Now given that the model $z_t = m + e_t$ is true, with $\sigma_e^2 = \mathrm{Var}(e_t)$ estimated from the sample variance of z_t , then $SE(\hat{m}_t) = ((1-\theta)/(1+\theta))^{1/2}\sigma_e$. Deviations of $\hat{m}_t$ from $\bar{z}$ by more than $2SE(\hat{m}_t)$ say, provide evidence that the mean is in fact changing.

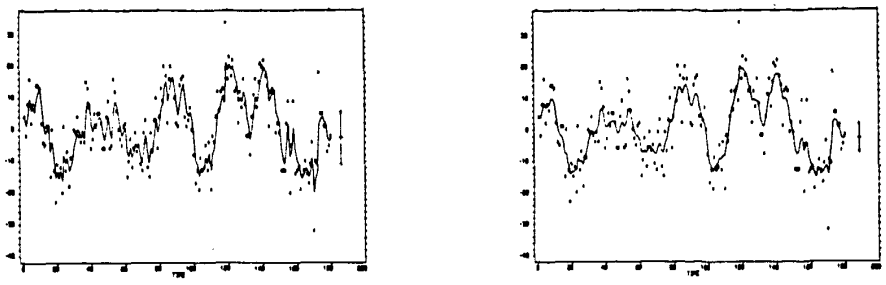
There are some concerns however. First is how to choose the smoothing constant θ , the effective time span of the filter being $1/(1-\theta)$. This may be chosen so as to detect changes on either a short or long time scale. Secondly, and related to the first, is how this procedure responds when the mean $m = m_t$ is time dependent. The statistical answer to these concerns is to model the time series in its entirety, including the variation of m_t . An excellent example and exposition of this is given by Ogawara (1980). A simple model for m_t is to assume that the changes $\delta_t = \nabla m_t = m_t - m_{t-1}$ are independent with constant variance. This can include the case when m_t remains constant for several time periods by having a large probability that $\delta_t = 0$. More commonly, these changes are taken to be Normal $(0,\sigma_\delta^2)$, so that m_t is a classical random walk signal :

$$m_t = m_{t-1} + \delta_t$$

and $z_t = m_t + e_t$ is the observed signal plus noise.

The consequences of such a model are that it may be fitted to the data - i.e. the parameters σ_e^2 , σ_δ^2 estimated, and the component m_t estimated by minimum mean square error signal extraction. In particular, if σ_δ^2 is estimated to be 0 , then the inference is that m is constant. Otherwise, the estimate $\hat{m}_t$ of m_t from present and past observations is given by the above Exponential smoothing formula, with θ now determined by $\sigma_\delta^2/\sigma_e^2 = (1-\theta)^2/\theta$. The *normal state of the system is now taken to be one with a changing mean level*, and the value of $SE(\hat{m}_t)$ is slightly increased from the previous formula, to allow for this. It is also possible to construct an estimate $\tilde{m}_t$ of m_t using all past present and future observations z_t . This is a 2-sided Exponentially Weighted mean, conveniently implemented by smoothing $\hat{m}_t$ in the reverse direction : $\tilde{m}_t = (1-\theta)\hat{m}_t + \theta\tilde{m}_{t+1}$. Now $SE(\tilde{m}_t)$ may be appreciably less than $SE(\hat{m}_t)$.

Figure 1 shows the SOI together with $\hat{m}_t$ and $\tilde{m}_t$ constructed in this way. The ±2 SE limits are indicated by the vertical arrows. The value of θ is 0.52 , giving a time span for the average of about 2. Evidently the mean level of the series is changing quite rapidly on this time scale. If the estimate σ_e^2 were 0 , then the model implies that z_t is a pure random walk, with a constantly changing level, observed free of error. No smoothing would then be appropriate.



(a) (b)

Southern Oscillation Index from 1964 to 1978 with values smoothed
(a) using a one sided Exponential smoother with parameter 0.52
(b) using a two sided Exponential smoother
The vertical arrows around a bar indicate ± 2SE limits to be applied to the smoothed series.

Figure 1.

2. ARIMA MODELS AND DATA PROCESSING

The previous model for $z_t = m_t + e_t$ has a representation in the form

$$\nabla z_t = z_t - z_{t-1} = a_t - \theta a_{t-1}$$

where a_t is an uncorrelated sequence which may be interpreted as the linear innovation in z_t, i.e. the error in predicting z_t using $z_{t-1}, z_{t-2}, \ldots$, which in this case happens to be $a_t = z_t - \hat{m}_{t-1}$. This representation is a special case of the class of ARIMA models which Box and Jenkins (1976) did much to promote, and which have received much attention in climatology, by Katz and Skaggs (1981) for example. These models are primarily orientated to prediction, although some of them are readily interpreted as signal plus noise structures, which invite smoothing applications.

The advantage of ARIMA modelling is that, done well, it provides an objective and verifiable summary of the statistical structure of the data. Further, the model has immediate implications regarding prediction. The interpretation of the model in terms of structural components, such as signal and noise, is not however always clear. The reason is that the dynamics of the series of interest may be quite complex, involving other unknown variables, and a simple interpretation of the structure of the single series cannot be expected.

Whilst ARIMA models can represent a wide variety of behaviour, such as trends, cycles and seasonality, they have limitations in that they are based on ideas of linear filtering and Normal distributions. Thus they may not represent well, series with nonlinear structure, or, for example, level changes which occur by occasional large jumps rather than a continuous random walk process. Extensions to the models would be necessary to represent such behaviour. Nevertheless, even in such circumstances ARIMA models are very valuable in promoting good statistical practice. To understand this, a few points concerning Data Structure, Data Processing and Data Modelling are helpful.

Data Processing - the general term in time series would be filtering - generally focusses on revealing particular features such as trends or cycles and eliminating noise and other features. The difficulty here is knowing just what is being estimated, and its statistical properties. In this respect a Data Model which is believed adequately to represent the true Data Structure, can be valuable. Conversely, a good Data Model may have implications about the Data Structure, and will determine certain Data Processing operations such as the linear predictor, and signal extraction filters. Nevertheless, this model should not constrain other Data Processing activities which might be more sensitive at revealing some features.

To consider again the SOI series, there is some evidence that the signal is not a random walk, but a highly autoregressive process about some mean value μ say :

$$(m_t - \mu) = \phi(m_{t-1} - \mu) + \delta_t \ .$$

This evidence arises from fitting a corresponding ARMA model to z_t :

$$(z_t-\mu) = \phi(z_{t-1}-\mu) + a_t - \theta a_{t-1}$$

for which $\hat{\phi} = .87$ (SE .05), $\hat{\theta} = .43$ (SE .09), $\hat{\mu} = .54$ (SE 2.71).

Thus although m_t represents a local varying mean, it could be asked whether the long term mean μ is changing. To detect any such change Exponential smoothing could still be used, with the smoothing parameter no longer taken as $\hat{\theta}$, but much closer to 1 so as to represent a long time span. In that case it would be wise to calculate the SE of the smoothed series on the basis of the above ARMA Structure - to assume independent variations of the series values about μ would give a most misleading SE value.

3. TIME DEPENDENCY OF TRENDS, CYCLES AND SEASONALITY

The foregoing ideas will be illustrated with respect to a study of the Mauna Loa Observatory (MLO) CO_2 index shown in Figure 2. Changes in the amplitude of the annual variation of CO_2 indices have been studied by Pearman and Hyson (1981), changes in the growth rate have been related to SOI by Komhyr et al (1985) and interannual variations of CO_2 have similarly been related to SOI by Thompson et al (1986).

A common tool in this work has been complex demodulation. This is closely related to the Exponential Smoothing process for estimating the mean, but premultiplying the series by sinusoidal factors allows estimation of the variations in the amplitude of any specified frequency component. These components can be removed from the series to permit clearer study of the underlying trend, or their amplitude variations may be of separate interest. An important reference for complex demodulation is Bloomfield (1976). Other approaches to estimating seasonal components are now advocated by Cleveland, Freeny and Graedel (1983). The aim here is to present a role for the seasonal ARIMA models of Box and Jenkins in this context.

These models allow for variations in both the seasonal pattern and the trend of a time series such as the MLO CO_2. In the notation of Box and Jenkins (1976) the model we used may be written

$$\nabla\nabla_{12}\, z_t = (1-\theta B)(1-\Theta B^{12})a_t$$

where ∇_{12} is the seasonal difference operator, B is the backward shift operator, and a_t the innovation series. The model is perhaps best viewed in its predictive role as follows: An Exponential Smoothing predictor is first applied on a year to year basis, i.e. to predict the January 1986 value, the previous January values only are used. Recursively, this gives a first (seasonal) prediction of z_t, say $\hat{s}_t$, according to

$$\hat{s}_t = (1-\Theta)z_{t-12} + \Theta\, \hat{s}_{t-12}.$$

The errors $\alpha_t = z_t - \hat{s}_t$ which result from this, apart from a constant bias when z_t contains a trend, are generally slowly varying. The forecast s_t , which makes no use of $z_{t-1}, \ldots, z_{t-11}$, may therefore be improved by predicting the error α_t from an ordinary Exponential Smoothing applied to past values :

$$\hat{\alpha}_t = (1-\theta)\alpha_{t-1} + \theta\hat{\alpha}_{t-1} .$$

This is used to construct a corrected value $(\hat{s}_t + \hat{\alpha}_t)$ which is now the predictor of z_t . The final prediction error a_t is thus $z_t - \hat{s}_t - \hat{\alpha}_t = \alpha_t - \hat{\alpha}_t$, and the above model is equivalent to the pair of equations :

$$z_t - z_{t-12} = \alpha_t - \Theta\alpha_{t-12} \,,\; \alpha_t - \alpha_{t-1} = a_t - \theta a_{t-1} .$$

There is one possible disadvantage in that the parameter Θ effectively controls the permitted variation both of the seasonal pattern and the trend. Thus if $\Theta = 1$, the predicted seasonal pattern remains constant, and so does the predicted trend. The model is then equivalent to a fixed trend + seasonal pattern, with an error component made up in part of a random walk, and in part of independent errors.

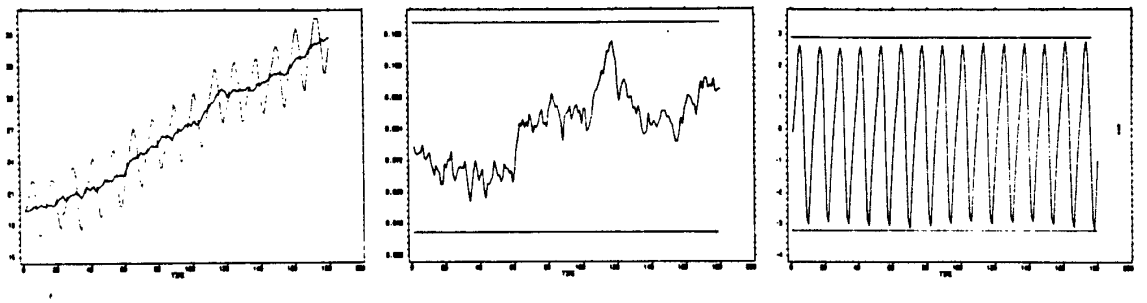
This model was fitted to the MLO CO_2 series of Figure 2a, resulting in estimates $\hat{\theta} = .301$ (SE .074) and $\hat{\Theta} = .998$ (SE .074) .
There is therefore no evidence of changes in trend or seasonal pattern over this period. The model was checked, first of all by observing that the innovation series a_t did appear to be uncorrelated with constant variance. Secondly, the parameter constraints were relaxed, by fitting the model

$$\nabla\nabla_{12} z_t = (1-\theta_1 B-\theta_2 B^2 - \ldots -\theta_{13} B^{13})\, a_t \ .$$

This allows considerable modification of the predictor weights to allow for varying seasonal pattern and trend, but there was no improvement in fit. However the value of $\hat{\theta}$ is much less than 1, implying a substantial random walk component. It is this which gives the visual impression of changing trends. A cumulative random component is not surprising in a series such as this, where some part of the measured increase from one year to the next, will be independent of previous increments. Intuition often misinterprets the pattern of random increments as trend changes.

4. ESTIMATING CHANGES IN TREND AND SEASONALITY

Unlike the signal and noise model described in Sections 1 and 2, the seasonal ARIMA model has no natural interpretation as the sum of independent components corresponding to trend and seasonal signals with added noise. Such interpretations can be contrived however, though they are not unique, and Burman (1980) presents the procedures for extracting the implied components using two-sided filters. He also describes how to handle the end effects which occur when applying filters.



(a) (b) (c)

(a) MLO CO_2 in ppm - 300 from 1964 to 1978 with estimated level.
(b) Estimated gradient of MLO CO_2 with ±2SE limits
(c) Estimated Seasonal component of MLO CO_2 with lower limits on months of minima and upper limits on months of maxima. The vertical arrows around a bar indicate ±2SE limits to be applied to this series.

Figure 2

There is however, a natural decomposition using one-sided filters. This derives from the predictive application of the model. At any time point the model predicts future values which are the sum of a trend line and fixed seasonal pattern - but these differ from point to point. However, at any point the level m_t and gradient g_t of the trend line, and the seasonal departure s_t from the trend may be computed. The original series may then be expressed as $z_t = m_t + s_t + e_t$ where e_t is white noise. The appropriate filters are defined by constructing first

$$u_t = (1+B+B^2+\ldots+B^{11})/\{(1-\theta B)(1-\Theta B^{12})\}\ z_t \text{ and } v_t = \nabla u_t$$

whence $m_t = \alpha u_t + \beta v_t$, $g_t = \alpha v_t$, $e_t = \gamma a_t$ and $s_t = z_t - m_t - e_t$ where $\alpha = (1-\theta)(1-\Theta)/12$, $\beta = (1-\theta)\Theta + \theta(1-\Theta)/12 + (1-\theta)(1-\Theta) \times 11/24$, $\gamma = \theta\Theta$. Now in the case when $\Theta = 1$, the result is that m_t consists of a random walk about a fixed trend, g_t is constant and s_t is exactly periodic. It is possible however, even when $\hat{\Theta}=1$, to apply this decomposition using other values of $\Theta < 1$, as a Data Processing procedure designed to reveal any possible variations in trend and seasonal pattern. This was done in the case of MLO CO_2 , taking $\theta = 0.3$ as estimated, but $\Theta = 0.8$, equivalent to a 4 year smoothing time span. The estimated level m_t is shown with the original series in Figure 2a. The estimated gradient values are shown with ± 2SE limits in Figure 2b, and the estimated seasonal pattern with ± 2SE limits in Figure 2c . The SE's were obtained by calculating the effects of the filters, assuming the estimated model with $\Theta = 1$ to be true. The estimated values do lie within the limits, which is consistent with the earlier conclusions that the true values are not changing, although the seasonal pattern approaches its outer limits towards the end.

To end this section reference must be made to the procedures of Harrison and Stevens (1976) which are designed to detect anomolous observations, and sudden changes in level and trend. Also Harvey (1984) has pursued the Kalman Filter approach to seasonal modelling and decomposition.

5. ON THE USE OF ESTIMATED PARAMETER CHANGES

A brief comment is appropriate, on the fact that estimated changes such as those of the gradient of MLO CO_2, are often compared with others such as estimated changes in the level of SOI, in order to search for causal relationships. The fact that the gradient estimates do not differ significantly from a constant value, does not mean that the variations in these estimates are inexplicable. They may appear random, but might still be explained by other environmental variables such as the SOI level. Such a comparison is therefore a valid use of Statistical Data Processing, and the work of Komhyr et al (1985) and Thompson et al (1986) focus on such investigations. The statistical problem however, is in validating the lagged regression equations which arise from these comparisons. Much cautionary comment has been written on this - see for example Pittock (1978) and Tunnicliffe Wilson (1984) . The safe answer is that having thereby revealed a possible relationship, one should go back to the original data, and see if the explanatory variable has any further predictive contribution to make to the dependent variable, over and above that which can be achieved from the seasonal ARIMA model. Thus with z_t as the MLO CO_2 series and x_t as the SOI, the model as advocated by Box and Jenkins (1976), for the relationship suggested in the above papers, is

$$\nabla z_t = v_0 x_t + v_1 x_{t-1} + v_2 x_{t-2} + \dots + y_t$$

where y_t encapsulates the seasonal and other features in a model $\nabla_{12} y_t = (1-\theta B)(1-\Theta B^{12})a_t$. The terms in x_t may be reduced to a single lagged value, or be more efficiently represented by a transfer function model which filters x_t . The essential point is that in the absence of x_t this equation reduces to the former seasonal ARIMA model, so that x_t has to prove itself by explaining some part of the remaining variation in z_t . Suffice it to say that by this means we were unable to verify the suggested relationship, apart from a weak effect at lag 0, with $\hat{v}_0 = -.0031$ (SE .0014).

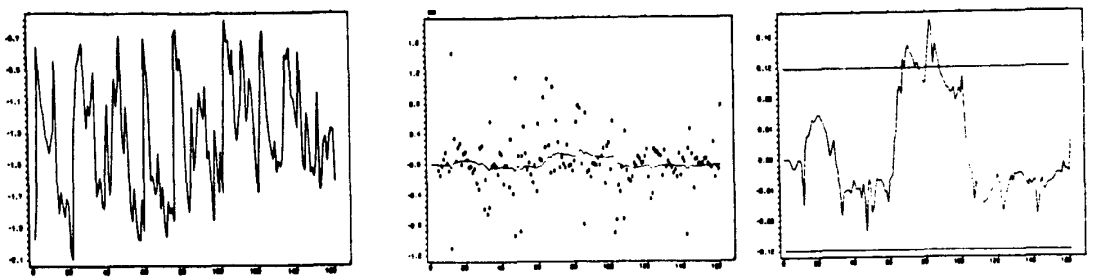
There may be some advantages in not going back to the original data, but staying with the processed data of estimated parameter variations. Thus the relationship discovered may be non-linear, and not easily represented in terms of the original data. Using complex demodulation, one of our Environmental Science researchers Mr. T. Young has correlated the amplitude of sunspot cycles with the amplitude of the 11 year cycles present within temperature and rainfall series. The argument is that non-linear relationships could confound the information in the phase of cycles, but remain apparent in amplitude changes. Nevertheless, such correlation should be subject to severe statistical scrutiny.

6. TIME DEPENDENT MODEL PARAMETERS

In many other contexts such as Engineering and Physics, time dependent parameters are thought of as regression coefficients (gains) in relationships between variables, or dynamic parameters measuring the speed of response of a system to some stimulus. Thus in the examples above, the error variance σ_e^2 may change, or the parameter θ in the ARIMA model. To investigate such changes the first step is to derive some kind of time varying estimate of the parameter, which can be subjected to visual inspection. An excellent presentation of such methods of recursive parameter estimation, whereby the estimate $\hat{\theta}_t$ is updated by each new observation z_t , is to be found in Young (1984). A method for detecting changes in regressions using residuals was given by MacNeill (1980), at a previous meeting of this conference.

A general method of investigating possible parameter change in an ARIMA model is to evaluate first the time series of derivatives of the innovations with respect to the parameters, e.g. $d_t = \partial a_t / \partial \theta$. If the value of θ used in calculating a_t and d_t is the least squares estimate, then $\Sigma a_t d_t = 0$, else a parameter correction to θ is given by $\delta\theta = -\Sigma a_t d_t / \Sigma d_t^2$. Time dependency of the parameters can then be investigated using $\delta\theta_t = -a_t d_t / (\Sigma d_t^2 / n)$ where n is the series length. Assuming θ to be constant, $\delta\theta_t$ is an uncorrelated zero-mean series with constant variance σ_θ^2 say. An exponential smoothing filter may then be applied to $\delta\theta_t$, and the result graphed with error limits of $\pm 2((1-\lambda)/(1+\lambda))^{1/2}\,\hat{\sigma}_\theta$, where now λ is the smoothing parameter. Excursions of the graph outside the limits provide evidence for significant parameter variation.

As an example, Figure 3c shows the smoothed parameter purturbations and error limits, for an autoregressive parameter in an ARMA (4,4) model for the time series in Figure 3a. The series is of measurements of an oxygen isotope ($^{18}0$), made at equally spaced depths of Pacific Ocean core, V28-238 and believed to reflect global ice volumes at intervals of approximately 2Kyrs. Variations in the sedimentation rate affect the time scale of the series, and it is hoped that these will show up as time-dependency of the autoregressive parameters. Some significant variation does occur, and consistent patterns of variation in two of the parameters can help to confirm that the cause is due to time scale changes. The aim is to estimate a new time scale, to allow correlation of the series with orbital variations of the Earth, as investigated by Hays et al (1976), and Imbrie et al (1984).



(a) (b) (c)

(a) Oxygen isotope measures in Pacific Ocean Core V28-238
(b) Instantaneous ARIMA model parameter perturbations and smoothed values
(c) Smoothed parameter perturbations on a larger scale, with ±2SE limits.

Figure 3

7. CONCLUSION

We have come full circle and must remind ourselves of the adage 'Big fleas have little fleas upon their backs to bite 'em, and little fleas have lesser fleas and so ad infinitum'. Why not allow parameter variations in the smoother used to estimate parameter variations in the.....? What we seek eventually is a fixed parameter model which represents the changes we observe according to some dynamic scheme. In practice we use models which are valid locally in time, and whose parameters we allow to change slowly. The important point is to adhere in so far as possible to an adequate statistical model which can be used to validate any inference. An advance on the simple use of ARIMA models as presented in this paper, is to base the signal extraction upon the Kalman Filter and recursive estimation of the model, as illustrated in Young and Young (1986). This work forms part of a cooperative research program in Environmental Time Series Analysis, between the Departments of Mathematics and Environmental Science at Lancaster University.

8. ACKNOWLEDGEMENT

We are most grateful to Mr. T. Young for helpful discussions on the climatological aspects of the paper. The SOI values were taken from Thompson et al (1986) table 5 and the oxygen isotope series was supplied to us by N.J. Shackleton, Godwin Laboratory for Quaternary Research, University of Cambridge. All the modelling and computations were carried out using the GENSTAT package distributed by the Numerical Algorithms Group, Oxford.

REFERENCES

Bloomfield, P. (1976) *Fourier Analysis of Time Series : An Introduction.* John Wiley, New York.

Burman, J.P., 1980. Seasonal Adjustment by Signal Extraction. *J.R. Statis. Soc. A.* **143**, 321-337.

Box, G.E.P. and Jenkins, G.M., 1976. *Time Series Analysis : Forecasting and Control*, 2nd Edition. Holden-Day, San Francisco.

Cleveland, W.S., Freeny, A.E. and Graedel, T.E., 1983. The Seasonal Component of atmospheric CO_2 : Information from new approaches to the decomposition of seasonal time series, *J. Geophys. Res.*, **88**, 10934-10946.

Harrison, P.J. and Stevens, C.F. Bayesian Forecasting. *J.R. Statist. Soc. B*, **38**, 205-247.

Harvey, A.C., 1984. A Unified View of Statistical Forecasting Procedures. *Journal of Forecasting*, 3, 245-283.

Hays, J.D., Imbrie, J. and Shackleton, N.J., 1976. Variations in the earth's orbit : pacemaker of the ice ages. *Science*, **194**, 1121-1132.

Imbrie, J., Hays, J.D., Martinson, D.G., McIntrye, A., Mix, A.C., Moreby, J.J., Pisias, N.G., Prell, W. and Shackleton, N.J., 1984. The orbital theory of Pleistocene climate : support from a revised chronology of the marine $^{18}0$ record, in *Milankovich and Climate*, eds. Berger, A., Imbrie, J., Hays, J., Kukla, G. and Saltzman, B., 269-305, D. Reidel, Hingham, Mass.

Katz, R.W. and Skaggs, R.H., 1981. On the use of Autoregressive-Moving Average Processes to Model Meteorological Time Series. *Mon. Wea. Rev.* **109**, 479-484.

Komhyr, W.D., Gammon, R.H., Harris, T.R. and Waterman, L.W., 1985. Global atmospheric CO_2 distribution and variations from 1968-1982 NOAA/GMCC flask sampling data. *J. Geophys. Res.*, **90D3**, 5567-5596.

MacNeill, I.B., 1980. Detection of changes in the parameters of periodic or pseudo-periodic systems when the change times are unknown, in *Statistical Climatology*, eds. Ikeda, S., Suzuki, E., Uchida, E. and Yoshino, M.M., 183-195, Elsevier, Amsterdam.

Ogawara, M. On the use of exponential smoothing for the estimation of climatic elements, in *Statistical Climatology*, eds. Ikeda, S., Suzuki, E., Uchida, E. and Yoshino, M.M., 217-228, Elsevier, Amsterdam.

Pearman, G.I. and Hyson, P., 1981. The annual variation of atmospheric CO_2 concentration observed in the northern hemisphere. *J. Geophys. Res.* **86**, 9839-9843.

Pittock, A.B., 1978. A critical look at long-term sun-weather relationships. *Rev. Geophys. Space Phys.*, **16**, 400-420.

Thompson, M.L., Enting, I.G., Pearman, G.I. and Hyson, P., 1986. Interannual variation in atmospheric CO_2 concentration. *Journal of Atmospheric Chemistry* **4**, 125-155.

Tunnicliffe Wilson, G., 1984. Problems in scientific time series analysis. *Questiio*, **8**, 9-19.

Young, P.C., 1984. *Recursive estimation and time series analysis*, Springer-Verlag, Berlin.

Young, T.T. and Young, P.C., 1986. Recursive estimation of time-varying environmental parameters. Report No. TR30. Department of Environmental Science, University of Lancaster, Lancaster, U.K.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

SOME PROBLEMS OF DESIGNING METEOROLOGICAL NETWORKS: APPLYING THE INFORMATION THEORY IN METEOROLOGY AND CLIMATOLOGY

by

Ferenc Rákóczi

Department of Meteorology, Eötvös Loránd University
1088 Budapest, Múzeum krt. 6-8

1. INTRODUCTION

The spatial and temporal distribution of climatological elements are of a stochastic character. Because of this we cannot describe its features on the basis of a single sample. It is necessary to have a series of samples to do it. The spatial and temporal statistical properties of climatological elements can be described by the moment of their distribution functions. Thus we can speak about the spatial or temporal distribution of mean values, or of standard deviations and so on.

In the case of stochastic fields, it is very important to know what points of the fields are to be chosen for measuring of the values of the elements, i.e. there is a general requirement to know details about the properties of meteorological networks. The national and international sets of meteorological networks. The national and international sets of meteorological networks have been developed partly spontaneously, partly the development was a result of conscious decisions. For both types of network construction there can be found examples.

The fact is that nobody can make appropriate assesments about the efficiency of a network without the knowledge of the statistical properties of meteorological fields. Factically the knowledge of the statistical character of meteorological fields is the milestone one can start from and has to return to.

2. THE STRUCTURE AND CORRELATION FUNCTIONS

The theoretical background of the investigations on meteorological networks was developed by L. Keller and A. Friedmann /1925/ who introduced the so called covariance function to describe the relations between two points in the turbulente media. G.A. Taylor /1937/ and Th. Kármán /1937/ have developed the concept of homogeneous and isotrop turbulences. A.N. Kolgomoroff /1941/ was the one who introduced the idea of local homogeneous and isotrop turbulences. An additional theoretical work in the development of the theory of turbulence was that of M. Wiener /1949/ who gave the theoretical basis for extrapolation, interpolation and smoothing of stationary time series.

The meteorological application of the results of the turbulence theory for the optimilation of the meteorological networks was introduced by O.A. Drozdov /1936/. O.A. Drozdov and A.A. Sepelevskiy /1956/ were the first to use the concept of linear interpolation. L.S. Gandin /1963/ in his work "Objective analysis of meteorological fields" gave the modern theoretical basis for the investigation of meteorological networks.

The theoretical background created the fundaments of additional investigations.

There are three way for solving the problems of meteorological networks. One of them is the using of structure function of meteorological elements:

$$b_f(d) = \overline{\left[f(M_i) - f(M_j)\right]^2} \qquad (1)$$

where b_f is the structure function of any meteorological element, $f(M_i)$ is the value of the element in point M_i, $f(M_j)$ is the same in point M_j and $d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ is the distance between the points M_i and M_j. The bar means the time average of quantity.

This structure function describes statistical behaviour of the differences in meteorological elements in space. From the structure function one can derivate the interpolation errors between the meteorological stations and, knowing this, we can determine the optimal design of meteorological networks.

The second way to characterise the statistical structure is the covariance function

$$m_f(d) = \overline{\left[f(M_i) - \overline{f(M_i)}\right]\left[f(M_j) - \overline{f(M_j)}\right]} \qquad (2)$$

where $m_f(d)$ is the covariance function, $f(M_i)$ is the value of the meteorological element in point M_i and $f(M_j)$ is the same in point M_j; $d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ the distance between the two points. The bar means the average in time. Instead of (2) often is used the correlation function

$$\mu_f(d) = \frac{m_f(d)}{\sigma_{f(M_i)}\,\sigma_{f(M_j)}} \qquad (3)$$

$\sigma_f(M_i)$ is the standard deviation of the meteorological element in point M_i, and $\sigma_f(M_j)$ is the same in point M_j.

A characteristic type of the structural and correlation functions is presented in Fig. 1. with d meaning the distance.

From Fig. 1. one can see that the structure function becomes saturated at the value of $2\sigma_f^2$, in the case f homogenous and isotrop fields. It can be seen, too that $\mu_f(o) = 1$. The value of $\mu(o)/e$ is often called as the radius of correlation.

Having the correlation function of the meteorological elements, one can calculate the interpolation weights, and from them the interpolation errors and the minima of errors is characterizing the optimal meteorological networks.

The third possibility way for the determination for rational density of the meteorological networks is to calculate the average quadratic error of the territorial average of elements. This method was developed by R.L. Kagan /1966/ and it is in a very close connection with the interpolation method. Kagan's method is used for the determination of the rational density of the precipitation network.

The problem of the rational planning of the meteorological network is closely connected with economical considerations. The solving of this problem is very difficult and only few artides deal with it. Among them remarkable are the works of L.S. Gandin and R.L. Kagan /1967/ and E. Freydank's /1986/ article.

If the amount of outlap is (A) which contains the sum of maintance and existance of a network (C), and the loss is (L) which is arising from insufficient knowledge, we have to minimalize

$$A = C + L$$

The first term of this economical balance, the amount term is very important, but it is very difficult to estimate it. In the term C is included the data processing, too. The of the term L is simitary hard to estimate, because we have little information about the economical profit of meteorological informations. In spite of these difficulties, in my opinion, this problem is fundamental for the designing of meteorological networks, and I hope that this conference will contribute to the solution of this problem.

3. THE INVESTIGATIONS ON TOPIC OF DESIGNING OF METEOROLOGICAL NETWORKS

After this short outline of the problems and of the methods to solve the optimal design of meteorological networks, one may put the question: why is this problem so important? The answer is very easy. Without solving this problem nobody can answer a lot of questions. For example, we can not scientifically evaluate a climatological map, we can not interpolate or extrapolate meteorological data when we have no information about the interpolation errors. Without the knowledge of interpolation errors we can not combine or interpret the measurements of two different system, e.g. the data of rainfall gauge with the values of measurement of the intensity of precipitation by radar. Moreover, the experience of any real meteorological network gives impulses and advices in the case of the establishment of a new network.

Both the structure function and the correlation function have been determined experimentally, and their shape and behaviour depend on the climate, therefore it is no wonder that a lot of investigations deal with this problem.

I myself have been following with great interest the publications on this topic from the year 1965. In my collection there are about 190 of such publications. I know very well that this collection is far from being complete, but I believe that it gives not a bad cross-section and it may be sufficient to make some generalizations.

On the basis of my collection, according to Table 1., the intensity of interest in the topic was very great during the investigated period.

From Table 1. one can see that the years 1971-75 and 1976-80 were very intensive periods of the investigations, but the problem has remained up to the present, too.

Maybe I can further increase this interest with the data in Table 2., where the statistics of the choise of this subject can be found on the basis of the data from my collection.

About 47 % of the publications deal with the problem of precipitation. This is not surprising, because this topic has been very important both economically and meteorologically. Hydrologists also are greatly interested in precipitation.

The theoretical interest in the topic is also great. This is indicated by the fact that 28,4 % of the articles deal with general problems with theoretical detailes. In Table 2. the importance of radar-rainfall is shown,

which can also be related to the problem of precipitation. If I may say so, it is interesting that a considerable part of works deal with the problem of rational density of the network of chemical properties of precipitation.

In addition, on the basis of the data in Table 3., we try to give a picture of territorial distribution of publications in terms of the languages they were written in.

Since English is a very important language of science, it is no wonder that more than half of the publications have been written in this language. The intensity of investigations on this topic in the Soviet Union is emphasized by the great number of publications in Russian. In the number of publications in other languages one can find works in Spanish, Hungarian, Bulgarian, Polish and so on. The data in Tabel 3. clearly show that the investigations on this topic have an international interest.

4. THE INFORMATION THEORY AND METEOROLOGY

The stochastic caracter of climatological elements means that their distribution in space and time show several uncertainties. The degree of uncertainty is varying, both in space and in time. There is a real demand to measure the degree of uncertainty. The concept of statistical entropy was introduced by C.C. Shannon /1948/ and with it's help we can measure the uncertainty of an experiment. In the information theory the concept of experiment is given in a totally general manner. For instance, in the terminology of this theory a distribution function of any meteorological element is called an experiment. This experiment gives us the p_i probability of i-th event.

The statistical entropy, or Shannon's entropy of the distribution is

$$H(p_1, p_2, \ldots, p_n) = -\sum p_i \ln p_i \qquad (4)$$

and it is a non negative number. It has a maximum value if $p_1 = p_2 = \ldots = p_n = 1/n$, when the experiment consists of n equivalent probable events. Accordingly, the definition of statistical entropy $H(p_1, p_2, \ldots, p_n) \equiv 0$, if $p_i = 0$ when $i \neq j$ and $p_j = 1$, namely when p_j is a surely occurring event. Both properties of the statistical entropy are evident, because in the case of a surely occurring event there is no uncertainty.

If we have two events

$$X\left[p(x_1), p(x_2), \ldots, p(x_n)\right] \quad \text{and} \quad Y\left[p(y_1), p(y_2), \ldots, p(y_n)\right],$$

we can calculate the connexion entropy.

$$H(X,Y) = -\sum_{i=1}^{n} \sum_{k=1}^{m} p_{ik} \ln p_{ik} \qquad (5)$$

This entropy of the discrete probability field $p(x_i, y_k) = p_{ik}$ is the degree of uncertainty, too. If the probability field has $n \cdot m$ event fields, and there is a case of an identical distribution, then all $p_{ik} = 1/m \cdot n$. The degree of uncertainty has in this case a maximum

$$H_{max}(X,Y) = \ln(n \cdot m) \qquad (6)$$

If between X and Y there is no stochastic relationship, that is $p(x_i, y_k) = p(x_i) \cdot p(x_k)$, from (5) we can write

$$H(X,Y) = - \sum_{i=1}^{n} \sum_{k=1}^{m} p(x_i)p(x_k) \left[\ln p(x_i) + \ln p(x_k) \right] =$$

$$= - \sum_{i=1}^{n} p(x_i) \ln p(x_i) - \sum_{k=1}^{m} p(x_k) \ln p(x_k) =$$

$$= H(X) + H(Y) \qquad (7)$$

consequently, the entropy H(X,Y) is the sum of the entropies of primary events.

The information quantity for two events is

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \qquad (8)$$

Since $H(X,Y) = H(Y,X)$ it is true that $I(X,Y) = I(Y,X)$.

If H(X) and H(Y) are in no stochastic relationship, then $I(X,Y) = 0$.

Equation (8) measures the degree of relationship between the events X and Y.

Both the H(X) entropy and the I(X,Y) information quantity are very important in the analysis of meteorological processes. It was mentioned earlier that because of the stochastic character of meteorological events, the results of measurements have uncertainties. The degree of uncertainty is different from time to time, therefore we have to calculate it systematically.

In my opinion J.L.Holloway and M.A.Woodbury /1955/ were among the first meteorologists to have applied this theory. They used it in the verification of forecasts. Another example for the application of the information theory in weather forecast can be found in Wahl's /1955/ work. In his publication van der Bijl /1959/ deals with the problem of weather forecast. Other applications have been given by E. Suzuki /1958/, Koo Chen-Chao /1957/, R. Roth /1963/ and by A. Hoffmann /1961/.

On the basis of the results of information theory predictors for statistical forecast have been searched f for by N.N. Tokayev /1971/ and Rákóczi F., Szidarovezky F. /1980, K. Kohlsche /1957/ proposed a relationship between the predictors.

From the climatological utilizations we may mention the works by Dady, Guy /1958/ and W. Fett /1973/. M. Olberg /1951, 1967/ described the statistical behaviour of wind vectors in Potsdam with the information theory. T. Hoshino /1961, 1968/ used the information theory for the evolution of weather locality. Rákóczi F. and Szidarovszky F. /1975/ determined the values of information quantities for the temperature and precipitation series of Budapest. Pelle L. /1980/ also determined the information quantities for temperature series in Europe.

N.A. Bagrov /1957/ showed that the concept of statistical entropy can be used for the determination of the degree of connection of meteorological phenomena.

N.A. Bagrov /1963/ also showed that statistical entropy can be utilized for measuring the similarities or differences of meteorological fields.

Rákóczi F. and Szidarovszky F. /1980/ used the information quantity to search for predictors on the basis of Manley's Central England temperature data, and developed a statistical forecast technique.

D.M. Sonetchkin /1966, 1970/ dealt with the classification of cloud types according to the information theory. The information theory has even been used for the discussion of the feature of turbulence by N. Romanof /1966/ and by J. Horák /1974/.

We can find work in which the results of information theory have been used for the designing of meteorological networks. Among them are the work of Rákóczi F. /1966/, that of Czelnai R. et al. /1966/ and work of T. Husain, M. Ukayli /1982/.

In recent times H. Weingärtner /1985/ discussed the question of connection between correlation and information and he also spoke about the connection of information and contingency coefficient. [Weingärtner /1986/].

M. Oldberg and Rákóczi F. /1984/ in a monography collected the results of meteorological applications of the information theory and showed that the structure function and the entropy-distance function are similar characteristics.

In the case of normal distribution of event X the statistical entropy is

$$H(X) = \ln (\sqrt{2\pi}\, e\, \sigma) \tag{9}$$

From this we can calculate the standard deviation σ for normal distribution.

If we deal with pressure field and the event is

$$X = p_i - p_j \tag{10}$$

where p_i is the value of pressure in point M_i and p_j is the same in point M_j, then the structure function is

$$b(d) = \overline{X^2} \tag{11}$$

The standard deviation of this event depends on d, too and one can write

$$\sigma^2(d) = \overline{(X - \overline{X})^2} = \overline{X^2} - \overline{X}^2 \tag{12}$$

or

$$\sigma^2(d) = b(d) - \overline{X}^2 \tag{13}$$

therefore

$$b(d) = \sigma^2(d) + \overline{X}^2 \tag{14}$$

If the probable value of event X is zero, the structure function and the entropy-distance function are the same characteristics, because $b(d) = \sigma^2(d)$. In a general case the relation (14) is true.

In our opinion the above mentioned examples prove the importance and useful applications of the information theory in meteorology. I hope that this theory will give a lot of new results in the future and I hope, too, that this conference will contribute to the acceleration of the investigations on this topic.

Table 1. The intensity of research work on meteorological networks in precentage of total number of articles in literature

	Number	%
1966 - 70	18	9,47
1971 - 75	82	43,16
1976 - 80	61	32,11
1981 - 86	29	15,26
	190	100,00

Table 2. Statistics of the choise of subjects

	Number	%
Precipitation, Hydrology	71	37,4
General, Principle	54	28,7
Temperature, Radiation	15	7,9
Radar-rainfall connection	12	6,3
Pressure	7	3,7
Automatic stations	7	3,7
Chemical properties	7	3,7
Mezo network, Buoy	7	3,7
Other	6	3,1
Wind, Strom	4	2,1
	190	100,0

Table 3. The language territory distribution of publications

	Number	%
English	112	59
Russian	39	21
German	19	10
Franch	6	3
Other	14	7
	190	100

REFERENCES

Bagrov, N.A., 1957: Statisticheskaya entropiya kak mera neopredelennosti i svyazannosti sluchayniyah yavleniy.
Meteorologiya i Gidrologiya, No. 9. 43-48.

Bagrov, N.A., 1963: Statisticheskaya entropiya kak pokazatel' skhodstva ili razlichiya meteorologicheskih poley.
Meteorologiya i Gidrologiya, No. 1. 9-15.

Czelnai, R., Dési, F., Rákóczi, F. 1966: The information quantity as a measure of the representativity of meteorological stations.
Idójárás, 70, 321-327.

Dady, Guy, 1958: La theorie de l' information dans problems meteorologiques.
Journ. Sci. de la Meteorologie, 10/39/, 121-131.

Drozdov, O.A., 1936: Metod postroeniya seti meteorologicheskih stancii ravninnoy mesnosti.
Trudy G.G.O. vyp. 12/3/.

Drozdov, O.A., Sepelevskiy, A.A., 1956: Teoriya interpolatsii v stohasticheskom pole meteorologicheskih elementov i yeyo primenenie k voprosam meteorologicheskih kart i ratsionalizatsii seti.
Trudy N.I.Y.G.Y.G.M.C. ser. I. vyp. 13. 65-115.

Fett, W., 1973: Ökonomische Messzeitverteilung meteorologisch abhängigen Größe /ein informationtheoretisch Modell/.
Annales der Meteorologie, No. 6. 131-138.

Freydank, E., 1986: Einige Fragen der Planung von Niederschlagsmeßstellen.
Zeitschrift für Met. 36, Heft 1. 15-26.

Gandin, L.S., 1963: Obyektivnyy analiz meteorologicheskih poley.
Hidrometeoizdat, Leningrad

Gandin, L.S., Kagan. R.L., 1967: Ob ekonomicheskom podkhode planirovaniya meteorologicheskih stantsiy.
Trudy G.G.O. vyp. 208.

Hoffmann, A., 1961: Über den Informationsgehalt von Vorhersagen.
Meteorol. Rundschau, 14/4/, 117-118.

Holloway, J.L., Woodbury, M.A., 1955: Application of information theory and discriminal function analysis to weather forecasting and forecast verification. Tech, Rep. No. 1. to Office of Naval Research, Univ. of Pennsylvania. Inst. for Cooperative Research, Philadelphia

Horak, J., 1974: Pouziti teorie informace k diskusi poloempirika roonice turbulent difuse /k-teorie/.
Met. Zpravy. 27/5/, 132-137.

Hoshino, T., 1961: On the use of the entropy of weather to evolution of weather locality. Pl. On the areal distribution of entropy of weather in Niigate Prefecture.
Journ. of Met. Research, Tokyo, 13/9/, 703-713.

Hoshino, R., 1968: On the use of entropy of weather of the evolution of weather locality. Pl. 4. Areal distribution of entropy of weather in Chibe Prefecture.
Journ. of Met. Research, Tokyo, 20/10/, 545-552.

Husain, T., Ukayli, M., 1982: Meteorological network expansion for Saudi-Arabia.
Journ. de Recherches Atmospheriques, Clemont-Ferrand, France, 16/4/, 281-294.

Kagan, R.L., 1966: K otsenke reprezentativnosti osadkomernyh dannyh.
Trudy G.G.O. vyp. 191. 22-34.

Kármán, Th., 1937: The fundamentals of the statistical theory of turbulence.
Journ. Aeronaut. Sci. No. 4.

Keller, L., Friedmann, A., 1925: Differentialgleichungen für die turbulente Bewegung einer kompressibilen Flüssigkeit.
Proceed of the 1st Congress for Applied Mechanics.

Kohlsche, K., 1957: Bemerkungen zur Verwendung des Begriffes der Informationsentropie für Vorhersagprobleme und Abteilung einer Beziehung zur Bewertung vor Prediktoren.
Zeitschrift für Met. 11/7/, 193-199.

Kolgomoroff, A.N., 1941: Lokalnaya struktura turbulentnosti v neszhimayemoy vyazkoy zhidkosti pri ochen bol'-soy chislah Reynol'dsa.
Doklady An S.S.S.R. Tom 30, No. 4. 299-303.

Koo, Chen-Chao, 1957: The evaulation and utilization of weather forecasting from the viewpoint of communication theory.
Acta Met., Peking, 28/4/, 256-263.

Olberg, M., 1951: Die monatliche Informationsentropie des horizontalen Windvektors in Potsdam für Jahr 1951. Gerl. Beitr. zur Geophys. 76/3/, 201-206.

Olberg, M., 1967: Die Informationsentropie und Beständigkeit des Windes.
Zeitschrift für Met. 19/7-8/, 195-201.

Olberg, M., Rákóczi, F., 1984: Informationstheorie in Meteorologie und Geophysik. Akad. Verlag. Berlin, 181 pp.

Pelle, L., 1980: Európai hómérsékleti mezók információtartalma.
Idójárás, 84/2/, 109-113.

Rákóczi, F., 1966: Die Eigentümlichkeiten des Temperatur- und Feuchtefeldes in Ungarn und die Fragen des Beobachtungsnetzes.
III. Konferenz für Karpatmeteorologie, Belgrad, 27. - 30.Mai 1965, 395-402.

Rákóczi, F., Szidarovszky, F., 1975: Informationsgehalt der Niederschlagreiche von Budapest.
Ann. Sec. Geologica, Tom XIX. 137-143.

Rákóczi, F., Szidarovszky, F., 1976: Budapest hómérsékleti sorának információtartalma.
Idójárás, 80/2/, 86-92.

Râkóczi, F., Szidarovszky, F., 1980: Some statistical parameters of Manley's Central England temperature data.
Idójárás, 84/5/, 255-260.

Romanof, N., 1966: Asupra utalizatii informatiei in difuza atmosfera Romania.
Institut Meteorologie, Bucharest

Roth, R.I., 1963: Futher application of weather information.
Bull. of the Am. Met. Soc., 44/2/, 72-74.

Shannon, C.C., 1948: A mathematical theory of communication.
Bell. Syst. Techn. J., New York, 27, 379-423., 623-656.

Sonetchkin, D.M., 1966: Nekatorye karakteristiki oblachnosti po dannym samoletnyh i nazemnyh nabludenie.
Trudy M.M.C. vyb. 11. 34-41.

Sonetchkin, D.M., 1970: Obrabotka i analiz televizionnyh i infrakrasnyh snimkov s tselu polucheniya informatsii ob oblachnosti.
Trudy Mezhdunarodnovo Sympoziuma po Sputnikoy Meteorologiya, Moscow

Suziki, E., 1958: Weather forecast and entropy in information theory.
Papers in Met. and Geophys., Tokyo, 9/2/, 51-61.

Taylor, G.A., 1937: Statistical theory of turbulence.
Proceed of the Royal Soc., A. 151.

Tokayev, N.N., 1971: K voprosu o vybore informativnyh prediktorov pri realizatsii diskriminatnyh i regressivnyh prognosticheskih skhem.
Trudy Ukr.N.I.G.M.I. vyp. 108. 41-46.

Van der Bilj, W., 1959: Das statistische Entropieverhältnis, ein Hilfsmittel zur Lösung von Vorhersageproblemen.
Meteorol. Rundschau, Berlin/Göttingen, 12/1/, 33-36.

Wahl, E.W., 1955: Das statistische Entropieverhältnis, ein Hilfsmittel zur Lösung von Vorhersageproblemen.
Meteorol. Rundschau, Berlin/Göttingen, 8/1/, 51-54.

Weingärtner, H., 1985. Korrelation und Information.
Meteorol. Rundschau, Berlin, 38/1/, 1-8.

Weingärtner, H., 1986: Zur Methode der Korrelationsanalyse in der Klimaforschung.
Ann. der Meteorologie, Nr. 23., 27-28.

Wiener, N., 1949: Extrapolation, interpolation, and smoothing of stationary time series with engineering applications.
Wiley, New-York

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

METEOROLOGICAL NETWORK DESIGN METHODS - A COMPARATIVE STUDY

Tahir Husain, Mustafa A. Ukayli and Hasin U. Khan

Water Resources and Environment Division
The Research Institute
University of Petroleum and Minerals
Dhahran, Saudi Arabia

ABSTRACT

For proper estimation of the regional variabilities of meteorological parameters using data measured at station locations, it is necessary to check the status of the meteorological network density as well as its distribution using multivariate statistical and stochastic methods. Various available network design methods are, therefore, reviewed. In order to optimize meteorological network configuration in the Kingdom of Saudi Arabia, network design methods using Shannon's Information Theory and Gandin's Objective Analysis are formulated. Daily meteorological data for a period of ten years from the stations within the Kingdom and also from the bordering stations in neighboring countries on both Arabian Gulf and Red Sea sides are processed.

To apply Shannon's Information Theory, the whole study area is divided into feasible triplets by joining the stations. Using the information decay relationship, the total information at station as well as non-station points are computed. Gandin's objective analyses is simplified to determine the maximum spacing between meteorological stations as a function of permissible standard errors in the meteorological fields. Methodology based on interpolation methods is developed and applied to identify low and high density networks. The applicabilities and limitations of these two methods are discussed to improve the network design formulation.

INTRODUCTION

A number of objective analysis methods have been developed in the past and used in various countries. Panofosky [1949] developed the first method to be outlined for objective analysis. This method, which is based on purely analytical interpolation, involves least square fittings of a polynomial to the data. Doos [1955] developed another method of objective analysis based on "first guess" field. A different approach of objective analysis, based on statistical principles under the name of optimum interpolation has been adopted in the Soviet Union [Gandin, 1963; Gandin and Kagan, 1974; Drozdov, 1936; Drozdov and Shepelevski, 1946]. Gandin and Lugina [Gandin and Lugina 1969], claimed that "optimum interpolation analysis" is the best method in comparison to all other methods of objective analysis. The main advantage of the scheme is in its explicit and simple dependence on network density and its accuracy of the observational data. However, the validity of the above scheme will be the function of homogeneity and isotropy of meteorological fields. Another major problem in the application of any objective analysis scheme is the problem of data reliability and detection and the elimination of errors.

In this paper method of objective analysis using linear and optimum interpolation schemes has been formulated and applied to study the performance of the existing meteorological network. Shannon's information theory has also been used to derive information at various grids in the region and applied to identify locations for network expansion. The details of the methodology and application to meteorological network design are discussed in the following sections.

FORMULATION - OBJECTIVE ANALYSIS

The meteorological network design using objective analysis is based on the application of statistical characteristics of random fields. These include the estimation of the terms such as covariance structure and the correlation functions in space domain.

Gandin [1963] describes the relationship between the standard error ES and interpolation error ζ as follows

$$ES = \sigma(\zeta)^{\frac{1}{2}}$$

or $$\zeta = (Es/\sigma)^2 \quad (1)$$

where σ is the true standard deviation of the meteorological variable.

Interpolation error which is a function of correlation field is estimated using linear and optimum interpolation methods. In case of linear interpolation, information about the statistical structure is used only for estimating the accuracy of interpolation and the coefficients by which the results of interpolation at points are changing, i.e. the interpolation weights are defined beforehand. In case of optimum interpolation, the interpolation weights are not laid down beforehand but sought on the basis of information concerning the statistical structure and the accuracy of observations. The interpolation weights in this case are estimated by minimizing the standard error. Therefore, the standard error of optimum interpolation is generally less than the standard error of linear interpolation.

The estimation of interpolation errors for various geometric events using both optimum and linear interpolation concepts have been discussed in detail by Gandin [1963]. Here the interpolation error at the middle of station pair only is listed.

Linear interpolation error 'ξe' at the middle of station pair distance 'd' apart is:

$$\xi e = 1.5 - 2r(d/2) + 0.5\, r(d) + 0.5\, \eta \quad (2)$$

Similarly optimum interpolation error 'ζopt' at the center of the station pair is given by

$$\zeta opt = 1 - [2\, r(d/2) \,/\, (1 + \eta + r(d))] \quad (3)$$

where

$$r(d) = ro\, e^{-d/D} \text{ and } r(d/2) = ro\, e^{-d/2D} \quad (4)$$

$r(d)$ correlation coefficient as function of distance 'd'

η noise index $= \bar{E}i^2/\sigma i^2$

$\bar{E}i^2$ mean square error in measurements

σi^2 true variance of the meteorological variables at ith location.

The coefficient ro and correlation radius 'D' are calculated as follows:

$$ro = \overline{\ln r} + \bar{d}/D \quad (5)$$

where

$$\overline{\ln r} = \sum_{i=1}^{m} \ln ri/N = \text{mean value of logarithmic of correlation coefficient ri}$$

$$\bar{d} = \sum_{i=1}^{m} di/N = \text{mean value of the distance be station pairs}$$

$$1/D = \sum_{i=1}^{m} (di-\bar{d})\ (\ln ri - \overline{\ln r}) \ / \sum_{i=1}^{m} (di-d)^2$$

ri = ith pair of correlation coefficient between two points di apart

But we know from Eq. (1) that $\zeta = (Es/\sigma)^2$

Substituting the value of ζ, r(d) and r(d/2) and solving for 'd' we get

(a) Linear interpolation

$$d = - D \ln [\{1+\eta(\sigma^2-Es^2)/\sigma^2\}/\{2\ ro^2 - ro(\sigma^2-Es^2)/\sigma^2\}] \quad (6)$$

(b) Optimum Interpolation

$$d = - 2D [\ln\sqrt{(4-C)} - 2] \quad (7)$$

where

$$C = [\eta + 3 - 2\ Es^2/\sigma^2] \ / \ ro \quad (8)$$

FORMULATION - SHANNON'S INFORMATION THEORY

Consider three stations (I,J,K) forming a subregion as shown in Figure 1. The distance between the stations I and J is 'D' and the distance from the station K to I and J are 'd_1' and 'd_2' respectively. The entropy H(Xi), at station I using station pair I and J can be rearranged as follows:

$$H(Xi) = [H(Xi) - H(Xi,Xj) + H\ (Xj)] + H(Xi,Xj) - H(Xj) \quad (9)$$

where

H(Xi) and H(Xj) are the entropies at station I and J respectively and H(Xi,Xj) is the joint entropy for station pair (I,J).

The entropy terms are typically defined as follows [Shannon and Weaver, 1949]:

$$H(Xi) = - \sum_{k=1}^{n} p(x_i^k) \log_2 p(x_i^k) \quad (10)$$

and

$$H(Xi, Xj) = - \sum_{k=1}^{m} p(x_i^k, x_j^k) \log_2 p(x_i^k, x_j^k)$$

where

$p(x_i^k)$ is the probability of occurrence of kth event for variable 'Xi'; and $p(x_i^k, x_j^k)$ is the joint probability of kth event for variable Xi and Xj but the information transmission between station pair (I,J), denoted by T(Xi;Xj), is equal to H(Xi) - H(Xi,Xj) + H(Xj). The relationships between information transmission and its decay with the distance is shown in Figure 2.

Hence

$$H(Xi) - T(Xi;Xj) = H(Xi,Xj) - H(Xj) \quad (11a)$$

and

$$H(Xj) - T(Xi;Xj) = H(Xi,Xj) - H(Xi) \quad (11b)$$

Similarly between station pairs (I,K) and (J,K), the entropies can be rewritten as follows:

$$H(Xi) - T(Xi;Xk') = H(Xi,Xk') - H(Xk') \quad (12a)$$

$$H(Xk') - T(Xi;Xk') = H(Xi,Xk') - H(Xi) \quad (12b)$$

$$H(Xj) - T(Xj;Xk') = H(Xj,Xk') - H(Xk') \quad (13a)$$

$$H(Xk') - T(Xj;Xk') = H(Xj,Xk') - H(Xj) \quad (13b)$$

In the above equation the right hand side terms are unresolved information and is assumed to decay exponentially with the distance. In order to interpolate information at different locations along the line segment joining two station pairs, the relationship has been developed [Husain, Ukayli, and Khan, 1986] and is summarized below:

$$H(Xi) = T(Xi;Xj) + [H(Xj) - T(Xi;Xj)]e^{-b(D-d_1)/d_1, d_2} + [H(Xi) - T(Xi;Xj)]e^{-bdl/(D-d_1)} \quad (14)$$

where H(Xi) is the total interpolated information at a location 'I' on line I-J at a distance 'd_1' from I.

where,

$$b = [d_2/(D-d_2)] \ln [H(Xi) - T(Xi;Xj)] / [T(Xi;Xk') - T(Xi;Xj)] \quad (15)$$

Using the above equation, the interpolated information at various locations in the study area can be computed by taking various combinations of three station sites.

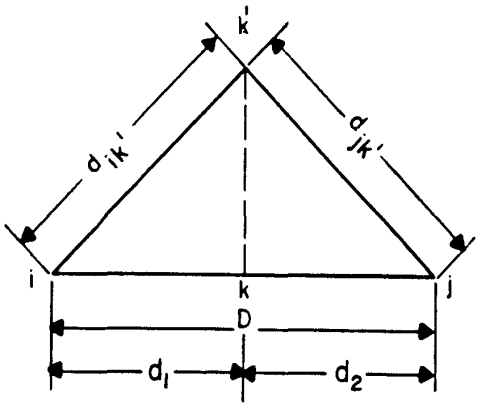


Figure 1. Representative Set of Three Stations

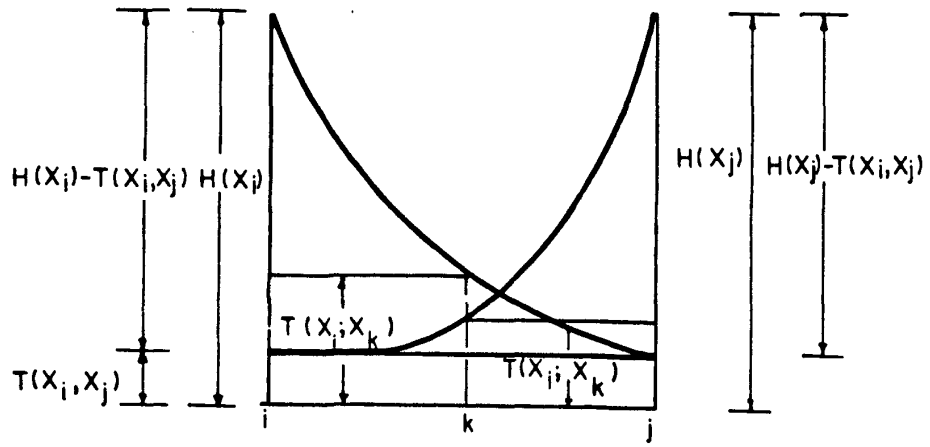


Figure 2. Schematic Diagram Showing Entropy, Information Transmission and Decay Curve (Husain, Ukayli, and Khan, 1986)

ANALYSIS

Meteorological data on a daily basis for a period of nine years for Saudi Arabian stations and bordering meteorological stations in Jordan, Syria, Iraq, Iran, Kuwait, Bahrain, the United Arab Emirates, Qatar, North Yemen, South Yemen, Egypt, Sudan, Ethiopia and Somalia were used to assess network density. The selected meteorological stations are shown in Figure 3. Data on mean sea level pressure, temperature, sky cover, horizontal visibility, precipitation, and wind speed were processed. The whole study area was divided into three subregions. Set of feasible station pairs were selected to estimate correlation field, standard deviation and mean square of the measurement errors for each meteorological variable independently. The values of ro, 1/D and coefficient of correlation for each meteorological variable in different regions are listed in Table 1. The mean square error of measurement was also computed for a known standard errors. Using the relationship for estimating the minimum distance required for a network within the specified standard error Eqs. (7 and 8) the distances at various grids of the study area were computed for different meteorological variables. The contour maps were drawn to show the variabilities in the minimum distance throughout the study area. The critical zones thus identified for different variables were superimposed to identify the locations for network intensifications. These locations are shown in Figures 4 and 5.

Variables	Region I			Region II			Region III		
	ro	1/D	r^2	ro	1/D	r^2	ro	1/D	r^2
Pressure	0.829	-.0003	.046	0.931	-.0007	0.1850	0.931	-.0048	0.445
Temperature	0.895	-.0008	.141	0.919	-.0013	0.290	0.699	-.0057	0.296
Sky Cover	0.749	-.0082	.404	0.894	-.0093	0.643	0.520	-.0127	0.468
Hor. Visibility	0.319	-.0124	.295	0.532	-.0142	0.563	0.127	-.0061	.097
Precipitation	.068	.0067	.068	0.140	-.0007	0.187	0.147	-.0038	.095
Wind Speed	.460	.0133	.438	0.673	-.0250	0.696	0.141	-.0079	.323

Table 1. Values of ro, 1/D and Coefficients of Correlation for Meteorological Variables in Each Region

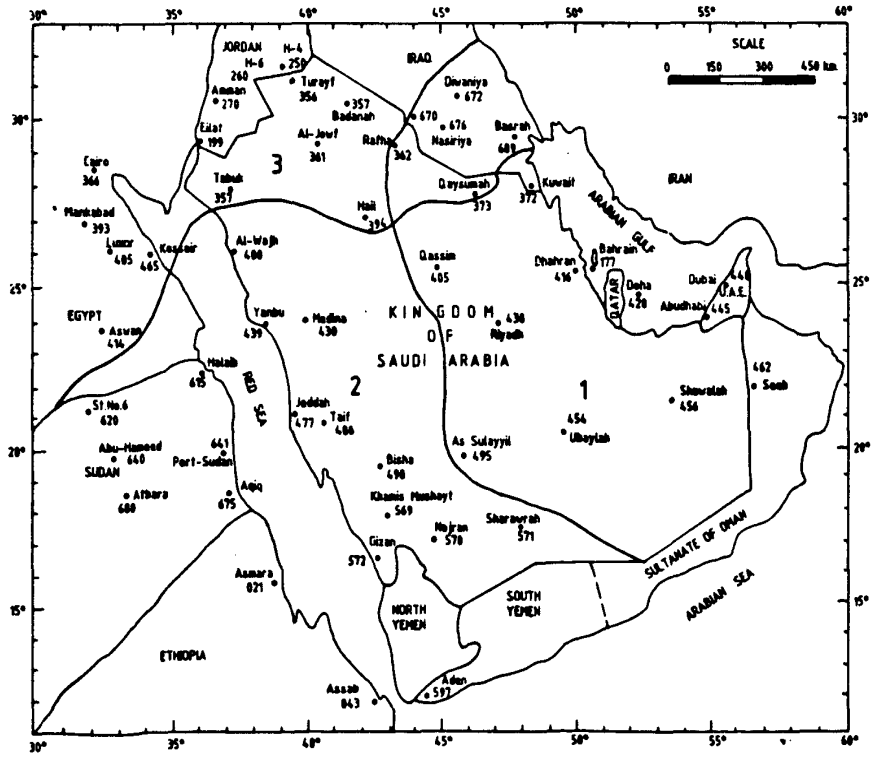


Fig. 3. Map Showing Locations of Meteorological Stations (WMO) in Neighboring Countries of Saudi Arabia

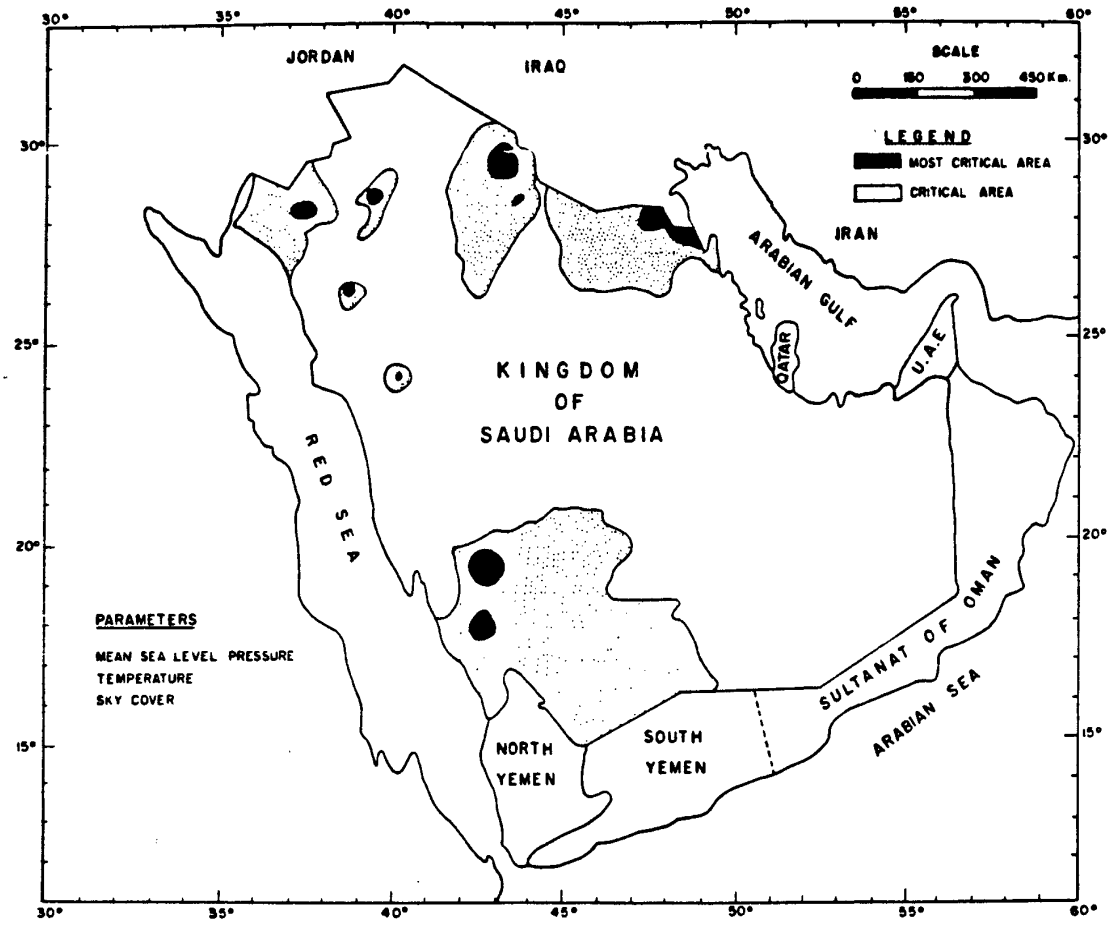


Figure 4. Map Showing Critical Areas Using Gandin's Approach (Linear Interpolation - Regional Data)

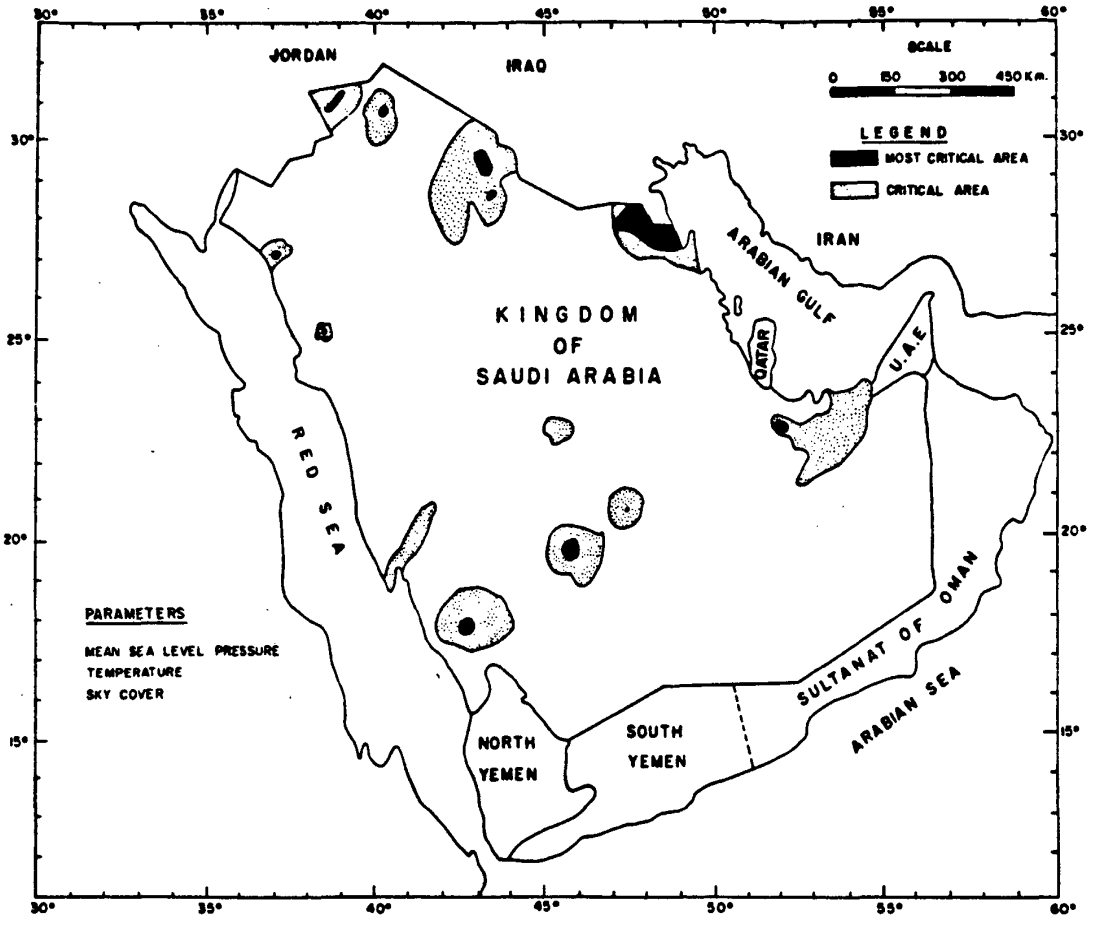


Figure 5. Map Showing Critical Areas Using Gandin's Approach (Optimum Interpolation - Regional Data)

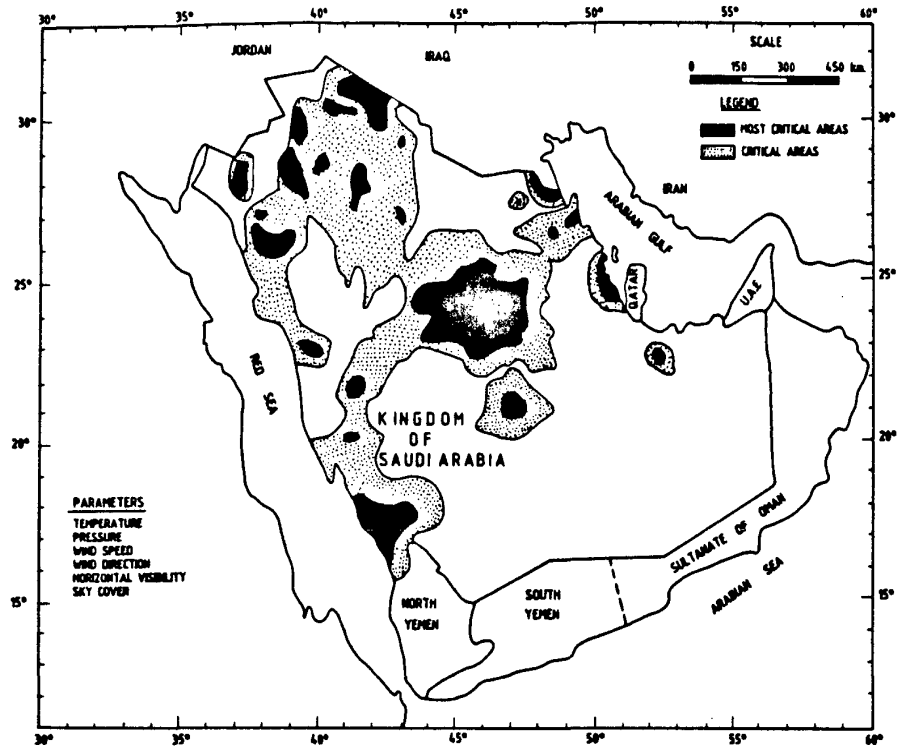


Figure 6. Map Showing Critical Areas with Least Information Transmission (Discrete Entropy - Regional Data)

In order to analyze the data using the Eq. (14) based on Shannon's information theory, each region was represented by triplets by joining three stations. The meteorological records on different variables were used to compute single and bivariate entropies. The information at both station and non-station locations were calculated. The interpolated information values were later used to draw contour maps. Low information zones for all variables were identified and transferred on to a common map as shown in Figure 6.

By comparing Figures 4 and 5, it is noticed that objective analysis approach is able to identify the critical areas for network intensification purposes in southern and northern regions only while the analysis based on Shannon's information concept gives better resolution. It may be due to linearity, homogeneity and normally assumptions used in deriving the relationship which in real sense is not valid. On the other side, the information interpolated at various grids using Shannon's information concept is not restricted to any of the above assumptions.

Based on the identification of low information zones, the potential monitoring sites for network intensification are proposed. The locations of the existing and proposed stations are shown in Figure 7. The priorities to intensify the network are grouped according to the occurrence of a common site. The second priority stations are those which are identified only once.

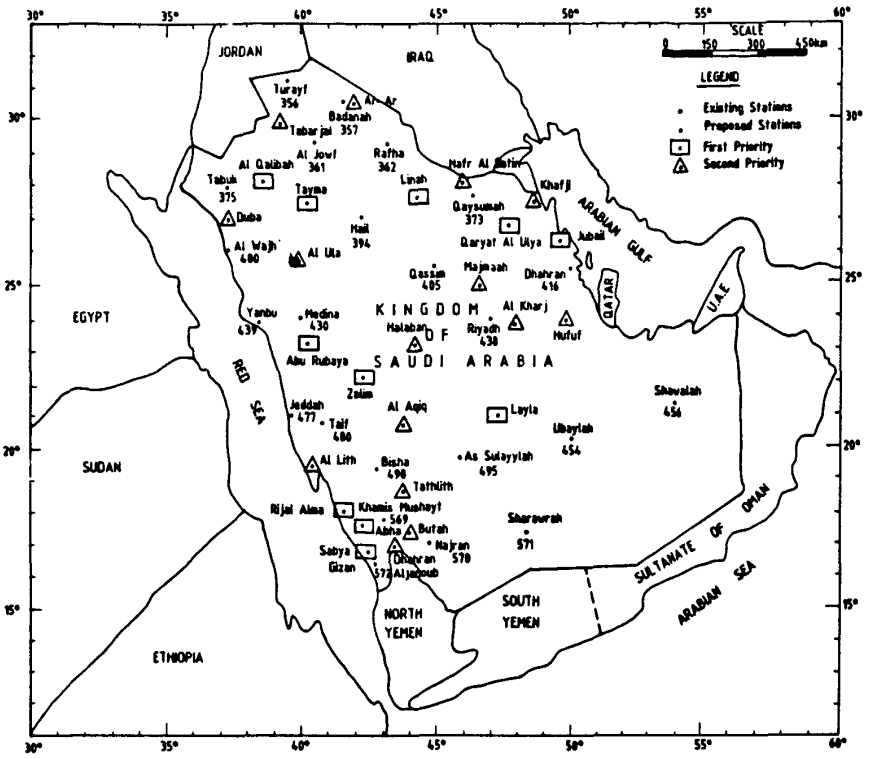


Figure 7. Map Showing Locations of Existing and Proposed Meteorological Stations in Saudi Arabia

CONCLUSIONS

Shannon's information concept is a powerful multivariate technique to evaluate regional uncertainties and variabilities associated with data in both space and time domains. The technique is not restricted to linearity and normality assumptions. On the other hand, the applicability of Gandin's objective analysis which is based on homogeneity assumptions is limited particularly in a big region with sparse network locations.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Abdulbar Al-Gain, Vice President, Meteorology and Environmental Protection Administration (MEPA), and Dr. Abdallah E. Dabbagh, Director, Research Institute, University of Petroleum and Minerals for their permission to publish the paper. This work was funded by the Meteorology and Environmental Protection Administration, Jeddah, Saudi Arabia under grand number 554007.

REFERENCES

Doos, B.R., 1956: Automation of 500 mb forecasts through successive numerical map analyses. Tellus, 8, No. 1.

Drozdov, O.A., 1936: A method of establishing a network of stations in a flat area. Trudy GGO 12.

Drozdov, O.A., and Sepelevski, A.A., 1946: The theory of interpolation of meteorological elements in a stochastic field and their application to questions of weather maps and network rationalization. Trudy NIU GUGMS Series 1, No. 13.

Gandin, L.S., 1963: Objective analysis of meteorological fields: Hydrometeorological Publishing House, Leningrad.

Gandin, L.S., and Lugina, K.M., 1969: Comparison of the accuracy of objective analyses produced in various countries. Proc. WMO/IUGG on Numerical Weather Prediction, Tokyo, 3. VI-VII.

Gandin, L.S., and Kagan, R.L, 1974: Construction of a system for objective analysis of hetrogenous data based on the method of optimum interpolation and optimum agreement. English Translation, Joint Publication Research Service, Arlington, VA.

Husain, T., Ukayli, M.A. and Khan, H.U., 1986: Meteorological network expansion using information decay concept, J. Atmospheric and Oceanic' Technology, 3, 27-35.

Panofsky, H.A., 1949: Objective weather map analysis. Journal of Meteorology, Vol. 6, No. 6.

Shannon, C.E., and Weaver, W., 1949: Mathematical theory of communication. University of Illinois Press, Illinois.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

OPTIMIZATION IN MESOSCALE METEOROLOGICAL NETWORK DESIGN

Stephen E. Jacobsen
James Brucker

Electrical Engineering Department
University of California at Los Angeles
Los Angeles, CA 90024

Francis Fujioka

Forest Service, U.S. Department of Agriculture
4955 Canyon Crest Drive
Riverside, CA 92507

Introduction

Forest managers responsible for fire programs typically are concerned with the effectiveness of current fire weather station locations. The development of remote automatic weather stations (RAWS) has spurred a trend toward reconfiguring fire weather networks, but increasing budgetary constraints and cost of meteorological services and equipment underscore the importance of designing networks to make effective use of resources.

This paper describes how optimization may be used in meteorological network design, and reports some results of a design study for a network of remote weather stations used to monitor fire weather conditions in Southern California. The study involved several phases, primarily: modeling and the selection of a suitable measure of network effectiveness; developing numerical methods for optimizing the network design with respect to the measure of effectiveness; and, comparing alternate designs resulting from different models, measures, and design parameters.

Models

The fire weather manager is interested in estimating the potential for fire in some region **R**. Several point measures of fire potential have been proposed, such as the Fire Weather Index (FWI) (Fosberg, 1978) which was used in this study. These point measures are functions of the values of weather variables (e.g. wind speed, temperature) and/or local terrain variables (e.g. slope, fuels) which yield a value at each point in **R** which attempts to summarize the fire potential. For example, the FWI at a point **x** is given by

$$\mathrm{FWI}(\mathbf{x}) = \frac{\nu(T, h)\sqrt{1+U^2}}{0.3002} \tag{1}$$

where U is the windspeed at **x** and $\nu(T, h)$ is a function of the tempurature T and relative humidity h at **x**, which is given by Fosberg. In general, then, the measure of fire potential is a function $f(\mathbf{x},\mathbf{P}_{\mathbf{x}})$ which may depend on the position **x** and a vector of weather and terrain variables $\mathbf{P}_{\mathbf{x}}$ at **x**.

The vector $\mathbf{P}_{\mathbf{x}}$ at each point **x** in the region will not generally be known; and hence must be estimated. By using a network of RAWS located at points $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, the values of $\mathbf{P}_{\mathbf{s}_i}$ at each RAWS location can be measured and used as a basis for estimates at other points. There are two ways in which the RAWS data can be used to form estimators:

(1) use the values of P_{s_i} at the RAWS locations to form an estimate P_x^e of the weather vector at x, and then compute $f(x,P_x^e)$.

(2) use the values of $f(s_i,P_{s_i})$ computed at the RAWS locations to form an estimate of $f(x,P_x)$.

The latter approach was used in our study. This approach has the advantage of alleviating the data requirements and produces a more tractible objective for optimization with respect to the measures of effectiveness discussed below.

The two models considered in the study were the Distance Weighted Interpolation (DWI) model (Fujioka, 1983) and a stochastic model. In the DWI model, a linear estimator of f at $\mathbf{x}$ is used, of the form

$$f_e(x,S) = \frac{\sum_{i=1}^{n} \mu(s_i,x) f(s_i,P_{s_i})}{\sum_{i=1}^{n} \mu(s_i,x)} \tag{2}$$

where S denotes the vector $(s_1, s_2, ..., s_n)^T$. Examples of distance weighting functions (Fujioka, 1983) are

$$\mu(s_i,x) = e^{-\alpha\|s_i - x\|^\beta} \tag{3}$$

and

$$\mu(s_i,x) = 1 + \alpha\|s_i - x\|^\beta \tag{4}$$

where α and β are positive constants.

The goal of the DWI model is to produce an estimator which is a good fit to a given set of target conditions, with the hope that the the estimator would then be good with respect to actual weather conditions as well. Given the target field $f(x)$, $x \in R$, the DWI estimator $f_e(x;S)$ is computed as in (2) and the value compared to $f(x)$. To get a measure of the goodness of fit over the entire region $\mathbf{R}$, the sum of squared errors over a mesh of points $x_1, x_2, ..., x_m$ is used:

$$SSE(S) = \sum_{i=1}^{m} (f_e(x_i;S) - f(x))^2 \tag{5}$$

For the DWI model with a given target field, a "good" network design is a vector of station positions S and a weighting function μ which minimizes the SSE. However, even for a given functional form of μ such as (3) or (4), the problem of minimizing the SSE with respect to both S and the parameters in μ (e.g., α and β) is computationally intractible; hence, a design is constructed by first selecting the weighting function μ and then minimizing the SSE with respect to the vector of station positions S.

The stochastic model differs from the DWI model in that the underlying idea is to model the values of $f(x,P_x)$ in a way that explicitly accounts for the inherent randomness (variability) in f. In the stochastic model, each $f(x,P_x)$ is a random variable, with expected value $F(x) = E[f(x,P_x)]$, variance $Var(f(x,P_x))$, and spatial covariance $Cov(f(x,P_x),f(y,P_y))$. A natural estimator for $f(x,P_x)$ based on the values of f at the station locations is a linear estimator of the form

$$f_e(x,P_x;S) = a + \sum_{i=1}^{n} \lambda_i f(s_i,P_{s_i}) \tag{6}$$

In this study the estimator was required to be unbiased, which determines the value of the leading constant a and leads to the expression

$$f_e(x,P_x;S) = F(x) + \sum_{i=1}^{n} \lambda_i (f(s_i,P_{s_i}) - F(s_i)) \tag{7}$$

The vector of weights $\lambda = (\lambda_1, ..., \lambda_n)$ depends on the point $\mathbf{x}$, and will in general be different for each point. In the stochastic model, a suitable measure of effectiveness of the network is the total mean squared error (MSE) of the estimators over the region $\mathbf{R}$. At any point $\mathbf{x}$ the expected (or mean) squared error is

$$E[(f_e(\mathbf{x},P_x;S)-f(\mathbf{x},P_x))^2] = \mathrm{Var}(f(\mathbf{x},P_x)) - 2\lambda^T c_{x,S} + \lambda^T C_S \lambda \tag{8}$$

where

$$c_{x,S} = (\mathrm{cov}(f(\mathbf{x},P_x),f(s_1,P_{s_1}), ..., \mathrm{cov}(f(\mathbf{x},P_x),f(s_n,P_{s_n}))^T$$

is the vector of covariances between $f(\mathbf{x},P_x)$ and the values of f at the station locations $\mathbf{S}$, and

$$C_S = [\mathrm{cov}(f(s_i,P_{s_i}),f(s_j,P_{s_j})]_{n \times n}$$

is an n by n matrix of covariances between the values of f at the stations. The vector of weights λ for the estimator at $\mathbf{x}$ which minimize the expected squared error can be solved for analytically, the solution being

$$\lambda = C_S^{-1} c_{x,S} \tag{9}$$

This choice of λ and the resulting estimator for $f(\mathbf{x},P_x;S)$ has been termed *optimal interpolation* by Gandin (Gandin, 1965, 1970) who developed its use for large scale meteorological forecasting in the Soviet Union. In the statistical literature this estimator is referred to as the Minimum Variance Unbiased Linear Estimator or Best Linear Unbiased Estimator (BLUE). Here we will use the latter term.

As a measure of effectiveness for the network, the MSE of the estimators at a mesh of points $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ is summed. Using the Best Linear Unbiased Estimator at each point, the total MSE is

$$\mathrm{MSE}(S) = \sum_{i=1}^{m} (\mathrm{Var}(f(\mathbf{x}_i,P_{x_i})) - c_{x,S}^T C_S^{-1} c_{x,S}) \tag{10}$$

An optimal design with respect to this formulation is a vector $\mathbf{S}$ which minimizes the MSE in (10).

The stochastic model does not place any restrictions on the variance and covariance functions, however, it is desirable that they be continuous functions of location. In many cases, sufficient data to appromixate these functions may not be available, and an expression for the variance and covariance functions must be hypothesized. Random processes found in nature often exhibit a variance which is roughly proportional to the mean, and a spatial correlation which decays exponentially with distance between the location of the processes; this motivates the selection of a covariance function of the form

$$\mathrm{cov}(f(\mathbf{x},P_x),f(\mathbf{y},P_y)) = k\ F(\mathbf{x},P_x)F(\mathbf{y},P_y)e^{-\alpha\|\mathbf{x}-\mathbf{y}\|^2}$$

where k is some proportionality constant. Analysis of available data for the region of Southern California under study indicates that the above covariance function provides a reasonable fit to the experimental correlation, using a decay constant $\alpha = 0.1\mathrm{km}^{-2}$. The proportionality constant k can be factored out of the expression (10) for the MSE, and hence for the purpose of optimization may be ignored (e.g., $k = 1$).

Methods

Optimization of the network design can be performed using the SSE objective (5) for the DWI model or MSE objective (10) for the stochastic model. In either model, the variables to be optimized are the x,y-coordinates of the station locations. This is a continuous optimization problem, with a differentiable objective provided that the target field $f(\mathbf{x})$ and weighting function μ (DWI model) or expected values function $F(\mathbf{x})$ and variance/covariance function (stochastic model) are differentiable. The most efficient algorithms for nonlinear optimization, such as the quasi-Newton methods, all require that the objective function be differentiable. For problems involving several variables (as in the network design problems here) the ability to actually compute the gradient of the objective is necessary for efficient solution of the problem.

The optimization problem resulting from these or other formulations of the network design problem are difficult optimization problems due both the large number of variables to be solved for and the nature of the objective function. Suppose that for some objective function the optimal configuration of stations is

$$S^* = (s_1^*, s_2^*, ..., s_n^*)^T$$

Then, by merely changing the indices on the stations (no change in actual station positions) we would also have that

$$S^{**} = (s_2^*, s_1^*, ..., s_n^*)^T$$

is an optimal configuration, as are any of the other n! possible permutations of the indices assigned to stations. In fact, the objective surface (as a function in the space of all possible S vectors) has an irregular topology with a large number of local minima, perhaps including many which are not optimal. This fact influenced the design of optimization methods for the problem: an interactive program was created which allows run-time changes in design parameters and optimization techniques.

A regularly-spaced grid of approximately 1250 points was superimposed on the region **R** covering part of Southern California, for use as the mesh points in evaluating the SSE and MSE objective functions. The grid point values of the Fire Weather Index for four high fire risk situations were analysed and used to develop smooth interpolated surfaces for the target field $f(\mathbf{x})$ and expected values field $F(\mathbf{x})$ covering the entire region.

Several network design problems were solved numerically, with networks of from 10 to 50 RAWS (i.e., 20- to 100-variable optimization problems). In addition, the presence of several existing weather stations in the region was recognized, and these stations were incorporated in the model as RAWS with fixed locations.

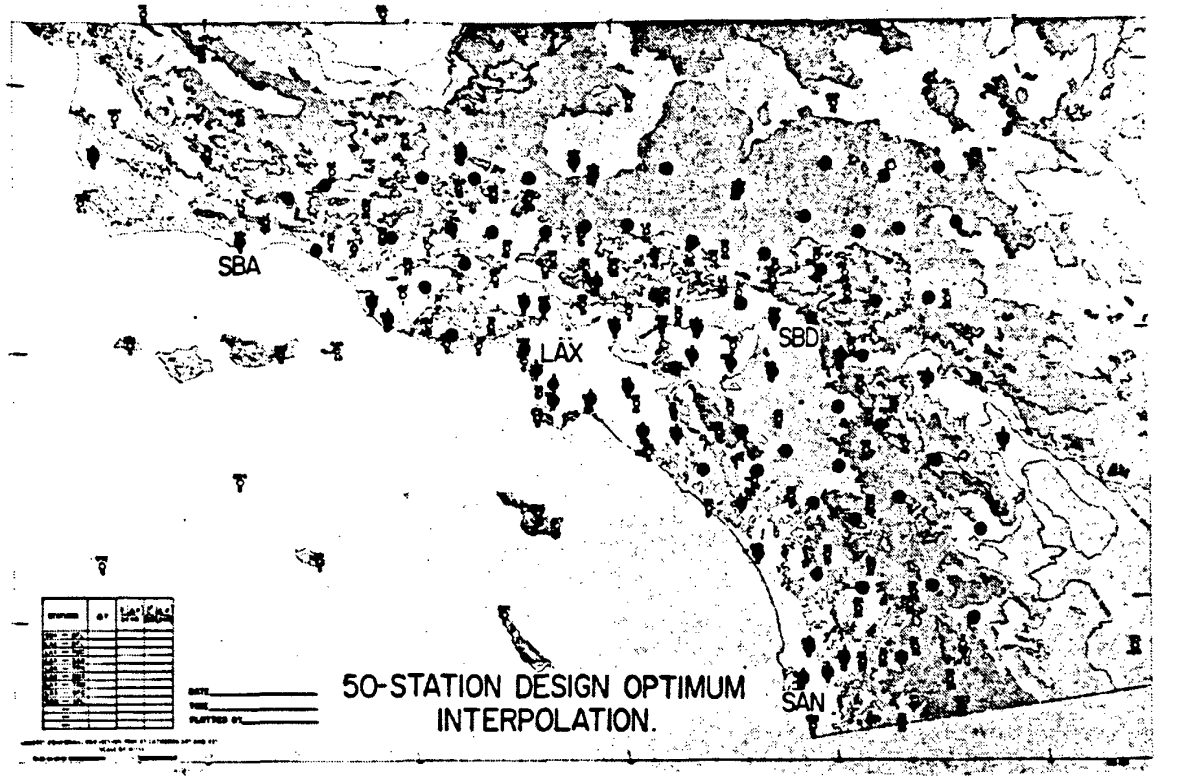
The target fields, initial station positions, optimization procedures, and numerical results are reported in (Jacobsen and Brucker, 1985).

Results

Figure 1 shows the weather station positions for two numerically optimized networks of 50 stations (100-variable problems) and 34 existing (fixed) stations. The solid black dots are the station positions. Figure 1a is the final configuration for the stochastic model, and Figure 1b is the configuration for the DWI model.

To compare the resulting network designs, the stochastic estimators and DWI estimators were computed for each of the four high fire risk weather conditions for which FWI values were available. The results are shown in Table 1. The table shows that the BLUE estimators used in the stochastic model consistently resulted in more accurate estimates.

Comparison of the optimal configurations for a variety of design conditions shows that the stochastic model produces a more even distribution of stations, which tend to accommodate the presence of fixed stations and to position stations where the uncertainty is greatest. An undesirable trait of networks based on the DWI model is the tendency for stations to be bunched together or moved to the boundary of the region in the optimal designs. The stochastic estimators were also found to be more adaptive to a variety of weather scenarios, whereas the DWI estimators typically do well for a weather scenario resembling the target field used in the network design, but produce large errors under other conditions.



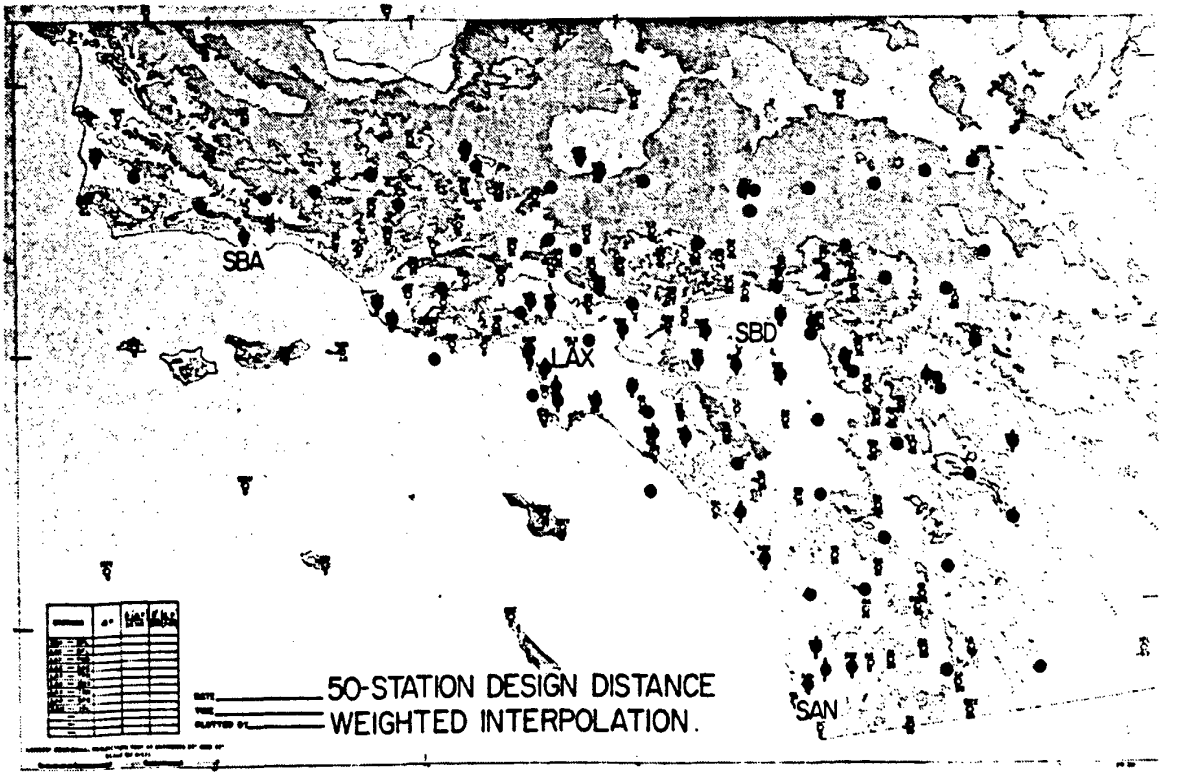


Figure 1: The weather station positions for two numerically optimized networks of 50 stations (100-variable problems) and 34 existing (fixed) stations. Figure 1a is the final configuration for the stochastic model, and Figure 1b is the configuration for the DWI model. The solid black dots are the station positions.

Table 1. Sum of squared estimation errors for optimized DWI and BLUE estimators: comparison using four sets of weather data.				
Model (configuration)	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Stoch. model (Figure 1a)	69963	54671	18660	19959
DWI model (Figure 1b)	143659	109283	57355	21163

Conclusions

This study has shown that optimization can be effectively applied to meteorological network design problems on a scale not previously reported. In addition to its use in locating "good" designs, optimization is valuable in evaluating different network models and exploring the effects of design alternatives. The ability to interact with the optimization process was found to be important in locating a good network design and in subsequent sensitivity analysis which leads to confidence in the solution.

A comparison of the DWI model and a stochastic model with linear unbiased estimators revealed that the stochastic model produces a more acceptable network design, with more adaptive estimators. These results suggest that further research on stochastic models may produce an even better design.

References

Fosberg, M. A., 1978: Weather in Wildland Fire Management: the Fire Weather Index. *Proc. Conf. on Sierra Nevada Meteorology*, 1-4.

Fujioka, F. M., 1983: Weighted Interpolation as an Interpretive Tool--Implications for Meteorological Network Design. *Proc. Seventh Conf. on Fire and Forest Meteorology*, 1-6.

Gandin, L. S., 1965: *Objective Analysis of Meteorological Fields*. Israel Program for Scientific Translations, Jerusalem.

Gandin, L. S., 1970: The Planning of Meteorological Station Networks. *Technical Note No. 111*, World Meteorological Organization.

Jacobsen, S. E. and Brucker, J., 1985: *Final Report on the Development of Nonlinear Optimization Procedures for Meteorological Network Design*. Dept of Electrical Engineering, UCLA, Los Angeles, CA.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

EXTREME VALUE ANALYSIS OF CLIMATOLOGICAL DATA

T.A. Buishand
Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

1 INTRODUCTION

Extreme-value analysis is concerned with the largest (or smallest) values in a record or in a set of several records. Knowledge of the probability distribution of these extremes is of great practical importance. Studies of extremes in climatological records mostly refer to annual or seasonal maxima. An example is given in Fig. 1. The figure shows a probability plot of annual maximum depths of the snow cover at Kleve (West Germany) for 47 years of observation. Two different theoretical distributions were fitted to the data. The straight line represents the Gumbel (or Fisher-Tippett Type I) distribution:

$$F(x) = Pr(X \leq x) = \exp\{-e^{-(x-\mu)/\sigma}\} \qquad (1)$$

and the curved line the two-parameter lognormal distribution:

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right) \qquad (2)$$

where Φ stands for the standard normal distribution function. Both distributions give nearly identical results over a wide range of return periods. However, large differences are found at high return periods. The choice between the two distributions becomes important when return periods of 50 or 100 years are of interest. It is, however, not easy to infer from data which distribution is most appropriate in this case.

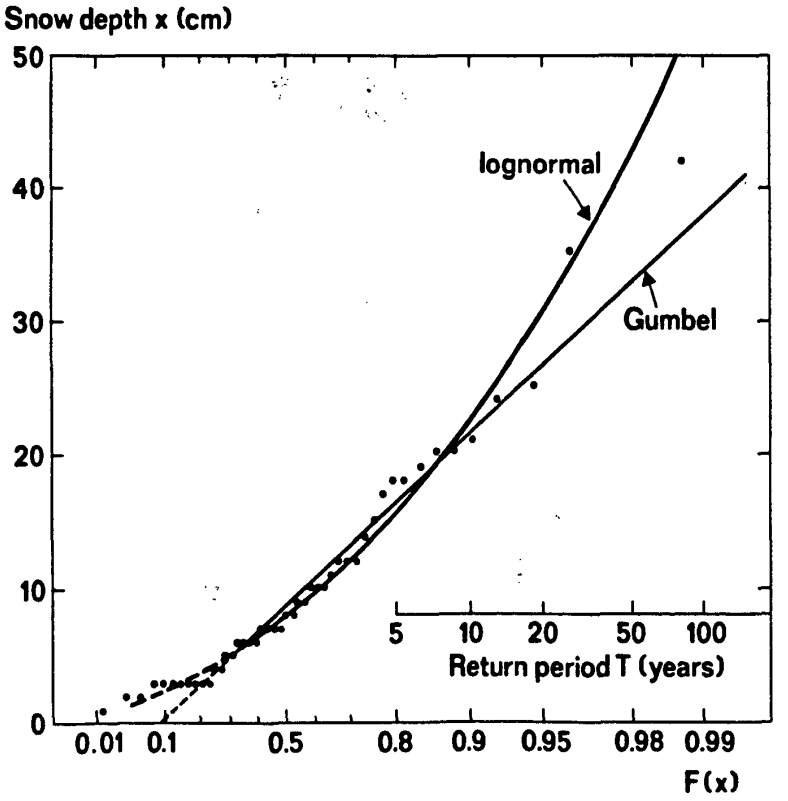


Figure 1 Gumbel probability plot of annual maximum depths of the snow cover at Kleve (51°46'N, 6°09'E) for the period covering the winters of 1934/35-1944/45 and 1947/48-1982/83. After Buishand (1986).

Nowadays there is a large literature on statistics of extremes, ranging from rather theoretical mathematics to engineering applications. On both sides much progress has been made in recent years. The aim of this paper is to review the current status. In

Section 2 some distributional properties of extremes are derived from a Poisson approximation to the stochastic process of high-level exceedances. Then the Generalized Extreme-Value (GEV) distribution is introduced in Section 3. After these two sections about distributions for extremes some recent results on parameter estimation are given in Section 4. Finally, the choice between probability distributions is discussed in Section 5. In this last section much attention is paid to extrapolation beyond the range covered by historical data.

2 THE PARTIAL DURATION SERIES AND THE POISSON APPROXIMATION

A useful method of analysing extremes in a time-series is to consider the exceedances of a high threshold. This technique is well-known in the hydrological literature under the names partial duration series method or peak over threshold modelling. Often a Poisson approximation is applied to obtain the distribution of maxima (NERC, 1975). In the meteorological literature the use of Poisson processes to model extremes was recently suggested by Revfeim (1983), see also Revfeim and Hessell (1984).

2.1 The basic model

The partial duration series consists of local maxima over a high threshold and their times of occurrence. As an example the daily maximum temperatures at De Bilt (The Netherlands) during July 1985 are considered in Fig. 2. Exceedances of the threshold x_0 are found to occur in three clusters. The largest observations in such clusters (peaks)

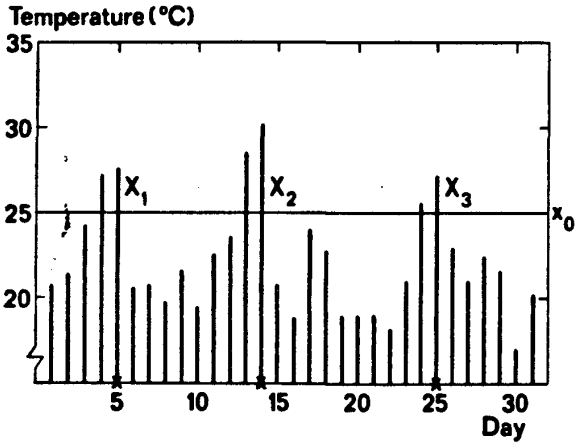


Figure 2 Daily maximum temperatures at De Bilt (52°06'N, 5°11'E) for July 1985. Local maxima (peaks) exceeding the threshold $x_0 = 25°C$ are denoted by X_1, X_2 and X_3, respectively.

and their date of occurrence are the elements of the partial duration series. Provided that the threshold x_0 is high enough, it is reasonable to assume that the X_i's are independent and that their occurrence-times follow a Poisson process. Then the number of peaks N in a year (or season) has a Poisson distribution:

$$\Pr(N=k) = e^{-\lambda(x_0)} \frac{\{\lambda(x_0)\}^k}{k!}, \quad k=0,1,\ldots \tag{3}$$

The parameter $\lambda(x_0)$ gives the mean number of peaks per year (or season). Further, the following holds for the number of peaks exceeding a higher level than x_0:

1. For all $x>x_0$ the number of peaks N_x greater than x has a Poisson distribution with parameter $\lambda(x)$. The quantity $\lambda(x)/\lambda(x_0)$ gives the probability that a peak exceeds x.

2. For $x_2>x_1>x_0$ the number of peaks in the interval $I = [x_1, x_2)$ has a Poisson distribution with parameter $\lambda(I) = \lambda(x_1) - \lambda(x_2)$.

3. Counting the number of peaks in two non-overlapping intervals I_1 and I_2 gives two independent Poisson variables with parameters $\lambda(I_1)$ and $\lambda(I_2)$, respectively (Weissmann, 1978).

The properties 1 and 2 follow from the fact that random deletion (thinning) of events in a Poisson process results again in a Poisson process for the remaining events (NERC, 1975; Leadbetter et al., 1983). Property 3 is a bivariate extension.

Small values can be treated in a similar way by analysing clusters of downcrossings of a low threshold or by adding a minus sign to the data. Because of this similarity small values and minima are not considered further.

2.2 The distribution of extremes

The annual (or seasonal) maximum Z_1 is the largest value of the N peaks occurring in that year (or season):

$$Z_1 = \max(X_1,\ldots,X_N).$$

The event $Z_1 \leq x$ is equivalent to the event that there are no peaks exceeding x or $N_x=0$. Therefore:

$$F(x) = \Pr(Z_1 \leq x) = e^{-\lambda(x)}, \quad x \geq x_0. \tag{4}$$

The Poisson approximation can also be used to derive the distribution of the m-th largest value Z_m and the joint distribution of $Z_1,\ldots,Z_m$. The method is illustrated here for the marginal and joint densities of the m-th extremes.

The density of Z_m is obtained from the probability of the event that there is one peak in the interval $I' = [x, x+\Delta x)$ and there are (m-1) peaks in the interval $I = [x+\Delta x, \infty)$. Since I' and I are disjoint this probability becomes:

$$p_{1,m-1}(I',I) = \{\lambda(x) - \lambda(x+\Delta x)\} e^{-\{\lambda(x)-\lambda(x+\Delta x)\}} \times \frac{\{\lambda(x+\Delta x)\}^{m-1}}{(m-1)!} e^{-\lambda(x+\Delta x)}$$

$$= \{\lambda(x) - \lambda(x+\Delta x)\} \frac{\{\lambda(x+\Delta x)\}^{m-1}}{(m-1)!} e^{-\lambda(x)}. \tag{5}$$

Dividing both sides by Δx and letting $\Delta x \to 0$, gives for the density of Z_m:

$$f_m(x) = -\lambda'(x) \frac{\{\lambda(x)\}^{m-1}}{(m-1)!} e^{-\lambda(x)}, \quad x \geq x_0. \tag{6}$$

For m=1 this expression reduces to the derivative of (4).



Figure 3 Partitioning of the interval $[x_m,\infty)$ into disjoint intervals to obtain the joint distribution of the m largest peaks.

For the derivation of the joint density of $Z_1,\ldots,Z_m$ in the point $(x_1,\ldots,x_m)$ the interval $[x_m,\infty)$ is partitioned into a number of disjoint intervals, as shown in Fig. 3. The joint density is then obtained, in a similar way as eq. (6), from the probability of the event that there is 1 peak in each of the intervals $I'_1,\ldots,I'_m$ and that there are no peaks in the longer intervals $I_1,\ldots,I_m$. The final result is:

$$f_{1,\ldots,m}(x_1,\ldots,x_m) = e^{-\lambda(x_m)} \prod_{i=1}^{m} \{-\lambda'(x_i)\}, \quad x_m \geq x_0. \tag{7}$$

2.3 The distribution of peak magnitudes

The distribution of extremes depends on the distribution of peak magnitudes through the value of $\lambda(x)$. In the literature (NERC, 1975) the latter distribution has

often been described by the following exponential distribution:

$$Pr(X_i>x) = e^{-(x-x_0)/\sigma}, \quad x>x_0. \tag{8}$$

The value of $\lambda(x)$ is obtained by multiplying (8) with the mean number of peaks per year (or season):

$$\lambda(x) = \lambda(x_0)e^{-(x-x_0)/\sigma} = e^{-(x-\mu)/\sigma} \tag{9}$$

where $\mu = x_0 + \sigma \ln\{\lambda(x_0)\}$. Substitution of (9) in (4) leads for $x>x_0$ to the Gumbel distribution given by eq. (1). For the density of Z_m the exponential form of $\lambda(x)$ results in:

$$f_m(x) = \frac{\sigma^{-1}}{(m-1)!} \exp\{-m(x-\mu)/\sigma - e^{-(x-\mu)/\sigma}\} \tag{10}$$

which is an alternative expression for eq. (2) in Section 5.3.1 of Gumbel (1958). Substitution of (9) in (7) gives for the joint density of $Z_1,\ldots,Z_m$:

$$f_{1,\ldots,m}(x_1,\ldots,x_m) = \sigma^{-m} \exp\{-\sum_{i=1}^{m}(x_i-\mu)/\sigma - e^{-(x_m-\mu)/\sigma}\}. \tag{11}$$

The use of this joint density for estimating the parameters μ and σ is discussed in Section 4.

When the peak magnitudes are identically distributed, the distribution of extremes does not depend on seasonal variation in the occurrence-rate of peaks. However, other distributional forms are obtained for the extremes when there is also seasonal variation in the parameter σ of the exponential distribution of peak magnitudes (the resulting $\lambda(x)$ is then no longer exponential). Models with seasonal variation in both the occurrence-rate of the underlying Poisson process and the parameter σ are considered by North (1980) and Revfeim (1982).

Several generalizations of the exponential distribution are possible. A generalization which has been suggested in the light of extreme-value analysis is the Generalized Pareto distribution (Smith, 1984; Van Montfort and Witter, 1985):

$$Pr(X_i>x) = \{1-\theta(x-x_0)/\sigma\}^{1/\theta}, \quad x>x_0. \tag{12}$$

For $\theta=0$ eq. (12) reduces to the exponential distribution (8). The use of the Generalized Pareto distribution for peak magnitudes results for the maximum Z_1 in the GEV distribution, which is discussed in the next section. If $\theta>0$ the Generalized Pareto variate is bounded from above by $x_0+\sigma/\theta$.

3 THE GENERALIZED EXTREME-VALUE (GEV) DISTRIBUTION AND THE DISTRIBUTION OF MAXIMA IN CLIMATOLOGICAL RECORDS

The GEV distribution was proposed by Jenkinson (1955) to have a single expression for the three Fisher-Tippett extreme-value distributions. Its distribution function may be written as:

$$F(x) = \exp[-\{1-\theta(x-\mu)/\sigma\}^{1/\theta}]. \tag{13}$$

For $\theta=0$ the GEV distribution reduces to the Gumbel distribution. The cases $\theta<0$ and $\theta>0$ correspond to the Type II (Fréchet) and Type III (Weibull) distributions, respectively. For $\theta<0$ the GEV variate has a lower bound $\mu+\sigma/\theta$ whereas for $\theta>0$ the variate is bounded from above by $\mu+\sigma/\theta$.

For calculating design values the inverse form is used:

$$x = \mu + \sigma \frac{1-e^{-\theta y}}{\theta} \tag{14}$$

where y is the standard Gumbel (or "reduced") variate:

$$y = -\ln\{-\ln F(x)\} = -\ln\{-\ln(1-1/T)\} \tag{15}$$

with T the return period. The relation between x and y (or T) is shown in Fig. 4 for $\theta<0$, $\theta=0$ and $\theta>0$, respectively. Large differences are found at high return periods. For

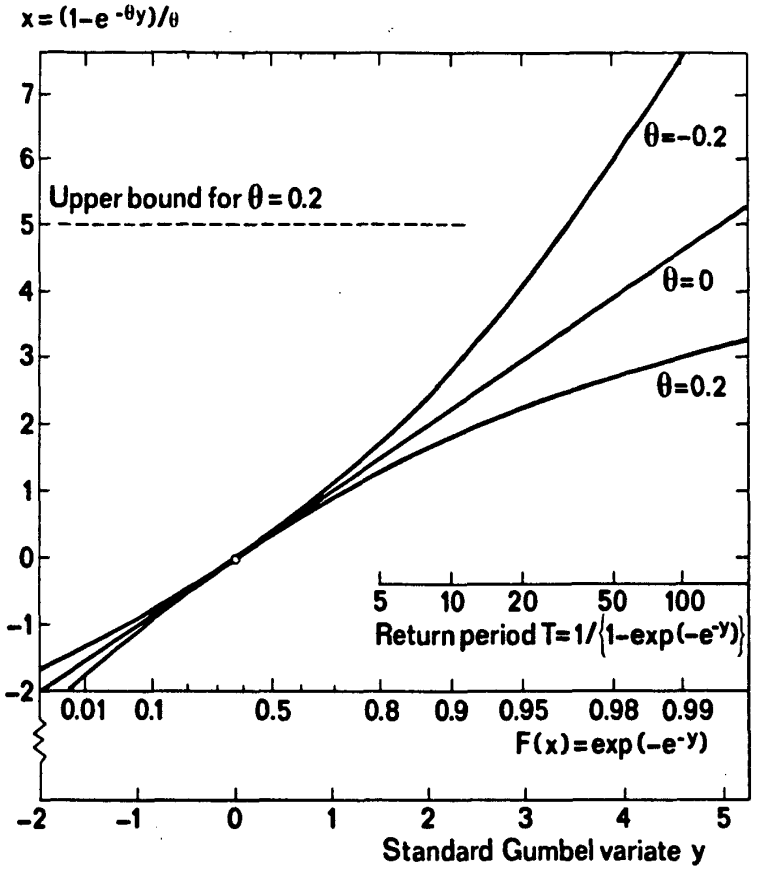


Figure 4 The three types of extreme-value variates as functions of the standard Gumbel variate y.

purposes of extrapolation the choice of θ becomes therefore extremely important. This problem is dealt with further in Section 5.

The Fisher-Tippett extreme-value distributions were originally obtained as the possible limiting distributions of linearly normalized maxima in sequences of independent and identically distributed random variables. Recent work on asymptotic properties of extremes in time-series shows that the same limiting distributions are obtained under much weaker conditions. For instance, for stationary sequences it is only necessary to require that strong long-term persistence is absent in the occurrences of high-level exceedances (a so-called mixing condition). This condition is also needed for the Poisson approximation in the previous section. For further details about extreme-value theory for time-series the reader is referred to the excellent work of Leadbetter et al. (1983).

The question remains, however, whether a year or season contains enough data to justify the use of limiting distributions for extremes. A sounder argument for choosing the GEV distribution is that this distribution provides in most cases an excellent fit to the data. Further, the numerical evaluation of both F(x) and its inverse is extremely simple, because the eqns. (13), (14) and (15) only contain logarithms and exponential functions.

In quite a number of cases a two-parameter distribution may be preferred to the three-parameter GEV distribution. Especially in situations where no extrapolations are required, it is hard to justify distributions with more than two parameters. Sometimes, even a normal distribution can be used. As an example, Fig. 5 shows that the normal distribution gives an excellent fit to the distribution of annual maximum air temperatures at De Bilt (The Netherlands).

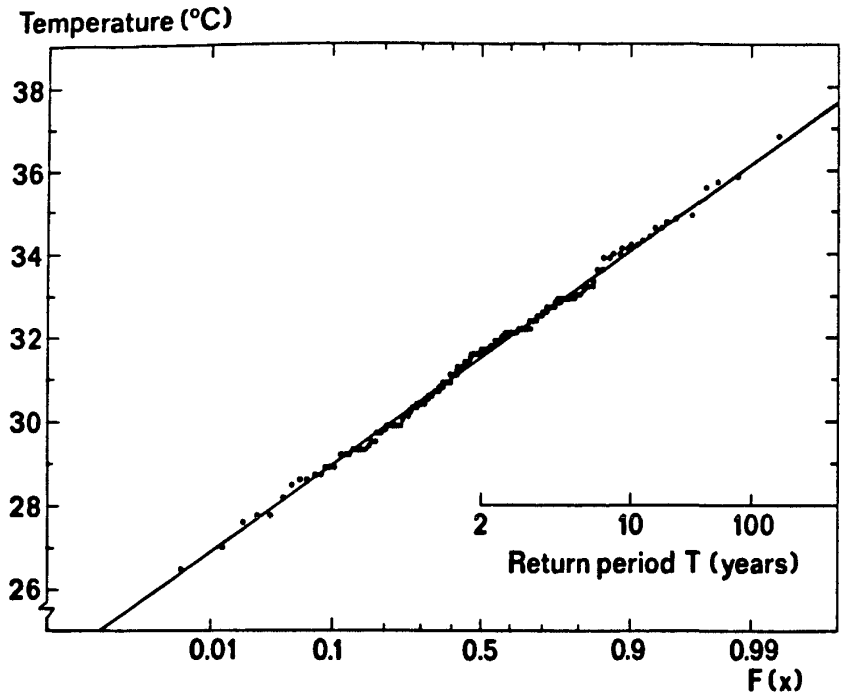


Figure 5 Normal probability plot of annual maximum air temperatures at De Bilt (52°06'N, 5°11'E) for the period 1849-1981. After Van Montfort (1982).

Fitting a single distribution to annual (or seasonal) maxima is sometimes discouraged when these maxima have originated from different meteorological phenomena. In some cases a separate analysis of the extremes for homogeneous subpopulations may be advantageous. The distribution function F(x) of the annual maxima is then obtained from the distribution functions $F_q(x)$ of the maxima for each subpopulation, by:

$$F(x) = \prod_{q=1}^{Q} F_q(x) \qquad (16)$$

with Q the number of subpopulations. Gomes and Vickery (1977/78) give an application of this technique to extreme wind gusts in regions with different storm-producing phenomena (extensive, extra-tropical pressure systems, thunderstorms and hurricanes). $F_q(x)$ is usually described by the Gumbel distribution. The method requires that a subdivision can be made on meteorological grounds. It is in general not appropriate to fit the mixed distribution F(x) in eq. (16) directly to annual maxima.

4 SOME NOTES ON ESTIMATION

Design values in engineering practice are often based on distributions fitted to annual maxima. This restriction to annual maxima is sometimes criticized because inclusion of other observations is expected to give better estimates of quantities like a 2-year or a 50-year event. This problem is discussed in Subsection 4.1. Another topic which has had much attention in recent years is the estimation of the parameters of the GEV distribution. A survey of techniques is given in Subsection 4.2.

4.1 Alternatives to annual maxima

For purposes of estimation it is possible to consider also the second largest extreme in a year, or more general the k largest extremes $Z_1,\ldots,Z_k$ simultaneously. For an N-year record the sample then consists of Nk observations. Assuming independence between the data for separate years, the likelihood of such a sample is obtained from the joint density (7) with m=k. The maximum likelihood estimates of the parameters of the Fisher-Tippett extreme-value distributions were derived by Gomes (1981). In her study special attendion was paid to the Gumbel/exponential case (11).

k	Relative standard error of		
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{x}_{10}$
1	1.00	1.00	1.00
2	0.85	0.77	0.88
3	0.82	0.65	0.81
4	0.80	0.58	0.77
5	0.79	0.53	0.73

Table 1 Asymptotic standard errors of maximum likelihood estimates for samples of the k largest extremes per year compared with the asymptotic standard error for samples of the annual maxima (k=1) in the Gumbel/exponential case. The quantity x_{10} denotes the 10-year event for the annual maxima: $x_{10} = \mu + 2.25\sigma$.

Table 1 compares the large sample standard errors of the maximum likelihood estimates of μ, σ and the 10-year event. Especially for the estimate of the scale parameter σ a substantial reduction in standard error is found when lower extremes than annual maxima are also taken into account. It is shown further in Gomes (1981) that the bias in σ decreases when the joint distribution of the largest two or three extremes in a year is considered.

Another alternative to annual maxima is to estimate the parameters and quantiles of the distribution from the number and magnitudes of peaks in a partial duration series. For the Gumbel/exponential case the annual maxima and partial duration series methods were compared by Cunnane (1973). For return periods T greater than about 50 years the partial duration series estimate of the T-year event has smaller standard error than the annual maxima estimate if the mean annual number of peaks $\lambda(x_0)$ is at least 1.65. Instead of taking all local maxima over a fixed threshold x_0, it is also possible to consider a fixed number of peaks. The likelihood function is then given by (7), with m the number of peaks. For the GEV/Generalized Pareto case Gomes (1984) compared this method with the previous mentioned method of taking the k largest peaks for each year. For a fixed sample size the choice between the two methods strongly depends on the shape parameter θ.

A third method to reduce standard errors of quantile estimates is to derive the distribution of annual maxima from the distributions of monthly maxima. Assuming independence between the maxima of successive months the distribution function F(x) of the annual maxima follows from:

$$F(x) = \prod_{j=1}^{12} F_j(x) \qquad (17)$$

where $F_j(x)$ is the distribution function of the maxima for the j-th month. An interesting discussion on this method is given in Tabony (1983) and Carter and Challenor (1984).

Beside a random error resulting from the fact that only a finite record of observations is available, there is also a systematic error because the underlying probabilistic model is only a crude approximation to the real world situation. Extending the sample with lower observations generally leads to an increase in systematic errors in the estimates of large quantiles. For those quantiles a standard error of estimation may give a too optimistic picture of the quality of the estimate. As a consequence adding lower observations to the sample does not necessarily result in better estimates for high return periods. These estimates may be even much worse. The limitation of eq. (17) for estimating a 50-year event of mean hourly wind speed in the United Kingdom has clearly been demonstrated by Tabony in his reply to a comment on his 1983 paper, see Carter and Challenor (1984). Moreover, even for annual maxima it is sometimes necessary to diminish the influence of observations in the left tail on estimates of upper quantiles. For this reason estimates of the parameters μ and σ of the distributions in Fig. 1 were obtained from a censored annual maximum series, in which snow depths less than 5 cm were considered as left-censored values.

4.2 Estimation of the parameters of the GEV distribution

The GEV distribution in eq. (13) contains three unknown parameters μ, σ and θ. The derivation of the maximum likelihood estimates of these parameters was first given by Jenkinson (1969). Some modifications of the numerical procedure were proposed by Otten and Van Montfort (1980). A computer program which follows the suggestions of these authors in a slightly different manner is given in Hosking (1985). The classical asymptotic properties of the maximum likelihood estimators only hold for $\theta<0.5$ (Smith, 1985). The limited validity of the classical theory is connected with the fact that the range of the GEV variate depends on its parameters.

In Jenkinson (1969) also a simple alternative to maximum likelihood estimation is described which is known as the method of sextiles. The range of the variate is divided into six intervals each containing one sixth of the data. Parameter estimates are derived from the means of the data in these intervals. The calculations can be done by hand. Monte Carlo studies show that this method has about the same performance as maximum likelihood estimation (Hosking et al., 1985).

A third method to obtain parameter estimates is the use of probability-weighted moments. The idea of probability-weighted moments was introduced in the hydrological literature by Greenwood et al. (1979). For the GEV distribution an algorithm has been developed by Hosking et al. (1985). Their algorithm is based on the following probability-weighted moments:

$$\beta_r = E\left[\, X\{F(X)\}^r \,\right], \quad r=0,1,2. \tag{18}$$

The quantity β_0 is just the population mean; the probability-weighted moments β_1 and β_2 are much less sensitive to the largest observation than the usual higher order moments. Estimates of β_0, β_1 and β_2 are obtained from the ordered sample $x_1 \leq x_2 \ldots \leq x_n$ as:

$$\hat{\beta}_r = \frac{1}{n} \sum_{j=1}^{n} (x_j F_j^r) \tag{19}$$

with $F_j = (j-0.35)/n$. The estimate of θ follows from:

$$\hat{\theta} = 7.8590\, c + 2.9554\, c^2 \tag{20}$$

where

$$c = \frac{2\hat{\beta}_1 - \hat{\beta}_0}{3\hat{\beta}_2 - \hat{\beta}_0} - \frac{\ln 2}{\ln 3}. \tag{21}$$

With this estimate of θ the parameters σ and μ are estimated as:

$$\hat{\sigma} = \frac{(2\hat{\beta}_1 - \hat{\beta}_0)\,\hat{\theta}}{\Gamma(1+\hat{\theta})\,(1-2^{-\hat{\theta}})} \tag{22}$$

$$\hat{\mu} = \hat{\beta}_0 + \hat{\sigma}\,\{\Gamma(1+\hat{\theta}) - 1\}/\hat{\theta} \tag{23}$$

where Γ stands for the gamma function. For moderate sample sizes (n = 50, 100) this method performs equally well as the method of maximum likelihood. For small samples the use of probability-weighted moments often leads to slightly better estimates than the method of maximum likelihood.

5 THE CHOICE BETWEEN PROBABILITY DISTRIBUTIONS AND THE NEED OF REGIONAL ANALYSIS

An important aspect of extreme-value analysis is choosing the form of the distribution of maxima. Physical knowledge of the underlying processes does generally not provide satisfactory information about the probabilities of rare events. It is also not correct to base the choice of a particular distribution merely on mathematical arguments concerning asymptotic properties of extremes. A careful analysis of historic data is needed to find the most appropriate distribution. Usually the fit of two or several candidate distributions has to be compared. This section gives two examples. The first example is concerned with the choice between the simple Gumbel distribution and the more

flexible GEV distribution. The second example deals with discrimination between the Gumbel distribution and the lognormal distribution. The latter example differs from the former so far that the two candidates belong to different families.

5.1 The Gumbel distribution versus the GEV distribution

Fig. 6 shows a probability plot of 1-day annual maximum precipitatin amounts at Scheveningen (The Netherlands) for a 55-year period. The Gumbel distribution seems to fit the data well. The only indication of a possible lack of fit consists of the fact that the largest observation lies rather far from the straight line representing this distribution.

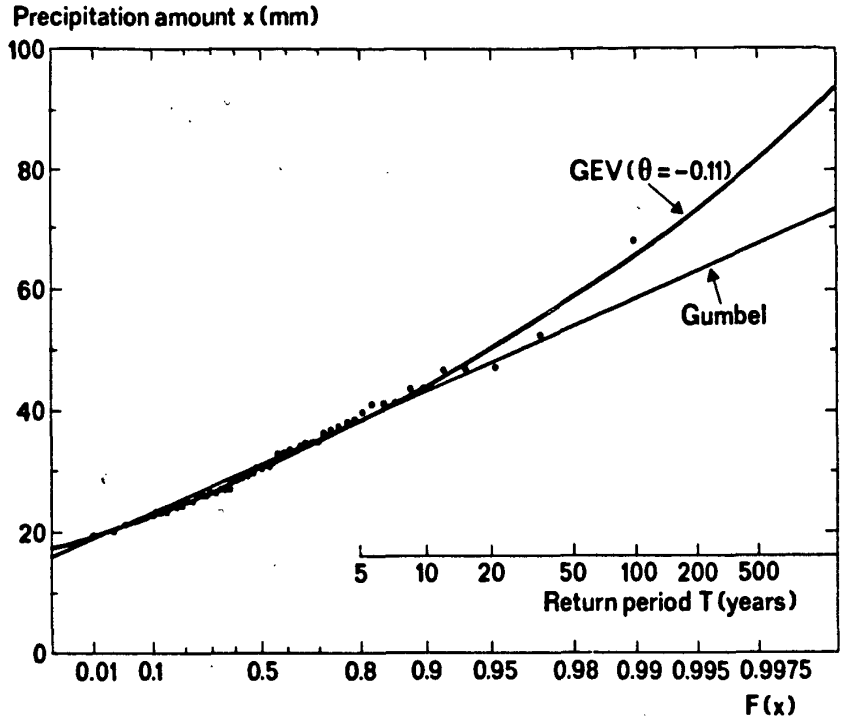


Figure 6 Gumbel probability plot of 1-day annual maximum precipitation amounts at Scheveningen (52°07'N, 4°18'E) for the period 1920-1974.

In Section 3 it was noted that the Gumbel distribution corresponds to the GEV distribution with $\theta=0$. Therefore, a useful method to check the adequacy of the Gumbel distribution seems to test the hypothesis $\theta=0$ against one of the alternatives $\theta<0$, $\theta>0$ or $\theta\neq0$. Several statistics have been proposed for this purpose (Hosking, 1984; Van Montfort and Gomes, 1985). For the data in Fig. 6 testing $\theta=0$ against $\theta\neq0$ with one of these statistics did not result in rejection of the null hypothesis at the 5%-level (Buishand, 1983). This result gives, however, not a valid argument to use the Gumbel distribution for estimating very large quantiles. In fact a rainfall record of 55 years is too short for making inference about the form of the upper tail. Reliable estimates of very large quantiles can only be obtained from combining information of several data sets (NERC, 1975; Jenkinson, 1977). In the GEV case this is necessary to improve the accuracy of the estimate of θ. The standard error of θ for the data in Fig. 6 is about 0.1. The two fitted distributions show that a change in θ of this magnitude leads to considerable differences in the upper tail.

Precipitation data from 14 other stations were also analysed for the same 55-year period as the Scheveningen data. The estimates of the shape parameter θ of the GEV distribution are given in Fig. 7. All stations in this figure are lowland stations with a mean annual rainfall of 740 to 840 mm. The estimates of θ show a large irregular variation over the country. This variation is a consequence of the large standard error of θ and the

low correlation between the $\hat{\theta}$'s for different stations, even at rather short distances. The latter can roughly be explained by the weak spatial association between very large observations (Buishand, 1984), which mainly determine the value of θ.

The fact that $\hat{\theta}$ is negative for 14 out of the 15 stations in Fig. 7 indicates that the upper tail of the Gumbel distribution is too short for this kind of data. It is much better to assume a GEV distribution with a common θ for the whole region. In Buishand (1983) the sample median (-0.08) of the $\hat{\theta}$'s in Fig. 7 was taken as an estimate of this regional θ. About the same estimate was obtained from a modified station-year analysis of 140 records over a 30-year period (Buishand, 1984).

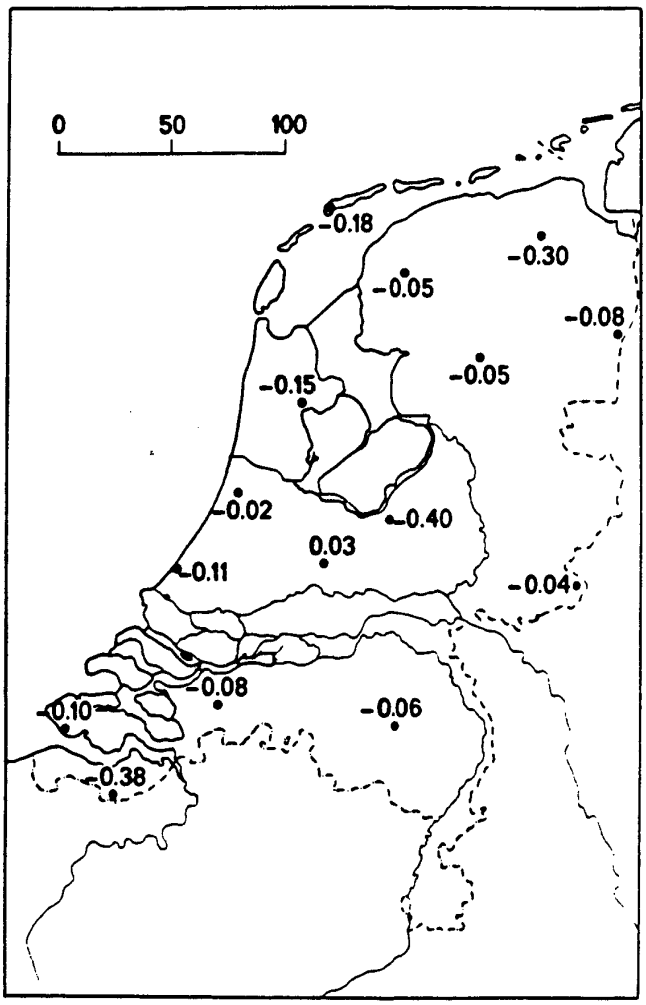


Figure 7 Maximum likelihood estimates of the shape parameter θ of the GEV distribution for 1-day annual maximum precipitation amounts in The Netherlands, based on data for the period 1920–1974.

5.2 The Gumbel distribution versus the lognormal distribution

Unlike the previous case the Gumbel distribution cannot be embedded in the lognormal distribution. Tests to distinguish between separate families of distributions have been developed by Cox (1961, 1962) and Atkinson (1970). The resulting test-statistics are based on the log likelihoods of the candidate distributions. These ideas have not

found wide applications in the literature. Yet, the choice between the Gumbel distribution and the lognormal distribution has sometimes been of interest for maxima of precipitation amounts (Sevruk and Geiger, 1981) and for maxima of snow loads on the ground (Ellingwood and Redfield, 1983). In this subsection special attention is paid to the latter publication.

The study of Ellingwood and Redfield refers to annual maximum water-equivalents of snow on the ground for weather stations in the northeast quadrant of the United States in the 28-year period covering the winters of 1952/53 through 1979/80. To make a choice between the two distributions the following two probability plot correlation coefficients were compared:

- r_G, the correlation coefficient for the plot of the annual maxima on Gumbel probability paper.

- r_{LN}, the correlation coefficient for the plot of the logarithms of the annual maxima on normal probability paper.

In both cases plotting positions based on medians of standard uniform order statistics were used. The Gumbel distribution was preferred if $r_G > r_{LN}$; otherwise the lognormal distribution was chosen. So in contrast with hypothesis testing the two distributions are treated on an equal level.

To evaluate the performance of this method the following probabilities are of interest:

$$p_G = \Pr (r_G > r_{LN} \mid \text{Gumbel distribution}) \tag{24}$$

$$p_{LN} = \Pr (r_{LN} > r_G \mid \text{lognormal distribution}). \tag{25}$$

These probabilities depend on the length of the record and the degree of similarity between the shapes of the two distributions. The latter is strongly determined by the coefficient of variation c_v. For the Gumbel distribution the coefficient of skewness equals 1.14, but for the lognormal distribution skewness increases monotonically with c_v. When $c_v=0.36$ the coefficient of skewness equals 1.13 and the two distributions are then practically indistinguishable. For water-equivalent data in the northeast quadrant of the United States c_v is about 0.8. For this value of c_v a fitted lognormal distribution has a longer upper tail than a fitted Gumbel distribution.

Ellingwood and Redfield performed a simulation experiment to determine the probabilities p_G and p_{LN}. In this experiment the parameters μ and σ were taken from fitted distributions to historic data. For a record of 28 years both p_G and p_{LN} turned out to be about 0.75. This value differs considerably from unity. Due to the limited record-length the probability is quite large that the wrong distribution is chosen. Therefore also in this case the discrimination between the two distributions has to be based on the joint evaluation of several data sets. Such a joint evaluation was made both for all weather stations in the region and for the 38 stations with a complete record over the 28-year period and a nonzero water-equivalent in each year. For the second set of stations r_{LN} exceeded r_G in 29 cases (76% of the stations). This observed percentage corresponds well with the theoretical probability p_{LN} for the lognormal distribution. Because of this result the lognormal distribution was preferred to the Gumbel distribution for estimating the 50-year snow load on the ground.

The probability plot correlation coefficient is sensitive to departures in the upper tail. Therefore the use of this statistic has to be preferred to Pearson's X^2 statistic or the Kolmogorov-Smirnov statistic, which have also been mentioned by Ellingwood and Redfield. It is not certain, however, whether the probability plot correlation criterion is an optimal, or a nearly optimal, criterion for discriminating between the two distributions.

5.3 Concluding remarks

In both examples homogeneity with respect to the type of distribution was assumed. Such an assumption may not always be justified. In the GEV case changes in the shape of the distribution can possibly be described by relating the shape parameter θ to geographical or climatological characteristics.

Spatial correlation between the data from different stations complicates statistical inference about regional distributions. A proper evaluation requires insight into the dependence structure of the data and perhaps also fitting a multivariate distribution. Extreme-value analysis has then shifted to solving difficult problems of multivariate analysis. Although the probabilistic theory of multivariate extreme-value distributions is well developed yet (Galambos, 1978; Deheuvels, 1984), it is not clear how far these theoretical limiting distributions can be used in regional analyses of meteorological extremes.

6 ACKNOWLEDGEMENT

Part of the data from the Kleve record were kindly provided by the Deutscher Wetterdienst.

REFERENCES

- Atkinson, A.C., 1970: A method for discriminating between models (with discussion). J. Roy. Statist. Soc., Ser. B, 32, 323-353.

- Buishand, T.A., 1983: Extremely high rainfall amounts and the theory of extreme values. Cultuurtechn. Tijdschr., 23, 9-20 (in Dutch). Corrigendum, 23, 81.

- Buishand, T.A., 1984: Bivariate extreme-value data and the station-year method. J. Hydrol., 69, 77-95.

- Buishand, T.A., 1986: Snow cover in The Netherlands. To appear as Sci. Rep., Roy. Neth. Meteorol. Inst., De Bilt (in Dutch).

- Carter, D.J.T. and Challenor, P.G., 1984: Comments on "Extreme value analysis in meteorology" by R.C. Tabony (with reply by R.C. Tabony). Meteorol. Mag., 113, 43-52.

- Cox, D.R., 1961: Tests of separate families of hypotheses. In: J. Neyman (Editor), Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability, 1. University of California Press, pp. 105-123.

- Cox, D.R., 1962: Further results on tests of separate families of hypotheses. J. Roy. Statist. Soc, Ser. B, 24, 406-424.

- Cunnane, C., 1973: A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. J. Hydrol., 18, 257-271.

- Deheuvels, P., 1984: Probabilistic aspects of multivariate extremes. In: J. Tiago de Oliveira (Editor), Statistical Extremes and Applications. Reidel, Dordrecht, pp. 117-130.

- Ellingwood, B. and Redfield, R., 1983: Ground snow loads for structural design. J. Struct. Div., Amer. Soc. Civ. Eng., 109, 950-964.

- Galambos, J., 1978: The asymptotic theory of extreme order statistics. Wiley & Sons, New York.

- Gomes, M.I., 1981: An i-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes. In: C. Taillie et al. (Editors), Statistical Distributions in Scientific Work, 6. Reidel, Dordrecht, pp. 389-410.

- Gomes, M.I., 1984: Statistical theory of extremes - comparison of two approaches. Paper presented at the European Meeting of Statisticians, Marburg.

- Gomes, L. and Vickery, B.J., 1977/78: Extreme wind speeds in mixed wind climates. J. Indus. Aerodyn., 2, 331-344.

- Greenwood, J.A., Landwehr, J.M., Matalas, N.C. and Wallis, J.R., 1979: Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour. Res., 15, 1049-1054.

- Gumbel, E.J., 1958: Statistics of extremes. Columbia University Press, New York.

- Hosking, J.R.M., 1984: Testing whether the shape parameter is zero in the generalized extreme-value distribution. Biometrika, 71, 367-374.

- Hosking, J.R.M., 1985: Algorithm AS 215. Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. Appl. Statist, 34, 301-310.

- Hosking, J.R.M., Wallis, J.R. and Wood, E.F., 1985: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics, 27, 251-261.

- Jenkinson, A.F., 1955: The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quart. J. Roy. Meteorol. Soc., 81, 158-171.

- Jenkinson, A.F., 1969: Statistics of extremes. In: Estimation of Maximum Floods. Tech. Note 98, World Meteorol. Org., Geneva, Ch. 5, pp. 183-227.

- Jenkinson, A.F., 1977: The analysis of meteorological and other geophysical extremes. Met 013 Tech. Note 58, Meteorol. Office, Bracknell (Unpublished, copy available in National Library, Bracknell UK).

- Leadbetter, M.R., Lindgren, G. and Rootzén, H., 1983: Extremes and related properties of random sequences and processes. Springer-Verlag, New York.

- NERC (Natural Environment Research Council), 1975: Flood studies report, 5 volumes. Whitefriars Press, London.

- North, M., 1980: Time-dependent models of floods. J. Hydraul. Div., Amer. Soc. Civ. Eng., 106, 649-655.

- Otten, A. and Van Montfort, M.A.J., 1980: Maximum-likelihood estimation of the general extreme-value distribution parameters. J. Hydrol., 47, 187-192.

- Revfeim, K.J.A., 1982: Seasonal patterns in extreme 1-hr rainfalls. Water Resour. Res., 18, 1741-1744.

- Revfeim, K.J.A., 1983: Stochastic process analysis of rainfall totals and extremes. In: Preprints 2nd International Meeting on Statistical Climatology. Inst. Nacional de Meteorol. e Geofysica, Lisbon, Paper 10.2.

- Revfeim, K.J.A. and Hessell, J.W.D., 1984: More realistic distributions for extreme wind gusts. Quart. J. Roy. Meteorol. Soc., 110, 505-514.

- Sevruk, B. and Geiger, H., 1981: Selection of distribution types for extremes of precipitation. Operational Hydrol. Rep. 15, World Meteorol. Org., Geneva.

- Smith, R.L., 1984: Threshold methods for sample extremes. In: J. Tiago de Oliveira (Editor), Statistical Extremes and Applications. Reidel, Dordrecht, pp. 621-638.

- Smith, R.L., 1985: Maximum likelihood estimation in a class of non-regular cases. Biometrika, 72, 67-90.

- Tabony, R.C., 1983: Extreme value analysis in meteorology. Meteorol. Mag., 112, 77-98.

- Van Montfort, M.A.J., 1982: Models for maxima and minima, estimates and confidence intervals, choice between models. Tech. Note 82-2, Dept. of Math., Agricultural University, Wageningen (in Dutch).

- Van Montfort, M.A.J. and Gomes, M.I., 1985: Statistical choice of extremal models for complete and censored data. J. Hydrol., 77, 77-87.

- Van Montfort, M.A.J. and Witter, J.V., 1985: Testing exponentiality against generalised Pareto distribution. J. Hydrol., 78, 305-315.

- Weissman, I., 1978: Estimation of parameters and large quantiles based on the k largest observations. J. Amer. Statist. Assoc., 73, 812-815.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

Bi-variate extreme value distributions in Climatology

K. Cehak
Institute of Meteorology and Geophysics, Universitäty Vienna and
Zentralanstalt für Meteorologie und Geodynamik in Wien

Summary. One-dimensional extreme-value distributions (Gumbel and Fisher-Tippett distributions) have found many applications in climatology and hydrology. Scientific interest and demands from the users have led to extensions to two and more dimensions with respect to extreme-value distributions, too. Investigating this problem area one has to treat questions of fit and of the interdependence of the single variables. In climatology one finds applications where-ever one has to assess the risk of two or more dangerous events, which may harm men or constructions, therefore, the two- or multi-dimensional extreme-value distributions are of great value for urban and regional planning as well as for the treatment of some aspects relating to climatic change.

1. Introduction

After many decades, in which climatological research was concentrated on the treatment of average values, since the end of the fifties of our century one finds a growing interest of climatologists in extreme value problems. The work of the hydrologists on the one hand and of Gumbel (1958) on the other hand have shown that the extreme value distributions can be a mighty tool in the hand of the researcher and, finally these distributions have found their way to the climatologists. Jenkinson (1955) has also helped a lot in this respect.

The main difficulties for the use of the extreme value distributions have been, first, that one has but a small number of observations just in that part of the range of values in which one is interested, and second, that exact distributions are rather complicated to use. This has led to use of the asymptotic distribution according to Fisher-Tippett (1928), which make use of a general relation between the distribution of the largest value ina sample of n observations and the asymptotic distribution of a reduced variable:

$$(1) \qquad F^n(u + a\,y) \to L(y) \quad .$$

L(y) describes one of the three Fisher-Tippett distributions and one can chose the best one of them according to Jenkinson (1955) by scetching the observed distribution in Gumbel's extremal probability diagram.

Climatological elements, for which a representation by means of the asymptotic extreme value distribution is now often used, include temperature maxima and minima as well as precipitation maxima.

2. The two-dimensional extreme value distribution

From the point of view of climatology it is interesting to investigate the joint probability of temperature maxima and minima of the same month. This is interesting for determinations of continentality and its change in the course of time. It is also interesting to combine temperature extremes and precipitation maxima. For Urban and Building Climatology it would be as interesting to look for the distributions of element combinations with wind speed.

The statistical problem involved is the investigation of the asymptotic distributions of pairs of variables in samples of size n. One looks for an asymptotic distribution function L(x,y), to which the exact distribution function converges for large samples. It should be stressed that, usually, the pair

(2) $(\max X_i, \max Y_i)$

is not realized in the samples. As in the one-dimensional case the exact distribution of the variables converges against the joint distribution of the reduced variates:

(3) $F^n(u_1 + a_1 x, u_2 + a_2 y) \to L(x,y)$.

The marginal distributions of this two-dimensional extreme value distribution are the one-dimensional Fisher-Tippett distributions. According to Jenkinson (1955) one can transform the variates so that variates distributed according type II or type III become distributed according type type I. Therefore, one can assume that the marginal distributions are Fisher-Tippett type I distributions or Gumbel distributions. In analogy to the normal distribution one sets

(4) $L(x,y) = (L(x).L(y))^{k(y-x)} = \exp(-(e^{-x} + e^{-y})k(y-x))$.

The function k(y-x) is called the dependence function, is stands for the coefficient of correlation in the two-dimensional normal distribution and is dependent on the difference between the reduced values of both variates. By integration one can derive a correlation coefficient from the dependence function:

(5) $$\rho = -\frac{L}{\pi^2} \int_{-\infty}^{\infty} \ln k(w)\, dw \quad .$$

According to the form of the dependence function k(w), with w=y-x, one can discern various forms of the two-dimensional extreme value distributions. Not all of these functions are differentiable.

Equ. (4) may be used to estimate k(w) from samples by comparing the two-dimensional distribution with the product of the marginal distributions:

(6) $$k(w) = -\frac{\ln L(x,y)}{e^{-x}+e^{-y}} \quad .$$

The rather sparse literature on two-dimensional extreme value distributions quotes five models (Tiago de Oliveira 1984) which are characterized by the dependence functions as follows.

Differentiable models:

Logistic model:

(7) $$k(w) = (1 + \exp(-\frac{w}{1-c})^{1-c}) \Big/ (1 + e^{-w}) \qquad 0 \leq c < 1 \quad .$$

Mixed model:

(8) $$k(w) = 1 - c\,\frac{e^w}{(1+e^w)^2} \qquad 0 \leq c \leq 1 \quad .$$

Non differentiable models:

Gumbel model:

(9) $$k(w) = 1 - c\,\frac{\min(1, e^w)}{1 + e^{-w}}$$

Bi-extremal model:

(10) $$k(w) = 1 - \frac{\min(c, e^w)}{1 + e^w}$$

Natural model:

$$(11)\qquad k(w) = \frac{1}{1 + e^{w}} \qquad\qquad w < a$$

$$= \frac{e^{b}-1+(1-e^{a})e^{b-w}}{(e^{b}-e^{a})\ (1 + e^{-w})} \qquad\qquad a \leq w \leq b$$

$$= \frac{1}{1 + e^{-w}} \qquad\qquad b < w$$

with $-\infty \leq a \leq 0 \leq b < \infty$.

The non-differentiable models show a singular behaviour at x = y, a jump ir the distribution function.

One has the problem to find the model with best fit to the sample data. This is rather difficult with small sample sizes. The methods of maximum likelihood or maximum entropy lead to numerical comparisons of sums of deviations from the model under the condition that one has fitted various models to the same data. A further method is the use of the principle of parsimony in chosing the model and testing that the model is not in contradiction against the data. Usually, this method leads to the mixed model on account of its simple dependence function. Posner (1969) has shown that for the practical work the selection of the model is not of paramount importance.

3. Two-dimensional extreme value distributions for climatological data in Vienna, Austria

The obsverations of temperatures extremes and maxima of daily precipitation totals in Vienna have been used to form the joint distributions of the temperature maxima and minima and of the temperature maxima and precipitation maxima. The dependence functions were calculated using equ. (6), in which the sample distribution functions were used instead of the unknown theoretical models. The joint distributions are often rather near to two-dimensional distributions with uncorrelated variables. Therefore, the mixed model was fitted, which offers the easiest equations as k(w) is a linear function of the parameter c.

Introducing a new variable x for the ratio in equ. (8) one gets

$$(12)\qquad k(w) = 1 - c\ x$$

and according to the method of least squares this gives

$$(13)\qquad c = -\frac{\Sigma\{k(w)-1\}x}{\Sigma x^{2}}$$

The values of c found for the different joint distributions and the various months are not far away from zero, the annual variation is small, too. Therefore, k(w) is close to one and thejoint distributions are not very different from the products of the marginal distributions or from the distribution of independent quantities. Nevertheless, the dependence is shown and it comes out more clearly in longer observation series.

Table 1 presents the results of the calculations effectuated in three different periods.The complete series of 113 years is partitioned 1910 and 1950 in order to get three equally long series. The differences of the results from one period to the other are of less importance than one had expected taking into account all the influences on extreme values of climatological elements.

The parameters a and u of the extreme value distributions are more similar to each other in the first and third period than to the second period, which was the time of the strongest warming in mean values. Statistical tests for these differences would probaly yield no significant results.

The differences between second and third period are much larger when the maxima of precipitation totals are considered. This is in good concordance with the experience ofa falling trend of precipitation totals in the eastern part of Austria.

Stronger differences are to be found in the correlations between the elements. In this case, the third period differs considerably from the other periods. There are even changes of the sign of the correlation coefficients, which gives also a change in the relation between the elements.

With the five parameters of the two-dimensional extreme value distributions one can give rather exhaustive descriptions of climatic variations in respect to these elements. This yields better informations than the use of the single series alone.

4. References

Fisher, R.A. and Tippett, L.H.C., 1928: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Phil. Soc. 24, 419-438.

Gumbel, E.J.,1958: Statistics of Extremes. Columbia Univ. Press, New York.

Jenkinson, A. F., 1955: The frequency distribution of the annual maximum (or minimum) values of meterorological elements. Quart. J.R.M.S. 81, 158-171.

Posner, E.C. et al, 1969: Applications of an estimator of high efficiency in bivariate extreme theory. J.Amer.Stat.Assoc. 64, 1403-1414.

Tiago de Oliveira, J. (Ed.), 1984: Statistical Extremes and Applications. Applications. NATO ASI Series. Reidel.

Table 1. Statistical parameters of the distributions of the temperature extremes and the maxima of precipitation in 24 hours in Vienna, in different periods: a 1873-1910, b 1911-50, c 1951-85.

Monthly maxima of temperature, °C

TXQ												
a	9,0	10,6	17,7	21,7	25,9	28,4	31,5	30,4	26,6	21,0	14,0	10,2
b	9,3	11,0	17,2	22,1	25,7	28,3	30,3	29,4	25,9	20,4	14,7	10,9
c	9,2	11,0	17,9	21,8	26,1	28,8	30,4	29,7	26,3	20,6	15,0	10,8
a												
a	2,7	3,4	3,3	2,6	1,9	1,6	1,6	2,2	2,3	2,9	2,8	2,8
b	3,5	3,3	2,7	2,2	1,8	2,0	2,2	2,2	2,6	2,4	3,1	2,9
c	2,7	3,1	2,3	1,9	1,6	1,4	1,6	1,9	1,7	2,5	2,4	3,4
u												
a	7,9	8,8	15,7	20,2	24,8	27,5	30,6	29,1	25,3	19,3	12,4	8,5
b	7,3	9,1	15,6	20,8	24,7	27,1	29,1	28,1	24,4	19,1	12,9	9,3
c	7,6	9,2	16,6	20,7	25,1	28,0	29,5	28,6	25,3	19,1	13,6	8,8

Table 1. Continuation. Monthly Minimum of temperature, °C

TNQ												
a	-12,3	-9,6	-6,4	-1,0	3,4	8,5	10,6	9,6	5,2	0,1	-5,6	-10,3
b	-10,4	-9,9	-4,7	0,0	3,8	8,6	11,1	10,4	5,9	0,3	-3,3	-8,8
c	-10,1	-8,4	-4,6	0,3	4,1	8,4	11,0	10,3	6,2	0,7	-3,3	-7,4
a												
a	4,5	3,1	3,6	1,6	2,1	1,1	1,3	0,9	2,0	2,0	2,7	3,3
b	3,9	3,7	2,8	1,7	1,7	1,2	1,0	1,0	1,8	1,8	2,0	3,4
c	3,9	3,7	2,8	1,8	1,6	1,5	1,1	1,0	1,5	1,8	2,4	3,2
u												
a	-9,6	-7,7	-4,3	-0,1	4,5	9,0	11,3	10,1	6,3	1,0	-4,1	-8 6
b	-8,1	-7,8	-3,0	1,0	4,7	9,3	11,7	11,0	6,9	1,3	-2,1	-6,8
c	-7,9	-6,2	-3,0	1,3	5,0	9,3	11,6	10,9	7,0	1,7	-1,9	-5,6
Parameter of the dependence function												
a	0,27	0,27	0,14	0,15	0,16	0,14	0,18	0,16	0,13	0,11	0,22	0,24
b	0,33	0,33	0,24	0,14	0,19	0,24	0,22	0,24	0,21	0,19	0,21	0,22
c	0,18	0,23	0,23	0,14	0,18	0,18	0,20	0,28	0,16	0,20	0,25	0,31

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE INTENSITY-DURATION-FREQUENCY CURVES OF THE ANNUAL MAXIMUM RAINFALL AT UCCLE (BRUSSELS).

G. DEMAREE AND R. SNEYERS

ROYAL METEOLOGICAL INSTITUTE OF BELGIUM
UCCLE (BRUSSELS)

1. INTRODUCTION.

In a first paper on the subject, Sneyers in 1977 gave, for Uccle (near Brussels), estimates of the fractiles of the distribution of the annual maximum rainfall in 1, 10, 20, 30, 40 min and in 1, 2, 4, 6, 9, 12, 18 and 24 hours. The data available for this study were the annual maximum rainfall in 1, 20 and 60 min read on the Hellman recording raingauge charts from 1938 to 1972, the annual maximum daily rainfall since 1901 and the annual maximum rainfall in 20, 30 min, 2 and 6 hours for the period 1956-1972. The distribution functions adjusted to the data are double exponential (Gumbel) distributions for the time intervals of 1 to 30 min, while for 1 to 24 hours, the distribution functions are logarithmic double exponential distribution functions.

Extending the determination of the annual maximum rainfall in 1, 10, 20, 30 min and 1, 2, 6, 12 and 24 hours from the registrations of the Hellman recording raingauge of Uccle to the whole period 1934-1983, Demarée in 1985 estimated fractiles of the annual maximum rainfall for time intervals ranging nearly continuously from 1 min to 1440 min, by adjusting Gumbel distribution functions to all the observed distributions.

The extension of the estimation of fractiles for intermediate time intervals was completed by Sneyers by interpolating the parameters of the distribution functions, while the results of Demarée were obtained through the use of a five parameter formula which is a generalization of the well known Talbot-Montana type formula used in hydrology.

The purpose of this paper is to test the fit of the Gumbel distribution functions to the observed data of the period 1934-1983, to intercompare the two methods of interpolation and to present a final general distribution function.

2. FIT OF THE GUMBEL DISTRIBUTIONS.

The general distribution function of the maximum value X is the three parameter distribution function (general von Mises-Jenkinson formula) :

$$P(Y \leqslant y) = \exp\left[-(1-\theta y)^{1/\theta}\right] \qquad (2.1)$$

with $X = \mu + \sigma Y$ and $1 - \theta y \geqslant 0$ for $\theta \neq 0$, where θ is the shape parameter and where μ and σ are respectively the location and the scale parameter.

This distribution has an upper bound $y < 1/\theta$ if $\theta > 0$ (Weibull type) or a lower bound $y > 1/\theta$ if $\theta < 0$ (Fréchet, Cauchy type).

Moreover, if θ tends to zero, the distribution function reduces to the two parameter double exponential function (Gumbel type) :

$$P(Y \leqslant y) = \exp\left[-\exp(-y)\right] \quad \text{with} \quad -\infty \leqslant y \leqslant +\infty \qquad (2.2)$$

Finally the logarithmic double exponential function gives the distribution of the Fréchet type with $X > 0$, that is : $\theta = -\sigma/\mu$.

To test the goodness of fit of the Gumbel type against the alternatives of the Fréchet or Weibull type, several statistics may be used, the simplest among them being the G–statistic suggested by Gumbel (1965). Computing the power of these test-statistics, Van Montfort and Gomes (1985) showed that the G–statistic has practically the same power as two other more sophisticated statistics.

If $X_1 \leqslant X_2 \leqslant \ldots \leqslant X_n$ are the observations ranked in ascending order the statistic G is defined by :

$$G = (X_n - X_m)/(X_m - X_1) \tag{2.3}$$

X_m being the median defined by :

$$X_m = (X_{[(n+1)/2]} + X_{[(n+2)/2]})/2 \tag{2.4}$$

where the integral part of the numbers within the brackets is merely considered.

Table 1

Annual maximum rainfall in t min. Values of G for the period 1934-1983.
Corresponding value of the standard normal variate $u(G)$.

t	G	$u(G)$
1	1.714	−0.65
10	1.473	−1.08
20	1.652	−0.74
30	1.883	−0.38
60	3.723	1.58
120	4.365	2.04
360	2.571	0.52
720	2.059	−0.12
1440	3.048	1.01

Under the assumption of a Gumbel distribution, log G has closely a normal distribution with mean log d_n, where

$$d_n = -\log n / \log \{ -\log 2 / \log [1-(0,5)^{1/n}] \} \tag{2.5}$$

the standard deviation being approximated by :

$$\sigma(\log G) = 1/[-0.49 + 0.86 \log n] \tag{2.6}$$

The values of the statistic G have been computed for each sample of 50 observations (1934-1983) of the annual maximum rainfall in respectively 1, 10, 20, 30, 60, 120, 360, 720 and 1440 min. Moreover, the corresponding values of the standard normal variate $u(G)$ has been computed. At the 5% level of significance the critical values of the latter statistic for rejection of the Gumbel distribution are thus ± 1.96. Table 1 shows that the assumption is rejected for the intensity in 120 min in favor of a distribution with lower bound (Fréchet type).

Now, if we consider this result as one among nine independent test results, the probability associated with the largest value is $(P)^9$, P being the value of the distribution function associated with the largest value. Having $P = 0.97932$ we have $(P)^9 = 0.829$ which is no longer significant at the 5% level.

Another remark has however to be made : it concerns the fact that the values of $u(G)$ are all negative for the intensities in 1, 10, 20 and 30 min and nearly all positive for the other intensities. More precisely, the average of the four first values of $u(G)$ is − 0.715 while for the five last values we have 1.006. The variances associated with these averages being respectively 1/4 and 1/5, the variance of the difference is 0.45. This difference being 1.721 ,

the corresponding value of the standard normal variate is $1.721/\sqrt{0.45} = 2.57$ with probability $P = 0.995$, which is highly significant.

The distinction which has been made by Sneyers (1977) between the two groups of intensities appears thus to be justified. Of course, in a first approximation, Gumbel distributions may be considered.

3. THE ADJUSTED GUMBEL DISTRIBUTIONS.

The adjustment of the Gumbel distribution function is based on the Kimball estimate of the scale parameter σ (Kimball, 1956) given by the approximate likelihood equation :

$$n\,\sigma^* = \Sigma\, X_i[1 + E(\log F_i)] \tag{3.1}$$

where n is the sample size, X_i are the observations ranked in ascending order and $E(\log F_i) = -(\frac{1}{i} + \frac{1}{i+1} + \ldots + \frac{1}{n})$. This estimate has a bias which may be corrected by a factor b_n which may be taken from a table or, if $n > 11$, given by the relation (Sneyers and Van Isacker, 1972) :

$$\log(b_n - 1) = -0.975652 \log (n-1) + 1.043532 - 1.950309/\log (n-1) + 3.574231/\log^2 (n-1) \tag{3.2}$$

The final estimate of σ is thus :

$$\hat{\sigma} = b_n\, \sigma^* \tag{3.3}$$

For the location parameter we have :

$$\hat{\mu} = \overline{X} - \gamma\hat{\sigma} \tag{3.4}$$

with $\overline{X} = \Sigma\, X_i/n$ and $\gamma = 0.577216 \ldots$ (Euler constant).

The empirical efficiencies of Kimball estimates have also been computed by Sneyers and Van Isacker (1972). They show that the asymptotic values of the covariance matrix of these estimates are given by the Cramer-Rao lower bounds which are found to be (Sneyers, 1975, p.55) :

$$\mathrm{var}_0\, \hat{\mu} = 1.10866\sigma^2/n \text{ , } \mathrm{var}_0\, \sigma = 0.60793\sigma^2/n \text{ and } \mathrm{cov}_0(\hat{\mu},\hat{\sigma}) = 0.25702\sigma^2/n \tag{3.5}$$

This leads to the estimates and to their quotients (table 2) of the scale and location parameters of the Gumbel distributions adjusted to the annual maximum intensities in 1, 10, 20, 30, 60, 120, 360, 720 and 1440 min.

Table 2

Annual maximum rainfall in t min (0,1mm). Kimball estimates of the scale parameter σ and of the location parameter μ. Values of the ratio $K = \hat{\sigma}/\hat{\mu}$.

t	$\hat{\sigma}$	$\hat{\mu}$	K
1	7.28297	17.2762	0.42156
10	26.2591	75.6228	0.34724
20	37.1850	98.5362	0.37737
30	43.4166	110.079	0.39441
60	51.3267	132.033	0.38874
120	57.8589	159.543	0.36265
360	71.4761	211.043	0.33868
720	83.5533	254.212	0.32868
1440	98.4991	323.525	0.30446

4. INTERPOLATION OF THE ANNUAL MAXIMUM INTENSITIES FOR INTERMEDIATE TIME INTERVALS.

4.1 The generalized Gumbel distribution.

The problem consists in searching for regressions of the location and of the scale parameters in function of the time interval t. If $\mu(t)$ and $\sigma(t)$ are those functions the general distribution is then :

$$P = \exp[-\exp(-y_P)] \text{ with } y_P = [x_P(t) - \mu(t)]/\sigma(t) \tag{4.1}$$

or

$$x_P(t) = \mu(t) + \sigma(t) y_P \text{ with } y_P = -\log(-\log P) \tag{4.2}$$

$x_P(t)$ and y_P being the fractiles of the general variate $x(t)$ and of the reduced variate y corresponding to the probability P.

Before searching for an adequate regression, it should be remarked that the annual maximum intensity in 1 min has not the same meaning as the intensities in other time intervals. In fact it is an estimate of the instantaneous intensity read on the tangent to the recorded curve. It has thus to be considered as an upper limit for all other intensities. It seems therefore justified to exclude the intensity in 1 min from the group of intensities.

Bearing this in mind, several regressions have been tried among which the best one appears to be of the form :

$$\log z = a \log t + b \tag{4.3}$$

for μ and σ. The values of the coefficients a and b and of the estimate ϵ^2 of the variance of the residual are given in table 3.

Table 3

Regression $\log z = a \log t + b$ for the location parameter μ and for the scale parameter σ of the Gumbel distribution of the annual maximum intensity in t min. Estimate ϵ^2 of the variance of the residual.

z	a	b	ϵ^2
μ	0.27951	3.7264	$(0.0288)^2$
σ	0.24189	2.8636	$(0.0835)^2$

To test the fit of these regressions it should be noted that $E(d \log z)^2 = (\text{var } z)/[E(z)]^2$.

Knowing from table 2 that the average of $K = \hat{\sigma}/\hat{\mu}$ is 0.355 and that with (3.5) we have :

$$\frac{\text{var } \mu}{\mu^2} = \frac{1.10866}{50} \cdot \frac{\sigma^2}{\mu^2} \cong (0.0529)^2 \text{ and } \frac{\text{var } \sigma}{\sigma^2} = \frac{0.60793}{50} = (0.1103)^2 \tag{4.4}$$

the comparison with the values of ϵ^2 in table 3 shows that the fit is excellent.

The general law of the annual maximum intensities in t min is thus, in 0,1 mm :

$$x_P(t) = 41.529 t^{0.27951} + 17.524 t^{0.24189} \cdot y_P \tag{4.5}$$

In order to make a comparison with the adjusted Gumbel distribution, formula (4.5) has been used to compute the secular maximum (P = 0,99) in t min (table 4) for t = 10, 20, 30, 60, 120, 360, 720 and 1440 min. The root mean square difference between the two estimates is 1.47mm.

4.2 The generalized hydrological formula.

Defining the return period T by $T = 1/(1-P)$, Demarée (1985) has adjusted to the data a formula of the form (intensity-duration-frequency function) :

$$i_T(t) = x_p(t)/t = \frac{a_1 (a_2 + T)^{a_3}}{a_4 + t^{a_5}} \tag{4.6}$$

which gives in mm/h the fractiles of the distribution of the annual maximum intensity of the rainfall corresponding to the time interval t min and to the return period T years.

The formula has been adjusted to the fractiles of the intensities in 1, 10, 20, 30, 60, 120, 360, 720 and 1440 min corresponding to the return periods 1.5, 2, 5, 10, 15, 20, 25, 50, 100, 200 and 250 years given by the adjusted Gumbel distributions (table 2). The applied procedure is a general least squares method which is a compromise between three different algorithms, namely Newton-Raphson, Steepest-Descent and Marquardt (see e.g. Abdy and Dempster (1974), Bard (1974) and Harwell Subroutine Library (1974), Subroutine VA05A). The estimates of the parameters are :

$$\hat{a}_1 = 477.27 \ , \ \hat{a}_2 = -1.3935 \ , \ \hat{a}_3 = 0.16229 \ , \ \hat{a}_4 = 2.3421 \text{ and } \hat{a}_5 = 0.78336 \tag{4.7}$$

It should be noted that $\hat{a}_2$ being negative, the formula is only of use for $T > 1.3935$. Actually a good fit is found already for $T \geqslant 1.5$.

The secular maxima ($T = 100$), computed with the formula for the intensities in $t = 1$, 10, 20, 30, 60, 120, 360, 720 and 1440 min, are given in table 4. The root mean square difference with the original Gumbel distributions is 2,60 mm.

Table 4

Secular maximum rainfall intensities in t min. Values given by the adjusted Gumbel distributions (I) and estimates given by the generalized Gumbel distribution (II), the generalized hydrological formula (III) and the final generalized Gumbel distribution (IV) (mm). Root mean square difference e between the estimates and the values yielded by the Gumbel distributions.

t	I	II	III	IV
10	19.64	21.97	19.91	19.40
20	26.96	26.23	26.20	26.23
30	30.98	29.10	30.10	30.37
60	36.81	34.75	37.16	37.50
120	42.57	41.50	44.81	44.67
360	53.98	55.00	58.61	56.51
720	63.86	65.71	68.77	64.60
1440	77.66	78.52	80.36	73.42
e		1.58	2.72	1.98

4.3 The final generalized Gumbel distribution.

Despite their general good fit, the preceding interpolation formulas have both some deficiencies. The generalized Gumbel distribution does not yield estimates of fractiles for $t \leqslant 10$ min, while the generalized hydrological formula is only suitable for $T \geqslant 1.5$ year. Moreover, in the latter case, it should be noted that the interpolation has been made with the assumption that the maximum intensity in 1 min corresponds effectively to the intensity for the time interval of 1 min, which is not the case as already stated.

This leads us to consider a final formula of the form :

$$x_p(t) = \mu(t) + \sigma(t) y_p \tag{4.8}$$

where $\mu(t)$ and $\sigma(t)$ are both of the form :

$$z = at/(1 + bt^c) \tag{4.9}$$

noting that a is the value of $\mu(1)$ and $\sigma(1)$ of table 2 according to the case.

The estimation procedure for the constants b and c is the same as for formula (4.6) with the condition of minimization of the statistics :

$$\Sigma[\frac{\mu_t - \mu^*_t}{\mu^*_t} /0.0529]^2 \text{ and } \Sigma[\frac{\sigma_t - \sigma^*_t}{\sigma^*_t} /0.1103]^2 \qquad (4.10)$$

where μ_t and σ_t are given in table 2 for $t \geq 10$ min, μ^*_t and σ^*_t are computed with formula (4.9) and where the constants 0.0529 and 0.1103 are given in (4.4). The fit of the estimates may be tested using the fact that the statistics (4.10) have a χ^2 –distribution with 8–2 = 6 degrees of freedom under the null hypothesis of good fit.

The values of the constants b and c determined in this manner are given in table 5, and the values of the statistics (4.10) are respectively 10.3 and 0.64 which are both less than the critical value 12.6 at the 5% significance level. The fit is thus acceptable.

Table 5

Final generalized Gumbel distribution. Adjustment of the location parameter μ and of the scale parameter σ with a function of the form : $z = at/(1 + bt^c)$.

	μ	σ
a	17.27620	7.28297
b	0.234379	0.243000
c	0.806292	0.840348

For the comparison the secular maxima have been computed with formula (4.8) for t = 10, 20, 30, 60, 120, 360, 720 and 1440 min. The root mean square difference with the estimates given by the original Gumbel distribution is 1.98.

5. CONCLUSION.

After having presented two models yielding estimates not without some restrictions of the fractiles of the distribution of the annual maximum rainfall at Brussels-Uccle for any time interval up to 1440 min, a third model has been constructed giving the same estimates, but avoiding the inconsistencies of the two preceding models. The advantage of the last model is that it takes into account the exact significance of the instantaneous intensity of the rainfall thus far known as the intensity in 1 min.

Accessorily, for practical applications it should be reminded that the recording raingauge generally shows a deficit of about 10% in comparison with the ordinary raingauge.

REFERENCES

Abdy, P.R. and M.A.H. Dempster (1974), Introduction to Optimization Methods. Chapman and Hall.

Bard, Y. (1974). Nonlinear Parametric Estimation. Academic Press.

Demarée, G. (1985). Intensity-Duration-Frequency Relationship of Point Precipitation at Uccle, Reference Period 1934-1983. Inst. Roy. Mét. de Belgique, Kon. Met. Inst. van België, Pub. A, nr. 116.

Gumbel, E.J. (1965). A quick estimation of the parameters in Fréchet's distribution. Rev. Int. Stat. Inst., 33, pp.349-363.

Harwell Subroutine Library (1974). Subroutine VA05A.

Kimball, B.F. (1956). The bias in certain estimates of the parameters of the extreme-value distribution. Annals of Math. Stat., 27, pp. 758-767.

Sneyers, R. et J. Van Isacker (1972). Sur l'ajustement de la loi de répartition de Fisher-Tippett du type I au moyen des estimateurs de Kimball. Revue Belge de Stat., d'Inf. et de Rech. Opér., 12, 1, pp.36-43.

Sneyers, R. (1975). Sur l'analyse statistique des séries d'observations. Note Technique n° 143, OMM-N°415.

Sneyers, R. (1977). L'intensité maximale des précipitations en Belgique. Inst. Roy. Mét. de Belgique, Pub. B, n° 86, 15p.

Van Montfort, M.A.J. and M. Ivette Gomes, (1985). Statistical choice of extremal models for complete and censored data. J. Hydrol., 77, pp. 77-87.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

ESTIMATION, REGIONALIZATION, AND SPATIAL AVERAGING OF EXTREME RAINFALL AT SMALL TIME-STEPS

by Ch. OBLED, T. LEBEL * and M. SLIMANI **

Institut de Mecanique de Grenoble, Groupe d' Hydrologie
B.P. 6838402 Saint Martin d'Heres Cedex FRANCE

A very practical problem in the design of hydraulic structures is that of extreme value flood analysis. Commonly based on long discharges series, the problem must be approached differently when such data is lacking. In such a case, a possible solution consists in using information derived from a more currently available variable, i. e. rainfall, together with other information, either hydrological or topomorphological.One such example is the Gradex method (Guillot and Duband 1967, Lebel and Guillot 1983), which extrapolates the discharge probability function according to that of an appropriate representative rainfall for the watershed considered. Other widely used methods of estimating given return period floods rely on rainfall statistics such as the 5- or 10-year daily rainfall (e.g. see the NERC Flood studies Report -1975-, or the French Ministry of Agriculture SOCOSE-CRUPEDIX method-1980-).

An argument for selecting rainfall rather than other flood generating factors is that it may be less sensitive to local effects such as land or vegetation cover, and thus better suited to regionalization.

These reasons have led to rising interest in the Climatology of Extreme Rainfall, at first mainly on a daily basis. However, concerns with smaller rural watersheds as well as needs in urban hydrology soon raised interest in shorter time-steps too, with the drawback of much less dense networks of recording raingages. This leads to the problem of regionalizing sparse information, as well as that of the representativeness of point values for spatially extensive watersheds.

---

* Now with ORSTOM - Montpellier - France
** Now with Institut National Agronomique Tunisien - Tunis - Tunisia

Thus, the following paper tries to address 3 sequentially encountered problems :

1 - estimation of the extreme value distribution of point rainfall for short time-steps.

2 - regionalization of the point statistics.

3 - statistics of spatially averaged rainfall.

An extensive case study was performed on a region subject to violent and sudden rainfall of Mediterranean origin in Central and South East of France (see Obled and Creutin -1982-, or Obled -1983-, for brief meteorological description of these regions).

The region sketched in Figure 1-a is equipped with a network of 47 long-standing recording raingages, with most record lengths in the range 13-22 years, a few exceeding 25 years and one reaching 62 years.

I - EXTREME VALUE DISTRIBUTION FOR POINT VALUES.

I - 1 Avalaible data and choice of a probability law.

Only half the stations were analyzed for all available clock hour time-steps (e.g. 8h 00 - 9h 00), while for the others, only the monthly maximum were selected for time steps of :

1 h, 2 h, 4 h, 6 h, 12 h, 24 h

A drawback of the procedure, coped with through a Weiss correction factor, is that the time period may only begin or end on a rounded clock hour.

The probability model widely accepted for extreme rainfall is the Gumbel model (Extreme Value Type 1 : E.V.1). This may be backed either by extensive sampling (Hershfield and Kohler 1960), or through analytical considerations (Lebel and Guillot 1983). In our case, the choice was mostly empirical.

The initial distribution of monthly maxima (3 per year during the rainy autumn season) displays significant departures from the Gumbel model through a curvature or broken line effect at the bottom tail of the distribution. However, the goodness of fit steadily improves when the maxima are taken over increasingly long periods (e.g. 1 per year or 1 per 2 years - see Figure 2).

I - 2 Estimation procedures .

The Gumbel model for a random variable X is :

$$F(x) = Pr(X < x) = \exp(-\exp(-\frac{x-b}{a}))$$

where the parameters a and b are related to the 1st and 2nd order moments μ and σ^2 of the population through :

$$a = 0.78\,\sigma \quad \text{and} \quad b = \mu - 0.577\,a$$

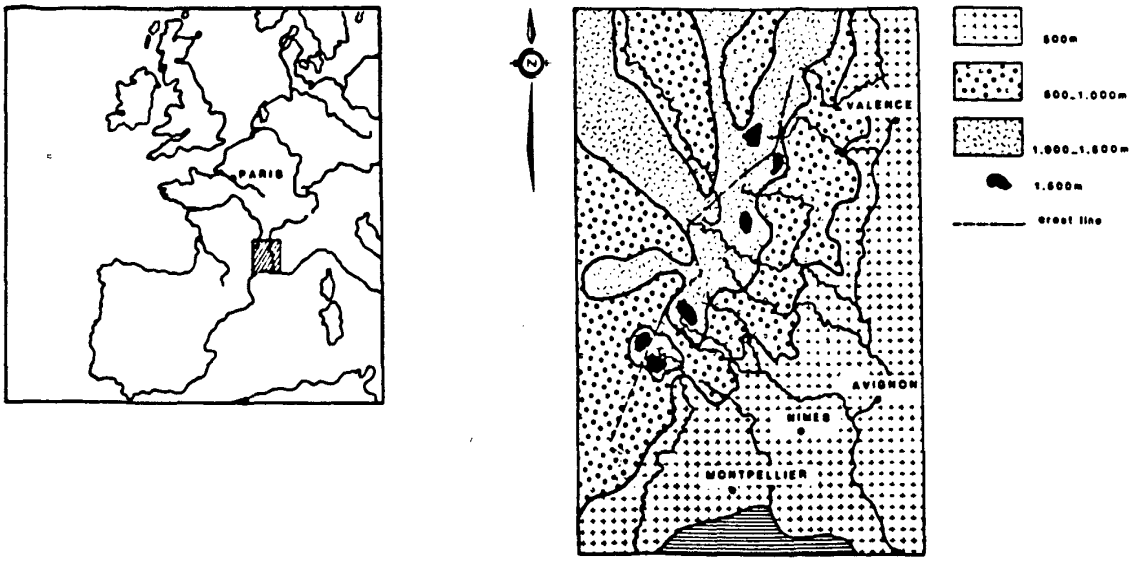
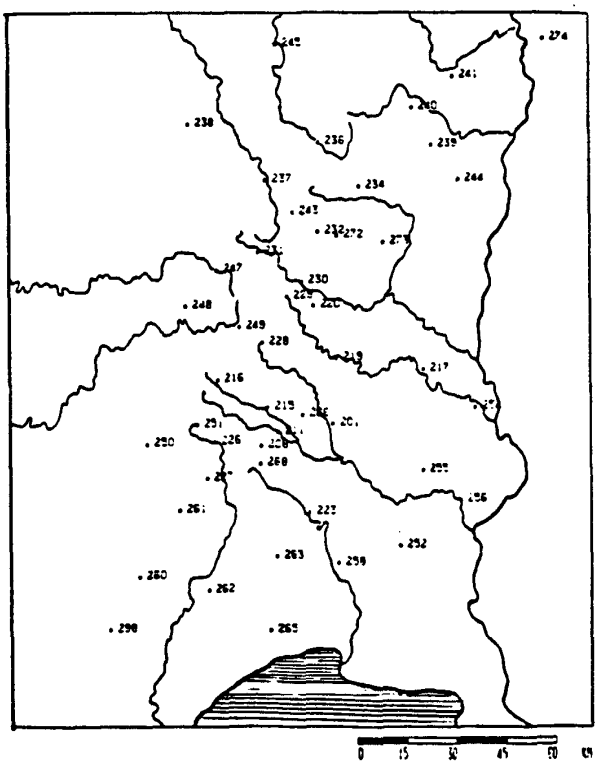


Figure 1 : Study Area

1 - a Location and relief
1 - b Drainage and Recording Raingages Networks



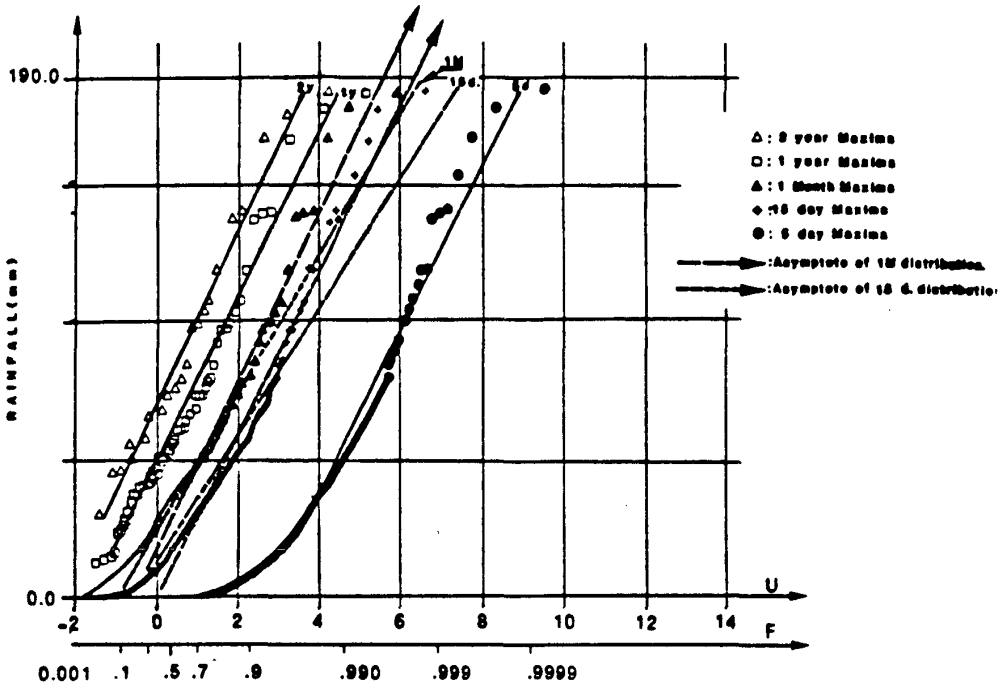


Figure 2 : Empirical Distribution of Fixed Duration Maxima for six-hour rainfall

The Gumbel scale parameter a is sometimes called the" Gradex" in the french literature (for gradient of the exponential).

The problem of correctly estimating a and b from a finite sample was approached in several ways :

a) through synthetically generating series of presumed extreme rainfalls following exactly a Gumbel distribution.

b) through generating series of continuous hourly rainfalls Y , or 2 h - rainfall etc..., according to the time- step considered. An appropriate model to do this is a mixture of 2 populations with a sum of exponentials as a probability function:

$$F(y) = Pr(Y < y) = 1 - \gamma \exp(-y/a) - \beta \exp(-y/c)$$

Here, the distribution of X, the maxima over fixed length periods is only assymptotically gumbelian.

c) through splitting the 62 year long series of hourly rainfall at Montpellier Bel Air station into subsamples of 20 years each.

For every dataset obtained from a), b) and c), several estimating procedures were applied, namely:

1) Moment Method (MO) , using sample mean and standard deviation, with a correction factor for small sample sizes (Gumbel 1954)

2) Maximum Likelihood Method (ML), as described by Clarke

(1973), and using a small sample size correction proposed by Fiorentino and Gabriele (1984).

3) Mean Square Fit (MS) of the upper part of the empirical distribution on Gumbel paper in case of pronounced curvature or break in the lower point swarm.

In fact, only the MO and MS methods were found acceptable in the experiments a), b), or c) above. The ML method, even with the proposed correction factors, always underestimates the Gumbel parameters, which is unacceptable for safety objectives.

Finally, the fixed origin effects, mainly sensitive for 1 h and 2 h rainfalls, were corrected, according to Weiss, by a coefficient of 1.14 and 1.07 respectively.

For each time-step Δt, 3 parameters were selected for a regional study and thus computed at each station :

Gumbel scale parameter	$a(\Delta t)$
10 year rainfall	$P_{10}(\Delta t)$
100 year rainfall	$P_{100}(\Delta t)$

along with their sampling estimation variances , as summarized by Laborde (1984) :

$$\sigma_a = 1.05\ a / \sqrt{n} \qquad \sigma_{P10} = 3.82\ a / \sqrt{n} \qquad \sigma_{P100} = 6.15\ a / \sqrt{n}$$

Although the choice of the Gumbel distribution seems hardly questionnable in our case, the authors agree that the estimation techniques may still be refined to obtain more robust estimates of quantiles , and even more importantly , better estimates of sampling variations (see e. g. Greenwood et al. - 1979- or Greis and Wood -1981-)

II - REGIONALIZATION OF POINT STATISTICS.

Here too, the approach was 3-fold.

II - 1 - Direct mapping of point values, through either spline surface fitting or kriging (Figure 3).

The first technique may be considered purely analytical and provides as smooth a surface as possible. The quantiles derived in section I show good organization for each time-step. However, this pattern is not unique and changes gradually from one step to annother (e. g. P_{10}(1 h) as compared to P_{10}(2 h) and so on...).

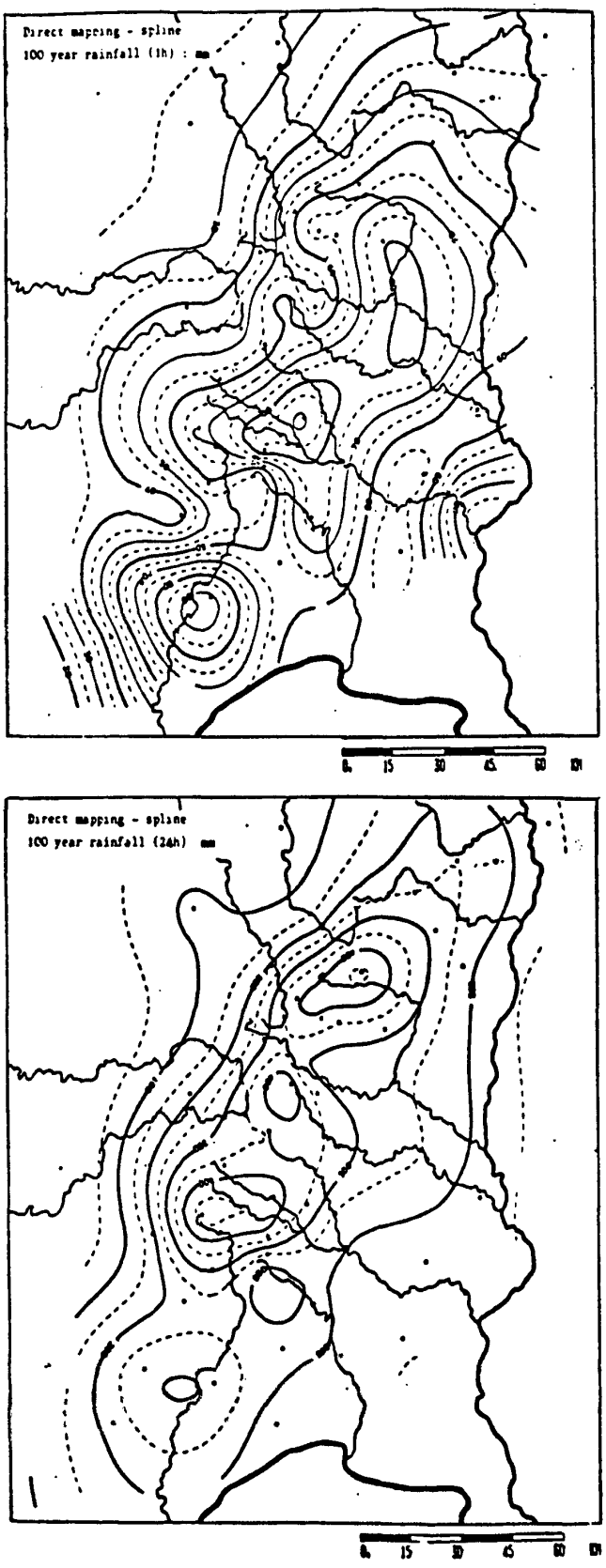


Figure 3 : Direct Mapping by Spline Surface Fitting of the 100-year rainfall for time-steps of 1 h and 24 h.

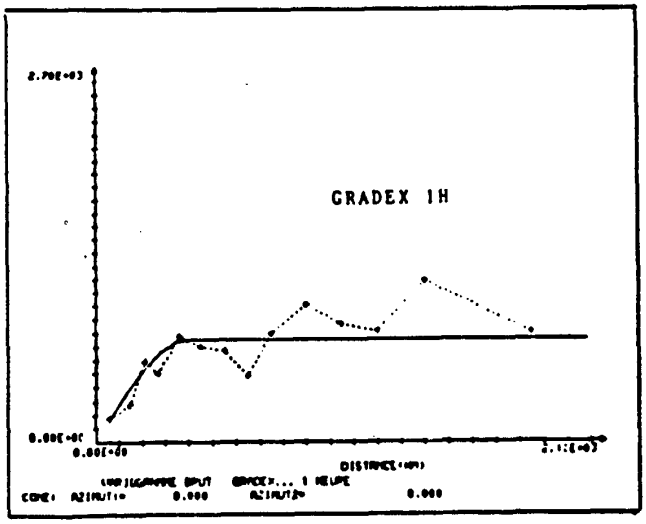
Kriging is a geostatistical technique that binds mapping to an underlying spatial random process, with its structure summarized in a variogram. (see Figure 4). In our case, the data were reasonably structured , with zero correlation length (or" range" in geostatistical terminology) increasing regularly from :

50 km for $P_{10}(1h)$ to 80 km for $P_{10}(24h)$

and 40 km for $P_{100}(1h)$ to 60 km for $P_{100}(24h)$

Although further data would be needed for a firm conclusion, the decrease in range with larger return periods seems physically acceptable, suggesting that strongest events are more spatially concentrated.

However, kriging tends to use outlying information conservatively and may thus underestimates extreme values, since it tends to fall back on the marginal field average in poorly sampled regions. Conversely, the purpose of such studies, namely design safety, leads to a preference for bias towards overestimation, so spline surface fitting was prefered in this 1st step.



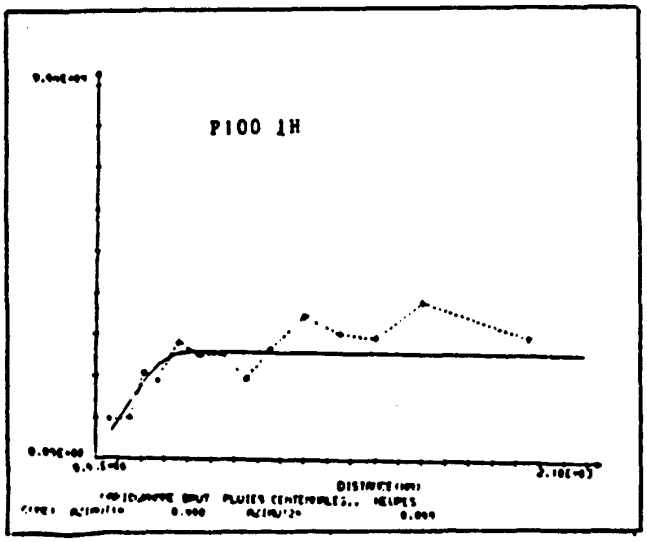


Figure 4:

Variograms of some 1 h-rainfall characteristics (note an approximate "range" of 40 km)

II - 2 Relationship with relief and morphological factors.

In Figure 3 , the relief sketched in figure 1 seems to play a major role in the well organized pattern of isolines. Previous studies by Laborde (1982,1984) in North East of France have shown that a significant part of the variation in extreme rainfall parameters may be found in topographic descriptors.

A large number of such variables , summarizing slope, altitude, exposure, embanking, etc ..., were thus computed for each of the 47 stations.

They were then regressed on each parameter (a, P_{10}, P_{100}) for each time-step. After screening, no more than 4 variables are needed (among 100 proposed), and exposure, distances , both to the sea and to the ridge of the mountain range, were almost always selected. The regression equation , calibrated over the sample of 47 stations x_i , may then be applied at any point x_0 were the morphometric variables are available, thus providing a regression estimate Z_R, where Z holds for a or P10 etc...

Here too, a gradually changing pattern emerges. Explained variance (R^2) ranges from 51 % to 68 % for time-steps of 1 h to 24 h, suggesting a stronger relation with the regional environment for totals over longer durations.

The relief of the whole region was then digitized over 500 m space-steps, and the regression equations obtained over the 47 stations were then extended to each gridpoint where the morphometric explanative variables could easily be computed.

This provided a very detailed map, generated by the relief alone, but that cannot be used as such because of the large part of unexplained variance.

II - 3 Cokriging.

Kriging in section II-1 can be considered as a regression of each unknown point on its measured neighbours only, i.e. the variable $Z(x_0)$ at point x_0 is regressed on the same variables $Z(x_i)$ at the i = 1, ... n observed stations . Here " observed" stands for " estimated from a sample of extreme rainfalls". Inversely, section II- 2 considered regressions over external relief variables only, but which were computed at the exact point x_0.

Cokriging (see for exemple Vauclin et al. 1984) is an attempt to merge the 2 steps described above in a kind of spatial ARMA model where the variable at the unmeasured point x_0 is estimated as :

$$Z(x_0) = \sum_{i=1}^{n} \alpha_i Z(x_i) + \sum_{j=1}^{k} \beta_j Y_j + \epsilon(x_0)$$

where Y_j could be the k original relief variables used in step II-2.

In fact, in order to reduce the number of degrees of freedom, we directly used in the Σ_1 the relief estimated value $Z_R(x_0)$, obtained from the regression of step II-2 , as the only external variable Y. However, this variable may be taken both at x_0 and at a few nearby surrounding mesh points. Figure 5 shows the estimating scheme.

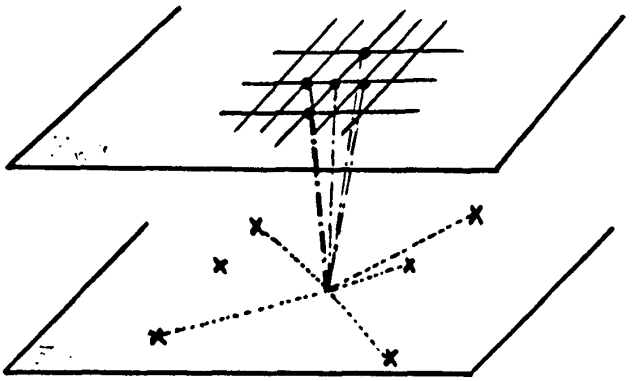


Figure 5 : Estimation Scheme for Cokriging , mixing "observed" values Z(x) with "external" information $Z_R(x)$

As in the kriging method, cokriging provides a map of the theoretical estimation variance, i.e. of

$$\sigma_\epsilon(x_0)$$

that may be used as a confidence interval index of the map obtained for $Z(x_0)$.

Without entering into the problem of estimating the required cross-structure functions, the conclusions that can be drawn from figure 6 are :

- introduction of the relief does not significantly change the pattern obtained in section II-1 for dense regions.
- the detail introduced by rapidly changing local relief may be somewhat unrealistic. Once combined with the map of estimation variance, it may be partly considered as noise and only large features need be considered.
- nevertheless, in poorly sampled zones, as well as at the edges of the map, the use of relief partially palliates the lack of nearby observed stations and significantly reduces border effects, as seen in figure 3.

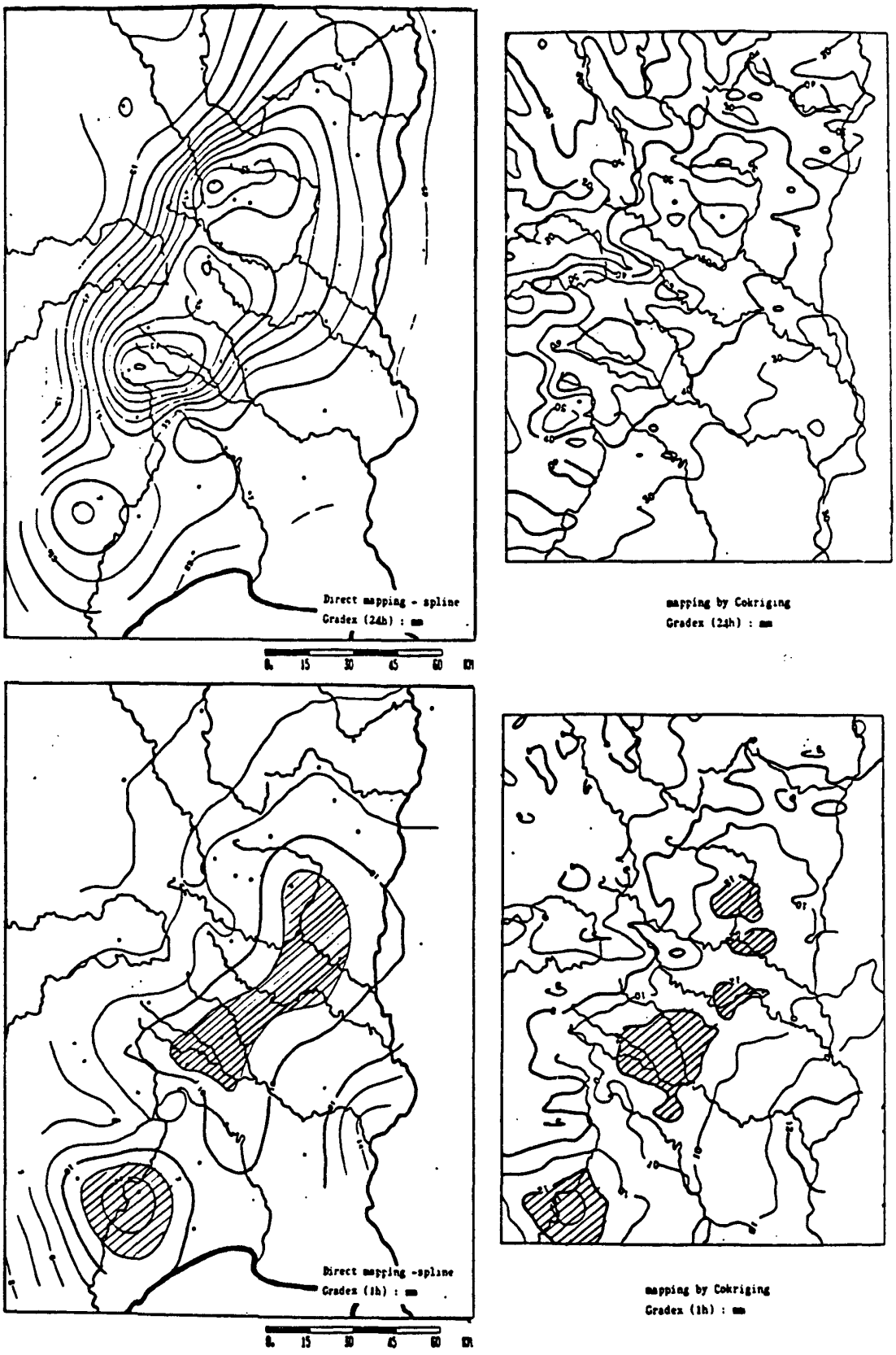


Figure 6: Comparison of direct Spline Mapping with Cokriging using Relief Estimation (Note that the relief has been digitized only over a part of the whole domain. The 1h-12 mm contour has been hachured to allow easier comparison)

A final but very interesting conclusion comes from the comparison of different time-steps : the zones of maxima for a given time-step smoothly shift from the ridge zone (maximum elevation) for long durations (24H, 12 h), towards the middle and the bottom half of the sloping relief (on the Mediterranean side) for short durations (2 h, 1 h). This backs up the very few results concerning the climatology of short duration rainfalls in mountains regions (Fletcher et al. 1981).

III - SPATIALLY AVERAGED RAINFALL .

In fact, the actual design problem is often set up in terms of a flood volume, itself related to a rainfall volume collected by the watershed. In other words, a given rainfall event may strongly vary in space, although picked out from a spatially homogeneous probability distribution, but only the basin average is of interest for the designer.

Furthermore, the extreme value distribution may also be heterogeneous, i.e. its parameters may vary significantly in space, as seen in section II for our region. This thus causes problems for the elongated watersheds of the Cevennes region, which spread through isolines 80 to 140 mm/h for the 2 hour - 100 year rainfall for exemple.

So, there is a strong need to obtain some kind of integrated representative value at the scale of a given watershed, as well as the distribution of its extremes. The appropriate time-step here is usually accepted to be the concentration time of the watershed, that may be roughly derived and fixes the time-step.

Hopefully, the designer should then be able to easily derive an estimate of the rainfall volume for any given return period without going back to the original data, for instance using the maps derived in section II.

Such a possibility was tested using the following 3 step approach:

III- 1 Studies of long series of spatially averaged rainfall.

For a few heavily instrumented watersheds (50 to 500 km ^{2}), series of basin average rainfall have been computed, using a kriging estimation process, at time-steps of 1 to 24 h (Lebel and Creutin 1983) over a period of about 12 years.

First, the Gumbel distribution was tested for this new variable and then, the Moment Method was used to estimate the appropriate parameters. It appears that this spatially averaged rainfall over small watersheds shows even more gumbelian behavior than point rainfall, especially for small time-steps.

These series were thus used as a kind of empirical, or sample-based, reference for more analytical approaches.

III - 2 Average statistics derived from point statistics.

In fact, the sample statistics required, at least for the Moment Method, were estimates of the mean and variance of the spatially averaged extreme rainfall :

$$P_M = \frac{1}{B}\int_B P(x)\,dx \qquad \text{over a basin area B.}$$

These may in turn be estimated from neighbouring point values as weighted estimators, since an event value may be obtained as :

$$P_M^* = \sum_{i=1}^{n} \lambda_i P(x_i)$$

then the population mean may be estimated as :

$$\mu_M = E\langle P_M\rangle = E\langle P_M^*\rangle = \sum \lambda_i \mu_i$$

and

$$\text{var}\langle P_M^*\rangle = \sum_i \sum_j \lambda_i \lambda_j \text{ cov}\langle P(x_i), P(x_j)\rangle$$

$$= \sum_i \sum_j \lambda_i \lambda_j (1 - \gamma_{ij}) \sigma_i \sigma_j$$

with γ_{ij} the scaled structure function (Gandin 1965) or variogram between observed points or stations x_i and x_j.

However, this variance var$\langle P_M^*\rangle$ underestimates the true variance since the estimate P_M^* has a lower variance than the true integrated value P_M.

The difference is the estimation variance, which may itself be evaluated as :

$$\sigma_E^2(P_M^*) = \sum_i \lambda_i \int_B \gamma(x_i,x)\,dx - \sum_i \sum_j \lambda_i\lambda_j \gamma_{ij} - \int_B \gamma(x,x')\,dx\,dx'$$

However, as a first approximation, σ_E^2 may be omitted. In this case, the basin average Gumbel parameter a_B may be estimated as a_B^* through the MO method taking:

$$a_B^* = 0.78\left[\sum_i \sum_j \lambda_i \lambda_j (1-\gamma_{ij}) \sigma_i \sigma_j\right]^{1/2}$$

instead of

$$a_B = 0.78\ \text{var}(P_M)^{1/2}$$

This may even be written :

$$a_B^* = \left[\Sigma_i \; \Sigma_j \; \lambda_i \lambda_j (1 - \gamma_{ij}) \, a_i a_j\right]^{1/2}$$

The λ_i's are the weights used in estimating the average rainfall and are given by the well known kriging system :

$$\begin{bmatrix} & & 1 \\ & \gamma_{ij} & \vdots \\ & & 1 \\ 1 \ldots & 1 \ldots 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_i \\ \mu \end{bmatrix} = \begin{bmatrix} \frac{1}{B} \int_B \gamma(x, x_i)\, dx \\ 1 \end{bmatrix}$$

And thus, an estimate of the Gumbel parameter a_B may be derived directly from the parameter point values a_i's.

The same holds for estimating μ_M, i.e.

$$E\langle P_M \rangle = E\left\langle \frac{1}{B} \int_B P(x)\, dx \right\rangle$$

Other studies show that γ_{ij} in fact depends only on the distance d_{ij} between x_i and x_j, and on the time-step duration Δt (Lebel and Creutin 1983).

Thus, the Gumbel distribution for basin averaged rainfall may be inferred from point statistics (or map), current rainfall geostatistical structure function (or an appropriate model $\gamma(d_{ij}, \Delta t)$), through the simple calculation of the kriging weights λ_i's.

III - 3 Simpler estimates.

However, this simple computation may itself be considered tedious, and a cruder estimate may be obtained from a simple averaging:

$$a_B^{**} = \Sigma_i \; \Sigma_j \; a_i \, a_j$$

with i,j screening the whole meshed domain B considered.

In fact, this estimate is not as crude as would first appear , in that it compensates the underestimation of $\mathrm{var}(P_M^*)$ viz. $\mathrm{var}(P_M)$ by considering that the λ_i's are all equal, i.e. that x_i, x_j are fully correlated at 1.0 .

Practically, such an averaging can easily be performed over the maps obtained in section II.

These 3 steps were compared over several watersheds with reasonnably good results.

Example: parameters a and P_{10} for

	$\Delta t = 1$ h		$\Delta t = 24$ h	
areal rainfall computations	8	35mm	53	220 mm
"kriging" of point parameters	7	41	43	195 m
crude averaging	12	52	61	245mm

for the Gardon d'Anduze watershed (550 km^2).

areal rainfall computations	9	37mm	58	240 mm
"kriging" of point parameters	9	43mm	49	220mm
crude averaging	12	51mm	65	240 mm

for the Gardon de Mialet (only 265 km^2 but higher on the mountain slope).

As expected, the kriging approach slighty underestimates, while full averaging overestimates. These 2 values, nevertheless, provide a reasonable range for the designer.

Note that these approaches arec only valid for watershed extension of 1 or 2 correlation lengths or "ranges" at the most. Outside this limit, the Gumbel hypothesis may no longer be acceptable.

CONCLUSIONS

Although the different steps have not yet been developed to the same extent, the overall approach is highly integrated and aims at merging as much information as possible in the form of easily handled maps, together with a simple technique for estimating basin average statistics. Real world applicaions to small size watersheds (<500 km^2) in the investigated region are underway.

ACKNOWLEDGEMENTS:

We are deeply indebted to the different services, EDF-DTG, SRAE's, DDE 's etc... which kindly provided the data . D. DUBAND, P.GUILLOT and J.P. LABORDE are greatly thanked for suggestions and helpful discussions. This study was partly funded by the Direction Departementale du Gard .

REFERENCES.

BERAN M. 1981 ' The "Gradex " method, letter to the Editor ' IAHS Bulletin , Vol. 26, 2, p. 223-224

CLARKE R.T. 1973 ' Mathematical models in Hydrology' Irrig. and Drain. Pap. n°19 . F.A.O. publ. United Nations . Rome.

COOPER C.F. 1967 ' Rainfall intensity and elevation in Southwestern Idaho' Water. Ress. Res. Vol 3, 1, p 131-137

CROFT A.R. and MARSTON R.B. 1950 ' Summer rainfall characteristics in Northern Utah' Trans. Amer. Geophys. Union , Vol 31, 1, p. 83-95

DUBAND D. 1967 ' Composition des lois de probabilite des pluies journalieres' Internat. Hydrol. Symp. Colorado State Univ., Proc. Vol 1, p . 516-526

FIORENTINO M. and GABRIELE S. 1984 'A correction for the bias of Maximum Likelihood estimators of Gumbel parameters ' J. of Hydrol., Vol. 39, 1, p.39-49

FLETCHER H.E. et Al. 1981 'Precipitation characteristics of summer storms at Straight Canyon Barometer watershed, Utah' Research Pap. INT-274, U.S. Dept of Agriculture, Nat. Forest Service - May 1981 - 11 p.

GREENWOOD J.A., LANDWEHR J.M., MATALAS M.C. and WALLIS J.R. 1984 ' Probability weighted moment: definition and relation to parameters of several distribution expressed in inverse form' Water . Ress. Res., Vol 15, 16, p. 1049-1054

GREIS N.P. and WOOD E. 1981 'Regional flood frequency estimation and network design' Water Ress. Res. , Vol. 17, 4, p. 1167-1177

GUILLOT P. and DUBAND D. 1967 ' The Gradex method to compute extreme flood probability from the observation of rainfall' Internat. Hydrol . Symp., Colorado State Univ., Fort Collins, Proc. Vol 1, p. 506-515

GUMBEL E.J. 1954 ' Statistical theory of extreme values and some practical applications ' Nat. Bureau of Standards, Appl. Math. Series 3-3

HERSHFIELD D. and KOHLER M. 1960 'An empirical appraisal of the Gumbel extreme value procedure' J. of Geophys. Res., Vol 65, 6, p. 1737-1746

LABORDE J.P. 1982 ' Cartographie automatique des caracteristiques pluviometriques: prise en compte des relations morphometrie-pluviometrie ' La Houille Blanche N°4, p. 331-338

LABORDE J. P. 1984 'Analyse des donnees et cartographie automatique en hydrologie - Elements d'hydrologie lorraine' These de doctorat d'Etat. I.N.P. de Lorraine. Nancy -France

LEBEL T. and CREUTIN J.D. 1983 'Areal rainfall estimation, forescast, and simulation over small watersheds subject to major flashfloods' 5th Conf. on Hydromet., Amer. Met. Soc., Tulsa, OK., Proc. p.291-298

LEBEL T. and GUILLOT P. 1983 '985 Statistical analysis of point and areal hourly rainfall: application to the design flood by the Gradex method' 5th Conf. on Hydromet., Amer. Met. Soc. Tulsa, OK. Proc. p. 17-25

NATURAL ENVIRONMENT RESEARCH COUNCIL 1975 ' Floods Studies Report' 5 vol. NERC London U.K.

OBERLIN G. 1980 'Une methode sommaire d'estimation des crues sur de petits bassins versants de metropole: la methode SOCOSE' La Houille Blanche N° 4/5, p. 271-273

OBLED CH. 1983 'Hydrometeorology in France : New results and developments ' 5th Conf. on Hydromet., Amer. Met. Soc. Tulsa, OK., Proc., p 1-8

VAUCLIN M. et al. 1983 ' The use of Cokriging with limited field soil observations' J. of Soil Science Soc. of Amer., Vol 127, 2, p 175-184

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

EXPONENTIALITY VERSUS GENERALIZED PARETO, QUICK TESTS

M. Ivette Gomes
Faculty of Science of Lisbon, Lisbon (Portugal)

and

Martin A.J. van Montfort
Agricultural University, Wageningen (the Netherlands)

Abstract: Two test statistics, G_n and $\hat{V}_n$, are proposed for testing Exponentiality versus Generalized Pareto distribution, where G_n is the ratio of the maximum and the median and $\hat{V}_n$ is a function of the coefficient of variation. Asymptotic approximations for G_n and $\hat{V}_n$ are given; those approximations turn out to be given by a Generalized Extreme Value and a Normal random variable respectively. Such statistics are compared, either for finite n or asymptotically. For small n (less than, say, 100) both statistics nearly have the same power, while $\hat{V}_n$ is more powerful asymptotically.

Keywords and phrases: Generalized Pareto distribution; hypothesis testing; asymptotic distributions.
AMS 1980 subject classification: Primary: 62F03; Secondary: 62E20, 62E25.

1. INTRODUCTION

Model choice in statistics of extremes is essential in applications where the attention is focussed at rarely occuring events, such as an annual maximal flood exceeding the dykes, or a seasonal minimal temperature below the critical value for crop production.

The simplist model for univariate extremes is the type I distribution of extremes, promoted by Gumbel. In practise longer and shorter tailed distributions appear, which are modelled by the General Extreme Value distribution (GEV) with three parameters (location, scale, shape). The type I distribution of extremes belongs to the GEV-family and has a fixed shape parameter.

Inference in model choice can be based on realized extremes or on counts of exceedances (or peaks) over a given threshold and their values, so called Peaks Over Threshold (POT-)data. Here attention is paid to the second case. A Poisson distributed number of Exponentially distributed exceedances gives rise to a type I maximum; generalizing the Exponential distribution to the Generalized Pareto distribution gives a GEV-maximum. So evidence in the exceedances for a Generalized Pareto distribution supports a GEV distribution for the maximum.

Alternatives for the Exponential distribution in the POT model preferably have a positive density for the lower bound of the carrier of the distribution, in which case a Gamma alternative is not adequate. An adequate statistic for this alternative is the ratio of the geometric and the arithmetic mean, which statistic especially pays attention to deviations in the left tail.

Two quick and easy statistics, measuring lack of fit of the Exponential distribution in favour of the Generalized Pareto distribution, are proposed; the one depends on the ratio of the maximum and the median of a sample, and the other on the coefficient of variation. Their power functions are studied, both for finite sample size (simulation) and by approximations based on their asymptotic behaviour.

In Statistical Theory of Extremes, when analysing Peaks Over Threshold (POT)-data, or when dealing with conditional exceedances of high levels [Pickands, 1975; Smith, 1984], we

are often faced with Generalized Pareto (GP)-data $\underset{\sim}{X} = (X_1, \ldots, X_n)$, an independent sample from a population with distribution function (d.f.)

$$F_\theta(x) = \begin{cases} 1 - (1-\theta x/\delta)^{1/\theta}, & 1-\theta x/\delta > 0,\ x > 0 \quad \text{if } \theta \neq 0 \\ 1 - \exp(-x/\delta),\ x > 0 \quad \text{if } \theta = 0 \text{ (Exponential d.f.)} \end{cases} \tag{1.1}$$

$\delta > 0$ an unknown scale parameter. Notice that the expression (1.1) for $\theta = 0$ is the usual continuous continuation of $F_\theta(x)$, $\theta \neq 0$, as θ approaches zero. We are then interested in testing the hypothesis H_0: $\theta = 0$ (Exponential behaviour), in GP-data, versus suitable one-sided or two-sided alternatives.

When dealing with a simular problem in a Generalized Extreme Value (GEV) model

$$G_\theta(x) = \begin{cases} \exp(-(1-\theta x)^{1/\theta}), & 1-\theta x > 0,\ x \in R \quad \text{if } \theta \neq 0 \\ \exp(-\exp(-x)),\ x \in R \text{ if } \theta = 0 \quad \text{(Gumbel d.f.)} \end{cases} \tag{1.2}$$

the statistic proposed by Gumbel, 1965 has proved to be powerful both for finite n and asymptotically [Gomes, 1982; Tiago de Oliveira and Gomes, 1984; Van Montfort and Gomes, 1985]. Gumbel, 1965 proposed the ratio of the upper range (maximum - median) and lower range (median - minimum) in a random independent sample in a context of estimation in Fréchet populations $\Phi_\alpha(x) = \exp(-x^{-\alpha})$, $x > 0$, $\alpha > 0$. This took us to consider here, analogously to what has been done by Van Montfort and Witter, 1985, but our interest lying now additionally in asymptotic properties (large n), the statistic

$$G_n = G_n(\underset{\sim}{X}) = \frac{\text{Maximum}(X_1,\ldots,X_n)}{\text{Median}(X_1,\ldots,X_n)} \tag{1.3}$$

whose d.f. is independent of the unknown scale parameter δ, i.e. $G_n = G_n(\underset{\sim}{X}) = G_n(\underset{\sim}{Z})$, $\underset{\sim}{Z} = \underset{\sim}{X}/\delta$, $\delta > 0$.

The behaviour of a test statistic based on G_n is compared with the behaviour of a test statistic based on an analogue of the Locally Most Powerful (LMP) test statistic in a standardized context.

Indeed, when working with $\underset{\sim}{Z} = (Z_1, \ldots, Z_n)$, $Z_i = X_i/\delta$, $1 \leq i \leq n$, the LMP test of H_0: $\theta = 0$ versus one-sided alternatives is provided by the statistic

$$V_n = V_{n,\theta}(\underset{\sim}{Z}) = \sum_{i=1}^{n} v_\theta(Z_i)/n \tag{1.4}$$

where

$$v_\theta(z) = \left.\frac{\partial \log f_\theta(z)}{\partial \theta}\right|_{\theta=0} = z - z^2/2, \tag{1.5}$$

$f_\theta(z) = \frac{\partial}{\partial z} F_\theta(z)$, $F_\theta(z)$ given by (1.1).

Following an approach similar to the one used in Tiago de Oliveira, 1981, we estimate δ, the unknown scale parameter, through its Maximum Likelihood (ML) estimator under H_0: $\theta = 0$, i.e., by $\hat{\delta}_0 = \bar{X}$, the sample mean, and consider the analogue of the LMP statistic

$$\hat{V}_n = \hat{V}_{n,\theta}(\underset{\sim}{X}) = \frac{1}{n}\sum_{i=1}^{n} v_\theta(X_i/\bar{X}), \quad v_\theta(.) \text{ given by (1.5)} \tag{1.6}$$

The statistic $\hat{V}_n$ could be written as $(1-S^2/\bar{X}^2)/2$, where S^2 is the sample variance $\sum_{i=1}^{n}(X_i-\bar{X})^2/n$; so the discrepancy between the sample standard deviation and the sample mean shows the evidence against Exponentiality. Note that the coefficient of variation equals one for the Exponential distribution.

In section 2 of this paper, the asymptotic behaviour of G_n and $\hat{V}_n$ is discussed. The proof of these asymptotic results is presented in the Appendix. In section 3 we give simulated power and critical regions of suitable transformations of the test statistics G_n and $\hat{V}_n$. A comparison of asymptotic power of both statistics is made in section 4. Finally, in section 5, the rate of convergence of suitable transformations of G_n and $\hat{V}_n$ to their limits, as $n \to \infty$, is investigated experimentally.

2. ASYMPTOTIC BEHAVIOUR OF G_n AND $\hat{V}_n$; CHOICE OF STATISTICS

We present here the asymptotic properties of G_n and $\hat{V}_n$ given by (1.3) and (1.6) respectivily; proofs are presented in the Appendix.

2.1. Asymptotic behaviour of G_n.

Regarding G_n, and under a general $F_\theta(.)$ model, we have that, for $\theta < 1/2$,

$$\{G_n - \mu_n(\theta)\}/\sigma_n(\theta) \xrightarrow[n\to\infty]{w} Y, \tag{2.1}$$

Y a random variable (r.v.) with d.f. $G_\theta(.)$ given by (1.2), where

$$\left.\begin{aligned} &\mu_n(\theta) = \{1-n^{-\theta}\}/\{1-2^{-\theta}\} \xrightarrow[\theta\to 0]{} \mu_n(0) = \log(n)/\log(2) \\ &\text{and} \\ &\sigma_n(\theta) = \theta n^{-\theta}/\{1-2^{-\theta}\} \xrightarrow[\theta\to 0]{} \sigma_n(0) = 1/\log(2) \end{aligned}\right\} \tag{2.2}$$

i.e, the d.f. of G_n is, for large n, well approximated by a r.v. with d.f. $G_\theta((x-\mu_n(\theta))/\sigma_n(\theta))$.

This took us to consider the test statistic

$$G_n^* = \log(2)\, G_n - \log(n) \tag{2.3}$$

which is asymptotically, under H_0, a standard Gumbel r.v., being, under H_1, a r.v. with d.f. $G_\theta((x-\mu_n^*(\mu))/\sigma_n^*(\theta))$, with

$$\left.\begin{aligned} \mu_n^*(\theta) &= \log(2)\mu_n(\theta) - \log(n) \\ \sigma_n^*(\theta) &= \log(2)\sigma_n(\theta) \end{aligned}\right] \qquad (2.4)$$

$\mu_n(\theta)$ and $\sigma_n(\theta)$ given by (2.2).

2.2. Asymptotic behaviour of $\hat{V}_n$.

In what concerns $\hat{V}_n$, and under a general $F_\theta(.)$ model, $\theta > -\frac{1}{4}$, we have the asymptotic Normality of $\hat{V}_n$, i.e.,

$$\sqrt{n}\,\frac{\hat{V}_n - b(\theta)}{c(\theta)} \xrightarrow[n\to\infty]{W} W, \text{ a standard Normal r.v.} \qquad (2.5)$$

where

$$b(\theta) = \theta/(1+2\theta)$$

and

$$c^2(\theta) = (1+\theta)^2(1+\theta+6\theta^2)/((1+2\theta)^3(1+3\theta)(1+4\theta)) \qquad (2.6)$$

This took us to consider the test statistic

$$\hat{V}_n^* = \sqrt{n}\,\hat{V}_n \qquad (2.7)$$

which is asymptotically, under H_0, a standard Normal r.v., and under the alternative hypothesis H_1, a Normal r.v. with mean value $b_n(\theta) = \sqrt{n}\,b(\theta)$ and variance $c^2(\theta)$.

2.3. Some numerical values of standardizing coefficients.

In table 1 we present for each entry (n,θ), $n = 20, 100, 500$ and $\theta = -.2(.05).2$, the numerical values for $\mu_n^*(\theta)$ and $\sigma_n^*(\theta)$ in the first and second lines respectively. The scale parameter $c(\theta)$ for $\hat{V}_n^*$ is also provided in the last line of the table.

θ:	-.20	-.15	-.10	-.05	.00	.05	.10	.15	.20
n= 20	.829	.593	.377	.180	.000	-.165	-.316	-.455	-.582
	1.697	1.487	1.303	1.142	1.000	.876	.767	.672	.588
n=100	2.442	1.691	1.043	.484	.000	-.420	-.785	-1.104	-1.382
	2.342	1.893	1.531	1.237	1.000	.808	.653	.528	.426
n=500	5.279	3.528	2.107	.948	.000	-.780	-1.424	-1.959	-2.405
	3.231	2.410	1.798	1.341	1.000	.746	.556	.415	.309
$c(\theta)$	6.206	3.071	1.902	1.325	1.000	.800	.668	.577	.512

Table 1. Values of normalizing constants for G_n^* and $\hat{V}_n^*$ (see text).

3. SIMULATED POWER AND CRITICAL REGIONS FOR G_n^* AND $\hat{V}_n^*$.

Thinking first on finite values of n, we present table 2 simulated critical regions and simulated power functions of the statistics G_n^* given by (2.3) and $\hat{V}_n^*$ given by (2.7) for sample sizes n = 20, 100, 250 for test sizes α = .05 and α = .10, and for values of θ = -.3, -.2(.05).2, .3 (999 simulations).

n	α		θ values											Critical region for	
			-.3	-.2	-.15	-.10	-.05	.00	.05	.10	.15	.2	.3	$H_1:\theta<0$	$H_1:\theta>0$
20	.05	G_n^*	.31	.20	.16	.13	.09	.05	.06	.07	.09	.12	.19	4.04↑	-1.19↓
		$\hat{V}_n^*$	.30	.28	.23	.17	.08	.05	.08	.10	.12	.15	.23	-1.41↓	1.22↑
	.10	G_n^*	.39	.28	.23	.17	.14	.10	.13	.17	.20	.25	.37	3.15↑	-.89↓
		$\hat{V}_n^*$	.41	.29	.23	.18	.13	.10	.12	.15	.18	.22	.34	-.89↓	1.11↑
100	.05	G_n^*	.73	.50	.38	.26	.14	.05	.11	.19	.34	.49	.84	3.40↑	-1.21↓
		$\hat{V}_n^*$	.79	.55	.40	.28	.14	.05	.10	.21	.33	.49	.81	-1.79↓	1.50↑
	.10	G_n^*	.83	.61	.52	.36	.24	.10	.18	.32	.46	.65	.94	2.55↑	-.94↓
		$\hat{V}_n^*$	.86	.66	.52	.37	.24	.10	.20	.33	.48	.67	.91	-1.23↓	1.25↑
250	.05	G_n^*	.93	.70	.53	.35	.16	.05	.16	.38	.64	.89	1.0	3.37↑	-1.13↓
		$\hat{V}_n^*$	.98	.85	.69	.45	.20	.05	.20	.45	.70	.94	1.0	-1.80↓	1.46↑
	.10	G_n^*	.96	.80	.65	.46	.26	.10	.27	.51	.78	.95	1.0	2.57↑	-.86↓
		$\hat{V}_n^*$	1.0	.93	.80	.59	.33	.10	.30	.58	.85	.97	1.0	-1.28↓	1.24↑
Asymptotically:															
	α=.05	G_n^*												2.97↑	-1.09↓
		$\hat{V}_n^*$												-1.64↓	1.64↑
	α=.10	G_n^*												2.25↑	-.83↓
		$\hat{V}_n^*$												-1.28↓	1.28↑

Table 2. Simulated power and simulated critical regions of G_n^* and $\hat{V}_n^*$.

Notice here that the statistic G_n^* provides asymptotically and at significance level α,

a right sided critical region $[-\log(-\log(1-\alpha)), +\infty)$ for alternatives $\theta<0$, and a left sided critical region $(-\infty, -\log(-\log(\alpha))]$ for alternatives $\theta>0$. On the other hand and with ζ_p denoting the quantile of probability p of a standard Normal r.v., the statistic $\hat{V}_n^*$ provides asymptotically, and at the significance level α, a left sided critical region $(-\infty,\zeta_\alpha]$ for alternatives $\theta<0$, and a right sided critical region $[\zeta_{1-\alpha}, +\infty)$ for alternatives $\theta>0$. Hence the reason for the entries associated to the asymptotic critical regions; $x\uparrow$ and $x\downarrow$ denote values greater than and smaller than x, respectively.

We thus conclude that the power of both statistics is nearly the same for sample sizes $n = 20$ and $n = 100$. For $n = 250$, $\hat{V}_n$ turns out to be more powerful than G_n, which gives an indication of the eventual asymptotic superiority of $\hat{V}_n$ over G_n.

4. COMPARISON OF ASYMPTOTIC POWER OF G_n AND $\hat{V}_n$.

With $\mu_n^*(\theta)$ and $\sigma_n^*(\theta)$ given by (2.4), the asymptotic critical regions for G_n^* mentioned before, in section 3, together with the asymptotic behaviour of G_n^* under the alternative hypothesis, imply that the power function of a statistical choice test of H_0: $\theta = 0$ versus H_1: $\theta < 0$, based on G_n^* (or on G_n), and with size α, is well approximated, for large values of n, by

$$P_{n,a}(\theta|G_n,\theta<0) = 1 - G_\theta\{(-\log(-\log(1-\alpha))-\mu_n^*(\theta))/\sigma_n^*(\theta)\}, \tag{4.1}$$

G_θ given by (1.2).

For alternatives H_1: $\theta > 0$, we get

$$P_{n,a}(\theta|G_n,0<\theta<.5) = G_\theta\{(-\log(-\log(\alpha)) - \mu_n^*(\theta))/\sigma_n^*(\theta)\}, \tag{4.2}$$

G_θ given by (1.2).

Similarly, now for a statistical choice test of the parameter θ, based on the statistic $\hat{V}_n^*$ (or $\hat{V}_n$), we get the approximations

$$P_{n,a}(\theta|\hat{V}_n, -\tfrac{1}{2}<\theta<0) = \Phi((\zeta_\alpha - \sqrt{n}\, b(\theta))/c(\theta)) \tag{4.3}$$

and

$$P_{n,a}(\theta|\hat{V}_n, \theta>0) = 1 - \Phi((\zeta_{1-\alpha} - \sqrt{n}\, b(\theta))/c(\theta))$$

$b(\theta)$ and $c(\theta)$ given by (2.6), and with $\Phi(.)$ denoting the standard Normal d.f., ζ_p, as before, the p-th quantile of such d.f.

The comparison of asymptotic power functions is now undertaken: for different values of sample size n, we fixed a level, say .5, for both $P_{n,a}(\theta|\hat{V}_n,-\frac{1}{2}<\theta<0)$ and $P_{n,a}(\theta|\hat{V}_n,\theta>0)$, getting then as solutions of such equations $\theta_1 = \zeta_\alpha/(\sqrt{n} - 2\zeta_\alpha) < 0$ and $\theta_2 = \zeta_{1-\alpha}/(\sqrt{n} - 2\zeta_{1-\alpha}) > 0$ respectively. The computation of $P_{n,a}(\theta_1|G_n, \theta<0)$ and

$P_{n,a}(\theta_2|G_n,\ 0<\theta<.5)$ is then undertaken.

Table 3 shows, in case of applicability of asymptotic formulas for $\hat{V}_n$ and G_n, that θ values ($\theta < 0$) with power .5 for $\hat{V}_n$ have a lower power for G_n. For alternatives $\theta > 0$ the Gumbel statistic turns out to be more powerful for $n < 400$.

	$H_1:\theta<0$		$H_1:\theta>0$	
n	$\theta:P_{n,a}(\theta\|\hat{V}_n,\theta<0)=.5$	$P_{n,a}(\theta\|G_n,\theta<0)$	$\theta:P_{n,a}(\theta\|\hat{V}_n,\theta>0)=.5$	$P_{n,a}(\theta\|G_n,\theta>0)$
50	-.159	.328	.435	.933
100	-.124	.331	.245	.844
200	-.094	.316	.152	.644
300	-.080	.300	.117	.556
400	-.071	.286	.098	.503
500	-.064	.274	.086	.467
1000	-.047	.236	.058	.370
2000	-.034	.198	.040	.296
3000	-.028	.177	.032	.260
4000	-.025	.164	.027	.238
5000	-.022	.155	.024	.222

Table 3. Asymptotic power comparison of $\hat{V}_n$ and G_n for one-sided alternatives, $\alpha = .05$, $P_{n,a}(\theta|\hat{V}_n,.)$ fixed at level .5.

Figure 1 (a continuous version of the discretized graph really obtained) is a visual representation of table 3.

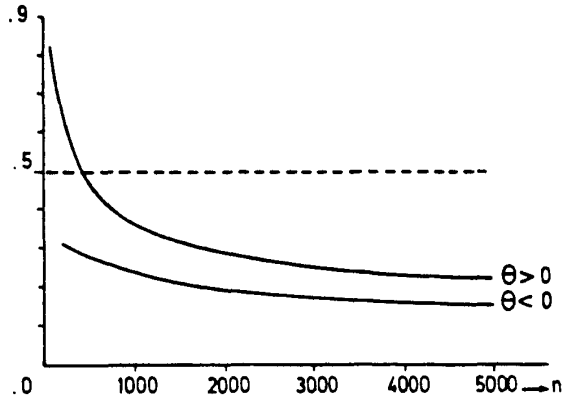


Figure 1. Power of G_n whenever the power of $\hat{V}_n$ is .5 (asymptotic approximations)

Similar results were obtained for other fixed values of power of the test statistic $\hat{V}_n$.

Let us think also on the ML estimator of θ, $\hat{\theta} = \hat{\theta}_{ML}$, in a population with underlying d.f. $F_\theta(.)$ given by (1.1), for test statistic of the null hypothesis H_0: $\theta = 0$. Using the asymptotic result given by Smith, 1984, according to which

$$\sqrt{n}\,\{\hat{\theta}_{ML} - \theta\}/(1-\theta) \xrightarrow[n\to\infty]{w} W,$$

a standard Normal r.v., whenever $\theta < .5$, an asymptotic power of .5 for $\theta < 0$ is thus obtained for $\theta = \zeta_\alpha/(\sqrt{n} - 2\zeta_\alpha)$, which provides a power .5 for a test of H_0, based on $\hat{V}_n$, we notice here that for fixed n, α and power equal to .5, the quotient of θ values is $1-2\zeta_\alpha/\sqrt{n}$ in favour of $\hat{\theta}_{ML}$. Anyway this quotient goes to one, as $n\to\infty$.

5. RATE OF CONVERGENCE OF G_n^* AND $\hat{V}_n^*$ TO THEIR LIMIT LAWS.

We are here interested in the validation of the asymptotic results referred to in section 2, for finite values of n. We have thus undertaken a simulation (999 runs) of the statistics G_n^* and $\hat{V}_n^*$ for values of $\theta = -.2(.05).2$ and $n = 20$, 100 and 250.

Regarding G_n^*, which suitably normalized converges to a GEV r.v., we tried to fit for each pair (θ,n), a d.f. $G_\theta((x-\lambda)/\delta)$, λ, δ and θ estimated through maximum likelihood. The ML estimates of location λ, scale δ and shape θ are provided in the first line of each entry (θ,n) in table 4. Below these values we put, for comparison, the asymptotic $\mu_n^*(\theta)$, $\sigma_n^*(\theta)$ and θ respectively. The fourth column, in each entry (θ,n), corresponds to the Kolmogorov-Smirnov (K.-S) distance between the cumulative GEV d.f., either fitted through ML (first line) or asymptotically (second line), and the empirical d.f. associated to the 999 simulated values of the statistic G_n^*.

Table 5 is the equivalent of table 4, but for the statistic $\hat{V}_n^*$. We are then trying Normal fits to the d.f. of $\hat{V}_n^*$.

As conclusion, we may say that the lack of fit between the asymptotic $G_\theta((x-\mu_n^*(\theta))/\sigma_n^*(\theta))$ and the ML fitted $G_{\hat{\theta}}((x-\hat{\lambda})/\hat{\delta})$ depends heavily on θ. The asymptotic standard deviation $c(\theta)$ for $\hat{V}_n^*$ is a long way from the ML estimate for $\theta < 0$ and for $n \leq 100$. In what concerns the observed values for the Kolmogorov-Smirnov statistic, we may conclude that the GEV model for the statistic G_n^* seems to be a reasonable approximation, even for small sample size; Normality for the statistic $\hat{V}_n^*$ is not reached at $n = 250$ with the asymptotic parameters $\theta\sqrt{n}/(1+2\theta)$ and $c(\theta)$, but is "nearly" reached for $n = 250$ if we take for critical points of Kolmogorov-Smirnov statistic, the ones provided by Lilliefors, 1967.

Choice in practice between G_n and $\hat{V}_n$ statistics depends heavily on the sample size n. For n smaller than 100, G_n seems to be a good choice, because of its numerical simplicity, its reasonable approximation to the asymptotic distribution, and its competitive power. For large n, $\hat{V}_n$ performs better.

n	20				100				250			
θ	loc.	scale	shape	K.-S.	loc.	scale	shape	K.-S.	loc.	scale	shape	K.-S.
-.2	.81	1.76	-.28	.013	2.51	2.52	-.21	.024	4.01	2.97	-.211	.022
	.83	1.70	-.20	.036	2.44	2.34	-.20	.043	3.88	2.81	-.200	.044
-.1	.34	1.39	-.21	.017	1.08	1.69	-.11	.022	1.67	1.80	-.107	.025
	.38	1.30	-.10	.050	1.53	1.04	-.10	.055	1.60	1.68	-.100	.052
.0	-.05	1.10	-.15	.015	.01	1.15	-.02	.029	.03	1.10	-.008	.032
	.00	1.00	.00	.065	.00	1.00	.00	.070	.00	1.00	.000	.065
.1	-.38	.89	-.11	.015	-.79	.79	.06	.030	-1.13	.69	.076	.027
	-.32	.77	.10	.085	-.79	.65	.10	.084	-1.13	.60	.100	.068
.2	-.66	.72	-.08	.013	-1.41	.57	.11	.021	-1.96	.44	.132	.020
	-.58	.59	.20	.115	-1.38	.43	.20	.111	-1.94	.36	.200	.081

Table 4. ML fitted (first line) and asymptotic (second line) GEV distributions for G_n^*.

n	20			100			250		
θ	loc.	scale	K.-S.	loc.	scale	K.-S.	loc.	scale	K.-S.
-.2	-.61	1.64	.142	-2.85	3.21	.151	-5.03	3.94	.134
	-1.49	6.21	.343	-3.33	6.21	.235	-5.27	6.21	.206
-.1	-.14	1.14	.122	-1.08	1.69	.097	-1.91	1.82	.078
	-.56	1.90	.259	-1.25	1.90	.136	-1.98	1.90	.091
.0	.21	.82	.093	.08	1.02	.063	.03	1.02	.050
	.00	1.00	.199	.00	1.00	.095	.00	1.00	.064
.1	.50	.63	.078	.88	.71	.046	1.34	.69	.037
	.37	.67	.155	.84	.67	.080	1.34	.67	.053
.2	.71	.51	.071	1.46	.55	.035	2.28	.53	.034
	.64	.51	.128	1.43	.51	.067	2.26	.51	.052

Table 5. ML fitted (first line) and asymptotic (second line) Normal distributions for $\hat{V}_n^*$.

APPENDIX: DERIVATION OF ASYMPTOTIC BEHAVIOUR OF G_n AND $\hat{V}_n$.

Asymptotic behaviour of G_n.

We are first interested in the asymptotic sampling behaviour of G_n defined by (1.3). The fact that

$$U_n(\underset{\sim}{Z}) = n^{\theta} \{\text{Maximum}(\underset{\sim}{Z}) - (1-n^{-\theta})/\theta\} \tag{A.1}$$

converges weakly, as $n \to \infty$, to a r.v. with d.f. $G_\theta(.)$, and that

$$V_n(\underset{\sim}{Z}) = 2^{\theta}\sqrt{n} \{ \text{Median}(\underset{\sim}{Z}) - (1-2)^{-\theta})/\theta \} \tag{A.2}$$

converges weakly, as $n \to \infty$, to a standard Normal r.v., implies, for $\theta < \frac{1}{2}$, and by the use of the δ-method [Tiago de Oliveira, 1982], the representation

$$\tilde{G}_n = n^{\theta}(1-2^{-\theta}) \{G_n - (1-n^{-\theta})/(1-2^{-\theta})\}/\theta = U_n(\underset{\sim}{Z}) + o_p(1), \tag{A.3}$$

$o_p(1)$ denoting an infinitesimum in probability, as $n \to \infty$.

Consequently $\tilde{G}_n$ is asymptotically a GEV r.v., with d.f. $G_\theta(.)$, given by (1.2).

Asymptotic behaviour of $\hat{V}_n$.

Let us think now on the statistic $V_n = V_{n,\theta}(\underset{\sim}{Z})$ given by (1.4). Denoting by

$$\mu_{r\theta} = E[Z^r|\theta] = \prod_{j=1}^{r} j/(1+\theta j) \quad \text{if} \quad \theta > -1/r \tag{A.4}$$

we have obviously, for $\theta > -\frac{1}{2}$, the asymptotic reduced Normal behaviour of

$$\sqrt{n}\,\frac{V_{n,\theta}(\underset{\sim}{Z}) - \mu_{V(\theta)}}{\sigma_{V(\theta)}} \tag{A.5}$$

where

$$\mu_{V(\theta)} = \mu_{1\theta} - \mu_{2\theta}/2 \xrightarrow[\theta \to 0]{} 0$$

$$\sigma^2_{V(\theta)} = \mu_{2\theta} + \mu_{4\theta}/4 - \mu_{3\theta} - \mu^2_{V(\theta)} \xrightarrow[\theta \to 0]{} 2$$

We need now to derive the asymptotic behaviour of $\hat{V}_n$ defined by (1.6) either under H_0 or under alternatives $\theta \neq 0$.

The use of the δ-method (Cramer, 1946, Ch. 27; Tiago de Oliveira, 1982) gives asymptotically, for $\theta > -\frac{1}{2}$,

$$E[\hat{V}_n|\theta] = 1 - \mu_{2\theta}/(2\mu^2_{1\theta}) = \theta/(1+2\theta) \tag{A.6}$$

and

$$\text{Var}\,[\hat{V}_n|\theta] = c^2(\theta)/n = \frac{\mu_{1\theta}^2(\mu_{4\theta}-\mu_{2\theta}^2) - 4\mu_{1\theta}\mu_{2\theta}\mu_{3\theta} + 4\mu_{2\theta}^3}{4\mu_{1\theta}^6 n}$$

$$= \left\{ \frac{(1+\theta)^2(6\theta^2+\theta+1)}{(1+2\theta)^3(1+3\theta)(1+4\theta)} \right\}/n \qquad (A.7)$$

i.e,

$$\sqrt{n}\,\{\hat{V}_n - \theta/(1+2\theta)\}/c(\theta)$$

is asymptotically standard Normal.

Notice that we have asymptotically a standard Normal behaviour for

$$\sqrt{n}\,V_n|H_0$$

in contrast with the asymptotic standard Normal behaviour of

$$\sqrt{n/2}\,\hat{V}_n|H_0,$$

i.e., the introduction of $\bar{X}$ (ML estimator of δ) in the LMP statistic decreased its variance.

REFERENCES

- Cramer, H., 1946: Mathematical Methods of Statistics, Princeton University Press, 575 pp.
- Gomes, M.I., 1982: A note on statistical choice of extremal models. Actas IX Jornadas Matematicas Hispano-Lusas, II, Univ. Salamanca, 653-656.
- Gumbel, E.J., 1965: A quick estimation of the parameters in Fréchet's distribution. Revue Instit. Internat. Statist. 33, 349-363.
- Lilliefors, H.W., 1967: On the Kolmogorov-Smirnov tests for normality with mean and variance unknown. J. Amer. Statist. Assoc. 62, 399-402.
- Pickands, J., 1975: Statistical inference using extreme order statistics. Ann. Statist. 3, 119-131.
- Smith, R.L., 1984: Threshold methods for sample extremes. In J. Tiago de Oliveira (ed.), Statistical Extremes and Applications, 621-638, D. Reidel.
- Tiago de Oliveira, J., 1981: Statistical choice of univariate extreme models. C. Taillie et al. (eds.), Statistical Distributions in Scientific Work, Vol. 6, 367-387, D. Reidel.
- Tiago de Oliveira, J., 1982: The δ-method for the obtention of asymptotic distributions; applications. Publ. Inst. Statist. Univ. Paris, XXVII, 49-70.
- Tiago de Oliveira, J. and M.I. Gomes, 1984: Two test statistics for choice of univariate extreme models. In J. Tiago de Oliveira (ed.), Statistical Extremes and Applications. 651-668. D. Reidel.
- Van Montfort, M.A.J. and M.I. Gomes, 1985: Statistical choice of extremal models for complete and censored data. J. Hydrology 77, 77-87.
- Van Montfort, M.A.J. and J.V. Witter, 1985: Testing exponentiality against Generalized Pareto distribution. J. Hydrology 78, 305-315.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

NON-CLASSICAL EXTREME VALUE MODELS

M. Ivette Gomes and Dinis D. Pestana

Dept. Statistics, University of Lisbon
Centro de Estatística e Aplicações (CEAUL)

SUMMARY: The identical distribution hypothesis of classical extreme value theory is questioned, and a broad class M_∞ of limit distributions of normalized maxima of independent random variables is used to provide theoretical grounds for the success of non-extreme value models — such as Pearson III, lognormal, log-Pearson III — most favoured by hydrologists. Robustness results under weak dependence are commented upon. On the other hand, a limit generalized extreme value distribution still arises in case of periodic disturbances, under very mild assumptions.

1. A DISCUSSION OF EXTREME VALUE THEORY.

Classical extreme value theory deals with suitably normalized maxima (or minima) of i.i.d. random variables $\{X_k\}_{k\in N}$. If there exist appropriate attraction coefficients $a_n>0$ and $b_n\in R$ such that

$$[M_n-b_n] \,/\, a_n \;\to\; Y_\theta, \tag{1.1}$$

where Y_θ denotes a non-degenerate random variable, $M_n=\max(X_1,\ldots,X_n)$, then Y_θ is max-stable in the sense that the corresponding distribution function G_θ is a solution of Fréchet's functional equation (Fréchet stability postulate); in von Mises-Jenkinson form, or generalized extreme value (GEV) form,

$$G_\theta(x;\lambda,\delta) = \exp\left[-\left(1-\theta(x-\lambda)\,/\,\delta\right)^{1/\theta}\right] \tag{1.2}$$

(support $\{x:\ 1-\theta(x-\lambda)/\delta\geq 0\}$), $\theta,\lambda\in R$, $\delta>0$. The index θ depends on the upper tailweight of the parent distribution F.

Sometimes it is convenient to distinguish Gumbel type ($\theta=0$), Fréchet family of types ($\theta<0$) and Weibull family of types ($\theta>0$), for mathematical tractability reasons. For details cf. Gumbel (1958), Bury (1975), Galambos (1978) and Tiago de Oliveira (1984).

Under weak dependence hypothesis, Fréchet stability postulate still holds, and limit distributions are of the form (1.2); for details, cf Galambos (1978, ch. 3). Observe however that under more general dependence assumptions other limits do occur (Welsch,1972).

Several multivariate generalizations, dealing either with maxima of random vectors or with top i-order statistics (Gomes, 1981, 1984a) extend nicely univariate results, and provide powerful techniques for parameter estimation, e.g. BLUEs and concomitants of order statistics.

In the multivariate GEV model we assume to be working with a multivariate GEV vector (top observations available), with joint probability density function

$$\beta^{-m}\, h_\theta[(z_1-\gamma)/\beta,\ldots,(z_m-\gamma)/\beta],\ \theta,\gamma\in R,\ \beta>0, \tag{1.3}$$

where

$$h_\theta(x_1,\ldots,x_m) = g_\theta(x_m)\prod_{j=1}^{m-1}\left[g_\theta(x_j)/G_\theta(x_j)\right],\ x_1>x_2>\ldots>x_m, \tag{1.4}$$

with G_θ given by (1.2) in standard form, $g_\theta=G_\theta'$, is the density function corresponding to the limiting distribution of the top m order statistics suitably normalized in an i.i.d. scheme (Pickands, 1975, Weissman, 1978, Gomes, 1981).

The multidimensional GEV model (Gomes, 1981) is a compromise between Gumbel classical approach to statistical extremes and the multivariate GEV model — independent multivariate GEV vectors $\underset{\sim}{Y}_j$ of size i_j, $1\leq j\leq N$, are considered.

Statistical choice, in the sense of decision on the characteristic index θ, is easily made using simple statistics such as the balance of the upper tail to the lower tail, or properties of spacings (cf. van Montfort and Gomes, 1985, and references therein).

Statistical extremes have been used with moderate success in the study of geophysical phenomena, such as floods and droughts, heavy rains, windspeed, insolation, air temperature, magnetic field, waves and tides, etc. The main objection to the use of extreme value distributions is that recorded data hardly can be accepted as i.i.d. samples, and in many instances non-extreme value distributions provide much better fit to data than the GEV distribution (1.2) – cf. Yevjevich (1984), Pericchi and Rodríguez-Iturbe (1985). On the other hand, convergence in (1.1) can be rather slow and give rise to penultimate phenomena (Gomes, 1984b, Gomes and Pestana, 1985 a).

Pearson III (gamma), lognormal and log-Pearson III models (the last one recomended by U.S. Federal agencies) seem to be the favourite alternatives. For a discussion on parameter estimation cf. Pericchi and Rodríguez-Iturbe (1985).

The main reason for the relative insuccess of classical extreme value distributions in hydrological and geophysical data modelling is failure of the identical distribution hypothesis. As Pericchi and Rodríguez-Iturbe (1985) comment

> "The vital assumption is, in our opinion, one that has seldom been questioned as critical in the hydrological literature, namely the assumption that the parent distribution is made up of identically distributed random variables."

Normed maxima of independent, but not necessarily identically distributed random variables with an uniformity assumption have been characterized by Mejzler (1956): the corresponding limit distribution G has an associated M-function of the form

$$M(x) = \exp\,[-A(x)] \tag{1.5}$$

where $A(x)$, $x>0$, is a convex function, and

$$M(x) = \begin{cases} G(\omega - e^{-x}) & \text{if } \omega<\infty \\ G(x) & \text{otherwise} \end{cases} \tag{1.6}$$

($\omega=\sup\{x: G(x)<1\}$ is the right endpoint of the limit distribution).

Observe that Pearson III, lognormal and log-Pearson III distributions are indeed in Mejzler's class.

Mejzler's class didn't have the impact it deserves in statistical literature, due perhaps to the fact that general statistical techniques (namely density estimation) aren't developed. Recently Graça Martins and Pestana (1985a, 1985b) and Gomes and Pestana (1985b) refined Mejzler's M class, in particular characterizing M_∞ class of distribution functions G with associated M-function of the form

$$M(x) = \exp\,[-K(x)] \tag{1.7}$$

with $K(x)$, $x>0$, completely monotone (i.e., $K\in C^\infty$ and $(-1)^n K^{(n)}(x)\geq 0$, $n\in\mathbb{N}$, or, in other words, K the Laplace transform of a non-decreasing function). Observe that M_∞ is the smallest class closed under pointwise products and passage to limits that contains all max-stable distribution functions (1.2).

Observe however that in many practical situations the number of types or domains of attraction involved in the distribution function Φ_n of $\max(X_1,\ldots,X_n)$

$$\Phi_n(x) = \prod_{k=1}^{n} F_k(x) \tag{1.8}$$

(where $F_k(x)=P[X_k\leq x]$) is small, and we get a narrower class of limit distributions (this is related to Gnedenko's conjecture, cf. Graça Martins and Pestana, 1985a). Moreover, heavy-tailed factors in (1.8) play in general a prominent role, and in such circumstances the limit distribution of (1.8), with suitable norming constants, is still a GEV distribution (1.2).

Multivariate analogues of the extremal GEV process in (1.4), with G_θ and g_θ replaced by the limiting G and its derivative in (1.7) or (1.5) are easy to develop (Athayde et al, 1986).

2. AN EMPIRICAL APRAISAL.

Using the statistical package under development at CEAUL, Lisbon (Gomes and Pestana, 1986) rainfall and temperature daily data from the years 1958-1978 and 15 Portuguese meteorological statisons have been analysed, for the apraisal of the merits of

Gumbel model
GEV model
Lognormal model
Log-Pearson III model

with parameters estimated by maximum likelihood methods whenever possible, method of moments in all other cases.

The assessment was made in terms of Kolmogoroff-Smirnov, Cramer-von Mises and U_n^2-Stephens observed statistics.

In most cases any of the models considered fits reasonably well the data available; results cannot of course be regarded as definitive, due to the scarceness of recorded data, but this is the case for most Countries.

We found no evidence that either log-Pearson III or lognormal provide a better fit than the GEV model, for these types of data. Hence, due to the availability of simple and standard methods of parameter estimation for the GEV model (there is still some controversy in what concerns parameter estimation in the log-Pearson III model, and in particular on indications of U. S. Water Resources Council), we would favour GEV model.

We made no attempt of fitting the general form (1.7) to data, even in the more restricted assumption $\omega=\infty$, and hence $G(x) = \exp[-K(x)]$, K completely monotone, since appropriate density estimation techniques are far from developed.

3. ROBUSTNESS RESULTS.

As in the classical extreme value case, it is expected that under weak dependence hypotheses the limiting distribution of $\max(X_1,\ldots,X_n)$, suitably normalized, is still a member of Mejzler's M class, or of the narrower class M_∞.

This is indeed the case with spacings, under very general assumptions on the parent distribution (Deheuvels, 1986; Athayde, 1986; Athayde et al, 1986), and when periodic disturbances are considered.

Effects of seasonality can be most relevant for meteorological data (Buishand, 1983), and under quite mild hypotheses the limit distribution G is a member of M_∞; more than that, G is a GEV distribution.

Indeed, when thinking on periodic disturbances, we may formulate the following simplified model (Tiago de Oliveira (1976)): the sequence available $\{Y_k\}_{k\in N}$ may be expressed as $Y_k = X_k+\rho_k$, $k\in N$, where $\{X_k\}_{k\in N}$ is a sequence of i.i.d. r.v.'s with d.f. F attracted to G_θ in (1.2), and with attraction coefficients $\{a_n\}_{n\geq 1}$ $(a_n>0)$ and $\{b_n\}_{n\geq 1}$, and $\{\rho_k\}_{k\in N}$ is a periodic sequence of real constants with period P.

If we further assume that for every $h\in R$

$$(1-F(x+h))/(1-F(x)) \xrightarrow[x\, x_0^F]{} A(h),\quad 0<A(h)<1\ ,\quad x_0^F=\sup\{x:F(x)<1\} \tag{3.1}$$

then, with $M_n^Y = \max(Y_1,\ldots,Y_n)$,

$$P[M_n^Y \leq a_n y+b_n] \xrightarrow[n\to\infty]{} H_\theta(x) = \{G_\theta(x)\}^{\sum_{i=1}^{P} A^{-1}(\rho_i)/P}\ , \tag{3.2}$$

which is of the same type of $G_\theta(x)$.

More generally, if we think on m levels $\{u_n^{(i)}\}_{n\geq 1}$, $1\leq i\leq m$, so that $n(1-F(u_n^{(i)}))\to\tau_i$, as $n\to\infty$, it is still possible to derive a limiting Poisson character for the exceedances $S_{n,i} = \#\{Y_k,\ 1\leq k\leq n:\ Y_k>u_n^{(i)}\}$, $1\leq i\leq m$. Indeed

$$(S_{n,1},\ldots,S_{n,m}) \xrightarrow[n\to\infty]{w} (Z_1, Z_1+Z_2,\ldots, Z_1+Z_2+\ldots+Z_m)$$

where the Z_j are independent Poisson random variables with mean value $c(\tau_j-\tau_{j-1})$, $1\leq j\leq m$, $\tau_o=0$, $c=p^{-1}\sum_{i=1}^{p} A^{-1}(\phi_i)$.

The same multivariate limiting GEV(θ) structure in (1.6) is thus achieved for the top m order statistics in this scheme with periodic disturbances.

This is an argument more in favour of the GEV model. There is indeed strong evidence that the GEV model is inadequate in some situations, and M_∞ provides a broad class of alternatives. But until general statistical techniques are developed for this class (and namely statistical choice and density estimation techniques) it seems unwise to put the GEV model aside.

Aknowledgements:INIC (National Institute for Scientific Research, Lisbon) supported D. D. Pestana participation at the III International Conference on Statistical Climatology.

REFERENCES

Athayde, M. E. (1986). Note on extreme spacings when sampling from exponential distributions. CEAUL report No. 23/86.

Athayde, M. E., Gomes, M. I. and Pestana, D. D. (1986). Multivariate extremal processes under non-classical situations. VI Pannonian Symposium on Mathematical Statistics, Bad Tatsmandorff.

Buishand, T. A. (1983). The effect of seasonal variation and serial correlation on the extreme value distribution of rainfall data. in Proc. II International Conf. Statistical Climatology, 10.4.1-10.4.6.

Bury, K. V. (1975). Statistical Models in Applied Science. Wiley, New York.

Deheuvels, P. (1986). On the influence of the extremes of an i.i.d. sequence on the maximal spacings. Ann. Probability 14, 194-208.

Galambos, J. (1978). The Asymptotic Theory of Extreme Order Statistics. Wiley, New York.

Gomes, M. I. (1981). An i-dimensional limiting distribution function o- largest values and its relevance to the statistical theory of extremes. in C. Taillie et al, eds., Statistical Distributions in Scientific Work, vol. 6, 389-410.

Gomes, M. I. (1984a). Statistical Theory of extremes - comparison of two approaches. Statistics & Decision, special issue 3.

Gomes, M. I. (1984b). Penultimate limiting forms in extreme value theory. Ann. Inst. Statist. Math. 36A, 71-85.

Gomes, M. I. and Pestana, D. D. (1985a). Non-standard domains of attraction and rates of convergence. in M. Puri et al, eds., New Perspectives in Theoretical and Applied Statistics, Wiley, New York (in press).

Gomes, M. I. and Pestana, D. D. (1985b). Large claims - extreme value models. in M. Goovaerts et al, eds., Insurance and Risk Theory (in press).

Gomes, M. I. and Pestana, D. D. (1986). The extremal limit problem in data analysis - building a statistical package. in Proc, VIII International Symposium "Computer at the University", Dubrovnik, 5.02.1-5.02.9.

Graça Martins, E. and Pestana, D. D. (1985a). The extremal limit problem - extensions. in G. Mogyoródi and Wertz, eds., Proc. V Pannonian Symposium on Mathematical Statistics, North Holland (in press).

Graça Martins, E. and Pestana, D. D. (1985b). Non-stable limit laws in extreme value theory. in M. Puri et al, eds., New Perspectives in Theoretical and Applied Statistics, Wiley, New York (in press).

Gumbel, E. J. (1958). Statistics of Extremes. Columbia University Press, New York.

Mejzler, D. G. (1956). On the problem of the limit distribution for the maximal term of a variational series. L'vov Politechn. Inst. Naucn. Zp. (Fiz.-Mat.) 38, 90-109 (in Russian).

Pericchi, L. R. and Rodríguez-Iturbe, I. (1985). On the statistical analysis of floods. in A. C. Atkinson and S. E. Fienberg, eds., A Celebration of Statistics, Springer-Verlag, 511-541.

Pickands III, J. (1975). Statistical inference using extreme order statistics. Ann. Statistics 3, 119-131.

Tiago de Oliveira, J. (1976). Asymptotic behaviour of maxima with periodic disturbances. Ann. Inst. Statist. Math. 28, 19-23.

Tiago de Oliveira, J., ed. (1984). Statistical Extremes and Applications. D. Reidel, Dordrecht.

van Montfort, M. A. J. and Gomes, M. I. (1985). Statistical choice of models for complete and censored samples. J. Hydrology 77, 77-87.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. J. Amer. Statist. Assoc. 73, 812-815.

Welsch, R. E. (1972). Limit laws for extreme order statistics from strong mixing processes. Ann. Math. Statist. 43, 439-446.

Yevjevich, V. (1984). Extremes in hydrology. in J. Tiago de Oliveira, ed., Statistical Extremes and Applications, D. Reidel, Dordrecht, 197-220.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE STATISTICAL METHOD OF THE HEAVY RAINFALLS STUDY IN THE POLISH CARPATHIAN MTS

Tadeusz Niedźwiedź
Institute of Meteorology and Water Management
Kraków, Poland

One of the most important tasks in analysing heavy rains is to determine the relationship between the intensity or efficiency of downpours and the rainfall's duration of various probability of occurence. On preparing maps of heavy rains to the "Hydrological Atlas of Poland" a method of statistical analysis of that phenomenon was worked out at the Institute of Meteorology and Water Management. The methods of gathering data on heavy rains used hitherto are not adapted to a comuter analysis. Hence a proposal presented below, as based on the instruction of the World Meteorological Organization /Guide... 1981, Manual... 1969/ and on a paper by G.A. Alexeyev /1963/, permits a partial automatization of the analysis of data. In Poland the heavy rains /rain-storms/ are classified according the scale of K. Chomicz /1951/, when they are exceeding a lower limit of rainfall amount U_0:

$$U_0 = t^{0.5} \qquad /1/$$

where: t - duration of rainfall in minutes,
U_0 - amount of rainfall in mm.

A driving rains /violent rains/ are exceeding a lower limit of rainfall amount U_5:

$$U_5 = 5.66\ t^{0.5} \qquad /2/$$

The first, most laborius step consists in reading from pluviograms the duration values of a heavy rain as well as a total precipitation for that spell. Besides, within every downpour the steepest sectors of curves are chosen to assess in succession the greatest efficiency of rains for duration of 5, 10, 15, 30, 60 and 120 minutes. The matter set in adequate working tables is then transferred on perforated cards of 80 columns according to the following format:

columns:	elements:
1 - 2	day
3 - 4	month
5 - 8	year
9 - 10	initial hour of downpour /Central European Time/

11 - 12	initial minutes of downpour
	Maximum total of precipitation for time intervals:
13 - 16	5 minutes
17 - 20	10 minutes
21 - 24	15 minutes
25 - 28	30 minutes
29 - 32	60 minutes
33 - 36	120 minutes
37 - 40	total precipitation of downpour
41 - 44	duration of downpour in minutes
45 - 48	daily precipitation total according to pluviograph
49 - 52	daily precipitation duration in minutes
53 - 56	daily precipitation total according to raingauge
57 - 80	name of station and commentaries.

Precipitation totals are given with an accuracy to 0.1 mm, the comma being omitted. The lacking values are shown by putting a negative value /e.g. -1/. Series of data, after being checked are written down in the size presented on a magnetic tape as a base set of data. The data relating to daily precipitation totals both according to raingauge and to pluviograph are being recorded for purposes of comparison and of control.

The next step is a statistical analysis of data, this carried out with the help of the program HEAVY RAINS, written in the FORTRAN language on a digital computer of the system CDC CYBER-72 in the Kraków Computer Center "CYFRONET". The program analyses the logical correctness of data, supplies if possible missing data, omits precipitation of base values below U_0 on the Chomicz /1951/ scale. The minimum length of record was 10 years.

For all values of rainfall amount R_t /for each time interval t/ the data are being set in decreasing order and their empirical probability /frequency/ in the partial duration series /ploting position/ were calculated according to the formula:

$$p' = \frac{m - 0.3}{n + 0.4} \qquad /3/$$

where: n - total number of items,
m - order number of the values R_t arranged in descending magnitude.

Because of short 10-years period of observations during further calculations of the annual probability, all events of heavy rainfalls during each year, were taken into account. Formula for such type of data was prepared by Strupczewski /1967/:

$$p = 1 - e^{-\lambda p'} \qquad /4/$$

where: p - annual probability /0 - 1/,

$\lambda = \frac{n}{N}$ - mean number of events in year,

n - total number of events in data set,
N - number of the years,
p'- probability of an event in partial series, according to the formula /3/.

Empirical sequences of the orderly precipitation totals, together with relevant values of annual probability, were recorded on a magnetic tape and they serve for further calculations. The third step includes an adjustment of empirical data to a modified formula by Alexeyev /1963/:

$$R_{p,t} = \frac{A + B \ln\left[-\ln(1 - p)\right]}{t^n} \qquad /5/$$

where: $R_{p,t}$ - maximum total of heavy rainfall in mm of annual probability p and duration t,
A, B, n - coefficients

The estimation of coefficients A, B and n in the formula /5/ was performed by minimizing totals of squares of differences between experimental data and the cited theoretical formula describing these data. For purposes of calculations the program MINUIT from the library CERN was used /James, Ross 1976/. Applied at that was the Monte Carlo method as well as a method of minimum gradient after receiving the results of first approximation. For Zakopane over 1956 - 1975 the following values of coefficients were obtained: A = 3.33, B = -1.01, n = -0.40. Coefficient of rainfall intensity A /intensity of 1 minute duration rain with probability p = 0.63/ is changing in the Carpathian Mts from 2.3 to 7.1. Parameter B, which depends from probability, varies from -0.98 to -3.17. Exponent n, which expressed dependence of rainfall totals from their duration, changes from -0.24 to -0.46.

In the fourth step a program was set for graphical presentation of data according to the Chomicz /1951/ scale. Use was made at that of the library of program for plotters CENPLOT. After inserting the three coefficients into the computer the program is making calculations of the efficiency of heavy rains with an interval every 1 min, which is enough to drow round curves, as shovn in Figure 1 for Kraków and Figure 2 for Zakopane.

The obtainment of homogeneous and comparable data concerning heavy rainfalls for 28 main stations in Polish Carpathian Mts was made possible to show maps that phenomenon. That work has been done for this part of Carpathian Mts first time. The isohyets were created according the Slovak data, too /Samaj, Valovic 1973/. There were constructed 12 maps. As an example there were presented maps for the duration 15 minutes /Figure 3/ and 60 minutes /Figure 4/ simultaneously with the annual probability 10 % /recurence interval once in 10 years/. The investigation showed there is any connections of heavy rainfalls with the altitude and annual totals of precipitations.

REFERENCES

Alekseyev, G.A., 1963: Metodika rascota maximalnych dozdevych raschodov vody po krivym redukcii osadkov /The method of calculation of the maximum rainfall water runoff after the reduction curves of precipitations/. Trudy GGI, 107.

Chomicz, K., 1951: Ulewy i deszcze nawalne w Polsce /Rain-storms and driving rains in Poland/. Wiad. Służby Hydr. Met., 2,3.

Guide to hydrological practices, 1981: WMO No.168. TP.82, Geneva.

James, F., and Ross, M., 1976: MINUIT. A system for function minimization and analysis of the parameter errors and correlations. CERN Computer Centre Program Library, Long-write-up, D506, D516.

Manual for depth-area-duration analysis of storm precipitation, 1969: WMO No. 237, TP.129., Geneva.

Samaj F., Valovic, S., 1973: Intensity kratkodobych dazdov na Slovensku /The intensity of short-duration rains in Slovakia/. Zb. Prac Hydromet. Ustavu v Bratislave.

Strupczewski, W., 1967: Wyznaczanie prawdopodobieństwa rocznego niektórych zjawisk na podstawie wszystkich ich wystąpień /Determination of the probability annual distribution of some events on basis of all their occurences/. Acta Geophysica Polonica, 15, 3.

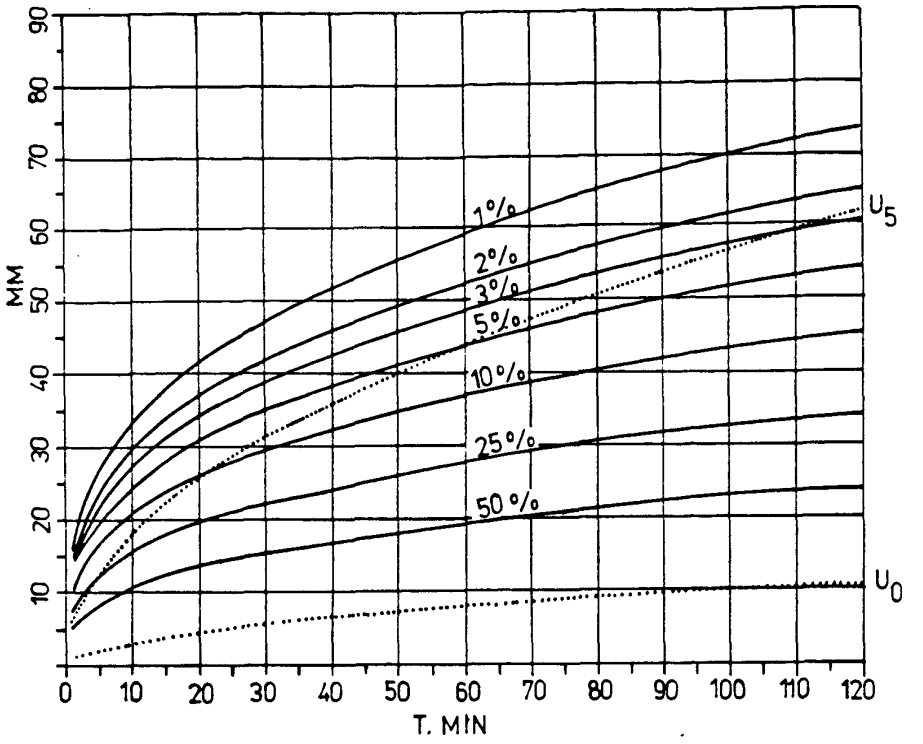


Figure 1. Maximal amount of heavy rainfall of different duration /t/ and probability of occurence /p/ in Kraków /1906-1975/

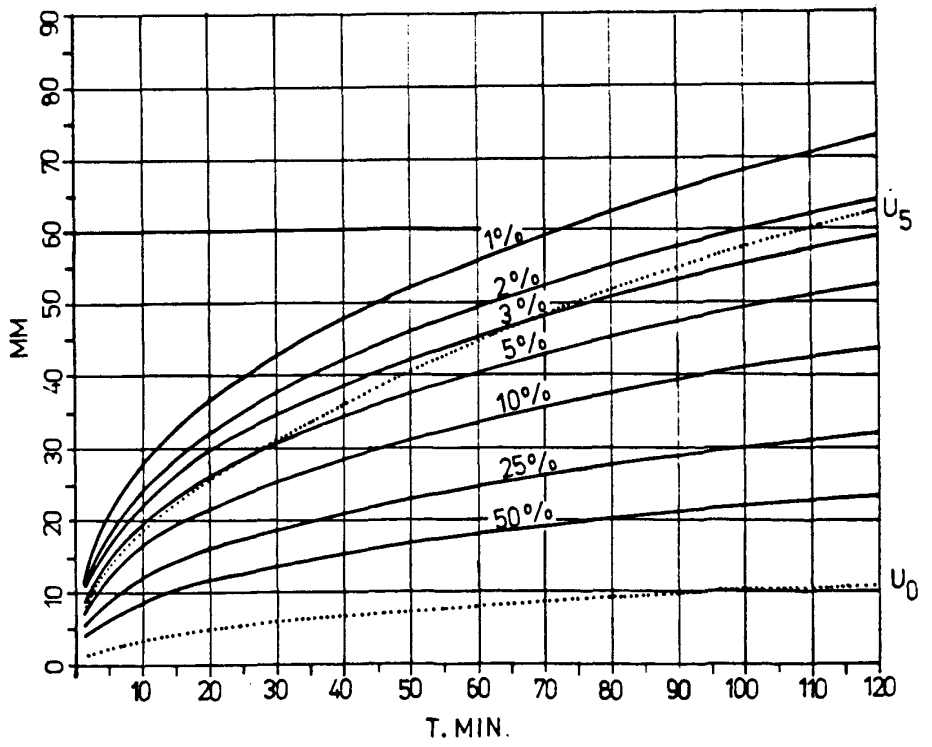


Figure 2. Maximal amount of heavy rainfall of different duration /t/ and probability of occurence /p/ in Zakopane /1930-31, 1933-1979/

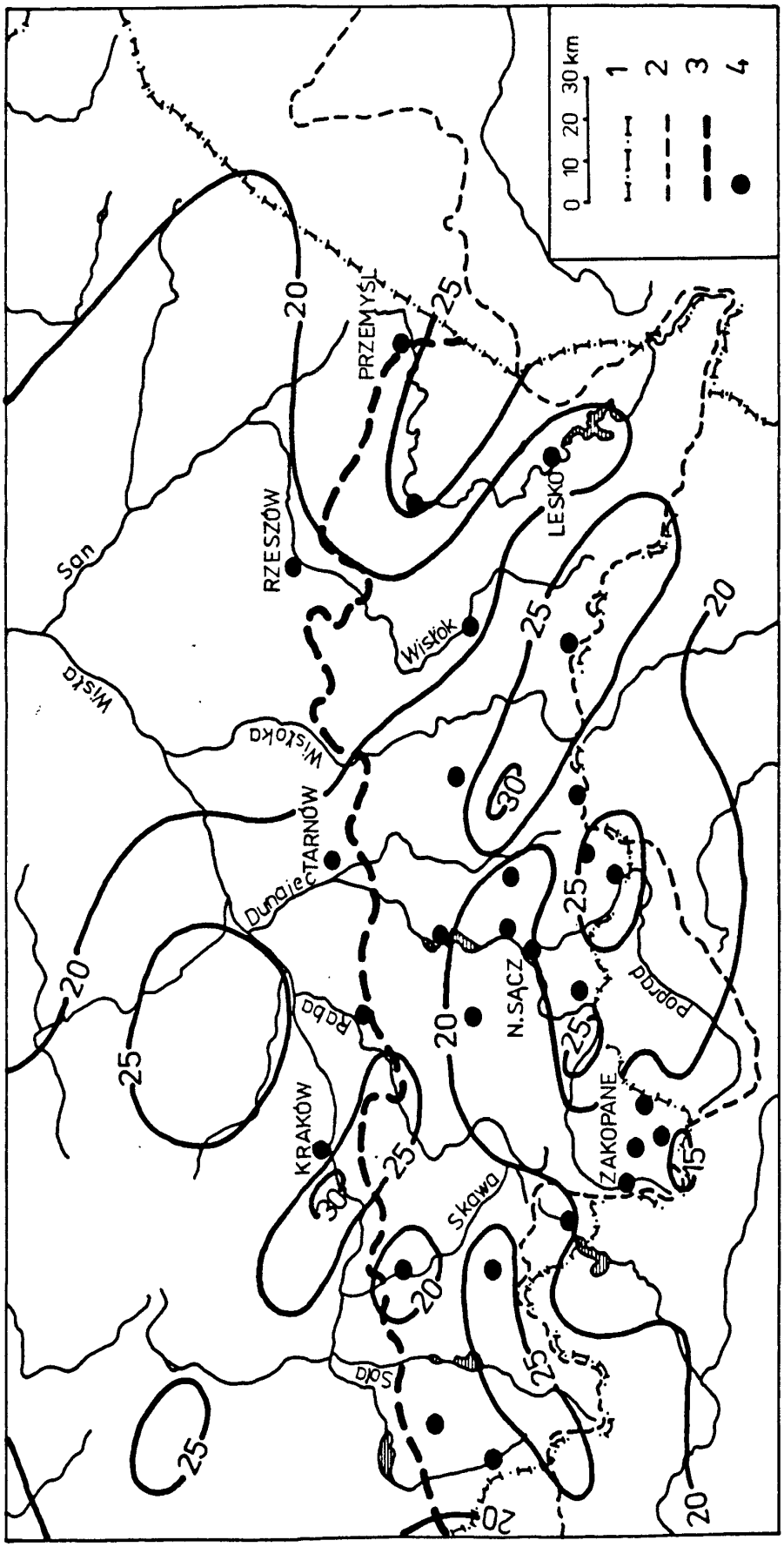


Figure 3. Isohyets of maximal amount of 15 minutes rainfalls in Polish Carpathian Mts appearing once in 10 years /p = 10 %/. 1 - state boundary, 2 - the European watershed, 3 - the northern edge of Carpathian Mts, 4 - pluviographic stations

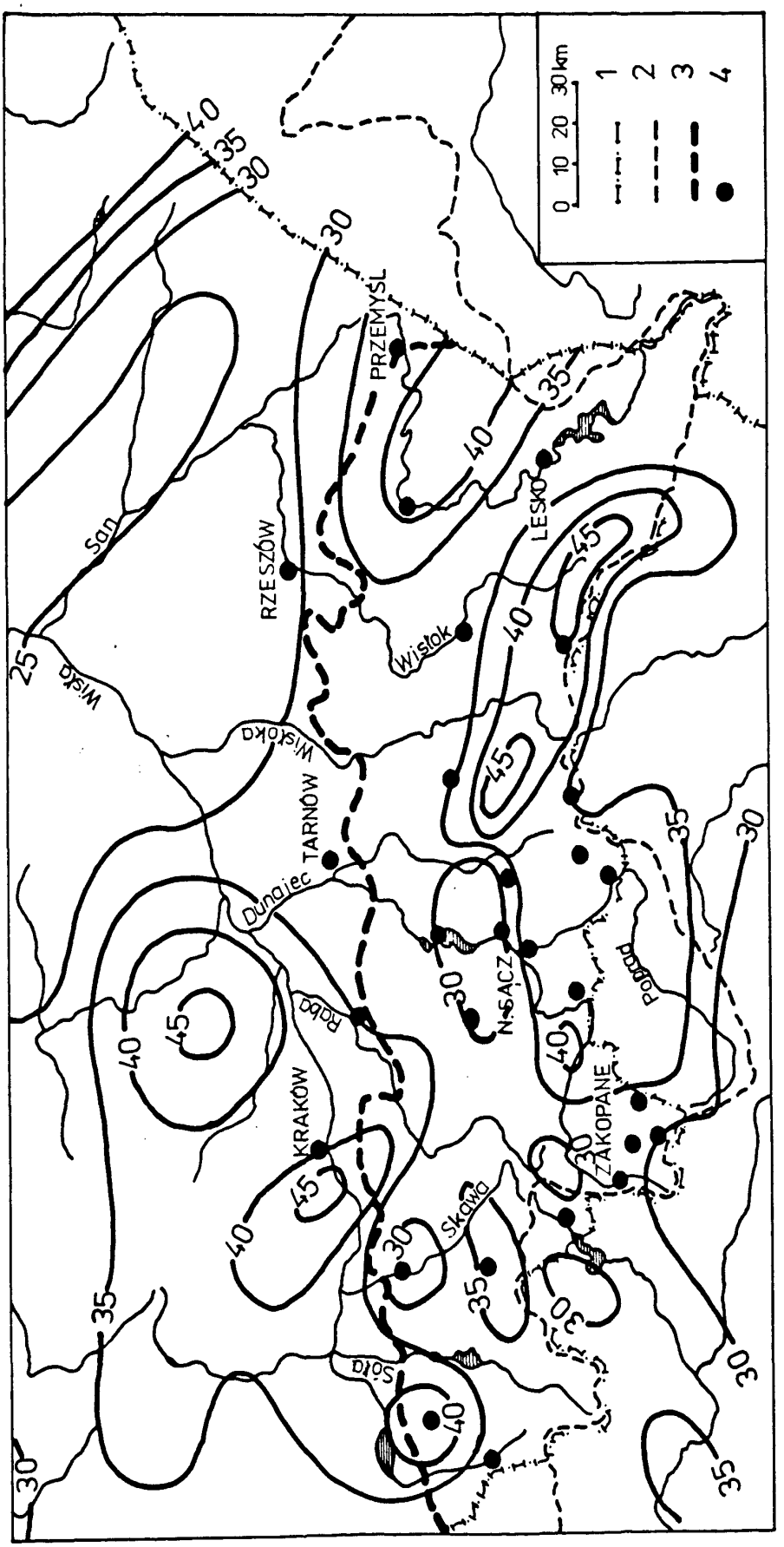


Figure 4. Isohyets of maximal amount of 60 minutes rainfalls in Polish Carpathian Mts appearing once in 10 years /p = 10 %/. 1 - state boundary, 2 - the European watershed, 3 - the northern edge of Carpathian Mts, 4 - pluviographic stations

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

ANALYSIS OF THE MAXIMUM PRECIPITATION DISTRIBUTION

Victor Oancea, Constanţa Boroneanţ
Institute of Meteorology and Hydrology
Bucharest - Romania

The paper presents the outcomes of the analysis performed on the type of maximum rainfall distribution curves over various time intervals in Romania and the way the maximum precipitation amounts are determined with certain probabilities of occurrence.

1. DETERMINATION OF THE THEORETICAL DISTRIBUTION FUNCTION

The major objective of determining the type of theoretical distribution function lies in the assessment of events with recurrence periods longer than the available sample.

The parameters of the distribution curve are generally determined on account of the data sample being analysed. These parameters together with the selected distribution curve help estimating the events of the desired probabilities.

So far, there is no general agreement on the type of distribution function for the maximum precipitations fallen at different time intervals (Sevruk, 1981). There is also no universally accepted method concerning the selection of a distribution function and the concordance statistical tests used to this purpose (Kite, 1977).

In a statistical analysis carried on for the determination of the maximum precipitations the option was for the comparison of the empirical distribution with the theoretical distributions, i.e. normal (Gauss), log-normal, three-parameter gamma, type-1 extreme (Gumbel) and exponential (Oancea, 1982).

The selection of a theoretical distribution function best describing the sample being analysed was based on the statistical tests of goodness of fit:

a) χ - square:

$$\chi^2 = \sum_{j=1}^{k} \frac{(Q_j - E_j)^2}{E_j}$$

where, Q_j is the number of events belonging to the class j;
E_j - the number of events forecast by the theoretical distribution;
k - the number of classes.

b) Smirnov - Kolmogorov:

The χ^2- square test, grouping the data into classes, misses a significant amount of information. This is avoided in Smirnov - Kolmogorov's test statistics:

$$Dn = \max/P(xj) - Po(xj)/$$

where, P(xj) is the probability of the j event belonging to the sample under study, and Po(xj) is the probability of the j event determined on account of the theoretical distribution

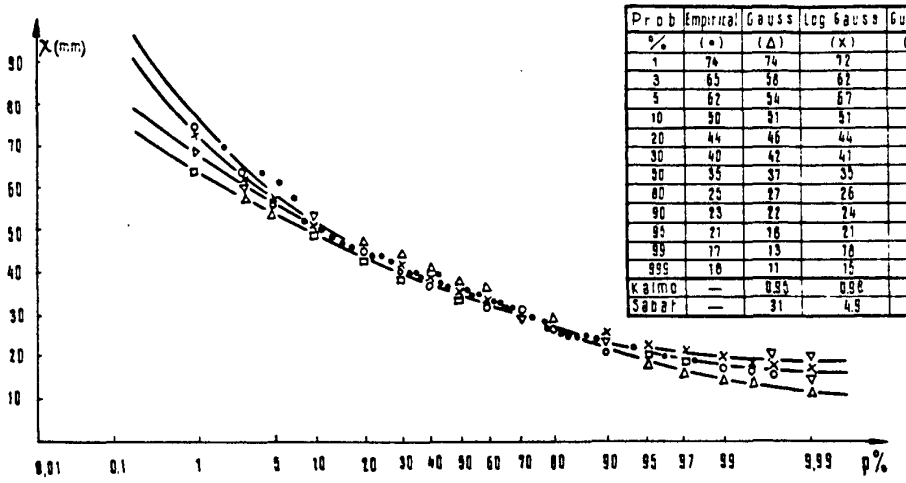
$$SABAT = \left[\frac{\sum_{i=1}^{N} (xi - yi)^2}{n - mi} \right]$$

where, xi are the sample values
yi are the values computed on account of the theoretical distribution

For all the goodness of fit tests: χ - square and Smirnov - Kolmogorov, the accepted confidence level was 95 % and 90 %.

The priority of the significance tests lies in:

- the selection of the theoretical distribution curves in terms of Smirnov - Kolmogorov and χ - square tests;
- the selection of those theoretical distributions that had the minimum value of the sum of square of the derivations from all the theoretical distributions that met the first criterion.



Prob	Empirical	Gauss	Log Gauss	Gumbel	Gama	Exponent
%	(•)	(Δ)	(×)	(○)	(□)	(▽)
1	74	74	72	63	63	60
3	65	58	62	63	60	61
5	62	54	57	58	56	57
10	50	51	51	51	50	52
20	44	46	44	44	43	44
30	40	42	41	40	38	40
50	35	37	35	35	34	35
80	25	27	26	25	26	26
90	23	22	24	23	21	23
95	21	18	21	22	20	21
99	17	13	18	17	17	19
999	16	11	15	14	16	16
Kalmo	—	0.95	0.98	0.91	0.93	0.98
Sabat	—	31	4.9	46	9.1	6.2

FIG. 1 Comparison of the distribution curves for maximum precipitations over 24 hours

Figure 1 shows comparatively the theoretical distribution functions as against the empirical distribution curve of the maximum precipitations within 24 hours recorded at one of the stations being analysed. The appended table provides the values of maximum precipitations over 24 - hour - intervals with various probabilities of occurrence and of the level of confidence for the Smirnov - Kolmogorov (KOLMO) test as well the sums of square derivations (SABAT).

The determination of the theoretical distribution function parameters was based on the samples analysed (the XMED average, the variation coefficient CV, the ratio between the variation coefficient and the asymmetry coefficient CS/CV, where the length of the series allowed it).

The statistical parameters thus determined were analysed from the standpoint of their space distribution over Romania's territory. Thus homogenous zones from the standpoint of maximum precipitation formation could be identified. These zones were also confirmed by the fact that the same distribution functions were obtained at the stations situated in the same zones.

The point analysis of the distribution functions for maximum precipitations was completed with the space-time analysis of the maximum precipitations fallen at various time intervals over the previously determined homogeneous zones (Diaconu, 1984).

At the basis of this method lies the theorem of ergodicity stating that the statistical time characteristics of a random quantity can be determined through simultaneous space measurements. The reciprocal also holds true (Kalinin, 1967).

The practical check-up of the conditions for the application of the ergodicity theorem is difficult and therefore a weaker theorem is resorted to, i.e. the conditioned ergodicity.

Figure 2 shows the space-time curve for southern Romania. The space curve was drawn on account of the simultaneous precipitation values recorded at 80 meteorological and raingauging stations situated in one of the previously established homogeneous zones, in south Romania. This series of values was based on the maximum precipitation amounts over 24 hours recorded in 1970, a year of maximum precipitation status in Romania.

The time curves were drawn for 5 stations in southern Romania with longer series than 75 years available. The results of this analysis are presented in figure two illustrating the good accordance between the temporal and spacial curves. The space-time curve was used in checking up the extrapolations at those stations having very short data series (2o - 3o years) available. Likewise, it may be considered as an average curve for the area being analysed.

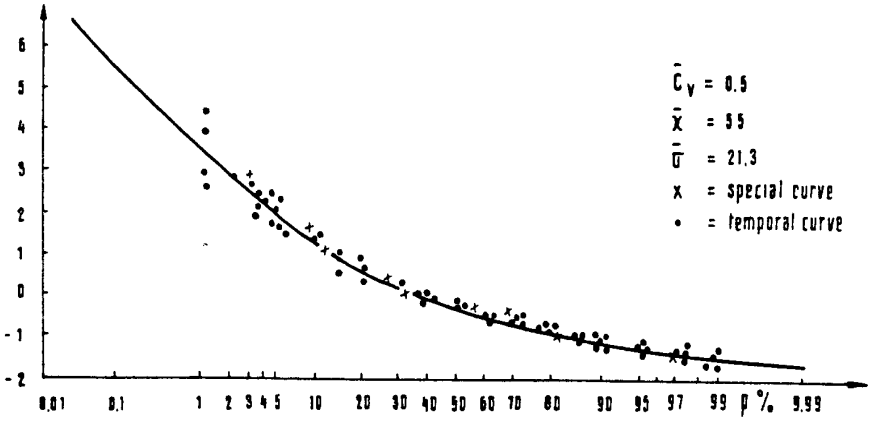


FIG. 2 Spacio - temporal curve of the maximum precipitations in southern Romania

2. AVAILABLE DATA

The paper is devoted to the study of maximum precipitations series fallen over varying time intervals at a set of 1oo meteorological stations, evenly distributed over the country's territory.

At each station, series of maximum annual precipitations fallen over intervals of 5; 1o; 15; 3o; 45; 6o; 9o; 18o; 36o; 72o;144o minutes and 1; 2; 3; 4; 5; 6; 7 days respectively, were established.

The values of maximum precipitations within 5 to 144o minutes were provided by measurements in the raingauging stations carried out on series of 3o - 98 years. The very large amount of data called forth the organization of two data files on magnetic tape corresponding to the primary source. The complex statistical analysis and the large amount of data, required the use of a specialized programme in FORTRAN - IV language compatible with the FELIX C - 1o24 Computer.

3. OUTCOMES OF THE STUDY

The point analysis of the distribution function at each of the 1oo stations over each of the rainfall durations mentioned before enabled the choice of some theoretical distribution functions for different specific durations on the areas being analysed.

The distribution functions over a certain rainfall duration were studied from the standpoint of their agreement in space distributions for the homogeneous zones of rainfall formation previously established. The outcomes of this analysis largely confirmed the zoning. In general, the type of distribution function varied both with the zone and with the maximum precipitation duration. However, consequence to the study of the entire territory the log-normal, Gumbel and 3 - parameter - gamma distribution functions proved best fit for the maximum precipitations fallen within 1 and 7 days and Gumbel exponential and 3 - parameter gamma, respectively, for precipitations fallen within intervals of 5 to 144o minutes.

The necessity of obtaining homogeneous values for the entire Romanian territory as well as the necessity of maintaining the good agreement between the extrapolated values over various time durations, called forth the choice of a unique theoretical distribution function for the whole territory. As Gumbel's distribution yielded satisfactory results to the goodness of fit tests imposed on all time intervals analysed, it was selec-

ted for a distribution function to work with. Using this function the maximum precipitations were computed with a 1 % and o.1 % excess probability for all stations being analysed and for all precipitations times.

The extrapolated values allowed of a new zoning of the country's territory function of the maximum rainfall depth. Checking up the consistency of the extrapolated values considered the space-time analysis in the previously established zones. The results yielded by the space-time analysis were in good agreement with the ones obtained in terms of statistical point analysis (Fig. 2).

4. CONCLUSIONS

The requirements of economic and safe design of various buildings for economic purposes as well as the studies for risk analysis need a through knowledge of the maximum precipitation amounts with low probabilities of occurrence (1 %, o.5 %, o.1 %). The analysis carried out in Romania to this purpose took into account the data provided by loo meteorological stations using maximum precipitation series along 18 time periods : from 5 to 144o minutes and from 1 day to 7 days which enabled the selection of Gumbel's distribution function as satisfactory from the standpoint of statistical agreement criteria χ^2, Smirnov - Kolmogorov and the deviation square sum.

This distribution function helped determining the maximum precipitation amounts of 1% and o.1% probabilities and led to the identification of 5 homogeneous zones from the viewpoint of maximum precipitation formation in Romania. The results of the point analysis were confirmed by the spatial-temporal analysis of maximum precipitations carried out in the homogeneous areas previously determined.

B I B L I O G R A P H Y

BRUNET-MORET (1969)

Etude de quelques lois statistiques utilisées en hydrologie. Cah. ORSTOM,ser.Hyd.

DIACONU C., OANCEA N., CIOCAN G. (1984)

Analiza spaţio-temporală a precipitaţiilor maxime (Spatial-temporal analysis of maximum precipitations) Rap. int. Bucureşti.

KALININ, G.P., N.V. NIKOLSKAIA, (1967)

Spatial and temporal analysis of maximum runoff, Proceedings of the Leningrad Symposium, 1967 (p. 174-193).

KITE, G.W. (1977)

Frequency and risk analysis in hydrology, Fort Collins, Colorado, USA.

OANCEA V., CIOCAN G. (1982)

Studiul cantităţilor maxime de precipitaţii căzute pe diverse durate de timp (Study of maximum precipitation amounts over varying time intervals), Rap. int., Bucureşti

SEVRUK B., GEIGER H. (1981)

Selection of distribution types for extremes of precipitation. Operational Hydrological Rep., no. 15, WMO.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

NEAREST NEIGHBOR SPACING DISTRIBUTIONS OF CUMULUS CLOUDS

Robert F. Cahalan
Laboratory for Atmospheres
NASA/Goddard Space Flight Center

1. INTRODUCTION

The statistical distribution of cumulus cloud spacings is important both in determining the resolution required for the remote sensing of cumulus clouds, and in the modeling of cumulus clouds and their associated radiation fields. As part of a program to develop statistical cloud models for Monte Carlo radiative computations, we are determining the nearest-neighbor spacing distributions of shallow (nonprecipitating) cumulus clouds from a number of digital images of the LANDSAT Multi-Spectral Scanner (MSS) and Thematic Mapper (TM) instruments, which have resolutions of 80 m and 30 m, respectively. Since typical shallow cumulus cloud spacings are a few hundred meters, they are not resolved by meteorological satellites, which have resolutions of 1 km or more. Our results are expressed in terms of a parameter δ which provides a measure of the tendency of clouds to be widely spaced (for $\delta > 0$) or to cluster together (for $\delta < 0$).

One may of course observe cloud clusters even in a completely random cloud field. The eighteenth century astronomer William Herschel worried about a similar problem--that perhaps the large number of "binary" star systems might be random line-of-sight coincidences. The Rev. John Mitchell, a contemporary statistician, showed that the observed number was significantly greater than that expected from random coincidences, and twenty years later Herschel found that many of them had moved in a way consistent with Kepler's laws (Zeilik, 1976).

A related question is: What is the nearest-neighbor spacing distribution of a set of points which have coordinates chosen independently from a uniform distribution? For points on a line, it is given by exp[-x], where x is the euclidean distance from each point to its nearest-neighbor. This decreases monotonically from its maximum value at zero separation. For points in a plane, the distribution is $2x\exp[-x^2]$, which vanishes at zero separation, rises to a maximum at $x^2 = 1/2$, and then drops rapidly to zero. When one generalizes to a space of arbitrary dimension d, one finds a Weibull distribution with shape parameter d (Onoyama et al., 1984). This has the property that large separations become more likely as the value of d increases.

In section 2 we further generalize this result by relaxing the requirement that the coordinates must be chosen independently. If we take the conditional probability C(x) that there is a point at a distance x from some chosen point, given that there are none closer, to vary as x^{δ}, then we obtain a Weibull distribution with shape parameter or "effective dimension" $d_{eff} = d + \delta$. When $\delta > 0$ then C increases with separation (cloud "repulsion") and we have $d_{eff} > d$, so that large separations become more likely than they are in the case of independently distributed clouds. Conversely, when $\delta < 0$ then C decreases with x (cloud "attraction" or "clumping") and then $d_{eff} < d$, in which case large separations are suppressed. Section 2 describes a simple (and well-known) method for determining the shape parameter d_{eff}.

In section 3 we describe the LANDSAT data, and the procedures used to extract the cloud coordinates and their nearest-neighbor distances. Section 4 gives initial results for the cloud fields analyzed so far. Finally, in Section 5 we summarize our results and conclusions.

2. SPACING DISTRIBUTION OF RANDOM POINTS

For comparison with the empirical nearest-neighbor distributions of cumulus clouds described in Section 4, we have constructed a one-parameter family of distributions, as follows. First, we ignore the geometry of individual clouds by replacing each cloud by a single point at its center-of-mass. For convenience we refer below to these center-of-mass points simply as "clouds". We also ignore any angular dependence, and assume that all probabilities depend only on the distances separating the various cloud points. Thus in this model there is no preference for cloud streets, for example, though they may appear in a given realization.

A. Derivation of spacing distribution

Although in the LANDSAT data the cloud coordinates are projected onto two dimensions and discretized, it is convenient for this derivation to consider d-dimensional continuous variables. Consider a field of points in a d-dimensional euclidean space, and imagine a spherical shell of internal radius x, thickness dx, and volume dv centered on one of these points. The nearest-neighbor distribution is determined by the joint probability that: (1) there are no clouds within radius x of the given cloud; and (2) there is a cloud in the volume dv between radius x and x + dx. If we assume that this is proportional to the volume element dv, then we can express it as $P_{nn}(x)dv$, where $P_{nn}(x)$ is the probability per unit volume that the nearest neighbor is at x.

The joint probability $P_{nn}(x)dv$ can also be expressed as the product of: (1) the probability $P_o(x)$ that there are no clouds closer than x; and (2) the conditional probability that there is a cloud in dv, given that there are no clouds closer than x. If we assume that the conditional probability is also proportional to dv, then we have

$$P_{nn}(x)dv = P_o(x) * C(x)dv \tag{1}$$

where C(x) is the probability per unit volume of finding a cloud at x, given no clouds closer. Now, the probability that there are no clouds closer than x equals the probability that the nearest neighbor lies between x and ∞, so that

$$P_o(x) = \int_x^\infty P_{nn}(x')\,dv' \tag{2}$$

Taking the derivative of (2) and using (1) gives

$$dP_o/dv = -P_{nn} = -P_o{*}C, \tag{3}$$

from which

$$P_o(x) = K \exp[-\int_o^x C(x')dv'], \tag{4}$$

and therefore we have

$$P_{nn}(x) = K\,C(x)\exp[-\int_o^x C(x')dv']. \tag{5}$$

Finally, we use the fact that

$$dv = Bx^{d-1}dx, \tag{6}$$

and

$$P_{nn}dv/dx = p_{nn}, \tag{7}$$

where p_{nn} is the probability per unit radius that the nearest-neighbor is at radius x, so that (5) becomes

$$p_{nn}(x) = AC(x)\,x^{d-1}\exp[-B\int_o^x C(x')\;x'^{d-1}dx'], \tag{8}$$

where A and B are constants determined by the normalization and first moments of p_{nn}.

If we allow the cloud coordinates to be chosen independently, the C(x) is independent of x, and p_{nn} reduces to the Weibull distribution with shape parameter d, a well-known result. (See for example Onoyama et al., 1984.) If instead we take

$$C(x) = k\,x^{\delta}, \tag{9}$$

where δ may be thought of as a "repulsion" parameter measuring the tendency of clouds to prefer large spacings, then (8) takes the form

$$p_{nn}(x) = A\,x^{d_{eff}-1}\exp[-B\,x^{deff}], \tag{10}$$

which is a Weibull distribution with shape parameter given by

$$d_{eff} = d + \delta\,. \tag{11}$$

B. Determination of the shape parameter

Here we review a well-known method for estimating the shape parameter d_{eff} by a simple least-squares procedure. For a review of estimation procedures for the generalized (threshold-dependent) Weibull distribution, and the related Frechet, Gumbel, and Fisher-Tippett distributions, see Mann (1984) and also Sneyers (1984).

The cumulative distribution associated with (10) has the form

$$\Phi(x) = \exp[-B\, x^{d_{eff}}], \tag{12}$$

so that

$$p_{nn} = -d\Phi/dx. \tag{13}$$

Here $\Phi(x)$ is the probability of finding a nearest-neighbor distance greater than or equal to x. It can be estimated empirically by ordering the distances from largest to smallest, then assigning the rank k = 1 to the largest, k = 2 to the second-largest, etc, and finally setting

$$\Phi \sim F = k/N, \tag{14}$$

where N is the total number of distances in the sample.

From (12) we have

$$\ln(-\ln\Phi) = d_{eff} \ln x + \ln B, \tag{15}$$

so that if we put

$$Y = \ln(-\ln\Phi), \tag{16a}$$

$$X = \ln x, \tag{16b}$$

then (15) takes the form

$$Y = aX + b, \tag{17}$$

where

$$a = d_{eff}, \tag{18a}$$

$$b = \ln B. \tag{18b}$$

According to (16) - (18), if the nearest-neighbor distances are distributed as in (10), then a scatter plot of the logarithm of a sample of N distances versus the double logarithm of the rank divided by N should lie close to a straight line with slope given by the shape parameter d_{eff}.

C. Monte Carlo simulations

In order to test our cloud analysis software, we have analzed some random points in a plane. Figure 1a shows a set of 512 points with coordinates chosen independently from a uniform distribution of integers from 1 to 512. Figure 1b shows each point connected by a straight line to its nearest-neighbor. Points which are closer to any of the four boundaries than to their nearest neighbor must be excluded from the sample. The remaining 456 distances were then sorted from largest to smallest, and a rank assigned. Figure 1c shows a scatter plot with ln(-ln(rank/456)) on the abscissa and ln(distance) on the ordinate. Since this is the inverse of (16), we obtain a least-squares fit line with slope approximately $1/d_{eff} = 0.5$. Distances of only a few pixels are excluded from the fit to minimize the effect of the discretization. The resulting slope is 0.50061. Figure 1d shows the distance histogram, along with the Weibull distribution for the estimated value of d_{eff}.

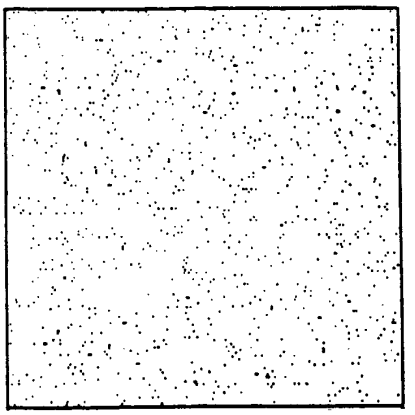


Figure 1a. Set of 512 points with coordinates chosen independently from a uniform distribution from 1 to 512.

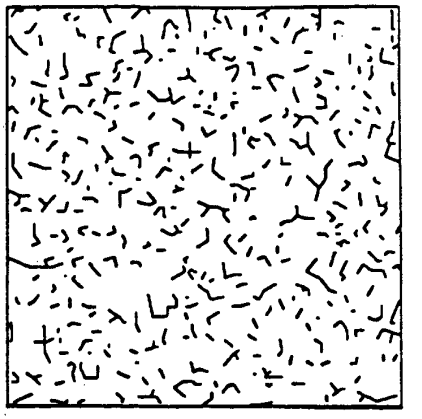


Figure 1b. Same as 1a, but with nearest-neighbors connected.

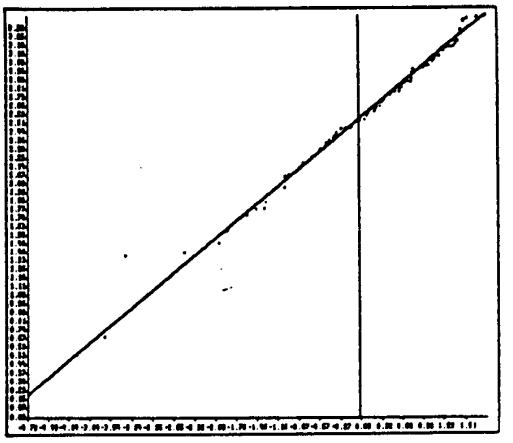


Figure 1c. Scatter plot of logarithm of distances versus double logarithm of rank over number.

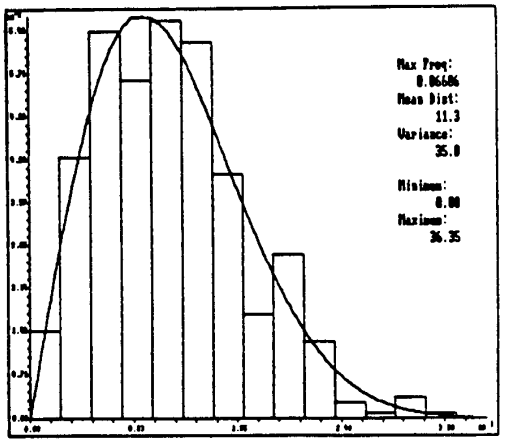


Figure 1d. Histogram of distances with Weibull distribution having shape parameter as in Figure 1c.

3. DATA AND ANALYSIS

Persistent large-scale regimes of shallow cumulus clouds are common in the subtropics over both the Atlantic and Pacific Oceans. (See for example Cahalan et al., 1981.) For our initial study of the spacing distribution, we have selected a LANDSAT TM scene of August 30, 1982 at 31°44'N, 78°54'W off the east coast of the United States. The TM instrument provides 3 reflected visible bands, 3 reflected near-infrared bands, and the water vapor window band of emitted infrared radiation. The reflected bands have a horizontal resolution of about 30 m, and a field-of-view about 185 km on a side. The reflectances and radiances are digitized to 8-bit values from 0 to 255. Since cloud reflectivity is often high enough to saturate the visible bands, we have chosen to study the near infrared band 7 with peak response at 2.05 microns.

The histogram of band 7 radiances shows a sharp peak of low reflectance values asociated with the dark ocean surface, followed by a long flat plateau of higher reflectance values due to clouds. We define the cloudy pixels by choosing a threshold just above the peak, so that the darker cloud edges are included. The digital image of 8 bit reflectances is then replaced by a binary image having 1 for cloudy pixels and 0 elsewhere. Contiguous areas are then identified, and each is given a unique integer identification number (ID). From the ID image, various cloud features such as area, perimeter, and center-of-mass coordinates are determined and stored in an ancillary file. We then compute the distance from each cloud to its nearest-neighbor, and to each of the four edges of the image. Clouds closer to an edge are excluded.

4. RESULTS

Figure 2a shows 294 center-of-mass cloud coordinates from a 1024 by 1024 pixel region in the LANDSAT scene described in the last section. This small sample clearly shows much more clustering than the random point set of Figure 1a. There are many small spacings within each cluster, but also large spacings associated with individual clouds not in a cluster. Figure 2b shows the nearest neighbors connected. Seventeen clouds are closer to an edge, and must be excluded. Figure 2c shows the scatter plot. Distances smaller than 5 pixels were excluded from the least-squares fit. The slope of 0.74611 gives a shape parameter of 1.34, and since $\delta = d_{eff} - 2$ we obtain

$$C(x) = x^{-2/3} \qquad (19)$$

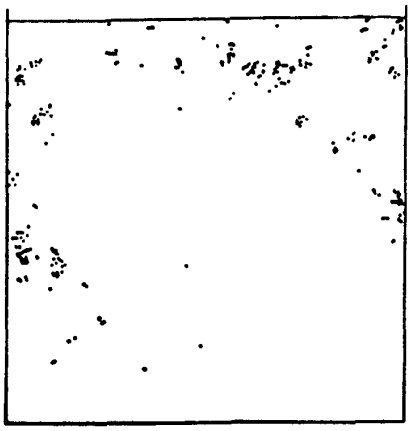


Figure 2a. Set of 293 points taken from LANDSAT scene described in section 3.

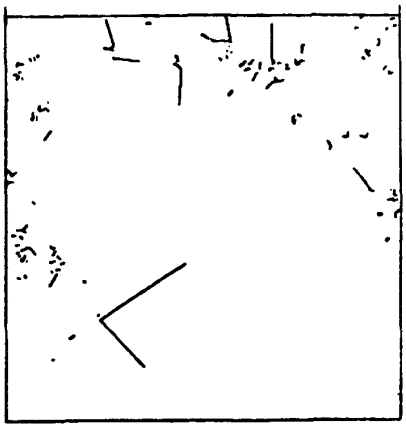


Figure 2b. Same as 2a, but with nearest-neighbors connected.

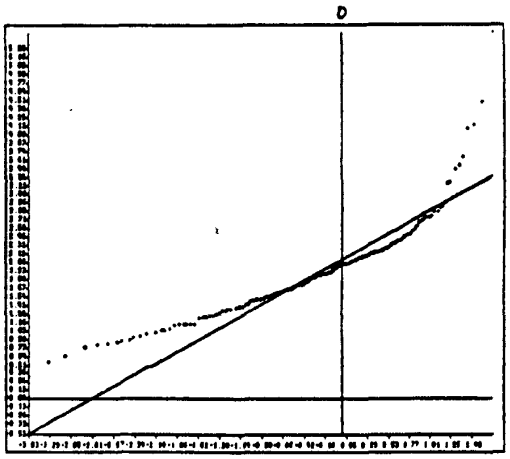


Figure 2c. Scatter plot of distances from Figure 2b versus double logarithm of rank over number.

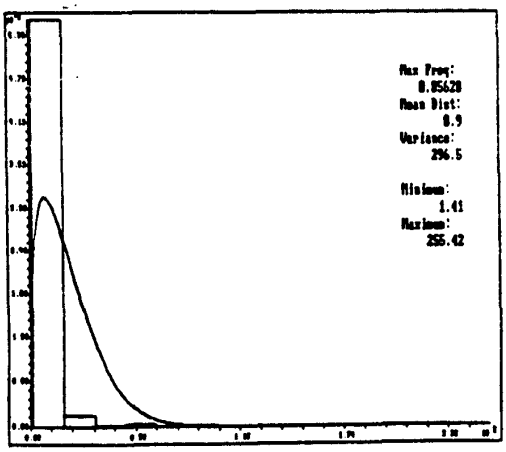


Figure 2d. Histogram of distances with Weibull distribution having shape parameter as in Figure 2c.

Because of cloud clustering, it is clearly necessary to perform this analysis on many samples with a large number of clouds, in order to obtain stable estimates of the large distances between clusters. Judging from Figure 2c, however, it may be that the monotonic form of C(x) assumed here will have to be generalized.

5. SUMMARY AND CONCLUSIONS

We have generalized the well-known result for the nearest-neighbor spacing of random points in d-dimensional euclidean space to allow for dependence between the points. The result (equation 8) is given in terms of a quantity C(x), which is the conditional probability per unit volume of finding a cloud at a distance x, given that there are no clouds closer. When $C \sim x^{\delta}$ then the nearest-neighbor distribution takes the form of a Weibull distribution with a shape parameter given by $d + \delta$ (equations 10 and 11).

The nearest-neighbor distribution of shallow cumulus clouds deviates from a Weibull distribution in situations when there is strong clustering. For such cases the simple power-law form of the conditional probability must be modified.

Randall and Huffmann (1980) have surveyed a number of possible physical explanations of cloud clustering. Their emphasis was on deep convection, and they studied a stochastic model in which cloud "seeds" were distributed independently. We may conjecture that the shallow cumulus cells studied here may be revealing a more appropriate way to distribute the seeds of deep convection.

Acknowledgments: I would like to thank Jon Robinson for developing the software described in sections 2C and 3, and Linda Pylant for typing the manuscript.

6. REFERENCES

Cahalan, R.F., D. A.. Short and G. R. North, 1981: Cloud Fluctuation Statistics, Mon. Wea. Rev., 110, 26-43.

Mann, N., 1984: Statistical Estimation of Parameters of the Weibull and Frechet Distributions, from Statistical Extremes and Applications, J. Tiago de Oliviera, ed., D. Reidel Publishing Co., c1984, pp. 81-89.

Onoyama, T., M. Sibuya, and H. Tanaka, 1984: Limit Distribution of the Minimum Distance Between Independent and Identically Distributed Random Variables, from Statistical Extremes and Applications, J. Tiago de Oliviera, ed., D. Reidel Publishing Co., c1984, pp. 549-562.

Randall, D. A. and G. J. Huffman, 1980: A Stochastic Model of Cumulus Clumping, J. Atmos. Sci., 37, 2068-2078.

Sneyers, R., 1984: Extremes in Meteorology, from Statistical Extremes and Applications, J. Tiago de Oliviera, ed., D. Reidel Publishing Co., c1984, pp. 235-252.

Zielik, M., 1976: Astronomy: The Evolving Universe, Harper and Row, Publishers, Inc., New York, 529pp.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

MISAPPLICATIONS OF EXTREME VALUE STATISTICS

Arnold Court, Emeritus Professor of Climatology,
California State University, Northridge CA 91330

1. DISTRIBUTIONS

Extreme value statistics were discovered in the 1920s, introduced into streamflow hydrology during the 1930s (Gumbel, 1940, 1941a, 1941b) and into climatology for rainfall, temperature, and pressure in the 1940s (Gumbel, 1942) and for wind a decade later (Court, 1953; Thom, 1954). In the ensuing three decades, their use has increased rapidly, and so have questionable and even erroneous applications to the various climatic elements. The theory is rigorously correct for variables meeting its basic assumptions, but its practical application to real-world data is fraught with many difficulties. Now an early proponent of extreme value analysis is climatology (Court, 1952, 1953) deplores recent over-precise computations, especially for small samples from non-stationary series, and objects to pointless discussion of how best to fit such inadequate data to theory.

Estimates of extreme value of various geophysical series are needed for the safe design of structures which must withstand the deepest flood, greatest rainfall, strongest wind, hottest (or coldest) temperature, etc., having some small probability of being exceeded during the desired lifetime T of the structure. To have a calculated risk of no more than U of failure from some environmental event, a structure should be designed to withstand the extreme having a return period of $-1/\ln(1-U)\ T$. For 10% risk, this multiplicative factor is 9.49, for 5% it is 19.49, for 1% it is 99.50 (Court, 1952): A structure intended to last 50 years with no more than 10% risk of premature failure should be able to withstand the extreme of any critical environmental element expected only once in 475 years, i.e. having a 0.21% chance of occurring in any one year.

Estimation of such small-probability extremes is the domain of extreme value theory. Even though the actual statistical distribution of the element is not known with sufficient precision to permit valid estimation of the events in its tails, the distribution of the largest values in each of m independent samples, each containing n independent observations from the original distribution, converges, as m and n increase without limit, to one of three forms. These are designated by the names of the statisticians more active in developing them, and by the sequence assigned them by Fisher and Tippett (1928):

FT-I	Gumbel	unbounded, thin tails
FT-II	Fréchet	bounded below, thick tail
FT-III	Weibull	bounded above

(Actually, the original Weibull distribution was bounded below, and hence the reciprocal of the FT-III, but the name is now applied generally to the latter.)

2. SAMPLE SIZE

Emphasized above are the two serious assumptions which limit drastically the application of extreme value procedures in climatology and other branches of geophysics. How large must m and n be for this asymptotic theory to be applicable? Gumbel wanted at least 30 for each; Galambos (1984, p. 351) concludes that "the asymptotic extreme value theory is quite accurate for the empirically determined sample size $n > 100$...". That each sample should be large enough to have a large probability of containing at least one observation from the extreme tail of the parent distribution has been suggested by Tabony (1983), who felt that "the tail ... may be loosely defined as the top 10-15% of the parent distribution."

The sample size n required for various probabilities $P_{n,q}$ that its largest value exceeds specific percentiles q of the parent distribution is $\ln(1-P_{n,q}) / \ln q$ (Table 1).

For 99% confidence that the sample contains at least one value from the uppermost 5% of the parent distribution, it should contain 90 independent observations, for 95% confidence nearly 60. Routine samples of 30 independent observations offer only 80% confidence of having an extreme from the uppermost 5%: one in five of such samples will have no values from the upper 5% of the distribution.

q% P:	.80	.82	.84	.86	.88	.90	.92	.94	.95	.96	.96	.98	.99
80	7	8	8	9	9	10	11	13	13	14	16	18	21
81													
82	8	9	9	10	11	12	13	14	15	16	18	20	23
83													
84	9	10	11	11	12	13	14	16	17	18	20	22	26
85													
86	11	11	12	13	14	15	17	19	20	21	23	26	31
87													
88	13	13	14	15	17	18	20	22	23	25	27	31	36
89													
90	15	16	17	19	20	22	24	27	28	31	33	37	44
91													
92	19	21	22	24	25	28	30	34	36	39	42	47	55
93	23	24	25	27	29	32	35	39	41	44	48	54	63
94	26	28	30	32	34	37	41	45	48	52	57	63	74
95	31	33	36	38	41	45	49	55	58	63	68	76	90
96	39	42	45	48	52	56	62	69	73	79	86	96	113
97	53	56	60	65	70	76	83	92	98	106	115	128	151
	64	68	72	78	84	91	100	111	118	127	139	155	182
98	80	85	91	97	105	114	125	139	148	159	174	184	228
	106	113	121	130	140	152	167	186	198	213	232	258	305
99	160	171	182	196	211	229	251	280	298	320	349	389	458

Table 1. Minimum sample size n for probability P that largest one $\geq$ qth percentile.

These computations, of course, are for independent observations, but in geophysics successive observations may have some serial correlation. This is well-known for daily values of most meteorological elements, leading to the general rule that such values represent only about one-third as many independent observations (Parker, 1984). Thus monthly extremes represent only about 10 independent values, so that one of every ten such extremes does not exceed the 80th percentile of the parent distribution. Such limitations are not considered in most analyses of monthly extremes, such as by Challenor and Carter (1983).

Because not all small samples contain values from the far tail of the parent distribution, they may behave as though they came from a bounded distribution, leading to an apparent following of a Weibull (FT-III) distribution; conversely, if the small sample has too many values from the far end, they will suggest a fatter tail and hence a Frechet (FT-II) distribution. Much of the confusion in current literature about the proper distribution to be used for extreme value analysis may arise from reliance on inadequate samples.

3. NON-STATIONARITY

Large samples are hard to obtain for climatic data, which are inherently non-stationary, not only climatically but also observationally and instrumentally. Besides the diurnal and annual cycles (and possible lunar cycles as well), weather and climate have aperiodic fluctuations of all scales, so that annual or longer-interval extremes do not necessarily come from the same distribution. In addition, but much less considered, observations rarely come from the same instrument in the same surroundings read at the same time and reduced in the same way for more than a few years at a time. A sample of 50 annual extremes, although perhaps free of serial correlation and from a climatically stable period, is still a melange of observations from different instruments by different people using different methods in different locations, and hence far from the set of "independent and identically distributed" ("i.i.d.") values on which the entire structure of extreme value statistics is based.

Yet, as already seen, 50 i.i.d. observations is perhaps a minimum of valid extreme-value analysis, giving 99% confidence that each extreme is from the upper 5% of the distribution. How can adequate samples be obtained for the needs of applied climatology? Grouping several sites, close enough to represent a single climatic regime but far enough apart to have slight observational and instrumental correlation, for a relatively short period, perhaps ten years, may avoid the difficulty of climatic non-stationarity, but at the cost of compounding observational and instrumental heterogeneity. Such a "station-year" approach was common half-a-century ago, and is now used in hydrology for estimating "regional skew" of the "log Pearson III" formulas used in rainfall and streamflow analyses by U.S. government agencies.

These complications could lead to the conclusion that the whole extreme-value approach to climatic data should be scrapped. This, however, is unacceptable to the engineers and others who need estimates of the strongest winds, highest water levels, hottest (and coldest) temperatures, and deepest rainfalls for which man's structures should be designed. At present, the best compromise seems to be to avoid hairsplitting over proper distribution types and eschew inordinate precision on results. Extreme value analysis of non-homogeneous data by any of the three distributions may give values reasonably correct to two significant figures for probabilities down to perhaps 5%, perhaps to 1% -- 100-year return period. This should be adequate for most applications, because the engineer will merely multiply the climatologist's estimate by a one-digit influence coefficient, then double the answer for safety.

4. REFERENCES

Challenor, P. G., and Carter, J. T., 1983: Monthly air temperatures at Rothamsted. J. Clim., 3, 395-404.

Court, A., 1952: Some new statistical techniques in geophysics. Adv. in Geophys., 1, 45-84.

-----, 1953: Wind extremes as design factors. J. Franklin Inst., 256, 39-56.

Fisher, R. A., and L. H. C. Tippett, 1928: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Cambridge Phil. Soc., 24(20), 180-190. (repr. in R. A. Fisher, Contributions to Mathematical Statistics, paper 15. New York, Wiley, 1950.)

Galambos, J., 1984: Rates of convergence in extreme value theory, pp. 347-352 in: Statistical Extremes and Applications, J. Tiago de Oliveira, ed. Proc. NATO Adv. Study Inst. on Stat. Extremes and Applications, Vimeiro, Portugal, 1983. Boston, Reidel, x + 692 pp.

Gumbel, E. J., 1940: Les crues du Rhône. Ann. Univ. Lyon, 3eser., fasc. 3, Sec. 4, 39-54.

-----, 1941a: The return period of flood flows. Ann. Math. Stat., 12, 163-190.

-----, 1941b: Probability interpretation of the observed period of floods. Trans. Am. Geoph. Un. III, 836-1850.

Parker, D. E., 1984: The statistical effects of incomplete sampling of coherent data series. J. Clim., 4, 445-449.

Tabony, R. C., 1983: Extreme value analysis in meteorology. Meteorol. Mag., 112, 77-98, disc. 113, 43-51.

Thom, H. C. S., 1954: Frequency of maximum wind speeds. Proc. Amer. Soc. Civil Eng., 80, separate 539.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

STATISTICAL MODELS FOR CLIMATIC VARIATION RESEARCH

C.-D. Schönwiese

Institute for Meteorology and Geophysics, Frankfurt University
Feldbergstr. 47, D-6000 Frankfurt a.M. 1 (FRG)

INTRODUCTION: CLIMATIC VARIATION AND VARIABILITY

Evidently, climate is varying in respect to space and time on all scales of space and time. This property of variation implies very different aspects of variation and it is necessary to use both precise definitions and appropriate-reproducible techniques of analysis. According to Gates (4) "a logical definition of climate" is "the complete statistical description of the state of the internal climate system over a specified time period, together with a description of the state of the external system or boundary conditions. Such a description may be called a climatic state."

Note that the reference period of such a climatic state is relative and arbitrary (say one month or 30 years or the Holocene), also the sample data interval T_S, and that, in turn, we may discern subscale and suprascale variations. Let T_P be the characteristic time of any phenomenon which shall be statistically described and T_R the reference period (record length), then must be

$$T_S \ll T_P \ll T_R ,$$

where subscale means $\leqslant T_S$ and suprascale $\geqslant T_R$. If, for example, the reference period or "scale" is 30 yr, then the daily cycle (e.g. of air temperature) is a subscale phenomenon and the difference of two climatic states (e.g. temperature change 1901-1930 versus 1931-1960) is a suprascale phenomenon, usually called climatic variation. (It is possible, of course, to define subintervals within T_R and, in turn, different climatic states within T_R.) Any climatic state and, in turn, climatic variations imply different climatic elements, field patterns and statististics (moments, data interdependence, spectrum etc.) and usually subscale variations of any climatic state are thought to be stationary or at least quasi-stationary. The variance among a number of climatic states of the same kind is called climatic variability (4).

THE DATA BASE: TYPES AND PROBLEMS

The basic information of climatic variation is provided by means of observational data. We may regard three, so to speak, dimensions of climatic data, see Fig. 1. First, the fundamental type or object of consideration: (a) observations of climatic phenomena (temperature, precipitation, etc.), phenomenal data; (b) data concerning any particular forcing mechanism (e.g. volcanism, CO_2 trend), causal data; (c) data concerning the climatic effects on the environment (e.g. vegetation, social affairs), impact data. Note that impact data may also be used for climate reconstructions (proxy data, e.g. tree ring widths or densities) and, therefore, transferred to phenomenal data (with a uncertainty of this transfer function).

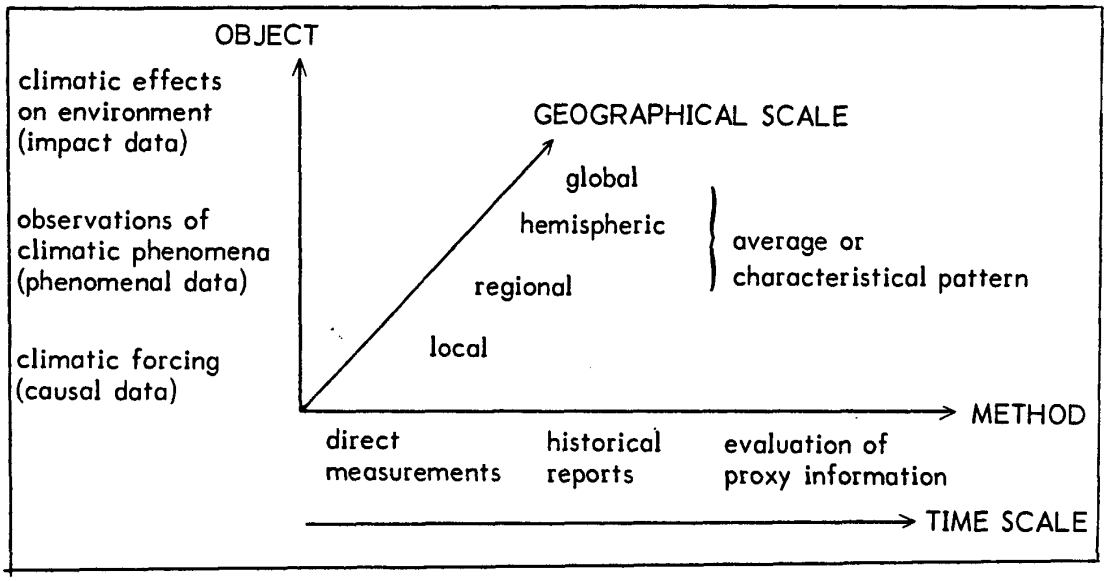


Fig. 1. "Dimension" scheme of climatic data types (explanation see text).

Second, the method of evaluation of observational data may be based on (a) direct measurements, (b) historical reports or (c) indirect methods (proxy data, in particular palaeoclimatology). The method of data evaluation is correlated with the time scale. Finally, we may consider local, regional (may-be hemispheric) and global data, with the exception of local data in terms of spatial averages or typical field patterns; this is the geographical (spatial) scale of climatic data.

There arise many problems which cannot discussed here in detail: the problem of accuracy; the problem of spatial correlations and representativeness (important for reasonable field definitions); the problem of time series homogeneity. (A homogeneous record reflects only meteorological variations and not those superimposed by change of instruments, station locality etc.) A typical problem in this context is the heat island effect of urbanization which is not representative for the surrounding. Homogeneity tests should use homogeneous reference series (relative

tests (3,12)), if possible average of a number of records, which are, however, seldom on a secular time scale. So, there is a frequent source of data uncertainty which cannot be neglected.

Fig. 2 presents examples of relative homogeneity tests, here Craddock's (3) test based on cumulative deviations form the reference series. Other methods use e.g. deviations from frequency distribution models (references see (12)). If proxy data are used, the stability of the transfer function may be a problem which can be considerd by means of moving correlation and coherence analysis (13), see also further contribution of the author in this volume, or by means of other calibration and verification techniques.

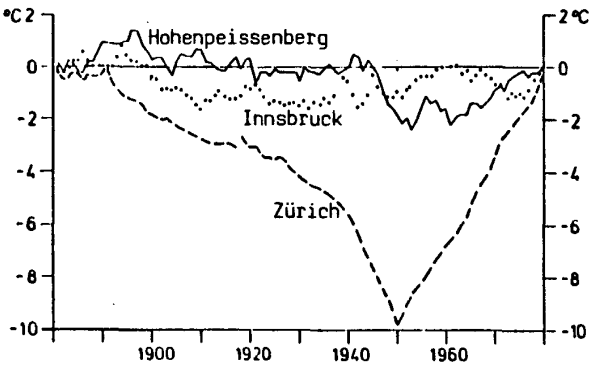


Fig. 2. Homogeneity test after Craddock (3) where the cumulative deviations of the Zürich temperature series from a reference series (spatially averaged) point to a inhomogeneity in 1950 (due to station displacement); adopted from (12).

THE ANALYSIS: SAMPLE STATISTICS

Climatic variations may concern variations in relation to space and/or time, spectral statistics and combinations of these aspects. The possible types of time series variations have often been described (continuous or abrupt, random or non-random, cyclical or non-cyclical, etc.). It should be realized that variations of variance and other (central) moments may be more important than the description of variations of the average; for an example see Fig. 3. Moreover, of course, the

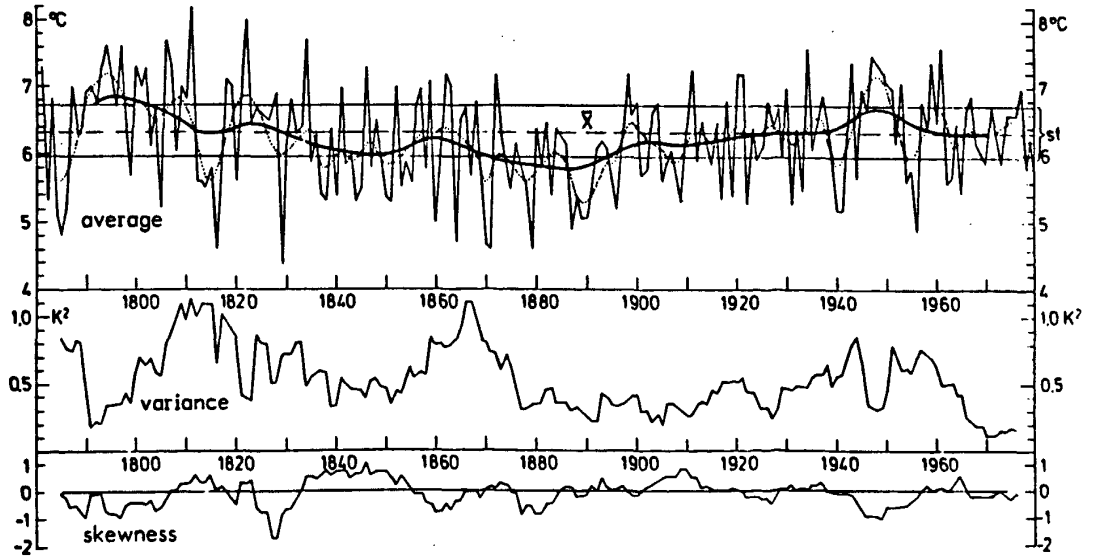


Fig. 3. Hohenpeissenberg annual means of air temperature 1781-1980, including low-pass filter (> 10 yr, dotted line, > 30 yr heavy line), upper plot, moving ten year variance and moving ten year moment coefficients of skewness; st: stationarity test of average (from (12)).

autocorrelation function and spectral statistics should be available. Cross correlation and cross spectral analysis enables both a description of spatial (also spatial-spectral) relations and assessments of possible forcing mechanisms if causal data are involved. A relatively new aspect is moving spectral analysis (10, 13) which is important in order to realize the stability of spectral statistics in time.

The sample statistics, however, are not more than the first approach to a reliable statistical analysis. One mut ask for the confidence of these statistics. This leads to a second approach: the evaluation of theoretical-statistical hypotheses which may describe any poulation processes. The last step will be: try to find any synthesis of sample statistics and population hypotheses.

STATISTICAL MODELS: REALISTIC OR NOT?

Theoretical assumptions or hypotheses, often simplified, are usually called "statistical models" if the evaluation follows a statistical way. In the field of time series variation analysis there are two important concepts independent from sample statistics: (1) the concept of frequency distributions and (2) the concept of data interdependence.

First, theoretical frequency distributions (probability density functions f(x;a,b,...)) define particular variation processes by means of particular parameters a,b,...; Fig. 4 presents the relations of some common frequency distributions. For additional distributions see Suzuki (14), moreover (5,6) and section IV of this volume.

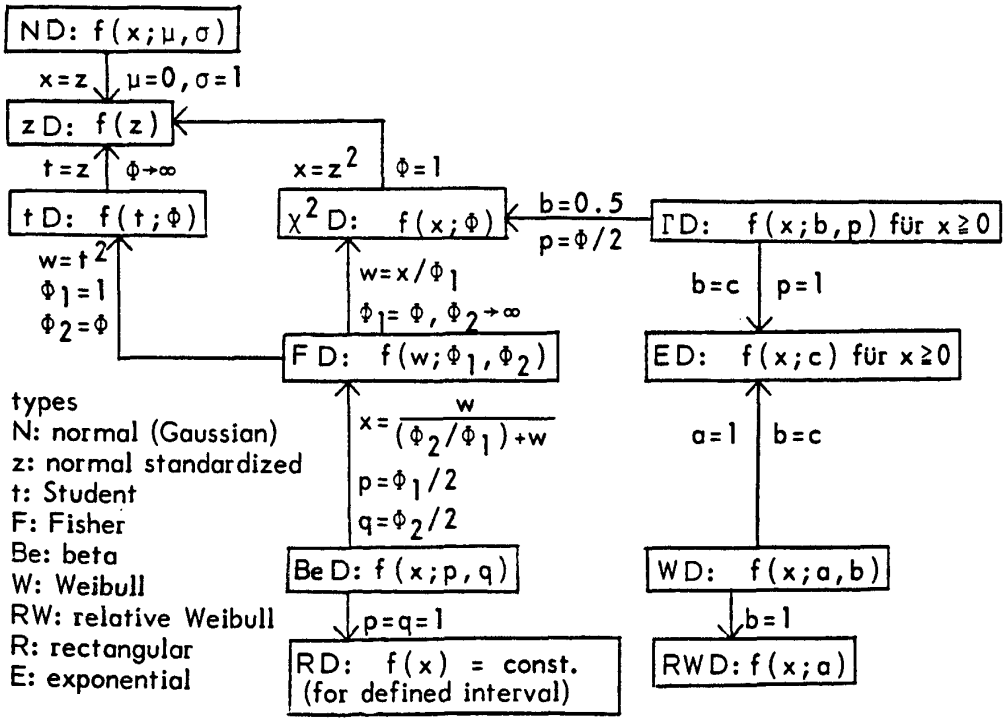


Fig. 4. Relations of some common frequency distributions "D" where the initial symbol denotes the type and the symbols following ";" the parameters (μ: average, σ: standard deviation, Φ: degrees of freedom, p: probability etc.).

Second, the interdependence of time series data is a matter of interest. Let $a(t, t+\Delta t)$ be the variation of any phenomenon occuring from time t to time $t+\Delta t$. Then a Poisson process is defined by independence in relation to time t, $t+k\Delta t$, moreover by orderliness (independence in relation to space) and stationarity. In contrast to that a Markov process describes a series of events where any event $a(t+\Delta t)$ depends exclusively on the immidiately preceeding event a(t); in a simple way the autocorrelations may be $r_A(t+k\Delta t) = r_A(t+\Delta t)^k$, k=2,3,...,M (first order autoregressive process (9,10)). More general, we may have a Renewal Process (2,8) where the interdependence of the time series data may be of any form.

In climatology Markov processes are often supposed to be realistic although this is not necessarily correct. Sample statistics of a finite time series which never can exactly reproduce a theoretical assumption, although used for the assessment of the necessary parameters (e.g. one lag autocorrelation for the definition of a Markov process), complicate the analysis. This is illustrated by Fig. 5 where the Frankfurt air temperature record (annual means 1881-1980) variance spectrum is compared with a corresponding spectrum based on Gaussian distributed random numbers. In both cases a Markov red noise theoretical spectrum is adapted (null hypothesis). It is found that the former spectrum (climatic quantity) indicates two peaks exceeding the 95 % confidence level (χ^2 test, alternative hypothesis) but the latter spectrum (random process, only sample available) indicates also one peak exceeding this level of confidence. In fact, the two spectra are very similar. It is recommended to use different, if ever possible independent algorithms and tests, in this case not only the Fourier transform of the autocorrelation function but also, for instance, maximum entropy spectral analysis where tests based on Monte Carlo techniques may be appropriate (7, 10, 13) in order to get more reliability and to be more independent from algorithmic artifacts.

This example demonstrates that even the statistical model cannot be strictly independent from the sample startistics in every case of analysis. Additional concepts or "models" for statitistical-climatological variation research are a priori based on sample statistics like correlation and regression techniques (although called "regression model"), coherence spectra, empirical orthogonal functions (EOFs) and others.

Now, are the statistical models realistic or not? This means, do the models coincide with the sample statistics so that any available sample may be supposed to be any possible realization of a defined population model? Often the question may simply be: Is the sample effect random (or Markovian, etc., null hypothesis) or not (alternative hypothesis)? An enormous variety of statistical tests, evaluated

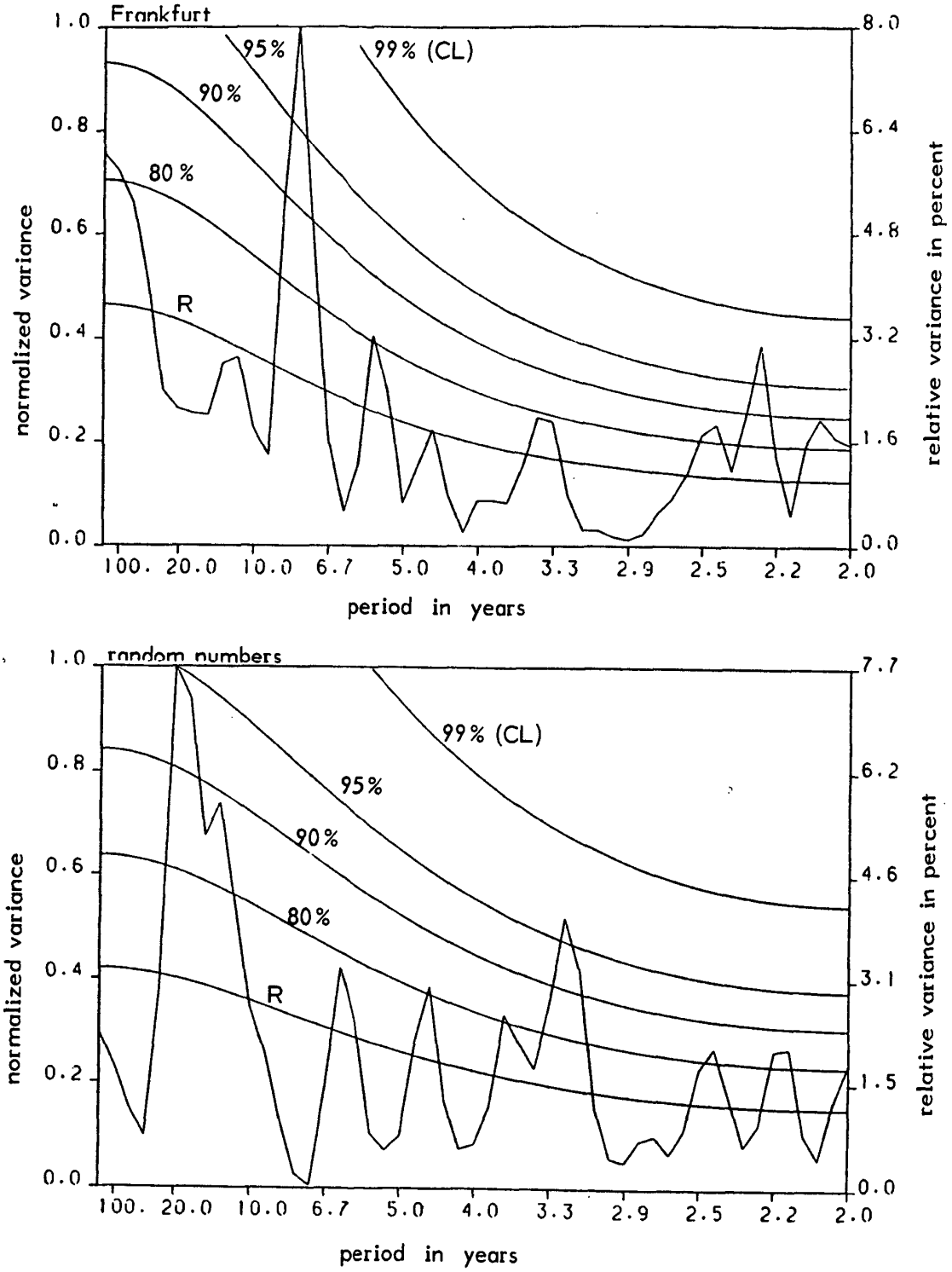


Fig. 5. Variance spectra (Fourier transform of autocorrelation function after (1)) of the Frankfurt (FRG) air temperature record, annual means 1881-1980, and a corresponding series of Gaussian distributed random numbers (below); maximum time lag M=50; R: Markov red noise theoretical spectrum (null hypothesis), adopted using the sample one time lag autocorrelation; CL: confidence levels (using χ^2 test (9); after Malcher and Schönwiese (unpublished).

for particular problems, is available to make any decision at any level of confidence (error probability). These methods may be applied not only on time series but also on spatial patterns (climatic anomalies, for example, in any particular region of the world within any particular time interval). The problem of statistical field anomaly significance, however, is not trivial and again the implication of Monte Carlo techniques is useful (7).

A very important problem in both time series and field analysis is the "detection" (separation) of any deterministic mechanism effect, the so-called signal (e.g. volcanic or anthropogenic effects) significantly different from the overall observed variability or "noise" at any climatic state. This "signal" may be the outcome of a (predominantly) deterministic model or the outcome of a statistical regression or EOF assessment. The matter of interest is the signal to noise ratio and again it is the problem to make any decision at any defined level of confidence whether this ratio is significant (effect "detected") or not.

Finally, the problem of predictability shall be mentioned which is of high interest in the public. From the statistical point of view this problem is matched with the signal and signal to noise problem, may be in the context of spectral variance or multiple regression or any other statistical analysis. In addition, however, interesting hints may be come from deterministic modelling.

In general, it is necessary to find an interaction between sample statistics and statistical models in the field of statistical research of climatic variations which lead to hypothesis decisions or other assessments, see Fig. 6, moreover an interaction between statistical decisions and the output of deterministic models for a verification (or falsification or modification) of the deterministic models

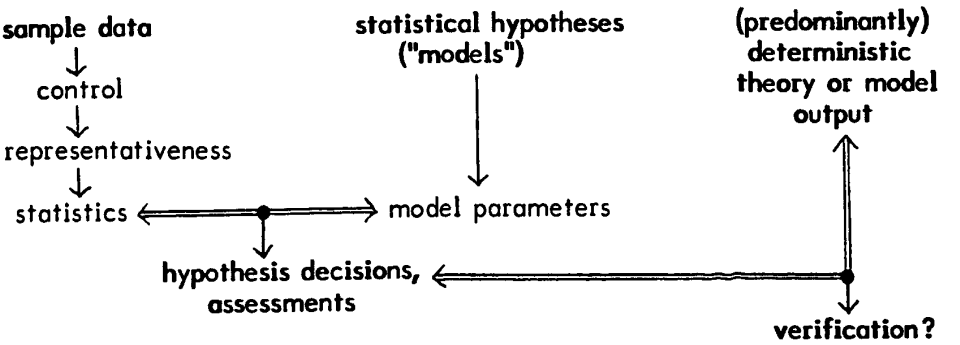


Fig. 6. The "interactions" of sample statistics, statistical hypotheses and deterministic models.

by statistical decisions and vice versa. This last aspect is important in the field of causal and impact statistical-climatological analysis because statistical methods can never prove such relationships but only point to some possible signs of these effects.

REFERENCES

(1) Blackman, R.B. and Tukey, J., 1958: The Measurement of Power Spectra. Dover, New York.

(2) Cox, D.R., 1962: Renewal Theory. Methuen, London.

(3) Craddock, J.M., 1979: Methods for comparing annual rainfall records for climatic purposes. Weather, 34, 332-346.

(4) Gates, W.L., 1981: The climate system and its portrayal by climate models: a review of basic principles. I. Physical basis of climate. In Berger, A. (ed.): Climatic Variations and Variability: Facts and Theories. Reidel, Dordrecht, 3-19.

(5) Ikeda, S., Suzuki, E., Uchida, E. and Yoshino, M.M. (eds.), 1980: Statistical Climatology. Proceedings of the I Internat. Conf. on Statistical Climatology. Elsevier, Amsterdam.

(6) Institito Nacional de Meteorologia e Geofisica, 1983: (Proceedings) II Internat. Conf. on Statistical Climatology. Liboa.

(7) Livezey, R.E. and Chen, W.Y., 1983: Statistical field significance and its determination by Monte Carlo techniques. Monthly Weather Rev., 111, 46-59.

(8) Lomnitz, C., 1974: Global Tectnonics and Earthquake Risk. Elsevier, Amsterdam.

(9) Mitchell, J.M., Dzerdzeevskii, Flohn, H., Hofmeyer, W.L., Lamb. H.H., Rao, K.N. and Wallén, C.C., 1966: Climatic Change. WMO Publ. No. 195 (Tech. Note No. 79), Geneva.

(10) Olberg, M. and Rakoczi, F., 1984: Informationstheorie in Meteorologie und Geophysik. Akademie-Verlag, Berlin.

(11) Panofsky, H.A. and Brier, G.W., 1958: Some Applications of Statistics to Meteorology. Pennsylvania State Univ, University Park (USA).

(12) Schönwiese, C.D. and Malcher, J., 1985: Nicht-Stationarität oder Inhomogenität? Ein Beitrag zur statistischen Analyse klimatologischer Zeitreihen. Wetter und Leben, 37, 181-193.

(13) Schönwiese, C.D., 1986: Moving spectral variance and coherence analysis and some applications on long air temperature series. Submitted to J.Clim.Appl.Met.

(14) Suzuki, E., 1980: A summarized review of theoretical distributions fitted to climatic factors and Markov chain models of weather sequences, with some examples. In Ikeda, S. et al. (eds., see above), 1-20.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

STATISTICAL STUDIES OF CENTRAL EUROPEAN PALAEOCLIMATE BASED ON BIOSPHERIC DATA OF THE RECENT MILLENNIUM.

Wolfgang Birrong
Institut für Meteorologie und Geophysik
Feldbergstraße 47
D - 6000 Frankfurt 1

1. Introduction and data base

A satisfactory climatic data base is available not before say 1850-1880. In order to get climatic information before this instrumental period on a year-to-year time scale in the first instance the recent millennium is of particular interest. Suitable proxy information is available primarily by dendrochronological data (derived from tree rings) and historical grape harvest records.

The high correlations between the maximum wood density (within each yearly tree ring) and the temperature of the second half of the vegetation period are well known. This effect has been demonstrated for a number of time series (among others for Central Europe, especially in cold and wet Alpine regions, see for instance [6]). Another remarkable example for Central Europe has been pointed out in [2] (beginning of the grape harvest versus April to July temperature).

Here another example is demonstrated. For climate temperate regions of Central Europe not much information exists about the relationship between proxy data and precipitation, another reason is that in general climatic variables can be matched only with one species of proxy variables. We use the following data for South West Germany: monthly precipitation totals, tree growth value [4] and vine quality index [5].

2. Basic methodology of biosphere-climate matching

The behaviour of the year-to-year biospheric proxy time series is that of quantities integrated in respect to time ("grown" quantities). The integration period covers these month for which the used climatic variables influence the corresponding proxy variables, that means mathematically (here simplified for one proxy and one climatic variable):

$$P_t = \sum_{j=0}^{M} A_j \cdot K_{t-j} + \text{Residual}_t \qquad (1)$$

where K = climatic time series (monthly values, predictors), P = proxy time series (yearly values, predictand), t = time index on a year-to-year base (referred to the last month of the actual vegetation period), j = monthly time index of lagging (running back M months from the last month of the actual vegetation period), A = transfer function. In general we consider the climatic variables temperature and precipitation.

Our climatological data has been averaged for two or three month, respectively. These variables are referred to: Apr/May, Jun/Jul, Aug/Sep, Oct/Nov/Dec (previous year) and Jan/Feb/Mar, Apr/May, Jun/Jul, Aug/Sep (actual year).
To obtain biosphere-climate relations we use the so-called response and transfer functions [3] in order to explain a predominant part of variance by the leading orthogonal components. The proxy time series often imply strong autocorrelation. In such cases we use a simple method adopted from [1], slightly modified. The relative year-to-year derivation is calculated and then the first two autoregressive components are eliminated. Modified proxy time series are then obtained:

$$Q_t = \boxed{\frac{1}{P_t}\cdot\frac{\Delta P_t}{\Delta t}} - \boxed{a_1\cdot\frac{1}{P_{t-1}}\cdot\frac{\Delta P_{t-1}}{\Delta t}} - \boxed{a_2\cdot\frac{1}{P_{t-2}}\cdot\frac{\Delta P_{t-2}}{\Delta t}} \qquad (2)$$

This is important not only for regression calculations but particularly for cross spectrum analysis.

3. Results

3.1 Time domain

Let us begin with two response functions (Fig. 1 a, b). In order to test stationarity these relations were computed not only according to the overall period but also with respect to overlapping subperiods (not shown here). These examples show the tree growth value and the vine quality index as predictands and monthly precipitation totals as predictors. The tree growth is controlled by the actual spring and mid summer precipitation (positive correlation) while the vine quality is mainly influenced by the actual mid and late summer precipitation (negative correlation). These two effects evidently overlap in June and July.

This enables us to reverse eq. (1) and allows to estimate a so-called transfer function (Fig. 2). Thereby mid summer precipitation is used as predictand and tree growth as well as vine quality are the predictors. In addition, these two predictors have to be shifted forward for one year (for the implication of the climatic influence of the previous year). These transfer functions have been computed for overlapping subperiods, too.

The demonstrated climatic signals (the relative explained variance is somewhat lower than 50 per cent) were now used to reconstruct mid summer precipitation (Fig. 3). Here we restrict to the instrumental period (no extrapolation), furthermore here we do not divide into calibration and verification period because the moving transfer functions show sufficient stability. It is seen that the fluctuations of the measured climatic data are similar to those of the estimated values. It should be remembered that this example reflects a certain phenomenological character and is not optimal in the view of explaining most of variance.

3.2 Frequency domain

At least we consider some aspects referring to spectrum analysis. The longest available regional proxy time series is that of the Saar/Mosel tree growth (730-1975 AD). The variance spectrum is shown in Fig. (4). The origins of the two distinct peaks at 7 and 5 years are not known. Possible climatic signals are, of course, disturbed by biological influences and site conditions. The 7-year peak proved to be relatively constant with time. Fig. (5 a, b) shows two variance spectra related to the instrumental period (mid summer precipitation and tree growth); similarities can be recognized. Fig. (6 a, b) demonstrates moving cross spectra referred to mid summer precipitation, tree growth and vine quality. We show the amplitude spectra (i.e. the absolute values of the complex cross spectra, or in other words, the spectral decomposed covariance without consideration of the phase angles). These computations proved to be more stable than the often used coherence spectra. Most of the significant areas are relatively constant with time, furthermore some similarities between the two moving spectra exist.

This work was supported by the German (FRG) Government Climate Research Program.

R E F E R E N C E S :

/1/ Box, Jenkins, 1976: *Time Series Analysis. Holden-Day.*

/2/ Burkhardt, Hense, 1985: *Arch. Met. Geoph. Biocl. B, p. 341.*

/3/ Fritts, 1976: *Tree Rings and Climate. Acad. Press.*

/4/ Hollstein, 1980: *Mitteleur. Eichenchron. Zabern, Mainz.*

/5/ Rudloff, 1984: *pers. comm.*

/6/ Schweingruber, 1983: *Der Jahrring. Haupt, Bern.*

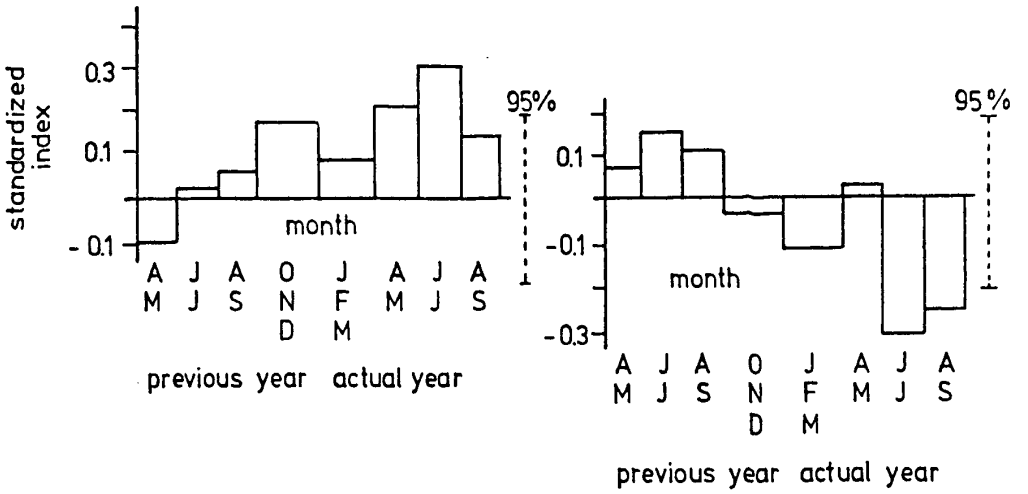


Fig. 1 a, b: Response functions, period 1850-1965
left: tree growth versus precipitation
right: vine quality versus precipitation

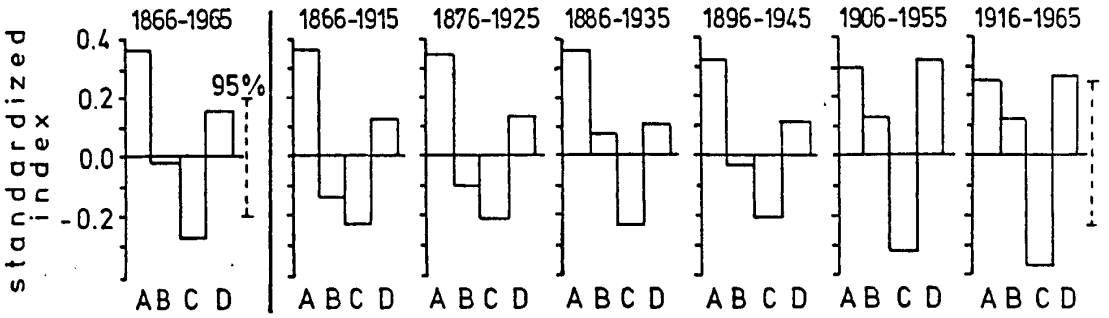


Fig. 2: Transfer functions, overall period and six moving subperiods
precipitation June/July versus
A tree growth (current) B tree growth (1 year shifted)
C vine quality (current) D vine quality (1 year shifted)

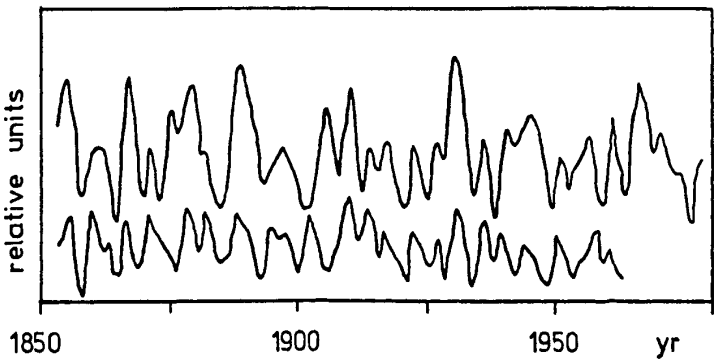


Fig. 3: Measured precipitation June/July Trier (above) and
reconstructed values (below), 5-year low-pass filter

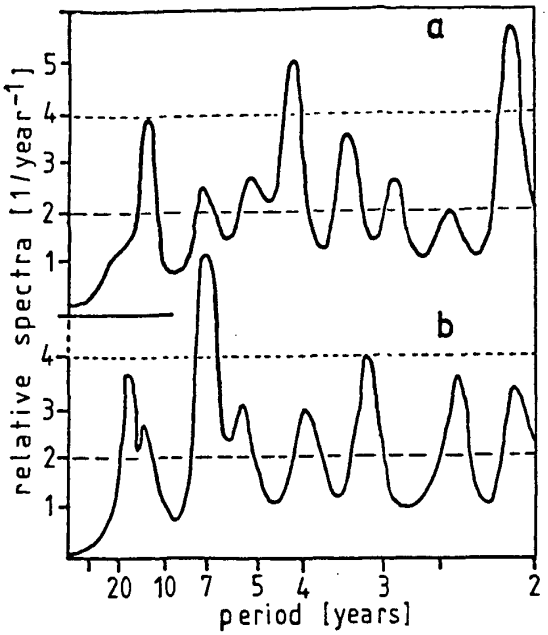


Fig. 5 a, b: Variance spectra
a precipitation June/July
b tree growth
period 1850-1965
MESA (Burg), 25 AR-coeff.
including white noise
and 95% confidence limits

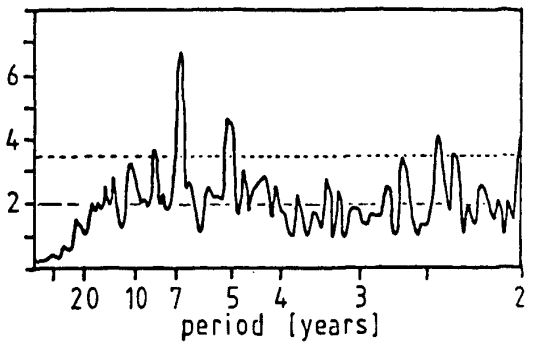


Fig. 4: Variance spectrum
tree growth, period 730-1975 AD
Blackman/Tukey, 16% max. lag
including white noise
and 95% confidence limit

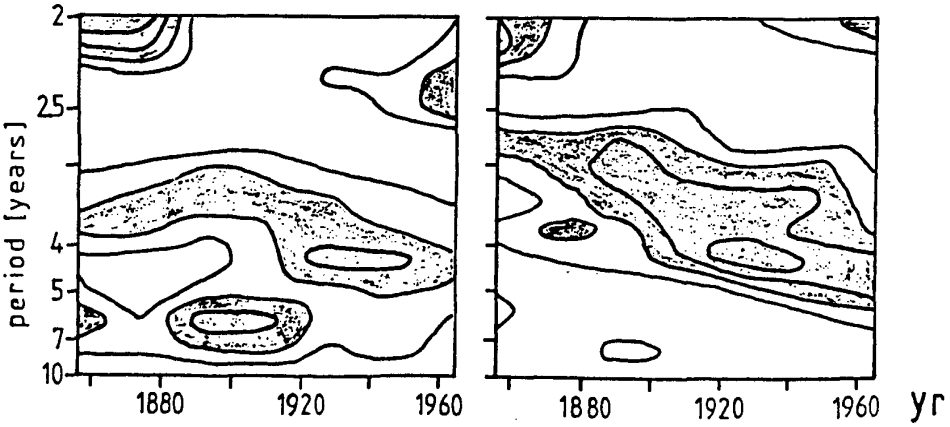


Fig. 6 a, b: Moving amplitude spectra, overall period 1850-1965
subperiods 1855-1905, 1865-1915, ..., 1915-1965
left: precipitation June/July versus tree growth
right: precipitation June/July versus vine quality
(hatched: 90% confidence area)

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

SOME STATISTICAL ASPECTS OF VOLCANISM-TEMPERATURE RELATIONSHIPS

C.-D. Schönwiese

Institute for Meteorology and Geophysics, Frankfurt University
Feldbergstr. 47, D-6000 Frankfurt a.M. 1 (FRG)

INTRODUCTION

Climate is produced by internal mechanisms and external forcing of the climate system (atmosphere, ocean, etc.). Considering external forcing, the influence of volcanism on the air temperature is one of the most significant phenomena and, in consequence, has been frequently investigated. These investigations concern both the climate modelling (e.g.(12)) and the statistical aspect (e.g. (8,9)) as well as the atmospheric boundary layer near surface (8,9) and the stratosphere (5). An increased level of volcanic activity within the recent 20-30 years, in particular the Mt. Agung (1963) and El Chichon (1982) eruptions, have triggered this scientific interest.

The most important problems arising in this context are as follows. (1) Which index ("parameter") describes correctly the year-to-year variations of volcanic activity during the recent centuries? (2) What about the time-space scale of the temperature response, especially within the atmospheric boundary layer near surface? (3) Moreover, it is necessary to account for non-stationarity effects (long-term fluctuations) of volcanism and, in turn, of this temperature response.

PARAMETERIZATION OF VOLCANIC ACTIVITY

The well-known "dust veil index" DVI of Lamb (6) parameterizes the volcanic particle loading of the stratosphere where northern and southern hemisphere averaged annual data are available. Here only the northern hemisphere data are considered, see Fig. 1, which may be correlated with corresponding temperature data time series. Because of some shortcomings (subjective index, temperature observations partly used for the index derivation) the DVI data must be compared with appropriate measurements. This may be done by means of the ice core acidity (H^+ ion concentration supposed to be proportional to the volcanic material deposited) measurements at station Crête (central Greenland) provided by Hammer et al. (2); this time series is called "acidity index" AI (Crête), data plot see (2,9).

The AI data, however, may include comparably weak volcanic eruptions which do not influence the stratosphere (and may not cool but warm the lower atmos-

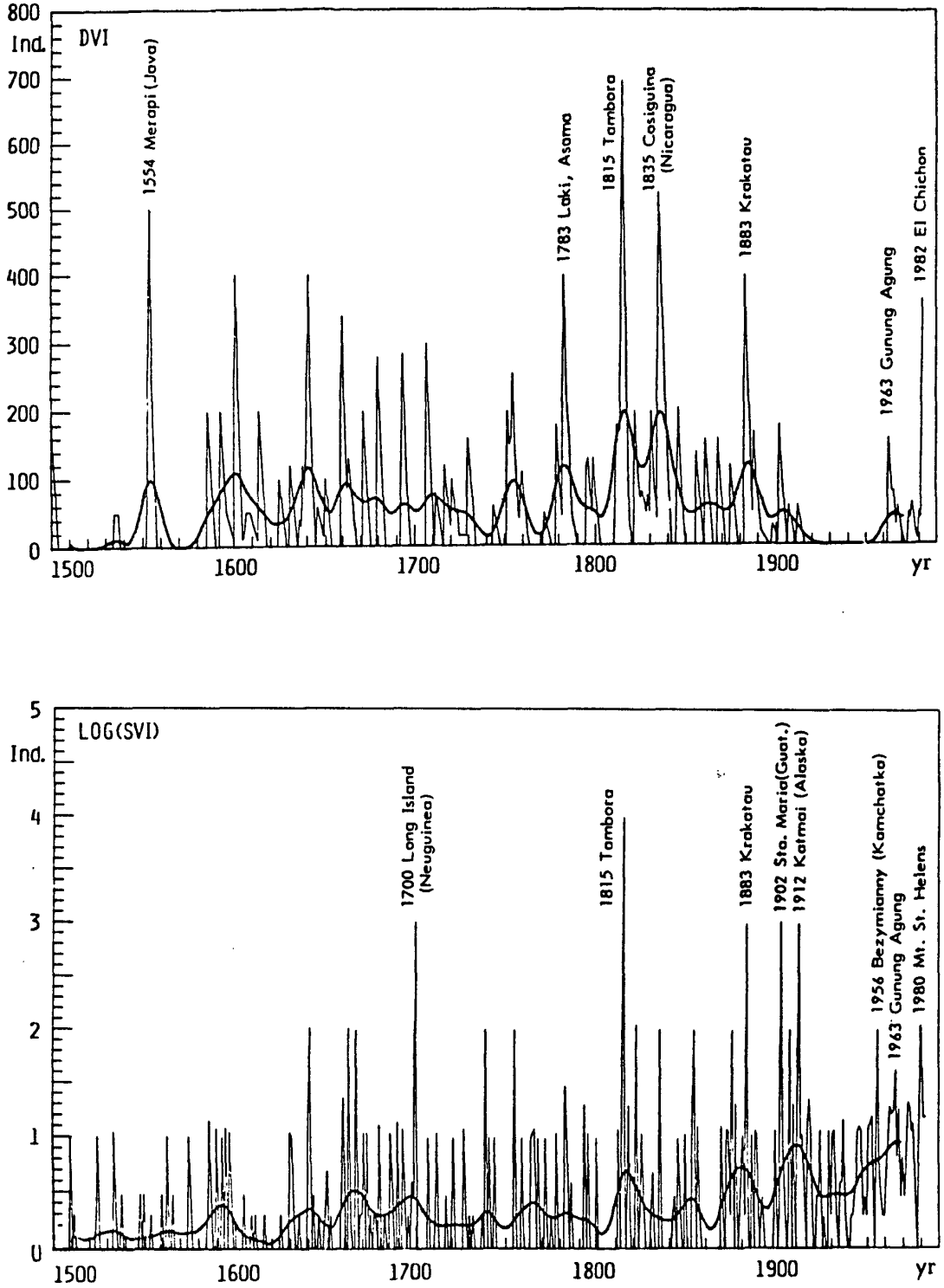


Fig. 1. Annual data of the volcanic activity index "DVI" after Lamb (6) and "SVI" (this study, see text) where some particular eruptions are specified. The smoothed heavy lines represent the relatively long-term fluctuations (Gaussian low-pass filter suppressing periods T < 10 years); SVI scale logarithmic. "AI" (Crête ice core acidity index) plots see (3,9,10).

phere) or even material which is not really volcanogenic. Therefore a third volcanic activity index is used based on the volcanoe chronology of Simkin et al. (11), Smithsonian Institution. This chronology specifies the volcanoe eruptions (all over the world) by means of the "volcanoe explosivity index" VEI covering numbers 0 to 8. (Since 1500 AD VEI=8 did not appear and VEI=7 was attributed only to the Tambora eruption in 1815.) Newhall and Self (7) recommend to use in climatology only VEI ≥ 3 values because VEI < 3 does mean eruption heights < 5 km. It is problematic, however, that VEI=3 covers eruption heights 3 - 15 km which may influence the stratosphere or not (VEI=4: 10-25 km, etc.). In this study an index is used (10),

$$SVI = \sum_j (10^{VEI})_j \quad ,$$

where the sum j=1,...,J means accumulation for every year (and for northern hemisphere assessments 90° N to 10° S). This "Smithsonian volcanic activity index" (1,10) is proportional to the volume of the volcanic material erupted. The data (since 1500 AD) time series is shown in Fig. 1 where the annual data (also DVI) are supplemented by Gaussian low-pass filtered data suppressing periodes $T < 10$ years (9-term filter) and representig, therefore, the long-term fluctuations.

ASSESSMENT OF THE TEMPERATURE RESPONSE

The assessment of volcanism-temperature relationships may be obtained by means of the usual methods of time series analysis or one may analyse particular events (superposed epoch method, see recent study of Angell and Korshover (1)). Furthermore, one may combine these techniques using moving statistics (moving correlation and signal analysis) which takes account for the fact that volcanism is a phenomenon varying considerably in time.

First, in Table 1 some time series correlation coefficients are listed up concerning regional or local air temperature variations and the volcanic activity indices

Table 1: Annual data correlations of some air temperature and volcanic index time series (abbreviations DVI and SVI as well as selected plots see Fig. 1 and 3). In addtion, related maximum temperature signals and observed temperature series standard deviations s (in Kelvin). The numbers in parantheses are based on low-pass filtered data suppressing periods $T < 10$ years. Correlation coefficients which exceed the 95% confidence level are underlined, heavy in case of 99% (autocorrelation implied).

region or station	time interval	correlations DVI	correlations SVI	signals(K) DVI	signals(K) SVI	s(K)
north. hemisphere*	1851 - 1984	-.27(-.62)	-.17(-.34)	-.38(-.57)	-.28(-.26)	.25
	1781 - 1984	-.34(-.51)	-.12(-.38)	-.62(-.57)	-.47(-.70)	.28
arctic (65-85° N)*	1851 - 1984	-.28(-.64)	-.18(-.34)	-.95(-1.37)	-.73(-.61)	.60
Philadelphia	1738 - 1984	-.32(-.48)	-.11(-.26)	-1.24(-.96)	-.99(-.91)	.57
Hohenpeissenberg	1781 - 1984	-.12(-.30)	-.06(-.22)	-.43(-.56)	-.63(-.69)	.76

* regional mean, data from Jones (4), before 1851 from Groveman and Landsberg(2).

DVI and SVI. With the exception of Hohenpeissenberg (FRG) the correlation coefficients exceed at least the 95 % confidence level (corresponding .5 error probability), but the SVI correlations are considerably lower which could be interpreted in the way that the DVI parameter possibly overestimates the volcanism-temperature relationships. In case of low-pass filtered data the correlation increases but also the time series autocorrelations so that the 95 % confidence level is not exceeded. Nevertheless, the correlations are established predominantly within the long-term domain of the variance spectrum which can be demonstrated by means of coherence analysis, see Fig. 2.

The maximum temperature signals computed from the regression equations confirm the assumption that DVI may overestimate the temperature response. In respect to the northern hemsiphere mean, for example, the number of -.28 K (reference period 1851-1984) is quite similar to -.31 K found by Angell and Korshover (1) for the Mt. Agung (1963) event. When the reference period 1781-1984 is considered, however, this signal amplifies (-.47 K, whereas Angell and Korshover estimated a maximum signal of not more than -.32 K, related to the Coseguina eruption in 1835). Note that the signals computed from the low-pass filtered data regression equations are not necessarily higher than in case of annual (unfiltered) data and that the ratio of the signals compared with the observational standard deviation in general is > 1 (with the exception of Hohenpeissenberg), for example 1.68 in respect to the northern hemisphere data 1781-1984 or even 1.74 in respect to Philadelphia (SVI, annual unfiltered data). The graphs plotted in Fig. 3 illustrate, for example, the low temperature level in the year of the Tambora eruption (only two temperature records selected here).

A moving correlation and signal analysis may be helpful for a more detailed analysis. In Fig. 4, related to the northern hemisphere average, 100 yr subinterval correlations (1 yr steps) with the DVI, SVI and AI parameters are presented. There

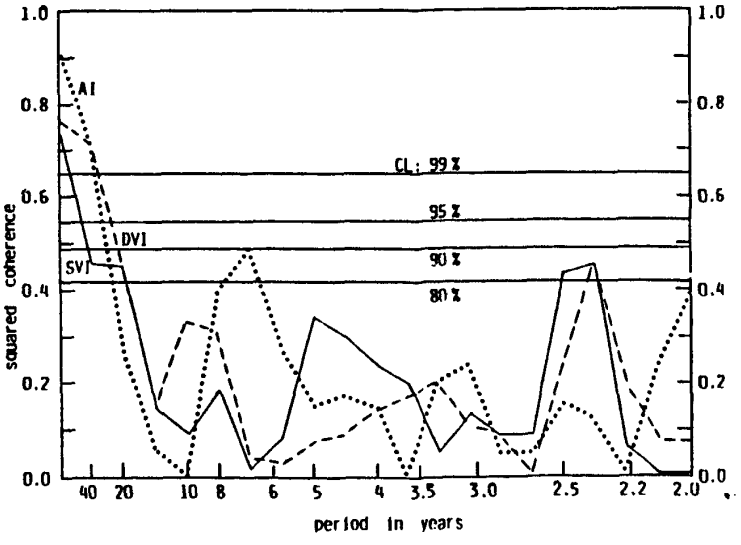
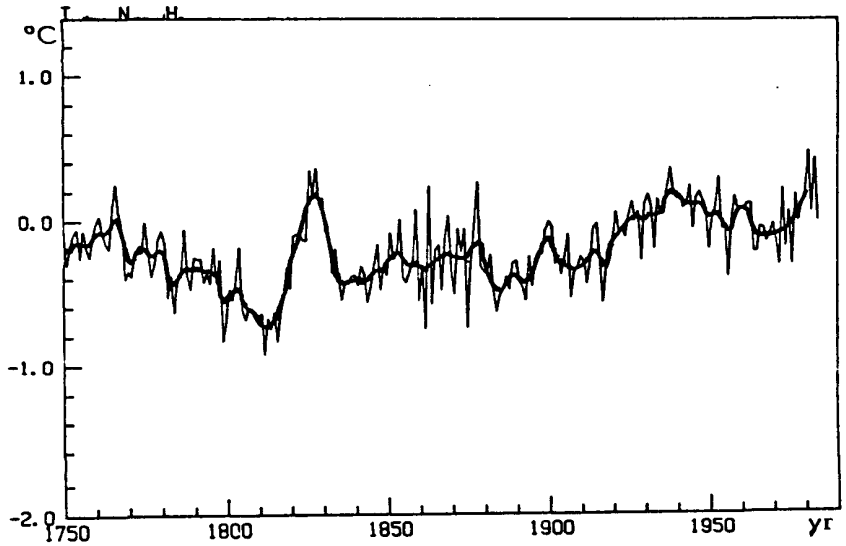


Fig. 2. Squared coherence spectrum of annual data of the northern hemisphere average temperatures and the volcanic activity indices AI, DVI and SVI. Reference period 1881-1980, M=20 lags. Only the long-term component exceeds the 95 % and 99 % confidence levels CL.



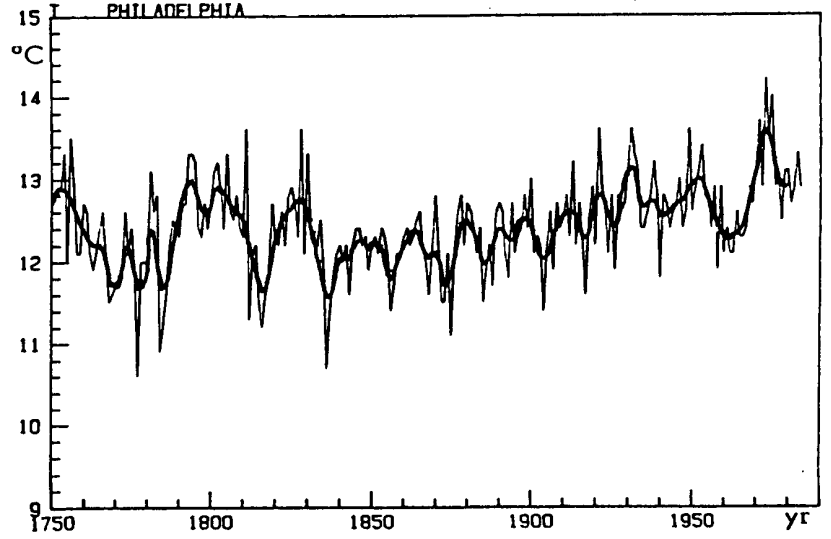


Fig. 3. Two examples of annual temperature time series: northern hemisphere average (deviations from the reference period 1951-1970), left-hand (upper plot), and Philadelphia, right-hand. The smoothed heavy lines represent low-pass filtered data similar to Fig. 1.

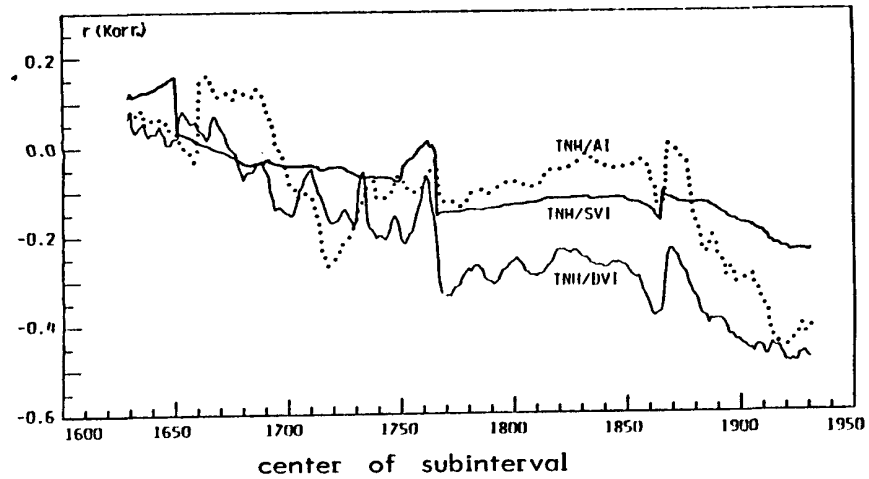


Fig. 4. Moving correlation coefficients, 100 yr subintervals and 1 yr steps, of northern hemisphere average annual temperatures (TNH) and the volcanic activity indices AI, DVI and SVI. Note the inhomogeneity c. 1770 (subinterval center) and the increasing trends within the recent c. 100 yr, in particular in case of TNH/AI.

is a evident inhomogeneity in c. 1770 (center of subinterval) which may be seen to be one more sign supposing that the temperature reconstructions (2) before that time are suspect. Later on, the correlation level TNH/SVI is relatively stable in time and more similar to TNH/AI than to TNH/DVI. Note that since c. 1880 (center of subinterval) there is an outstanding TNH/AI correlation increase which may be point to a systematic error (industrial acid deposition?).

Finally, in Fig. 5 moving 30 yr subinterval signals are shown representing the following examples: northern hemisphere average, Philadelphia and Hohenpeissenberg temperature records. Similar to the correlation coefficients, also the signals vary considerably with time. In this view the efficiency of particular volcanic activity episodes within different regions are reflected.

REFERENCES

(1) Angell, J.K. and Korshover, J., 1985: J.Clim.Appl.Meteorol., 24, 937-951.
(2) Groveman, B.S. and Landsberg, H.E., 1979: Publ. No. 79181/182, Dep. of Meteorol., Univ. of Maryland.
(3) Hammer, C.U. et al., 1980: Nature, 288, 230-235; priv. comm., 1983.
(4) Jones, P.D., 1985: Climat. Monitor, 14, 14-21; 43-50.
(5) Labitzke, K. and Naujokat, B., 1983: Contr.Atm.Physics, 56, 495-507.
(6) Lamb, H.H., 1970: Philos.Transactions R.Met.Soc. A 266, 425-533; priv. comm., 1983.
(7) Newhall, C.G. and Self, S.: J.Geophys.Res., 87, 1231-1238.
(8) Schönwiese, C.D., 1983: Arch.Met.Geoph.Biocl., Ser. B, 32, 337-360.
(9) - , 1984: Arch.Met.Geoph.Biocl., Ser. B, 35, 155-178.
(10) - , 1986: Meteorol.Rdsch., in print; see also Bissolli, P., 1985: Dipl. thesis, Inst. Meteorol. Geophys., Univ. of Frankfurt (FRG).
(11) Simkin, T., et al., 1981: Volcanoes of the World. Academic Press, New York.
(12) Toon, O.B. and Pollack, J.B., 1980: Amer.Sci., 68, 268-278.

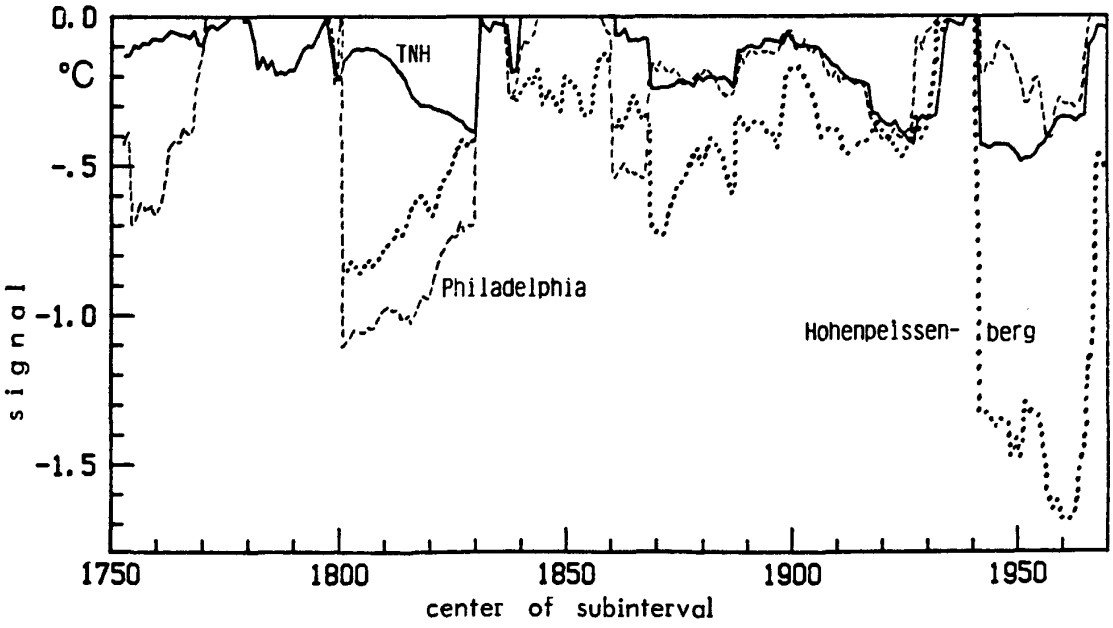


Fig. 5. Moving signal assessments (°C), 30 yr subintervals and 1 yr steps, of some annual temperature records and the volcanic activity index SVI. The maximum signals are (correlation coefficients in parantheses): TNH (compare Fig. 4) -.39 K (r=-.24), Philadelphia -1.09 K (r=-.35) and Hohenpeissenberg -1.69 K (r=-.48).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

A POTENTIAL PREDICTABILITY STUDY WITH A GENERAL CIRCULATION MODEL

F.W. Zwiers
Canadian Climate Centre
Downsview, Ontario
CANADA, M3H 5T4

1. INTRODUCTION

A potential predictability study is basically an analysis of variance conducted in the time domain. Its purpose is to determine whether the observed inter-annual variability of seasonal or monthly means is significantly larger than would be expected if the only source of variation in time averages is daily fluctuations in the observed time series. The literature now contains several reports of potential predictability studies of the atmosphere including Madden (1976), Shukla and Gutzler (1983), Trenberth (1985) and Chervin (1986). There have been no comparable studies of simulated climates. However, there have been a several studies of the inter-annual variability of climates simulated with atmospheric general circulation models (GCM's) including those of Manabe and Hahn (1981) and Lau (1981) who report on a 15 year climate simulation with the GFDL model and Volmer et al. (1983) who report on a 6 year climate simulation with the ECMWF model. Boer et al. (1984b) describe the climatology, but not the inter-annual variability, of the climate simulated with the Canadian Climate Centre (CCC) GCM as obtained from a 5 year simulation. The study of Chervin (1986) might also indirectly fall into this category. The main contribution of these studies is to illustrate that even in the absence of 'external forcing', such as provided by the oceans or cryosphere, atmospheric GCM's realistically simulate a substantial amount of inter-annual variability. However, these studies do not address the question of whether or not the inter-annual variability observed in the simulated climates is potentially predictable. The purpose of this paper is to address this question with the results of a 20.5 year integration of the CCC-GCM.

The CCC-GCM has been described in detail by Boer et al. (1984a, b) and thus only a brief summary of its features is given here. The model is a 10 level, global, primitive equations, spectral model which contains radiative processes including a diurnal cycle, a surface heat budget and a hydrological cycle including surface hydrology, soil moisture, and snow amounts. The sea-surface temperatures in the model are fixed as are the zonally averaged location and amount of clouds. The model is integrated numerically using a semi-implicit time stepping method with time step of 30 minutes. The integration was initiated from FGGE Jan 1, 1979 conditions and observations were saved every 18 hours. The analysis described in this paper is based on the 20 years following June 1 of the 1st year of integration.

2. METHODOLOGY

While there are some differences in the methodologies used in the studies which have appeared in the literature, all are predicated on the same set of basic assumptions. The idea which motivates these studies is that inter-annual variability of seasonal means derives from short time scale fluctuations which are inherently unpredictable and longer time scale variations which are potentially predictable because they may be associated with sources of 'external forcing', such as sea surface temperature anomalies. Madden (1976) refers to the inter-annual variability induced by short time scale variations as natural variability. We will refer to variability from this source as climate noise. Most studies have attempted to link any perceived potential predictability to long time scale quasi-periodic phenomena such as the Quasi-Biennial Oscillation (QBO) or the occurrence of El-Nino Southern Oscillation (ENSO) events.

In order to detect potential predictability it is assumed that:

(i) potentially predictable signals occur on time scales of a year or more, and that over shorter time scales, such as a season, observed time series are more or less stationary after adjustments have been for the annual cycle.

(ii) the potentially predictable signal, if it exists, does not interact with variations occurring on shorter time scales.

(iii) in the absence of a potentially predictable signal the observed ensemble of time series for a particular winter represents a random sample taken from the collection of all realisations of time series with the same statistical properties.

The mathematical model which expresses these assumptions is simply

$$Y_{ij} = U_j + \epsilon_{ij}, \quad i=1, \ldots n, \quad j=1, \ldots m \tag{1}$$

where index i runs over the observing times within a season and index j runs over years. The time series ϵ_{ij} represents some sort of noise process which has relatively short memory and whose realizations are independent from year to year and independent of any potentially predictable signal that might be present. The sequence of coefficients U_j, $j=1, \ldots m$, represents the potentially predictable signal and may be either random or deterministic. Except for the fact that the errors ϵ_{ij} have some sort of short term memory, this model is identical to that used in a standard 'one way analysis of variance' (see, for example, Sachs (1982)).

With the model and the assumptions made above, we see that the inter-annual variance of an N day seasonal mean

$$\bar{Y}_{\cdot j} = \sum_{i=1}^{N} Y_{ij}/N$$

is given by

$$\sigma^2_{\bar{Y}} = \sigma^2_U + \sigma^2_{\bar{\epsilon}} \tag{2}$$

and is estimated by

$$S^2_{\bar{Y}} = \sum_{j=1}^{m} (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2 / (m-1) \tag{3}$$

This estimate has m-1 degrees of freedom (df).

In order to test whether there is a significant contribution to this variance from potentially predictable sources we need an independent estimate of the variance of the time average of the noise process. The technique which we use for doing this was originally suggested by Jones (1975, 1976) and is the same as that used by Madden (1976). It is derived by regarding a time average as the output of a running mean filter. Thus the variance of $\bar{\epsilon}_{\cdot j}$ is given by

$$\sigma^2_{\bar{\epsilon}} = \int_{-\pi}^{\pi} |B(\lambda)|^2 f_{\epsilon\epsilon}(\lambda)\, d\lambda \tag{4}$$

where $B(\lambda)$ represents the transfer function of the running mean filter and $f_{\epsilon\epsilon}(\lambda)$ is the spectral density function of the noise. The gain of the filter, $|B(\lambda)|^2$, is given by

$$|B(\lambda)|^2 = \sin^2(N\lambda/2) \,/\, (N \sin^2(\lambda/2)) \tag{5}$$

For each season, the climate signal is estimated as

$$\hat{U}_j = \bar{Y}_{\cdot j} \tag{6}$$

and is subtracted from the observations Y_{ij} to produce an estimate of the noise sequence

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{U}_j \tag{7}$$

The spectral density function of the noise sequence is estimated by means of the periodogram so that

$$\hat{f}_{\varepsilon\varepsilon}(\lambda) = \left| \sum_{t=1}^{n} e^{i\lambda t} \hat{\varepsilon}_{tj} \right|^2 / 2\pi n$$

at non-zero frequences $\lambda = 2\pi i/n$, $i=\pm1, \pm2, \ldots, \pm n/2$. This estimate is extrapolated to the origin by applying the Low Frequency White Noise (LFWN) extrapolation

$$\hat{f}_{\varepsilon\varepsilon}(o) = \hat{f}_{\varepsilon\varepsilon}(2\pi/n) \tag{9}$$

Substituting the estimated spectral density function into (4), setting N=n so that the time average is taken over the entire period of observation, and approximating (4) with a sum results in

$$\hat{\sigma}^2_{\bar{\varepsilon}} = \hat{f}_{\varepsilon\varepsilon}(o) \,/n \tag{10}$$

as an estimate of variance of the time average of the noise process which has 2 degrees of freedom (df). Such an estimate is computed for each of the m years during which the climate is observed, and these estimates are pooled to form a single estimate of the variance due to climate noise, say $S^2_{\bar{\varepsilon}}$, which has 2m df.

The estimate of the variance due to climate noise has the property that it is asymptotically independent of the estimates U_j of the climate signal and hence of the estimate (3) of inter-annual variance. Therefore, when the null hypothesis that there is no climate signal is correct, the F-ratio

$$F = S^2_{\bar{y}} / S^2_{\bar{\varepsilon}}$$

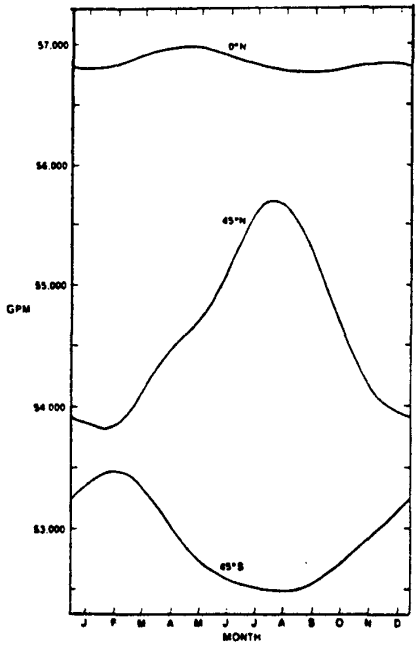
will be asymptotically distributed as an F random variable with (m-1) and 2m degrees of freedom. The null hypothesis is rejected at the 5% significance level when F is greater than $F_{0.05,\ m-1,\ 2m}$. In our case m=20 and the null hypothesis is rejected when F, the ratio of the estimated variance due to climate signal plus noise to the estimated variance due only to climate noise alone, is greater than 1.86. Monte Carlo experiments described by Zwiers (1986b) reveal that the true significance level of this test will be very close to the asymptotic significance level except where the observations have large persistence times in which case the climate noise is underestimated and the test becomes too liberal.

In order to apply the technique it is necessary that the time series Y_{ij} be approximately stationary, at least within a season. It is therefore necessary to standardize the observed 20 year time series by removing the annual cycle of the mean and dividing by the annual cycle of the standard deviation. The resulting 20 year time series is presumed to be more or less stationary, at least within seasons. The annual cycles are estimated using the same techniques used by Madden (1976). The annual cycle of the mean and variance are represented by a truncated Fourier series consisting of the first four harmonics with decreasing weights of 0.75, 0.5 and 0.25 placed on the last three harmonics. Figures 1 and 2 contain examples of the estimated annual cycle which are derived from our 20 years simulation.

3. RESULTS

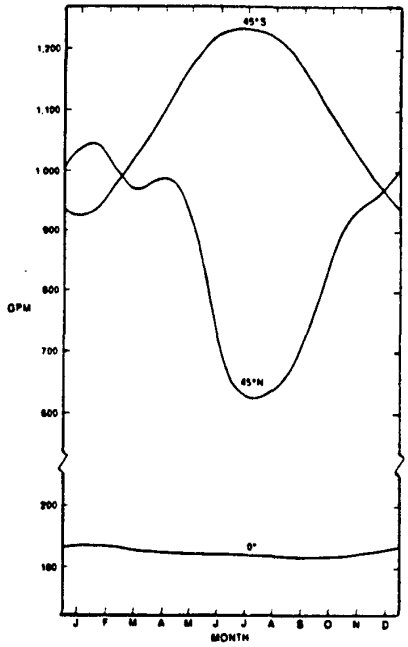
The results of our study for DJF are displayed in Figures 3 and 4 and are summarized in Table 1. Figure 3 displays the ratio of inter-annual variance to climate noise for surface pressure and regions in which this ratio is significantly different from one at the 10% significance level for DJF. Locations at which this ratio is significantly greater than one at the 5% significance level, that is, locations demonstrating significant evidence of potential predictability, are shaded. Locations at which the ratio is

Figure 1



Estimated annual cycle of the mean of 500 mb height at 45°N, 0° and 45°S on the Greenwich meridian.

Figure 2



Estimated annual cycle of the standard deviation of 500 mb height at 45°N, 0° and 45°S on the Greenwich meridian.

Table 1

	Proportion of Globe Significant	
Season	Surface Pressure	500mb Height
DJF	0.478	0.467
MAM	0.242	0.104
JJA	0.090	0.115
SON	0.103	0.098

Table 1 Proportion of F-ratios significantly different from one at the 10% significance level.

significantly less than one at the 5% significance level are cross hatched. When the null hypothesis that there is no potential predictability is true one would expect to see about 5% of the globe shaded and about 5% cross hatched if the estimate of climate noise is not greatly biased by the effects of serial correlation as was discussed above. The corresponding map for 500 mb height is displayed in Figure 4. Table 1 contains a summary of the proportion of the globe over which the null hypothesis is rejected.

In JJA and SON there is no evidence of potential predictability, either in the model's surface pressure field or its 500mb height field. In JJA the null hypothesis is rejected over approximately 9% of the globe for surface pressure and approximately 11% of the globe for 500 mb height. In the latter case there are approximately equal areas in which the F-ratio is significantly greater or smaller than one. In the former there is a tendency for the F-ratio to be less than one more frequently than expected indicating that the annual cycle of surface pressure may be somewhat more complex than is resolvable by the truncated Fourier series representation we have adopted. In SON the null hypothesis is rejected over approximately 10% of the globe for both fields. Again, there is a tendency for the F-ratio to be less than one more frequently than expected in the case of surface pressure.

In DJF there appears to be very strong evidence for potential predictability. The null hypothesis is rejected over approximately 48% and 47% of the globe for surface pressure and 500 mb height respectively. Significantly large F-ratios are seen in the tropics and the high latitudes of the SH. Indeed, the patterns seen in the SH are similar to those observed by Trenberth (1985). Moreover, the meridional structure of the computed F-ratios is consistent with that which has been observed in the atmosphere: the smallest F-ratios are seen at mid-latitudes in both hemispheres. Also, as in the atmospheric studies described above, the largest F-ratios are seen in DJF. We see much less evidence for potential predictability in the NH than in the SH. Although a global test of significance (Zwiers (1986a)) cannot be applied in the present case, the structure and large scale of the results would seem to indicate that they are globally significant.

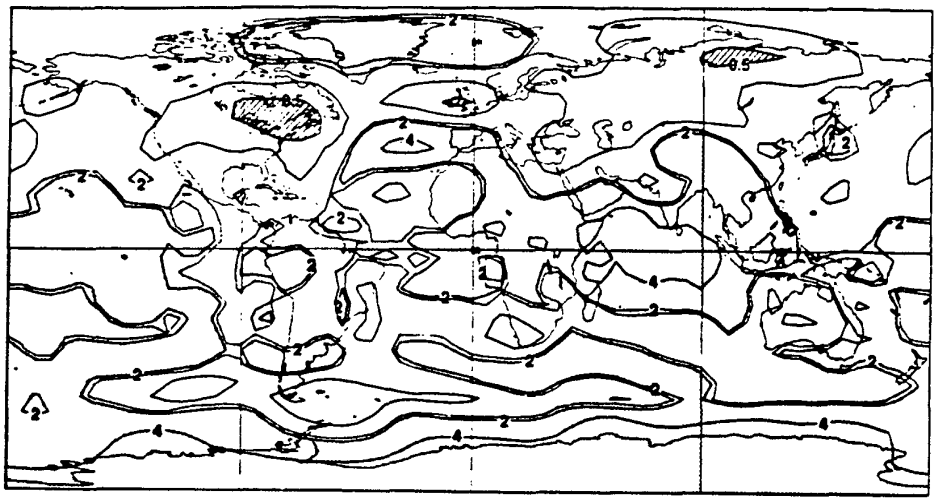
There is also some evidence for potential predictability in MAM. In the case of surface pressure the null hypothesis is rejected over 24% of the globe, primarily in the tropics and polar regions. There is weaker evidence for potential predictability in the 500 mb height field. The null hypothesis was rejected over 10% of the globe (equal to the significance level of the two-sided F-test) but almost all departures were associated with values of the F-ratio which are greater than one.

Potential predictability observed in the atmosphere has been linked with such long time scale phenomena as the QBO and ENSO. Such quasi-regular variability does not exist in climates simulated by atmospheric GCM's. Sea surface temperatures follow a prescribed annual cycle determined from climatology, and we do not observe the more or less regular reversal of stratospheric winds associated with the QBO. The only potential long time scale feedback mechanism which is modeled in the CCC-GCM involves its representation of surface hydrology. However, a temporal analysis of the model's surface moisture fields does not reveal the presence of any trends, periodicities, or unusually large or extended duration anomalies which might in turn impact upon surface pressure or height fields. Moreover, the potential predictability which we observed is largely a SH and tropical region phenomenon: areas in which surface hydrology does not play a large role in determining local climate.

In order to understand the nature of the phenomenon which lead to the conclusion that there is significant evidence for potential predictability in the climate of the CCC-GCM, we plotted a sequence of seasonal mean anomaly maps for surface pressure and 500mb height. These maps showed that by far the largest contribution to inter-annual variance of these fields in winter and spring comes from a large-scale anomaly in surface pressure and 500 mb height which occurred during the DJF season between the 6th and 7th year of the run and the MAM season of the 7th year and extended over a period about a season in length. Surface pressures at high latitudes in the SH begin to rise steadily about half way through the DJF season and reach a maximum shortly before the end of this season. They gradually decay to normal values during the subsequent MAM season. Figure 5 shows the effect of the extended anomaly on the mean surface pressure for this DJF season. We see that its structure is essentially zonal. There is a large shift of mass from the tropics and mid-latitudes to the high latitudes of the SH. Surface pressure over Antarctica is as much as 20 mb higher than normal, the corresponding mean 500 mb height is as much as 180 meters greater than normal and the the circumpolar trough is not as deep as normal. Figure 6 displays the same anomaly in units of inter-annual standard deviations. We see that there are larger regions in the tropics where the loss of mass to the SH amounts to more than two inter-annual standard deviations and that there are corresponding regions at high latitudes in the SH where the gain in mass is more than two standard deviations. This one anomaly contributes to more than 20% of the inter-annual variance at these locations.

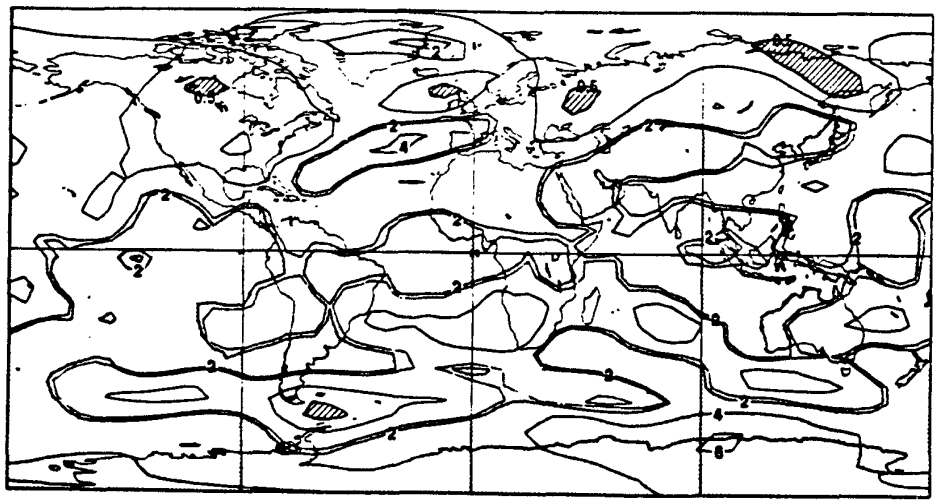
The first principal component of the zonally averaged seasonal mean surface pressure anomaly time series shown in Figure 7 provides a look at the temporal evolution of surface pressure during the 20 year climate simulation. The first principal component is the coefficient time series of the first empirical orthogonal function (eof) (also shown in Figure 7) of zonally averaged surface pressure anomalies. The latter explains 50% of the variance of these anomalies. The episode during the winter between years 6 and 7 of the climate simulation is clearly evident as is a similar event of smaller amplitude during the winter between years 8 and 9 of the simulation. The first principal component time series also suggests that for some reason, the model's climate became somewhat less variable during the latter half of the 20 year integration and that there was a shift in the mean of the principal mode of variation after about 9 years of integration. As far as can be ascertained, these are natural fluctuations which occurred within the simulated climate and are not due to changes in external conditions. Every effort was made to ensure that

Figure 3



The ratio of estimated inter-annual variance to climate noise for standardized surface pressure. Regions in which the ratio is significantly greater than one at the 5% significance level are shaded, and regions in which the ration is significantly less than one at the 5% significance level are cross hatched.

Figure 4

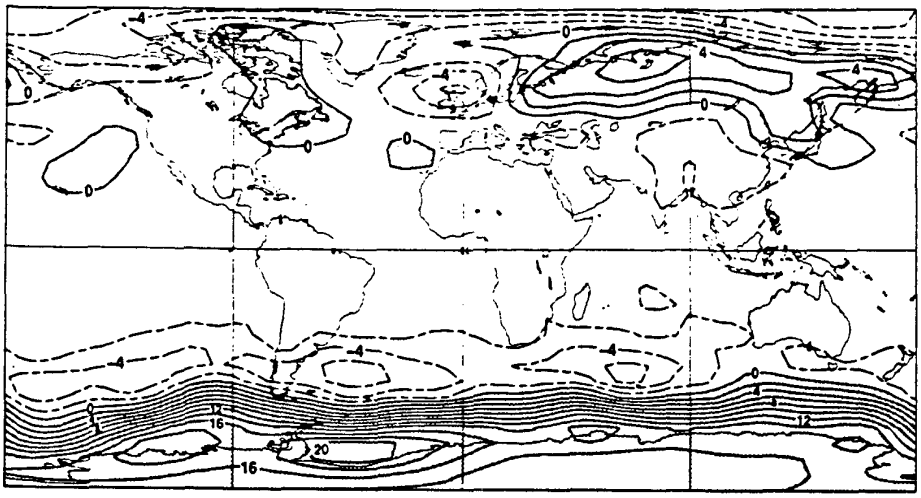


as Figure 3, except for 500 mb height.

the computing environment remained stable over the period in which the integration took place, and no changes were made either to the model or its supporting diagnostic and data management software. Standard diagnostics of the model's energetics during the anomalous periods show that the model remained numerically stable during these periods. Indeed, the model remained numerically stable during the entire 20 year integration.

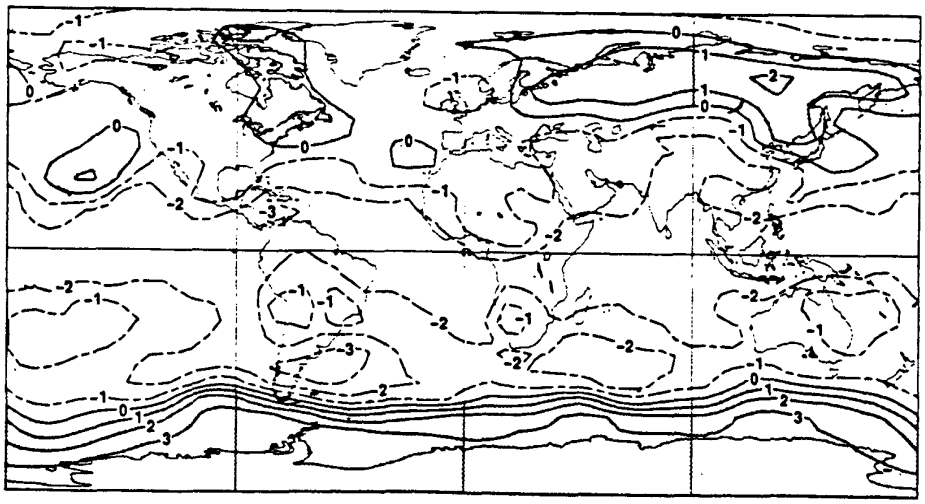
Our evidence, and that of other long simulations indicates that it is not unusual for an atmospheric GCM to display large variability on long time and space scales. Volmer et al. (1983), also observed a large anomaly of long duration in the same season over low-latitudes and the SH similar to the one which we have observed but of opposite sign. The principal component time series displayed by Lau (1981, see Fig. 12) also show occasional large anomalies and Lau's Fig. 26 of a SH look at the first EOF of daily variation of 500 mb height in NH winter hints that the principal mode of variation for seasonal anomalies in the GFDL model may not be much different from ours.

Figure 5



Mean surface pressure anomaly for the DJF season between years 6 and 7 of the 20.5 year integration of the CCC-GCM. The contour interval is 2 mb.

Figure 6

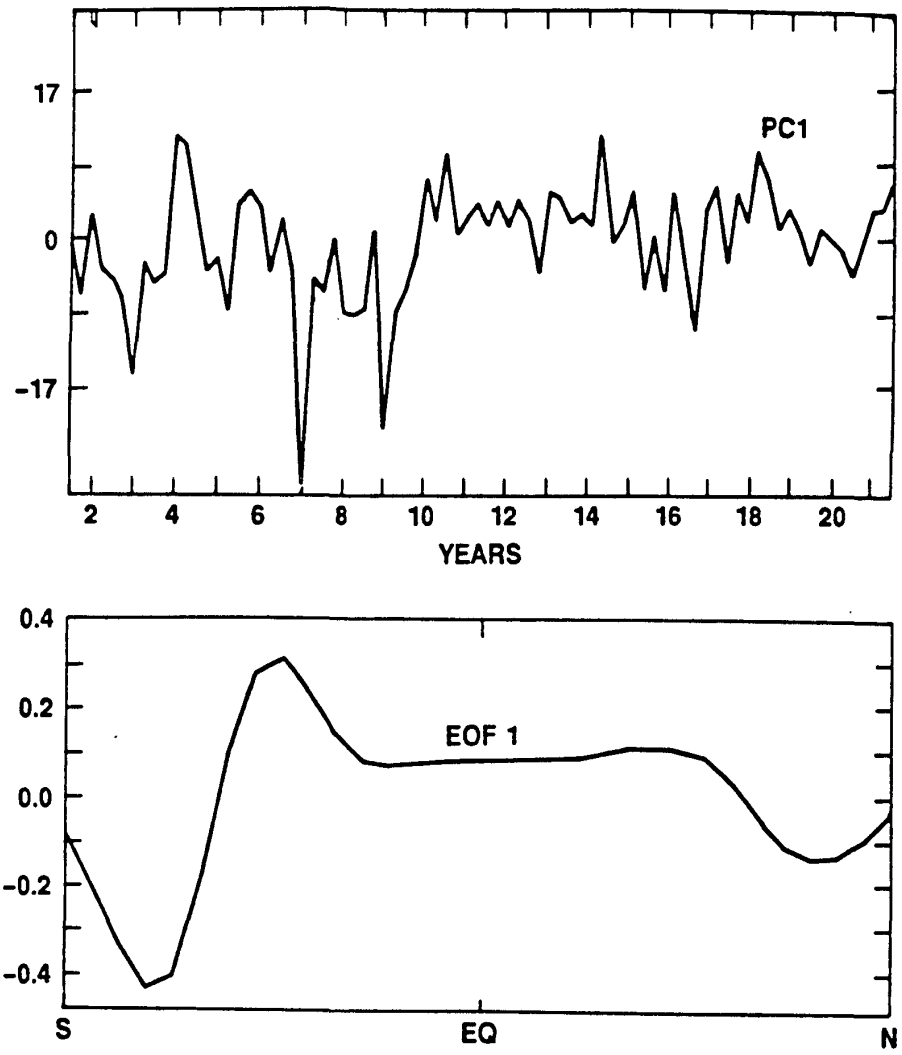


Mean surface pressure anomaly for the DJF season between years 6 and 7 of the 20.5 year integration of the CCC-GCM expressed in units of inter-annual standard deviation.

4. DISCUSSION

Studies of the potential predictability of the atmosphere which use techniques similar to those used in this study have been published by Madden (1976), Shukla and Gutzler (1983), Trenberth (1985) and others. These studies show, not surprisingly, that there is more evidence for a potentially predictability signal in the real atmosphere than in the climate simulated by the CCC-GCM. They also show that there is evidence for potential predictability in the real atmosphere in all seasons. On the other hand, our study reveals that the F-ratios computed for the simulated climate have the same general character as those computed for the atmosphere. F-ratios are larger in the SH than in the NH, and larger in the tropics and near the poles than at mid-latitudes.

Figure 7



The first principal component time series of zonally averaged seasonal anomalies of surface pressure together with the corresponding empirical orthogonal function

The differences between atmospheric studies and the present study are not at all surprising because the atmosphere experiences forcing from external sources which our simulated atmosphere does not experience. In simple terms, total interannual variance can be partitioned as in (2), so that

$$\sigma_T^2 = \sigma_S^2 + \sigma_N^2$$

where σ^2_T represents total variance, σ^2_S represents variance due to the climate signal and σ^2_N represents variance due to climate noise. The variance due to the climate signal can be further partitioned as

$$\sigma_S^2 = \sigma_{S(I)}^2 + \sigma_{S(B)}^2$$

where $\sigma^2_{S(B)}$ represents variance in the climate signal arising from external forcing and $\sigma^2_{S(I)}$ represents variance in the climate signal arising from internal dynamics. In the simulated climate $\sigma^2_{S(B)}$ is zero. Thus our results suggest, at least for the DJF season, that interannual variability arising from internal dynamics may be at least as large as that arising from climate noise in the tropics and high latitudes of the southern hemisphere.

Recent studies by Chervin (1986) and Lau (1986) provide hints about what the relative importance of the three sources of interannual variability might be in the real atmosphere. Chervin (1986) uses a 20 year integration of a version of the NCAR community climate model which does not contain interactive representations of soil hydrology or snow cover to estimate interannual variability due internal dynamics plus climate noise for mean sea level pressure and 700 mb height. He then compares this estimate with an estimate of total inter-annual variance derived from observations. Implicit in the calculation is the assumption that the modified NCAR model correctly simulates interannual variability from climate noise and internal dynamics. Chervin's diagrams only show results for the NH. They do not contain much evidence for significant interannual variance due to external forcing in the extra-tropics. However, the diagrams do suggest that external forcing is a significant source of inter-annual variance in the tropics.

The recent study by Lau (1986) examines inter-annual variability in the simulated climate when observed SSTs are used as lower boundary conditions. Lau (1986) reports on two 15 year integrations of the GFDL model where the 1962 to 1976 observed SSTs in the 30°S to 30°N band are prescribed. SSTs outside this region are the usual climatological values. While the primary purpose of the study was to examine the response of a simulated climate to observed SST anomalies, taking into account the form of onset and decay of anomalies and interactions with the annual cycle, Lau also compared inter-annual variability of 1000 mb and 200 mb height in these runs with that simulated in a 'control' run having climatological SSTs. The control run is the 15 year climate simulation described by Manabe and Hahn (1981) and Lau (1981). Results are only shown for NH winter (see Fig. 16, Lau (1986)). F-Ratios for both 1000 mb and 200 mb GPH are generally greater than one and the pattern of enhanced variability at 1000 mb is clearly linked with the ENSO phenomenon. Enhanced variability is apparent on a much broader scale at 200 mb, particularly in the tropics where F-ratios are significantly larger than one in the 20°S to 20°N latitude band between 60°E and 60°W. As is the case for 700 mb heights in Chervin (1986), the F-ratio is significantly large at only a few locations in the NH north of 20°N.

The work of Chervin and Lau suggest that interannual variability arising from boundary forcing may be at least as large as that arising from the other two sources in the tropics. That is, boundary forcing may account for 50% of inter-annual variability in the tropics. Our study suggests that internal dynamics might contribute 50% of the remaining interannual variability. So a very crude analysis of variance suggests that in the tropics something like 50% of variance might be due to boundary forcing, maybe 25% due to internal dynamics, and the remainder due to climate noise. In the high latitudes of the SH internal dynamics may be responsible for more than 50% of interannual variance at certain times of year.

The suggestion is that internal dynamics makes a considerable contribution to interannual variance which is in excess of climate noise. The question of whether any of the inter-annual variation arising from the climate signal due to internal dynamics is actually predictable is still open. Further work is required to determine (i) whether there are certain climate regimes, such as that prevailing during the large anomaly which was observed in the CCC-GCM simulated climate, during which there is a significant amount of predictability on a seasonal basis; (ii) the frequency with which such regimes occur; and (iii) whether such regimes can be identified so that useful forecasts with GCMs can be made.

The results of this study may be summarized as follows: (i) atmospheric GCMs with a prescribed annual cycle of SSTs simulate large amounts of inter-annual variability; (ii) a proportion of the inter-annual variability simulated by the CCC-GCM is identified as being potentially predictable, particularly in DJF; and (iii) other studies (eg Chervin (1986), Lau (1986)) together with our own suggest that a large proportion of the observed interannual variance which is in excess of that due to climate noise may be attributable to the atmosphere's internal dynamics. The extent to which variation from this source is predictable remains to be seen.

5. REFERENCES

Boer, G.J., McFarlane, N.A., Laprise, R., Henderson, J.D., and Blanchet, J.-P., 1984a: The Canadian Climate Centre Spectral Atmospheric General Circulation Model, ATMOSPHERE-OCEAN, 22, 397-429.

Boer, G.J., McFarlane, N.A., and Laprise, R., 1984b: The climatology of the Canadian Climate Centre General Circulation Model as obtained from a Five-Year Simulation, ATMOSPHERE-OCEAN, 22, 430-475.

Chervin, R.M., 1986: Interannual Variability and Seasonal Climate Predictability, J. of Atm. Sci., 43, 233-251.

Jones, R.H., 1976: On estimating the variance of time averages, J. of App. Met., 15, 514-515.

Jones, R.H., 1975: Estimating the variance of time averages, J. of App. Met., 14, 159-163

Lau, N.-C., 1981: A diagnostic study of recurrent meteorological anomalies appearing in a 15-year simulation with a GFDL General Circulation Model, Mon. Weath. Rev., 109, 2287-2311.

Lau, N.-C., 1986: Modeling the Seasonal Dependence of the Atmospheric Response to Observed El Ninos in 1962-76, private communication.

Madden, R.A., 1976: Estimates of the natural variability of time averaged sea level pressure, Mon. Weath. Rev., 104, 942-952.

Manabe, S., and Hahn, D.G., 1981: Simulation of Atmospheric Variability, Mon. Weath. Rev., 109, 2260-2286.

Sachs, L, 1982: Applied Statistics, a handbook of techniques, translated by Z. Reynarowych, Springer-Verlag, New York, NY. pp. 707.

Shukla, J., and Gutzler, D.S., 1983: Interannual variability and predictability of 500 mb geopotential heights over the northern hemisphere, Mon. Weath. Rev., 111, 1273-1279.

Trenberth, K.E., 1985: Potential Predictability of Geopotential Heights over the Southern Hemisphere, Mon. Weath. Rev., 113, 54-64.

Volmer, J.-P., Deque, M., Jarraud, M., 1983: Large-scale fluctuations in a long-range integration of the ECMWF spectral model, Tellus, 35A, 173-188.

Zwiers, F.W., 1986a: Statistical Considerations for Climate Experiments: Part II (Multivariate Tests), submitted to J. of Clim. and App. Met.

Zwiers, F.W., 1986b: A Potential Predictability Study of a Climate Simulated by an Atmospheric General Circulation Model, unpublished manuscript.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

CHARACTERISTICS OF SUMMER CLIMATE IN WESTERN EUROPE AND THE DETECTION OF THE CO_2-EFFECT

C.J.E. Schuurmans and A.J. Coops
Institute of Meteorology and Oceanography,
State University Utrecht, Utrecht, The Netherlands

1. INTRODUCTION

Worldwide modification of climate is to be expected as a result of the increasing concentration of CO_2 and other trace gases in the atmosphere. Theoretical estimates of the effects, based on mathematical climate model simulations show a large range of uncertainty, but even in the most conservative case the CO_2-induced change, most probably will manifest itself as a clear signal, out of the noise of natural climate variability. An early detection of the CO_2-climate signal therefore is important from the point of view of calibration of computer models of climate, apart from other reasons. Over the past ten years several attempts have been published, but in none of these papers the CO_2-climate signal was isolated convincingly. An extensive and authorative review of the subject, which recently has been publhised (MacCracken and Luther, 1985) agrees with this conclusion, but adds among other things that "Trend analysis of long-term records of land and ocean temperatures and sea level are qualitatively consistent with the climate changes projected by modeling studies".

This statement, however does not mean that temperature trends of different locations on earth all point in the same direction. For Western Europe for instance, we found (Schuurmans and Coops, 1984) that seasonal mean temperatures did not significantly increase over the 50-year period 1925-1975, except for the southern European stations Madrid (winter) and Lisbon (summer). Over a longer period however, 1875-1975, it turned out that a significant warming had occurred at nearly all stations and in all seasons. This corresponds to the well-known 20th-century wamring, which occurred nearly worldwide and had its maximum in the 1940-ties, before the large increase of the CO_2-concentration started. This is one of the complications in the effort of detecting the CO_2-climate signal in temperature records.

For this reason we decided to take also other variables like precipitation and air pressure into consideration. Unfortunately for these other variables the theoretical estimates of change due to the CO_2-effect are practically unknown. We therefore had to rely on empirical relations between these variables and temperature.

Restricting ourselves to the summer season, in which the temperature signal-to-noise ratio is optimal (Wigley and Jones, 1981) we than may expect that higher temperatures due to the CO_2-effect will be parallelled by less precipitation and higher air pressures. On the basis of this expectation we investigated the available records of temperature, precipitation and air pressure for the summer season at De Bilt, the Netherlands.

2. CHARACTERISTICS OF SUMMER CLIMATE

In analysing the records of summer climate at De Bilt, The Netherlands, an interesting feature became apparent, which most probably is a general characteristic of summer climate in Western Europe. The characteristic may be referred to as a jump-like behavior: periods of cool and wet summers alternating with periods of relatively warm and dry summers. Changes from one state to the other are rather abrupt, causing the time series of temperature, precipitation and also air pressure probably to be fitted better by sharp square waves than by smooth sine waves.

In Fig. 1 this behaviour of summer climate at De Bilt, the Netherlands is depicted for the total length of the available time series. In the picture the jumps are exagerated,since for each sub-period only the mean deviation from the overall mean values in indicated.

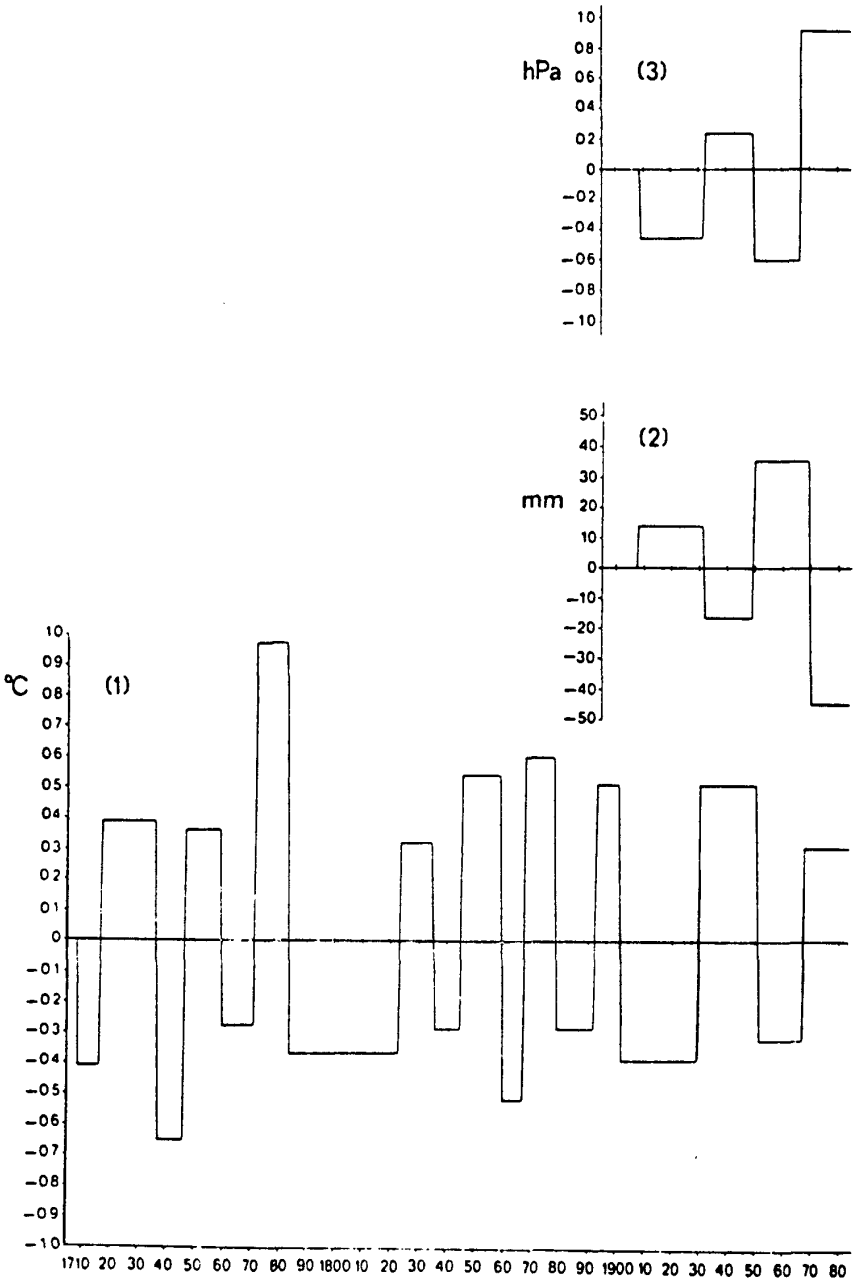


Figure 1 Block diagram of the mean deviation of temperature (1), precipitation amount (2), and air pressure (3) from the overall mean values for the subperiods indicated. Data are for summer (June, July, August) at De Bilt, The Netherlands and the length of the series is respectively, for temperature (1709-1984), precipitation amount (1896-1984) and air pressure (1902-1984).

The years of transition are determined objectively by taking the year of the maximum (or minimum) of the curve of the cumulative deviations of the yearly values from the overall mean. Clearly the years of transition are the same or nearly the same for the 3 variables considered. In itself this gives some confidence in the reality of the phenomenon, since it may be expected that any abrupt change of summer climate affects all variables, the common cause of it being a change in intensity and/or location of the Azores high.

Still, it may be of interest to investigate the nature and causes of the jumplike behavior further. However, in this paper we take it for granted and in the following will concentrate on the use of the phenomenon for the purpose of the detection of the CO_2-effect. The hypothesis we are going to test is the following.

Any CO_2-effect is superimposed upon the climate variations present by nature. If on the basis of climate model simulations we have to expect an improvement of summer climate in Western Europe, this would mean that in a climate of alternating good and bad periods, the jumps in the direction of a climate optimum become gradually larger than those in the opposite direction.

Admittedly, the indications from climate model simulations for an improvement of summer climate in Western Europe are very preliminary. Most probably in a CO_2-rich world summer temperatures in Western Europe will be higher than at present but there is still disagreement about the changes in precipitation climate and in air pressure. A recent study by Manabe and Wetharald (1986) however, supports the view that summer soil wetness in Western Europe will be reduced as a result of an increase in atmospheric CO_2. (which, of course, is not necessarily the same as a decrease in precipitation amount).

3. SIGNIFICANCE OF RECENT CLIMATE JUMPS

The jumps in temperature T, precipitation amount RR and air pressure p at De Bilt, occurring in the present century (see Fig. 1), have been tested for the validity of the above hypothesis.

This means that for each jump and for each variable separately the signal-to-noise ratio has been computed and its significance has been determined in a statistical sense. The signal in this case is defined as the difference between the mean values of the variable in two consecutive subperiods. In Table I the value of the signal is indicated for the year of transition. The noise, being the standard deviation or year-to-year variability of the variable, has been computed for the total length of the time series.

Year	1929	1950	1966
ΔT (°C)	+1.0	-0.9	+0.7
P(%)	<0.01	0.08	5.6
Year	1931	1949	1969
ΔRR (mm)	-30	+50	-80
P(%)	4.7	0.96	0.16
Year	1931	1949	1966
Δp (hPa)	+0.7	-0.8	+1.5
P(%)	16	3.1	0.55

Table I Year, amplitude and statistical significance of a jump in temperature T, amount of precipitation (RR) and air pressure (p). Results for summer (June, July, August) at De Bilt, The Netherlands, 1900-1984.

This noise level has been assumed to be constant. The significance level P of a jump is not only a function of the signal-to-noise ratio, but also of the length of the subperiods concerned. It was determined by using the distribution-free, two-sample Wilcoxon test.

The results of this exercise are given in Table I, from which it is immediately clear that the hypothesis stated in paragraph 2 is certainly not fulfilled for summer temperature. Here the last jump upward in 1966 is still far less significant than the jump downward in 1950 and really not comparable in

intensity with the warming around 1929. For precipitation and air pressure the situation looks better. The transition to dryer summers in 1969 and higher air pressures in 1966 turns out to be the most significant change since the beginning of this century. The intensity of this change is indeed remarkable: a drop of 80 mm in precipitation, with a general level of summer precipitation of the order of 220 mm really may be called a change to a period of summer droughts!

4. COMBINATION OF VARIABLES

As has been referred to above, it appears to be a reasonable assumption that a jump in summer climate is simultaneous in temperature, precipitation and air pressure. This would enable us to combine these variables and see whether or not the signal-to-noise ratio of the combination confirms our hypothesis.

In combining the variables we followed a method proposed by Bell (1982). A linear combination of the variables leads to a signal $S = \sum \alpha_i S_i$ (S_i being the signal of the variable X_i) whereby the α_i's are chosen in such a way that the signal-to-noise ratio of a particular jump is a maximum. The noise is defined here as the standard deviation of the combination, i.e. the square root of $\mathrm{var}\{\sum \alpha_i X_i\}$. When we describe that $\sum \alpha_i S_i = \sum S_i$, which remains constant, the method is equivalent with a minimization of the variance.

The solution for the α's is:

$$\alpha_i = \{\sum_{i=1}^{3} S_i\} \frac{\sigma_2\sigma_3\{S_1\sigma_2\sigma_3(1-\tau_{23}^2) + \sigma_1 \sum_{1\neq i\neq j\neq 1} S_i\sigma_j(\tau_{1j}\tau_{ij}-\tau_{1i})\}}{\sum_{i\neq j<k\neq i} S_i^2\sigma_j^2\sigma_k^2(1-\tau_{jk}^2) + 2\sigma_1\sigma_2\sigma_3 \sum_{i\neq j\neq k>i} S_i\sigma_j S_k(\tau_{ij}\tau_{jk}-\tau_{ik})} \qquad (1)$$

where α_2 and α_3 are found by exchange of indices. In the formula σ_i is the standard deviation of X_i and τ_{ij} is the correlation between X_i and X_j. For further details we refer to Coops and Schdurmans (1986).

The correlation coefficients which are supposed to be independent of time are the weighted (with respect to time) averages of the rank correlation coefficients of Kendall for the 4 subperiods in this century, within which no jumps occured. With 1) temperature, 2) precipitation and 3) air pressure these correlation coefficients turned out to be:

$$\tau_{12} = -0.3$$
$$\tau_{13} = +0.2$$
$$\tau_{23} = -0.5$$

It is difficult to say what the optimum value of the correlation coefficients for our purpose would be. It can be shown that even a linear combination of uncorrelated variables has some advantage as compared to the use of the variables separately (except when the signal-to-noise ratios are exactly equal). On the other hand variables with a correlation coefficient exactly equal to one need not to be combined.

Anyhow following the procedure outlined above we arrive at a signal-to-noise ratio Q for each of the 3 jumps during this century. It can be shown that the significance level P of the Q-values can be estimated by:

$$P = 1 - \Phi[\{2\Phi(Q\sqrt{\tfrac{2}{\pi}})-1\}\sqrt{\frac{3D_1D_2}{D_1+D_2+1}}] \qquad (2)$$

with D_1, D_2 being the number of years of two consecutive subperiods and Φ is the standard normal distribution function.

The final result is shown in Table II. From this Table we may conclude that at least for the data set of De Bilt, The Netherlands our hypothesis is confirmed, which means that the behavior of summer climate at this station is not in contradiction with a supposed CO_2-effect superimposed upon the natural climate variations.

Year	1930	1950	1967
ΔS	+1.31	-1.13	+1.35
P%	0.06	0.14	0.07

Table II Year, amplitude and statistical significance of a jump in summer climate S, taken as a linear combination of temperature, precipitation amount and air pressure (see text). Results for De Bilt, The Netherlands, 1900-1984.

Note added: Interestingly, the first summer after conclusion of our study, 1985, turned out to be cool and relatively wet. The present summer, however (1986) seems to fit completely the relatively warm, dry and high air pressure scenario. The next few years to come will make clear whether the summer of 1985 was the first of a new subperiod of cool and wet summers, or just a negative deviation in a continuing seires of warm and dry summers. But even when the first possibility is found to be the case, we expect that the new subperiod of cool and wet summers, will be of weak or moderate intensity, due to the CO_2-effect which tends to improve summer weather in Western Europe.

5. ACKNOWLEDGEMENT

This research was supported by the Ministry of Housing, Physical Planning and Environment.

REFERENCES

- Bell, T.L., 1982: Optimal weighting of data to detect climatic change: application to the carbon dioxide problem. J. Geo. Res., 87, pp. 11161-11170.

- Coops, A.J. and C.J.E. Schuurmans, 1984: Detection of the CO_2-effect upon the climate of Western Europe. Theoretical and Applied Climatology (in press).

- MacCracken, M.C. and F.M. Luther (ed.), 1985: Detecting the climatic effects of increasing carbon dioxide. DOE/ER-0235. U.S. Department of Energy, Washington D.C.

- Manabe, S. and R.T. Wetherald, 1986: Reduction in summer soil wetness induced by an increase in atmospheric carbon dioxide. Science, 232, pp. 626-628.

- Schuurmans, C.J.E. and A.J. Coops, 1984: Seasonal mean temperatures in Europe and their interannual variability. Monthly Weather Rev., 112, pp. 1218-1225.

- Wigley, T.M.L. and P.D. Jones, 1981: Detecting CO_2-induced climatic change, Nature, 292, pp. 205-208.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

USING NONPARAMETRIC RANDOMNESS TESTS FOR CLIMATE CHANGE ANALYSIS

V.T.Radiuhin, Ya.E.Mikhlin

All-Union Research Institute of Hydrometeorological Information - World Data Centre
USSR

I. INTRODUCTION

Constructing a model explicitly incorporating parameters of the signal (nonstationarities in the probability distribution parameters of a climate system characteristic) and noise as well as estimating the parameters is one of the approaches most commonly used in empirical-statistical analysis of the climate system. Such type model is described in (Gruza,1982), incorporating monotonic and periodic signals both in the mean and in the variance. When analysing empirical data, climatologists most often restrict themselves to detecting the signal in the mean.

Before suggesting a model and estimating its parameters it is useful to test the hypothesis of the presence of determined nonstationarities in the series. This is the way they do it in a number of studies devoted to climate change analysis (Aristova, 1980; Goossens, Berger, 1983; Harley, 1980; Hoyt, 1981; Madden, Sadeh, 1975; Ratcliff, et al., 1978; Woodcock, Jones, 1970) where a great variety of statistical tests is used without substantiating the choice of a criteria. Therefore it is often not clear how good is the test used and why it was given preference. Also, if the null hypothesis is not rejected, the question remains unanswered as to the data goodness of fit to the competing hypotheses (i.e. the question of the probability of a second kind error).

From this viewpoint a set of nonparametric randomness tests most commonly used for detecting trends in the means of the climatic time series is considered in Section 2.

Using the inversion of order tests, most efficient for detecting linear trends, time series of Northern Hemisphere gridded monthly mean surface air temperatures for the period 1891-1983 are analysed in Section 3. The degree of correspondence of the data under consideration to the null hypothesis of the time series randomness is taken as a potential predictability characteristic of the climate characteristic changes. The less probable the randomness hypothesis is on real data (the higher is the significance level at which it is rejected) the more probable is the presence of a potentially predictable signal in such a time series.

2. NONPARAMETRIC RANDOMNESS TEST EFFICIENCIES

Six randomness tests based on statistics given in Table 1 were investigated with the help of statistical modelling. The mathematical expectation and the variance of the statistics distribution considered under the null hypothesis are given as E/H_0 and V^2/H_0, respectively. H_0 implies the assumption that the time series under consideration is a sequence of n independent similarly distributed random values. An alternative to the null hypothesis (hypothesis H_1) is a time series which is a sum of the trend value and discrete white noise for each point. The other symbols in Table 1 are:

Q - the number of inversions in the time series, i.e. the number of such

point pairs in the series of which the value at the first point is strictly larger than that at the second one.
Z_i - the indicator of a "peak" or a "pit" at the i-th point of the time series ($2 \leq i \leq n-1$). $Z_i=1$ if $X_{i-1} < X_i$ and $X_{i+1} < X_i$ (the case of a "peak") or if $X_i < X_{i-1}$ and $X_i < X_{i+1}$ (the case of a "pit"), in the rest of the cases $Z_i=0$. Time series values at points i-1, i, i+1 are designated as X_{i-1}, X_i, X_{i+1}, respectively. In counting the number of "peaks" and "pits" two or more points in a row, with similar values of the time series are considered as one point;
Y_i - the indicator of an increase immediately after the i-th time series point ($1 \leq i \leq n-1$). $Y_i=1$ if $X_i < X_{i+1}$, otherwise equal to zero;
U_i - the indicator of the presence of "an upper record" at the i-th point of the time series ($2 \leq i \leq n$). $U_i=1$ if $X_j < X_i$ $\forall j < i$, otherwise $U_i=0$;
L_i - the indicator of the "lower record" at the i-th point of the time series ($2 \leq i \leq n$). $L_i=1$, if $X_j > X_i$ $\forall j < i$ and otherwise $L_i=0$.

Table 1. Nonparametric randomness tests

Statistics	Computation Formulae	Mean Value	Variance	Alternative
Inversion coefficient	$t=1-\frac{4Q}{n(n-1)}$	0	$\frac{2(2n+5)}{9n(n-1)}$	Monotonic trend
Number of "peaks" and pits"	$p=\sum_{i=2}^{n-1} Z_i$	$\frac{2(n-2)}{3}$	$\frac{16n-29}{90}$	Cyclic trend
Number of increase points	$c=\sum_{i=1}^{n-1} Y_i$	$\frac{n-1}{2}$	$\frac{n+1}{12}$	Monotonic trend
Total amount of upper and lower records	$s=\sum_{i=2}^{n}(U_i+L_i)$	$2\sum_{i=2}^{n}1/i$	$2\sum_{i=2}^{n}(\frac{1}{i}-\frac{2}{i^2})$	Monotonic trend
Difference between upper and lower record amounts	$d=\sum_{i=2}^{n}(U_i-L_i)$	0	$2\sum_{i=2}^{n}1/i$	Monotonic trend
Number of series	r	$\frac{n}{2}+1$	$\frac{n(n-2)}{4(n-1)}$	

The runs test (see for example, (Bendat, Pirsol, 1974)) shown here is for the median case. What is meant by the run here is the sequence of the time series terms in a row whose values all lie on one side of the sample median, while the values of the preceding and the subsequent terms of the time series lie on the other side of the median. The powers of tests taken from Table 1 were investigated to test H_0 against H_1 for certain specific kinds of the alternative hypothesis. The test powers were estimated from the results of repeated modelling of a random series realizations independent of one another satisfying the conditions of the corresponding alternative hypothesis H_1. The relative frequency of the

statistics values computed from the series elements and modelled under H_1 falling within the critical region was taken as an estimate. The critical region limits for three different significance levels were determined from the corresponding normal distribution approximations of the statistics in question under the null hypothesis.

When modelling the alternative hypotheses, we assumed that random independent deviations from the trend at the time series individual points are normally distributed with a null mean and standard deviation V while the alternative hypotheses differ by the trend type. Two trend types have been analysed: linear and cyclic.

Trends of the type $X_i^*=K\cdot i$ have been considered where i is the number of the trend series point and X_i^* is the value of the trend series at the same point. We considered such values of the parameter K with which the signal-to-noise ratio, i.e. the value $K\cdot n/V$, varies from zero to four with a step of 0.1. Note that the value $A=K\cdot n$, a determined change of the time series during the whole period, was taken as a signal in the case of the linear trend.

Cyclic trends were considered for eight different values of the frequency C. The frequency was taken in such a way that 0.25,0.5,1.0,1.5, 2.0,3.0,4.0 and 10 full cycles fit the time series in question (n=50). The cyclic trends from the family corresponding to the frequency C have the form $H_i^*=A\cdot\sin(C\cdot i)$. We considered such values of the parameter A for all the eight cyclic trend families that the value A/V (taken here as the signal-to-noise ratio) assumed values from zero to four with a step of 0.5.

The power functions of some of the above tests when testing H_0 against different type alternatives are given in Figures 1-4 as a function of the signal-to-noise ratio (Plots are given for three significance levels).

Efficiency characterictics of the first and sixth tests are shown in Fig.1 from which it follows, that under the linear trend alternative the first test is much more efficient than the sixth one. Even at a 10% significance level the sixth test distributions under the null hypothesis and the alternative one only start differing significantly beginning with the signal-to-noise ratio exceeding three while with the first test the competing hypotheses start differing essentially with the signal-to-noise ratio as low as 1.5. The tests based on four other statistics in the case of the alternative linear trend are quite useless, their distributions being almost similar for H_0 and H_1 and powers close to nill.

The test power based on the second statistic is shown in Fig.2 as a function of A/V against the cyclic trend alternative for frequencies corresponding to 2,3,4 and 10 periods throughout the whole lenght of the trend series in question (curves a,b,c and d in Fig.2 correspond to the above frequencies). Due to the smallness of the power value similar curves for lower frequency values are not given here. From the curves in Fig.2 it is seen, that the test power value based on the second statistics increases with frequency.

The test power based on the first statistic is shown in Fig.3 as a

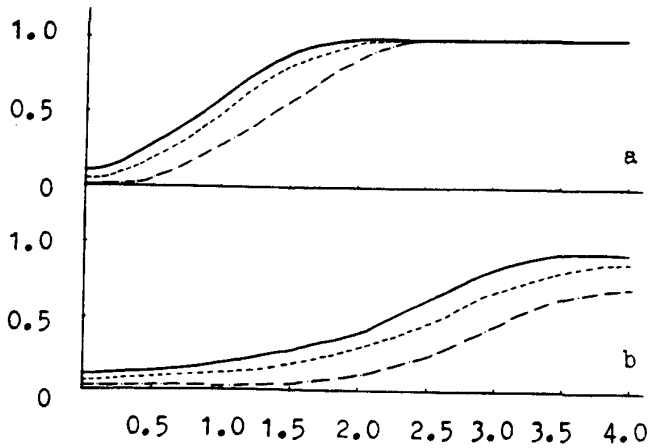


Fig.1. Test Power Function for Linear Trend Alternative:
a) first test;
b) sixth test:
1,2,3 - for α=0.01,0.05 and 0.01, respectively.

function of A/V against the cyclic trend alternative for frequencies corresponding to 0.25,0.5, 1.0,1.5,2.0,3.0,4.0 and 10 periods throughout the length of the trend series in question (curves a,b,c, d,e,f,g and h in Fig.3 correspond to the frequencies pointed out).From the curves in Fig.3 it is seen, that the test power value based on the first statistic decreases with frequency. It is also to be noted here that in the case of 0.25 period in the series in question (monotonic increase of the trend component) the test power behaviour is essentially the same as in the case of the linear trend (See Figs. 1a and 3a).

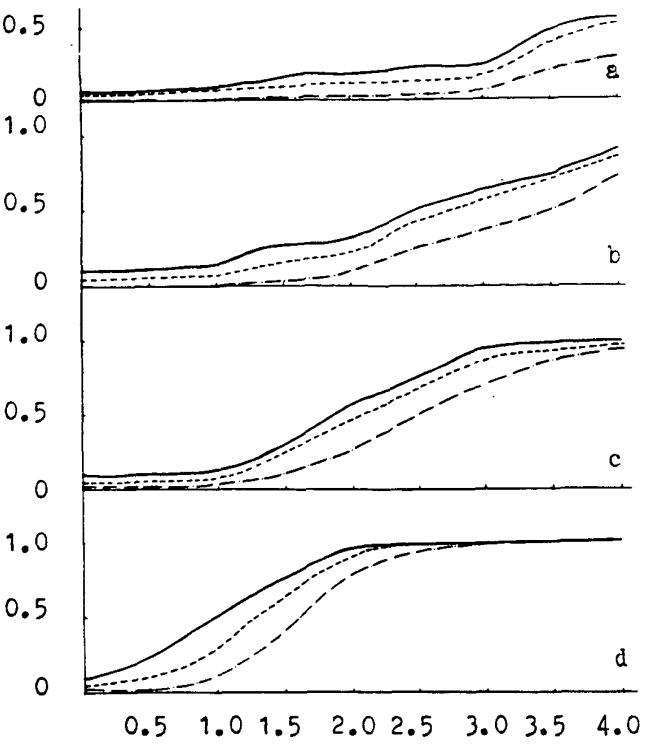


Fig.2. Second Test Power Function for Cyclic Alternative:
a-d - for 2,3,4 and 10 periods, respectively.

The test power based on the sixth statistic is shown in Fig.4 as a function of A/V against the cyclic trend alternative for frequencies corresponding to 0.25,0.5,1.0, 1.5,2.0,3.0,4.0 and 10 periods over the lenght of the trend series (n=50) (curves a,b,c,d,e,f,g and h in Fig.4 correspond to the frequencies pointed out). The test power value based on the sixth statistic is seen to increase first with frequency increase and then decrease.

The test powers based on the third to the fifth statistics appeared to be small for all frequencies considered, therefore the curves are not given here.

3. USING INVERSIONS OF ORDER TEST FOR ESTIMATING POTENTIAL PREDICTABILITY OF MONTHLY MEAN SURFACE AIR TEMPERATURES

Recall briefly what is meant by climatic noise: statistical estimate fluctuations of climatic system parameters occurring even in case of an unchangeable climate due to upredictable weather fluctuations (Leith,1978; Madden,1984). Generally speaking, the climatic signal characterizes time variations of the climatic ensemble state probability distributions. In practice the first two moments of the climate ensemble are commonly considered (most often the first one, or mean). The ratio of the climatic signal to the natural climate variability characteristic or climatic noise is commonly taken as the potential predictability characteristic of climate variations. The larger is the ratio, the greater is the potential for the predictability of variations. In this section, the potential predictability of climate element variations is characterized by the probability of absence of determined variations in the series. With this in view we resort to the inversions of order test discussed in Section 1. The less probable on real data is the hypothesis of the time series randomness (the higher is the confidence level it is rejected at), the more probable is its potential predictability, i.e. the presence of the signal.

The results thus obtained relate to revealing the potential

predictability of the one - dimensional time series first moment since the tests considered in Section 1 respond to the signal in the mean. Note that since the null hypothesis, under which the randomness test statistical properties were considered, was formulated as an assumption that the time series in question is a series of independent similarly distributed values, the statistical dependence between the subsequent values of the series even if there is no determined signal may distort the resulting inferences.

Monthly mean temperature series (for central months of seasons: January, April, July and October) and gridded annual mean surface temperatures for the Northern Hemisphere obtained from the monthly mean temperature anomaly archive (Borzenkova et.al., 1977; Gruza, Rankova, 1979, Jones et.al., 1982; Woodcock, Jones, 1970; Zhabina et.al., 1982) were analysed using the first of the six tests considered above.

Over the whole series available (the record taken was through 1983) for each i-th point the statistic t_i was computed and using this value $P(t_i/H_0)=P(t \geq t_i/H_0)$ i.e. the probability was defined of obtaining on the random series of statistics whose values are not lower than those obtained on a real series. The probability characterizes the degree of real data agreement with the null hypothesis, small probability values (e.g., < 0.1) testifying to the fact that the null hypothesis can be rejected at a significance level of $\alpha = 0.1$ in favour of the monotonously increasing trend hypothesis and large values (e.g., > 0.9) testifying to the fact that the null hypothesis can be rejected at the same significance level in favour of an opposite alternative - the existence in the data of a monotonously decreasing trend thus characterizing the degree of potential predictability in the series under consideration.

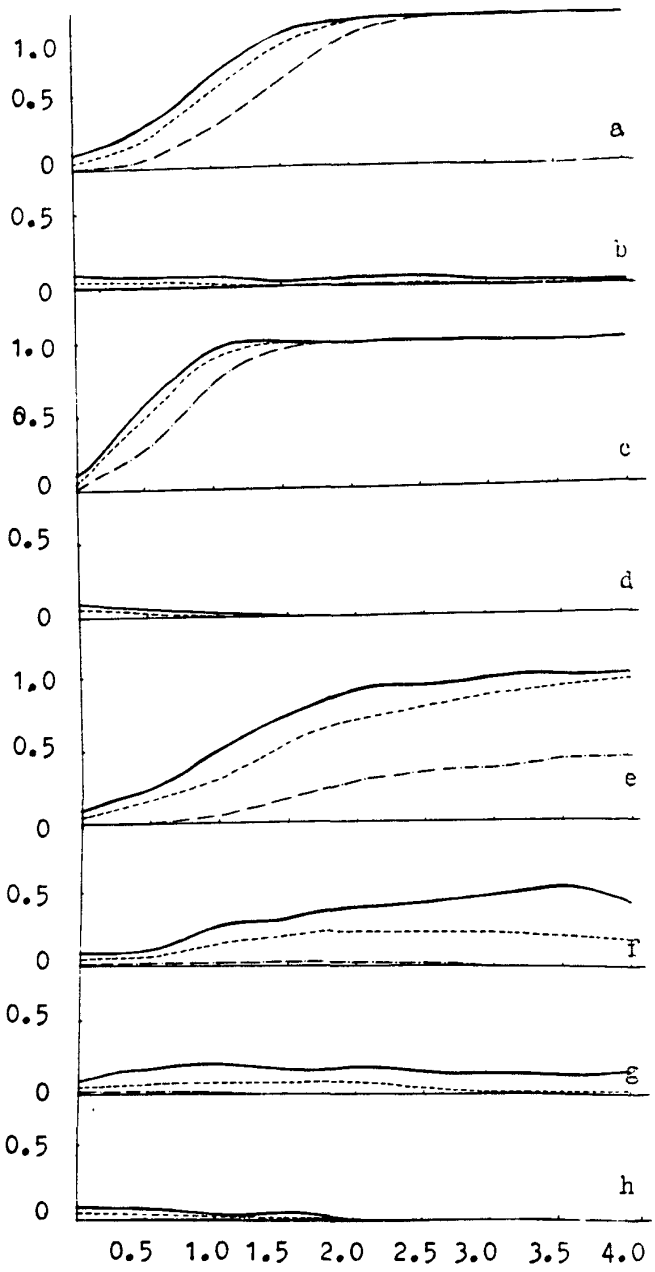


Fig.3. First Test Power Function for Cyclic Alternative:
a-d) for 0.25, 0.5, 1.0 and 1.5 periods, respectively;
e-h) for 2, 3, 4 and 10 periods, respectively.

Northern Hemisphere maps are given in Fig.5 with hatched areas inside the isolines $P(t_i/H_0)=0.1$ and $P(t_i/H_0)=0.9$ for monthly mean temperatures (Fig.5, a-d) and annual mean temperatures (Fig.5e).

Quite large homogeneous areas stand out in the Northern Hemisphere for which the hypothesis of an absolutely random character of mean monthly and annual temperature variations is hardly probable, deviating towards an increasing trend in data in some areas but towards a decreasing one in other areas.

Long-term tendencies over the oceans do not actually change from season to season but are oppositely signed, the non-predictability hypothesis being rejected in favour of the positive trend in the Pacific, while in the Atlantic Ocean it is rejected in favour of the negative one (the inferences should be treated cautiously since the amount of data over the oceans is rather small).

Note, that near - polar areas in all seasons are characterized by a higher potential predictability, but the tendencies are different, being homogeneous positive ones in winter and summer and oppositely signed in transition seasons.

The long-term behaviour of the mean January and July temperatures varies over the North American continent while a positive trend is easily distinguished over about the whole of the USSR area on the April map. The results obtained as a whole agree with the maps in (Gruza, Rankova,1979) where areas of a prevailing negative trend are shown.

Inversions of order test were also computed

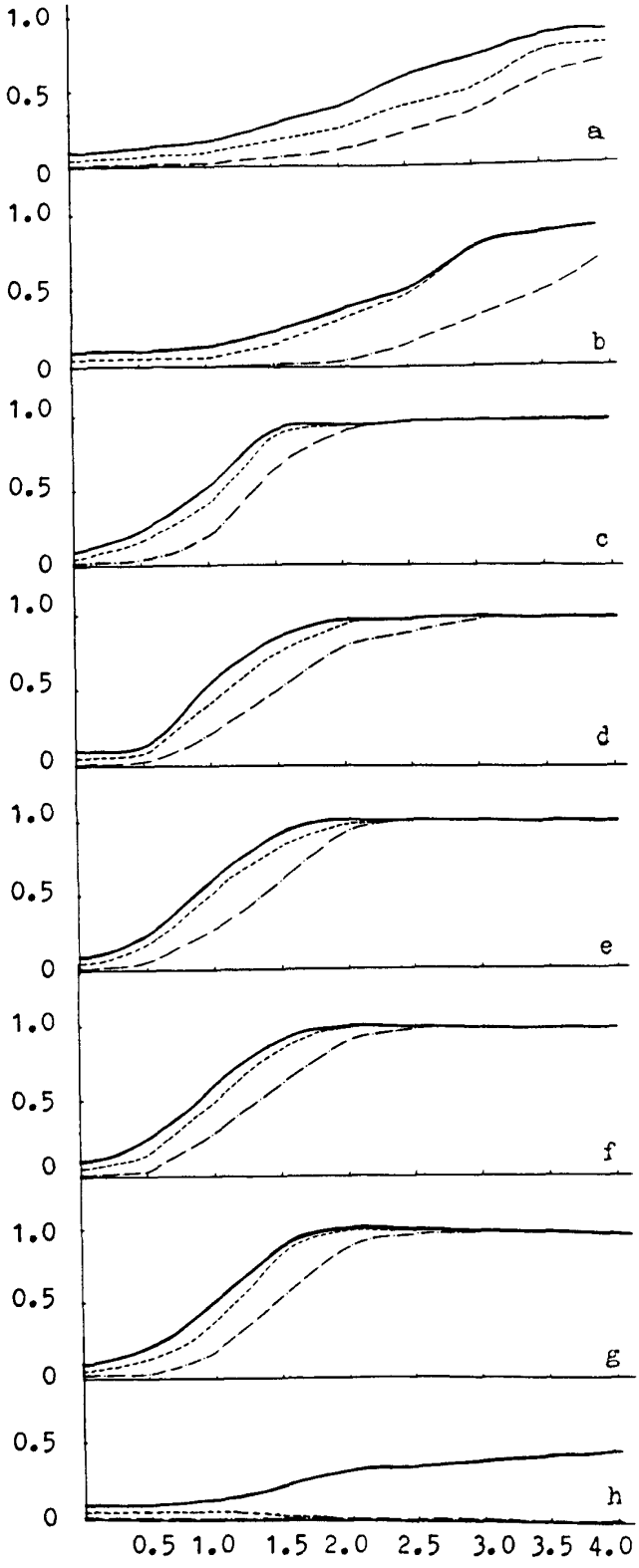


Fig.4. Sixth Test Power Function for Cyclic Alternative:
a-h) for 0.25,0.5,1.0,1.5,2,3,4 and 10 periods, respectively.

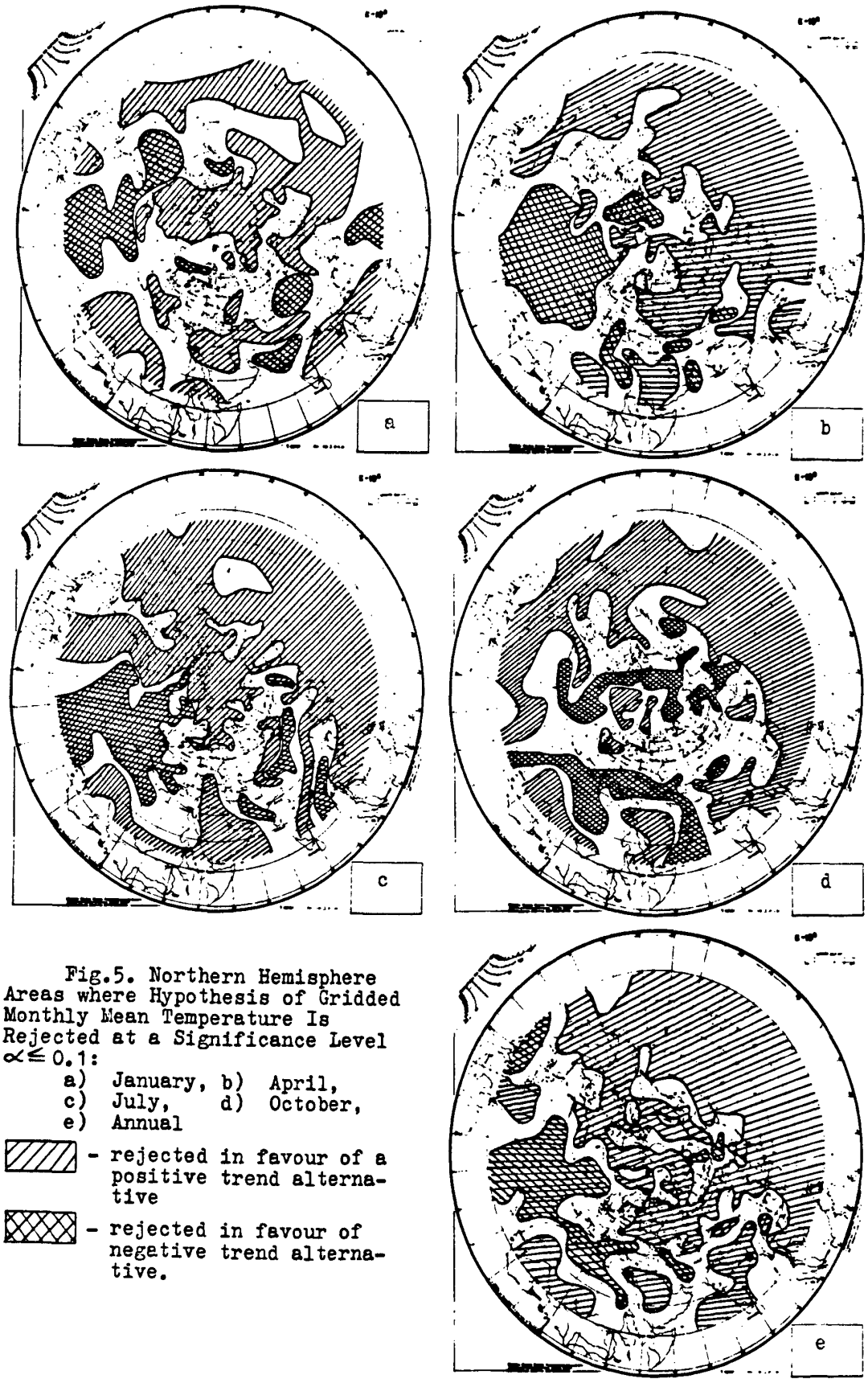


Fig.5. Northern Hemisphere Areas where Hypothesis of Gridded Monthly Mean Temperature Is Rejected at a Significance Level $\alpha \leq 0.1$:

a) January, b) April,
c) July, d) October,
e) Annual

▨ - rejected in favour of a positive trend alternative

▩ - rejected in favour of negative trend alternative.

for annual mean temperature value series averaged over the Northern Hemisphere as well as for a similar series given in (Jones et al.,1982). The statistics values for these series are 0.3966 and 0.4380, respectively, the probability of obtaining such statistics values from a random series being actually equal to zero. The hypothesis of the absence of trend in the global temperature series considered using a test based on the number of inversions in the series values is thus rejected at a very high significance level in favour of an increasing trend alternative (it should be noted, however, that the consecutive values of monthly mean hemispherically averaged temperatures are not independent which contributes to the rejection of the hypothesis of independence).

4. CONCLUSION

On the basis of studying the properties of the test under review the following can be inferred:

First, tests based on the third to fifth statistics proved to be inefficient under the alternatives considered and hence, useless in detecting climatic signals. Even if using any of the above tests the null hypothesis is rejected at a certain significance level, this in no way testifies in favour of the alternative hypothesis of the presence of the signal but can only signify a second kind error occurence whose probability is given by the significance level value.

Second, the test based on the first statistic appeared to be the most efficient one for testing H_0 against the linear trend alternative (or any monotonic change at all). Besides, when the null hypothesis is rejected, one can judge from the sign of the statistic of the monotonic trend direction determining the alternative hypothesis; negative values speak of the decreasing tendencies and vice versa.

Third, the second and the sixth tests are more efficient under the cyclic alternative, the sixth one bringing out the difference between the null and the alternative hypotheses more strongly when low-frequency variations are an alternative and the second one being more efficient for distinguishing a high-frequency signal.

It is proposed to estimate the potential predictability of variations in the time series means from the probability of any determined changes in them using randomness test which proved to be efficient. The less probable on real data is the time series randomness hypothesis (i.e. the higher is the significance level it is rejected at) the more probable is the potential predictability of variations in the series, the presence of the signal in it.

The technique has been illustrated by applying an inversions of order test to the gridded Northern Hemisphere surface air temperature series.

The technique suggested for estimating the potential predictability of variations in the mean value of the time series can be used as a first stage for defining objects (regions, series) for a deeper statistical analysis in future, for constructing a signal and noise defining model.

The authors are grateful to L.Ravina for preparing the English version of the paper and H.Devyatova for typing it.

5. REFERENCES

1. Aristova L.N.,1980: Secular variations of the temperature condition anomaly in the Northern Hemisphere and the assessment of its stationarty. Tr. of VNIIGMI-WDC, i.68,80-86 (in Russian).

2. Bendat J., Pirsol A.,1974: Measurement and analysis of the random processes, M.: Mir, 464pp. (in Russian).

3. Borsenkova I.I., Vinnikov K.Ya., Spirina L.P., Stekhnovsky A.I.,1977: Variations in the Northern Hemisphere air temperatures for the period 1881-1975, Meteorologiia i hidrologiia, N9, 106-109 (in Russian).

4. Goossens C., Berger A.,1983: Evaluation of the statistical significance of climatic changes Papers presented at the 2-nd Inter.Meet. on Stat. Climatology, Lisbon, Portugal.

5. Gruza G.V.,1982: Some general questions of the statistical analysis of the time-series in climatology, Tr. of VNIIGMI-WDC, i.83, 3-9. (in Russian).

6. Gruza G.V., Rankova E.Ya.,1979: Data on climate structure and variability. Sea level air temperature. Northern Hemisphere, Obninsk: Offset printing plant of All-Union Res.Inst. of Hydromet. Inf.-World Data Centre, 203pp (in Russian).

7. Harley W.S.,1980: The significance of climatic change in the Northern Hemisphere 1949-1978, Mon.Wea.Rev.,V.108,N3,235-248.

8. Hoyt D.V.,1981: Weather "records" and climatic change, Climatic Change, N3.

9. Jones P.D., Wigley T.M.L., Kelly P.M.,1982: Variations in surface air temperatures:part 1. Northern Hemisphere, 1881-1980,Mon.Wea.Rev.,V.110, N2,59-70.

10. Leith C.E.,1978: Predictability of climate, Nat.,V.276,N5686,352-355.

11. Madden R.A.,1984: Potential for long-range predictability of Northern Hemisphere temperatures Studies in Climate,NCAR Technical Note,92-114.

12. Madden R., Sadeh W.,1975: Empirical estimates of the standard error of time-averaged climatic means, Appl.Met.,V.14,N3.

13. Ratcliffe R.A.S., Weller J., Collison P.,1978: Variability in the frequency of unusual weather over approximately the last century,Quart. J.Roy.Met.Soc.,V.104,N440.

14. Woodcock A.H., Jones R.H.,1970: Rainfall trends in Hawaii,Appl.Met.,V.9, N8.

15. Zhabina I.I., Kastin O.M., Roschina S.A., Stolypina I.V., Khachaturova T.V.,1982: Operational implementation and usage of a temperature-pressure field and cloudiness data base, Meteorologiia i hydrologiia, N6,100-111 (in Russian).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

ANOMALIES IN THE TEMPERATURE PATTERNS OF COLD PERIODS ON THE USSR AREA

A.M.Vysotsky, M.K.Zdonova, V.N.Razuvaev
N.N.Shevchenko

All-Union Research Institute of Hydrometeorological Information - World Data Centre
USSR

1. INTRODUCTION

Extraordinary meteorological conditions have been observed in various regions of the globe in a number of recent years. The scale and duration of these phenomena are very important for the climate variability research. The 1982-83 El Nino and its associated effects as well as the droughts in Africa and Australia (WMO,1985) are vivid examples of such phenomena. In this regard, no exception is the USSR area where extreme conditions were repeatedly experienced during 1982-1985.

In this paper we restrict ourselves to considering the cold periods (November-March) of 1982-83, 1983-84 and 1984-85. The surface air temperature was chosen as a major parameter characterizing the state of the climatic system. The peculiarities of temperature patterns of individual periods of the year are often used to study the climate variability in various regions. The air temperature is also very important in assessing the extent of the climatic system state anomaly in climate monitoring (Gruza et al., 1984).

II. TECHNIQUES AND RESULTS

The assessments of the temperature pattern anomaly are usually based on various ways of comparing the current air temperature value and the long-term mean.

Monthly mean air temperatures at 31 stations scattered nearly evenly throughout the whole of the USSR were used for analysis. There are five areas selected in the USSR territory (see Fig.1), which have nearly an equal number of stations, to study the spatial distribution of the large-scale temperature anomalies. The series of monthly mean air temperatures from 1891 to 1985 was considered for the majority of stations.

The coefficient K suggested by N.A.Bagrov (1966) was used as an anomaly test. It is expressed by

$$K = \frac{1}{N} \sum_{i=1}^{N} \frac{(T_i - \bar{T}_i)^2}{\sigma_i} ,$$

where T_i, $\bar{T}_i$ are the current and long-term mean values of the monthly mean air temperature for the i-th station, respectively; σ_i is the standard deviation of monthly mean air temperature for the i-th station; and N is the number of stations considered.

The statistical characteristics of the anomaly coefficient are considered in the paper by N.A.Bagrov (1966), as well as in a number of subsequent papers (see, for example, Myakisheva, 1973). It is clear that the mean value of K is equal to 1.00 and the value N·K has χ^2- distribution with N degrees of freedom. In practice, however, even with a very sparse network of stations the values of the temperature anomaly at different stations would not be independent, Therefore the empirical function of the distribution of the values of K can be compared with the χ^2 - distribution with fewer degrees of freedom.

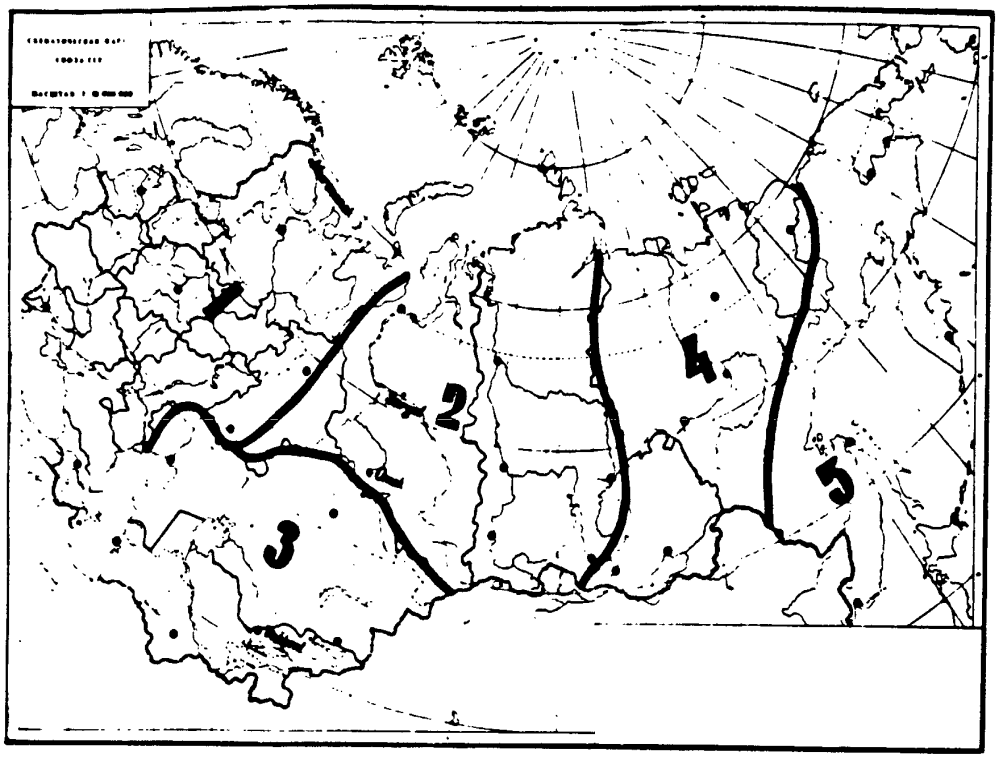


Fig.1. A schematic map of the distribution of stations, used in the analysis of temperature patterns of the cold periods of 1982-83, 1983-84 and 1984-85, over the USSR area (1 - the USSR European part; 2 - Western Siberia; 3 - Central Asia and Kazakhstan, 4 - Eastern Siberia; 5 - Far East).

The peculiarities of the temperature patterns of cold periods of the year (November - March) for 1982-83, 1983-84 and 1984-85 have been defined using the anomaly coefficient K. The values of K were computed for each month of the cold period from November to March for 1891-1985 for each of the five areas of the USSR (see Fig.1). The anomaly coefficients K for cold periods of 1982-83, 1983-84 and 1984-85 are listed in Table 1, which also shows the ordinal number of the anomaly coefficient in the ranked 1891-1985 series for each area (the ordinal number 1 corresponds to the largest value of the coefficient K). The temperature anomaly sign was defined on the basis of the air temperature anomaly field analyses for each month.

The analysis of empirical functions of the distribution of K showed that the boundary values for intervals holding the most probable values of K were essentially different for each of the cold period months in the USSR area, but the upper interval boundary, which corresponds to the accumulated probability of 66.7% does not exceed K=1.25. Hence, the $K>1.25$ values of K can be considered as anomalous. One can see from Table 1 that the number of such values of K for the cold period in 1982-83 was 47%, in 1983-84 - 44%, and in 1984-85 - 28%. Note that there occurred the values of K close to the maximum values for the 95-year series (i.e. K=5.68 for Central Asia and Kazakhstan in December 1984 and K=2.95 for Western Siberia in November 1983).

A peculiarity of the cold periods of 1982-83 and 1983-84 is

Table 1

Space - time distribution of the anomaly coefficient K over the USSR during the cold periods of 1982-83, 1983-84 and 1984-85 (see text for explanations)

Area	:November:	December	: January	: February	:March
			1982-1983		
1.USSR European part	1.34^{+}(24)	1.92^{+}(12)	2.06^{+}(9)	1.17^{+}(29)	0.34^{+}(80)
2.Western Siberia	1.37^{+}(22)	2.57^{+}(7)	2.22^{+}(11)	1.08^{+}(35)	1.46^{+}(17)
3.Central Asia and Kazakhstan	2.20^{-}(11)	0.35^{+}(66)	1.54^{+}(22)	2.34^{+}(10)	0.51^{+}(61)
4.Eastern Siberia	2.84^{-}(5)	0.97^{+}(40)	0.77^{+}(51)	0.35^{+}(81)	1.57^{+}(16)
5.Far East	0.42^{-}(78)	0.70^{+}(52)	1.24^{+}(27)	0.74^{-}(50)	0.88^{+}(40)
			1983-1984		
1.USSR European part	0.28^{-}(89)	0.95^{+}(33)	1.41^{+}(21)	0.66^{+}(54)	0.48^{+}(70)
2.Western Siberia	2.95^{+}(3)	2.34^{+}(8)	0.98^{+}(39)	0.60^{+}(56)	0.81^{+}(46)
3.Central Asia and Kazakhstan	1.65^{+}(14)	0.72^{+}(41)	0.59^{+}(46)	2.35^{-}(9)	0.48^{+}(62)
4.Eastern Siberia	2.70^{+}(7)	1.80^{+}(12)	1.30^{+}(28)	1.10^{-}(30)	0.28^{-}(81)
5.Far East	1.43^{+}(22)	1.97^{+}(11)	1.21^{+}(29)	1.39^{-}(22)	0.74^{+}(48)
			1984-1985		
1.USSR European part	1.02^{-}(35)	1.11^{+}(29)	2.08^{-}(8)	3.38^{-}(4)	1.27^{+}(21)
2.Western Siberia	1.85^{-}(13)	0.71^{-}(53)	0.49^{-}(58)	2.12^{-}(11)	0.35^{+}(84)
3.Central Asia and Kazakhstan	0.53^{+}(59)	5.68^{-}(3)	0.22^{+}(83)	0.70^{+}(43)	0.89^{-}(32)
4.Eastern Siberia	0.72^{+}(52)	0.38^{-}(84)	0.53^{-}(63)	2.56^{+}(6)	0.07^{+}(95)
5.Far East	0.59^{+}(66)	0.09^{-}(92)	0.61^{-}(58)	0.54^{+}(62)	1.22^{-}(23)

associated with the formation of an extensive zone of the positive air temperature anomalies covering the whole of the USSR area (see Table 1). This zone is most distinct during the two central months of the season (December and January). The zones of the positive anomaly maxima for both periods are in Western Siberia (area 2), where the anomaly coefficients in December amount to 2.57 and 2.34. Note the unusually large duration positive temperature anomalies in indivilual areas of the USSR. For example, there were positive anomalies observed from November to March, 1982-83 and from December to March, 1983-84 in the European part of the USSR (area 1) and Western Siberia (area 2). The anomaly coefficient values averaged for November - March for Western Siberia are close to the maximum values for the period 1891-1985.

The formation of a vast zone of positive air temperature anomalies over the USSR in a cold period of the year is not unique. According to G.Ya.Vangengeim's (Vangengeim, 1940; Girs, 1974) classification of the atmospheric circulation forms, the increase in air temperature in a cold period of the year over the USSR due to the intensification of the western air mass transfer is one of the characteristic features of the "western" (W) circulation type. However, the formation of such vast zones of positive temperature anomalies in two consecutive cold periods is a rare occurence and is first observed in the last 95 years.

The record monthly mean air temperature anomalies, reported for December 1982 for stations located in the zone of maximum anomaly coefficients (Western Siberia, area 2), are also demonstrations of anomaly. Figs.2a and 2b give the monthly mean air temperature anomaly series obtained at stations Omsk in January and Surgut in December for 1891-1985. The temperature anomaly value reported for January 1983 for Omsk is 8.9°C, which is the largest one and exceeds 2σ.

The cold period of 1984-1985 is essentially different from the two previous periods by the nature of temperature variations over the USSR

individual areas. An extensive zone of negative temperature anomalies was observed over Kazakhstan and Central Asia (area 3) in December 1984 with the anomaly coefficient K=5.68 which is the third-inmagnitude value for 1891-1985. The monthly mean temperature anomalies reported for individual stations amounted to -14.0°C with the long-term mean being -18.0°C. After cooling in 1944 there was an increase in the temperature anomalies observed at the stations in Central Asia up to 1984 when a record low temperature was reported. Fig.2c shows the long-term air temperature anomaly variations for the station Taskent. The cold air intrusion into Central Asia in December 1984 was the most intensive one over last 50 years.

In January 1985 the zone of significant negative temperature anomalies moved to the northern European part of the USSR (area 1). The anomaly coefficient was 2.08 in January and 3.38 in February 1985. Fig.2d gives the air temperature anomalies for Murmansk for January 1920-1985. The temperature anomaly registered in January 1985 was -11.1°C which is the largest one for the observation period considered for this station.

The intensive cold air mass intrusion into the European part of the USSR in January - February 1985 can be compared to the unusual development of the Siberian High during the cold period of 1984-85. Fig.3 shows the long-term mean surface pressure field in the North-Eastern zone of the Northern Hemisphere for December (dashed line), as well as the structure of the surface pressure field for December 1984 for the same area.The main difference lies in the north-western ridging of the Siberian High which is unusual as compared to the north-eastern one typical of the long-term mean pressure field. Such an anomaly and its effect on the formation of climatic conditions in the USSR European part have been observed before (Girs, 1974). However, the essentially anomalous conditions in the USSR European part in January 1985, resulting in the record negative temperature anomalies in the northern European part of the USSR, tells of a necessity to study in more detail the rela-

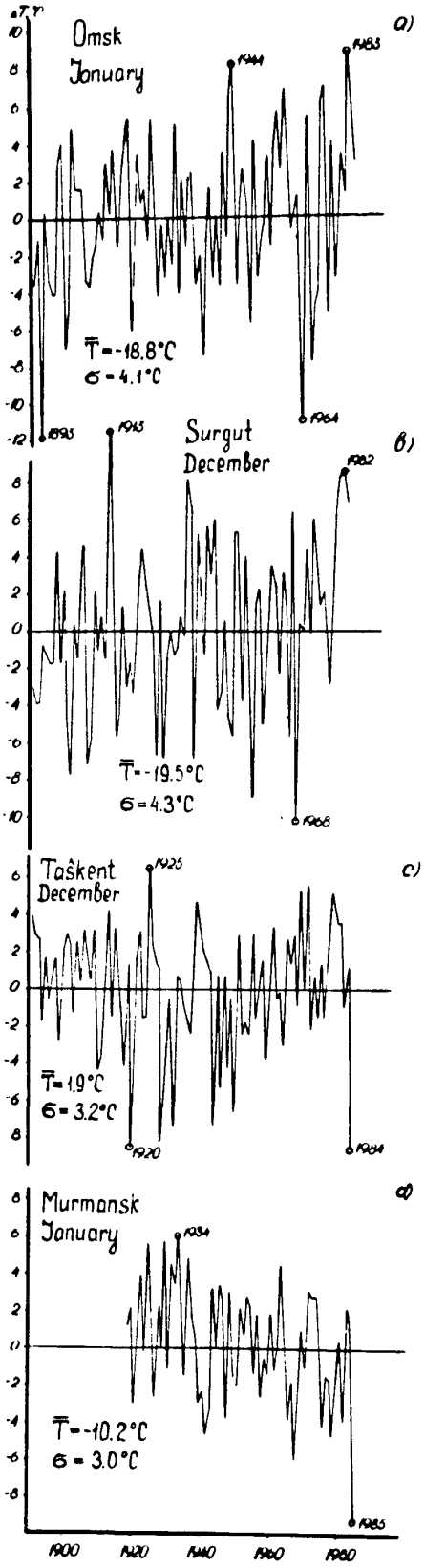


Fig.2. Long-term series of monthly mean air temperature anomalies for individual stations of the USSR.

tionship between the Siberian High formation and development and the air temperature anomaly distriburion over the USSR.

III. CONCLUSION

The climatic perculiarities of the winters of 1982-83, 1983-84, 1984-85 individual regions of the globe have been discussed before (Quiroz, 1983;

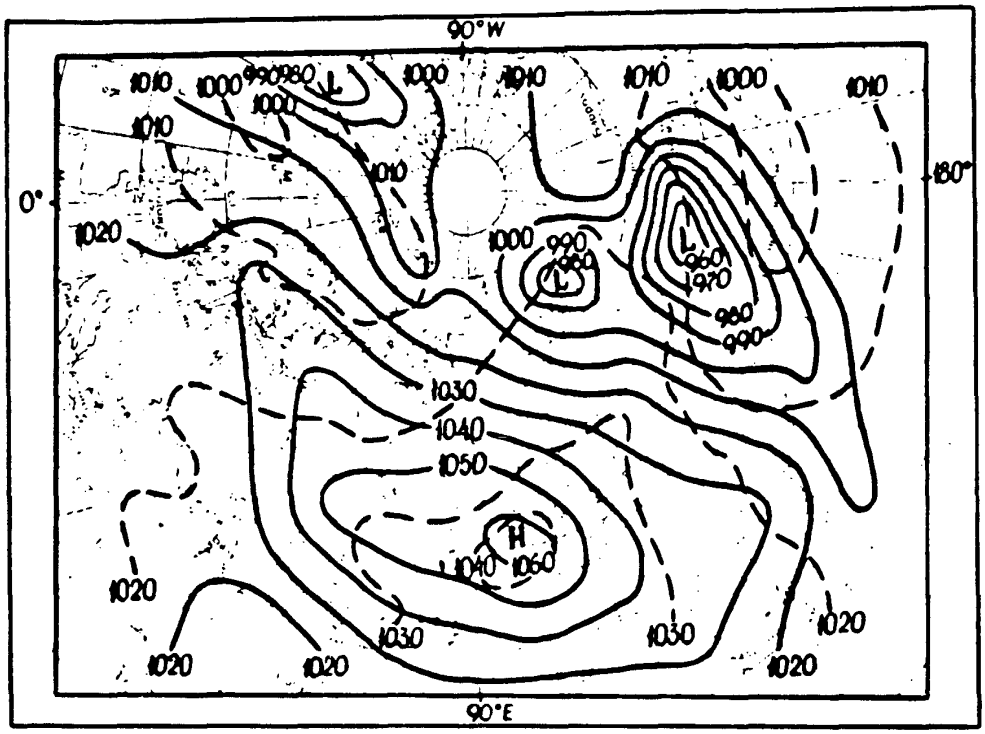


Fig.3. Surface pressure field (hPa) for the eastern Northern Hemisphere in December: long-term means for the period 1881-1935 (dotted line) from Sokhrina et al., 1959; a structure of the surface pressure field (solid line) for 15 Dec'84.

Quiroz, 1984; Kousky, 1985). In this paper we consider the anomalies in the temperature patterns of the cold year periods in the USSR, The coefficient K suggested by N.A.Bagrov (1966) has been used as an anomaly test. The statistical analysis of the long-term values of the coefficient K enables us to see the extent of anomaly of the air temperature anomaly field in the USSR individual areas. Each of the cold periods considered shows certain anomaly signs: while during the cold periods of 1982-83 and 1983-84 zones of positive temperature anomalies were observed over the great part of the USSR with the record positive monthly mean air temperature anomalies registered in individual locations (Omsk, Surgut); during the cold period of 1984-85 extraordinary low temperatures were reported for the area of Kazakhstan and Central Asia (December 1984) as well as for the northern European part of the USSR in 1985. The anomaly coefficients in some cases amounted to values close to the maximum ones for the whole of the period of record (1891-1985) under consideration.

The analysis of peculiarities of the cold period temperature patterns in the USSR gives additional information for studying climate variability in individual regions of the globe and is essential for the development of methods used in monitoring the climate system state.

The authors are grateful to N.Kazachkovskaya for preparing the English version of the paper and H.Devyatova for typing it.

REFERENCES

1. Bagrov N.A., Myakisheva N.N., 1966: Some characteristics of the monthly mean air temperature anomalies, Tr. of World Met. Centre, i.9,3-17 (in Russian).

2. Girs A.A., 1974: A macrocirculation method of the long-range meteorological forecasts, Gidrometeoizdat, Leningrad, 486pp (in Russian).

3. Gruza G.V., Birman B.A., Rubinshtein K.L., Razuvaev V.N., 1984: Main problems of climate monitoring (diagnoses), Izvestiya AN USSR, Phizicka atmosphery i okeana, N11, 3-25 (in Russian).

4. Kousky V.E., 1985: The Global Climate for December 1984 - February 1985: A Case of Strong Intraseasonal Oscillations, Mon. Wea. Rev., v.113, 2158-2172.

5. Myakisheva N.N., 1973: The anomaly of monthly mean temperature fields and their intermonthly variability, Tr. of Hydromet. Centre, i.106, 32-50 (in Russian).

6. Quiroz R.S., 1983: The Climate of the "El-Nino" Winter of 1982-83 - A Season of Extraordinary Climatic Anomalies, Mon. Wea. Rev.

7. Quiroz R.S., 1984: The Climate of the 1983-84 Winter - A Season of Strong Blocking and Severe Cold in North America, Mon. Wea. Rev., v.112, 1894-1912.

8. Sokhrina R.F., Chelpanova O.M., Sharova V.P., 1959: Air pressure and temperature for the Northern Hemisphere, Gidrometeoizdat, 474pp (in Russian).

9. Vangengeim G.Ya., 1940: Long-range air temperature forecast and debâcle, Tr. of State Hydrolog. Inst., i.10, 207-236 (in Russian).

10. WMO, 1985: Climate System Monitoring. A Critical Review of the Climate System during 1982-1984, WMO Edition, 52 pp.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

MONTE CARLO METHOD FOR RISK ANALYSES OF CLIMATIC DAMAGES AFFECTING THE YIELDS OF CROPS

Yu Xi-min Pen Wei-bin
Institute of Meteorology, Liaoning Province
No.1, Lane 10, Sec. 4, Peace St., Shenyang, China

ABSTRACT

First of all, the authors obtained four equations on the climatical damages that affect the yields of crops. Then, different independent variables in these equations were simulated by random numbers generated in computer. The results indicated drops of yields under combinations of various climatic damages in the past historical period. Both histograms and curve diagrams for risk analyses were plotted and yield reductions on various accumulated probabilities were given. Finally, the authors discussed the perspectives of some applications of Monte Carlo Method to agrometeorology, in concern with the basic law of the subjective method inevitably developing to objective one in meteorology.

1. INTRODUCTION

Climatic conditions are of great varieties in the past historical period. Yields reduced climatically that were computed from meteorological data and actual yields are not typical results. Some classical statistical methods hitherto used in agrometeorology are statistical inference techniques making use of some observed data. It is our task to lay bare the objectively actual states that can not be revealed by the classical statistical methods. In other words, the riskiness of climatic damages should be pointed out so as to make a correct decision in agricultural production.

However,it is difficult to simulate the climatic damages affecting the yields of crops: both actual field experiments and artificial climatic studies are expensive. At present, it can not be accomplished. Consequently, considered were the various possibilities of the riskiness studies of climatic damages affecting the yields of crops with Monte Carlo Method (Gold, 1977).

2. BASIC DATA AND PREPARATION OF PROBABILITY TABLES

The meteorological data were obtained from the observed results in Jianping, Fuxin, Yixian and Beipiao meteorological stations during 1953—1982. The phenological data were agrometeorological results observed on maize crops in these counties (In China, the phenological stages were recorded according to "the observational standard on agrometeorology" issued by the Central

Meteorological Administrition). In Jianping, the equation of climatically reduced yields of maize crop is as follows:

$$Y=-3.5+5.5X_1+2.5X_2+1.4X_3-0.5X_4 \quad (1)$$

where Y is the yield reduction caused by climatic damages, X_1 is the number of days in mean air temperature below 16°C in maturation stage of maize, X_2 is the drought during the whole growing period, X_3 is the beginning date of the Autumn frost, X_4 is the rainfall shortage from the previous winter to the spring when the maize grew.

In order to analyse the riskiness of these climatic damage effect on crops with Monte Carlo Method, the subjective probabilities of the four independent variables X_1, X_2, X_3, and X_4 must be obtained. These probabilities were shown in Table1 —Table4. For the other three counties, similar regression equations and corresponding tables were obtained respectively.

3. MATHEMATICAL MODEL AND COMPUTING STEPS

First, random numbers were obtained in computer FELIX C-512 by means of random number generator (Enslein et al., 1977). Then, risk analysis was done on the basis of random numbers. For subjective probability values, see Table1—Table4.

It is well known that the four independent variables are uncertain. In fact, in meteorological and biological fields, there are many different view-points on climatic variation and its influence on biological characteristics of crops (Takahashi, 1983). Different factors used to describe agrometeorological damages are stochastic. The values of accumulated probabilities P' of independent variables were given in Table1—Table4. A random number was taken and then compared with the values of accumulated probabilities shown in Table1. This random number would certainly fall into a proper line in that table, experimental value X_1 corresponding to this P' being obtained, and X_2, X_3, and X_4 would be obtained in similar manner. X_1, X_2, X_3, and X_4 were substituted into expression (1) and a value of Y was obtained.

The whole process mentioned above was done repeatedly on the basis of other random numbers. Since these random numbers are stochastic, the percentage frequencies of combination of different uncertain independent variables would approximate to their actual probability distributions when these processes were completed.

Immediately following these simulation computations, the yield reduction Y_i caused by various combinations of climatic damages was obtained (where subscript i is the order sign of random numbers). Y_i was divided into some equal-length increment, with Y=2.5 kg/mu as its step length (1mu=0.0667 ha). Then, the frequencies of Y_i in each Y printed by computer were found, and the objectively accumulated probabilities were computed. On the basis of these results, both histogram of frequencies and diagram of the objectively accumulated probabilities were plotted. The ordinates of these histograms were regarded as the experimental values of density functions of their distributions, and the ordinates of accumulated probability diagram

X_1	-4	-3	-2	-1	0	1	2	3	4
Frequency	3	1	6	5	4	4	2	3	2
Probability	0.10	0.03	0.20	0.17	0.13	0.13	0.07	0.10	0.07
Accumulated Probability P'	0.10	0.13	0.33	0.50	0.63	0.76	0.83	0.93	1.00
1000 P'	100	130	330	550	630	760	830	930	1000

Table1 Probabilities of the number of days of low temperature in maturation stage

X_2	-7	-4	-3	-2	-1	0	1	2	3	5	6
Frequency	1	1	1	2	1	1	2	1	2	1	5
Probability	0.03	0.03	0.03	0.07	0.03	0.03	0.07	0.03	0.07	0.03	0.17
Accumulated Probability P'	0.03	0.06	0.09	0.16	0.19	0.22	0.29	0.32	0.39	0.42	0.59
1000 P'	30	60	90	160	190	220	290	320	390	420	590

X_2	7	8	9	10	11	12	14	18	23	31	32
Frequency	2	1	1	2	1	1	1	1	1	1	1
Probability	0.07	0.03	0.03	0.07	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Accumulated Probability P'	0.66	0.69	0.72	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1.00
1000 P'	660	690	720	720	820	850	880	910	940	970	1000

Table2 Probabilities of the drought during the growing period

X_3	-21	-18	-16	-13	-12	-10	-9	-8	-7	-5	-4
Frequency	1	1	1	2	1	2	2	1	2	2	1
Probability	0.03	0.03	0.03	0.07	0.03	0.07	0.07	0.03	0.07	0.07	0.03
Accumulated Probability P'	0.03	0.06	0.09	0.16	0.19	0.26	0.33	0.36	0.43	0.50	0.53
1000 P'	30	60	90	160	190	260	330	360	430	500	530

X_3	-3	-2	-1	0	1	2	4	5	6	8
Frequency	2	1	1	3	1	2	1	1	2	1
Probability	0.07	0.03	0.03	0.10	0.03	0.07	0.03	0.03	0.07	0.08
Accumulated Probability P'	0.60	0.63	0.66	0.76	0.79	0.86	0.89	0.92	0.99	1.00
1000 P'	600	630	660	760	790	860	890	920	990	1000

Table3 Probabilities of the beginning date of the Autumn frost

X_4	-52.3	-48.1	-3.09	-30.6	-26.5	-24.4	-220
Frequency	1	1	1	1	1	1	1
Probability	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Accumulated Probability P'	0.03	0.06	0.09	0.12	0.15	0.18	0.21
1000 P'	30	60	90	120	150	180	210

X_4	-20.7	-20.4	-19.2	-17.2	-14.5	-12.9	-7.2	-6.9
Frequency	1	1	1	1	1	2	1	1
Probability	0.03	0.03	0.03	0.03	0.03	0.07	0.03	0.03
Accumulated Probability P'	0.24	0.27	0.30	0.33	0.36	0.43	0.46	0.49
1000 P'	240	270	300	330	360	430	460	490

X_4	-5.7	-2.3	-2.2	1.5	3.6	8.3	8.5	8.5
Frequency	1	1	1	1	1	1	1	1
Probability	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Accumulated Probability P'	0.52	0.55	0.58	0.61	0.64	0.67	0.70	0.70
1000 P'	520	550	580	610	640	670	700	700

X_4	10.6	12.7	28.3	28.7	30.2	34.8	37.0	38.2
Frequency	1	1	1	1	1	1	1	2
Probability	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.07
Accumulated Probability P'	0.73	0.76	0.79	0.82	0.85	0.88	0.91	1.00
1000 P'	730	760	790	820	850	880	910	1000

Table4 Probabilities of the rainfall shortage from the previous winter to the spring when the maize grew

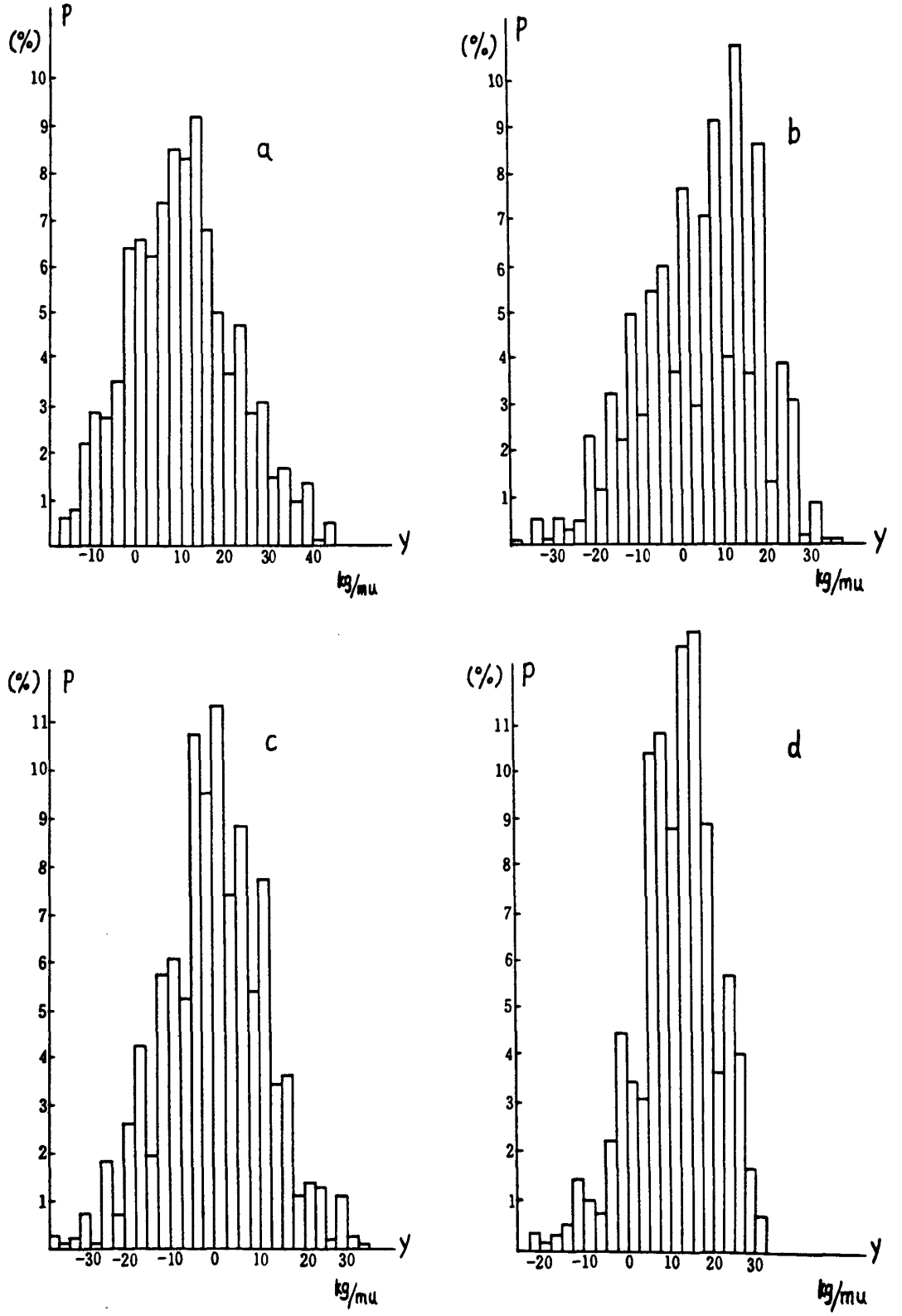


Fig.1 Histograms of frequencies of yield drop caused by climatic damages: a—Jianping b—Fuxin c—Yixian d—Beipiao

as the experimental values of the distribution fuctions. In order to verify the statistical simulation results based on Monte Carlo Method, still more repeated computations had to be conducted until both more satisfactory histograms (Fig. 1) and corresponding diagrams of distribution functions (Fig. 2) were prepared.

In our simulation riskiness analyses with Monte Carlo Method, computer was an indispensable tool. As random numbers were obtained from random number generator by means of physical method, standard FORTRAN-4 was used to accomplish the simulation work on computer. It should be noted that the maximum accumulated probabilities of different independent variables must be greater than any random number used, otherwise "the dead cycle" caused by roundoff errors would occure.

4. RESULTS AND SOME POSSIBLE APPLICATIONS

Monte Carlo Method is a very useful in practical agroclimatic work. Now, let us analyse its possible applications.

(1)Monte Carlo simulation process is a process in which objective probabilities can be inferred from frequencies. In agroclimatological analyses of crop damages on basis of this method, various objective percentages of yield drop caused by climatic damages can be determined, and the riskiness of crop yield reduction can be mastered according to climatic characteristics in many years. For example, with the accumulated probability at 80.0%, the yield reduction is less than 21.25 kg/mu. In other words, the riskiness of yield reduction above 21.25 kg/mu is 20.0%. For many counties located in the same climatic region, comparison was made among yield drops, in order to make decision in planning geographical distributions of crop growing. The yield drops within the four counties under different accumulated probabilities were shown in Table 5.

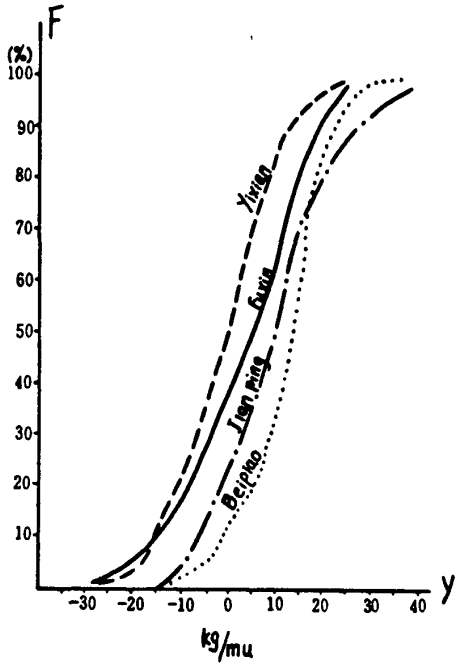


Fig.2 The distribution function of yield reduction deduced from Monte Carlo Method

Yield drops in all the counties within the same climatic region are comparable with an identical accumulated probability. For example, with the accumulated probability at 80.0%, the yield drop in Jianping is less than 21.25kg/mu, whereas that in Fuxin is less than 16.50 kg/mu. It is clear that the severity of damage effect on maize crop in Jianping is greater than that in Fuxin. This fact indicates that the riskiness of maize planting planning in Fuxin is less that in Jian-

ping in normal years.

Names of Counties	Accumulated probability (%)						
	65	70	75	80	85	90	95
Jianping	14.50	16.00	18.50	21.25	23.75	26.75	32.50
Fuxin	11.50	13.50	14.50	16.50	18.50	21.50	25.00
Beipiao	15.75	16.50	17.50	19.00	21.00	23.75	26.00
Yixian	5.00	6.50	8.00	10.50	12.00	14.50	19.50

Table 5 Yield drop deduced from Monte Carlo Simulation

(2)The formula determining the agro-economic output loss W caused by climatic damages in the whole county is as follows:

$$W=mn\,F(0.80) \qquad (2)$$

where m is selling price of maize (in RMB yuan), n is the total area (in mu) of land to grow maize in a whole county, F(0.80)is yield drop with distribution function at 80.0% (the riskiness is 20.0%). For different subscripts of F values in expression (2), the computation results were shown in Table 6. Once such a table was prepared, the actual agro-economic loss of crops caused by climatic damages can be easily estimated.This will be a valuable work for ditermination of agricultural taxation, computation of insurance premium, and distribution of relief fund.

(3)With Fig.2 and given yield drop, recheck of riskiness R_j of maize growing in the four counties could be conducted. For example, given the value of abscissa as 12.5 kg/mu, the corresponding value F_j of ordinate can be spotted. Here the subscript j of F_j and R_j is a denoting sign of county name. Thus we have

$$R_J=1-F_j \qquad (3)$$

The values of F_j and R_j in the four counties are as follows:

Jianping F_J=57.5% R_J=42.5%
Beipiao F_B=49.0% R_B=51.0%
Fuxin F_F=66.5% R_F=33.5%
Yixian F_Y=87.0% R_Y=13.0%

That is to say, the riskiness of loss above 12.5kg/mu caused by climatic damages in Beipiao county is greater than the other three counties,with its value being 51.0% to ensure a yield drop by climatic damages less than 12.5kg/mu, it would be the best decision to plant maize in Yixian, because the riskiness there is only 13.0%.

(4)Analyses of extreme values and extreme differences:From Fig.1, range U for yield drops (in Kg/mu) in all the counties concerned were found as follows:

Jianping U_J=65.0 from-20.0 to 45.0
Beipiao U_B=55.0 from-22.5 to 32.5
Fuxin U_F=77.5 from-40.0 to 37.5
Yixian U_Y=72.5 from-37.5 to 35.0

Extremum analyses showed that the most remarkable yield drop had occured in Jianping(45.0kg/mu), with the least yield drop in Beipiao (32.5kg/mu). In order to avoid extremly low yield, it is most approprite to plant maize in Beipiao. From the results of extreme diffe-

rence analyses, it was obvious that the extreme difference in Fuxin is the greatest of the four counties, its value being 77.5kg/mu.This fact shows that the difference value between high yield and low yield is very remavkable in Fuxin, while it is very small in Beipiao, its value being only 55.0kg/mu. The absolute range of extreme difference marks the stability of yield. In the process of decision making for crop planting, this should be taken into full consideration.

(5) In planning crop planting in reference to prediction of various meteorological elements in the year to come, it is important first of all to forecast the meteorological elements related to different independent variables in expression(1), then,with them as predictors, forecast dependent variable, and infer the riskiness and make a rational arragement for crop-area in each region.

(6) Quantitative agroclimatic impact assessment during one year period: In order to assess the effect of climatic damages on agricultural production, it is necessary to substitute the observational values of the meteorological elements in this year for the independent variables of expression (1). For example,values of the four independent variables were shown in Table 7,while the yield drop deduced from these independent variables was shown in the final column of the table.

From Fig.2,the yield drop caused by climatical damages were divided into several classes and then Table 8 was prepared. The value in 1983 was compared with those shown in Table 8, indicating that the classes of yield drops in the four counties are different: Jianping in class 4, both Fuxin and Beipiao in class 1, whereas no meteorological damages observed in Yixian. It is obvious that yield reduction in Jianping was caused by drought in the whole growing period and Autumn frost. However, in Yixian, the drought was offset by the other three factors. It should be emphasized that the yield reduction caused by climatic conditions is not the actual yield drop. If the actual yield drop in Jianping and Yixian were almost same, it means that other factors (e. g. chamical fertilizer, social conditions et.al.) brought about the increase in yield.

5. CONCLUSION AND DISCUSSION

The simulation analysis of riskiness of climatic damages affecting crops is a preliminary experiment of subjective method shifting to objective one in meteorological damage analyses. Because this work is quite new in meteorology,and it was restricted by orginal data,objectivity is limited.In the first place, so far as the form of equation is concerned, linear equation is not the sole form. In the process of developing meteorological studies, it is inexorable trend to develop some techniques from subjective to objective one, namely from qualitative method to cuantitative one. For example, the objective analysis of synoptic chart was introduced step by step into operational work. The process developed from Yang Jian-chu's curve in early 1950'S to time series analyses in China's meteorological work is also a good

Names of Counties	Subscript 0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
Jianping	4.3	4.6	5.0	5.3	5.6	6.0	6.3	6.6
Fuxin	8.6	9.3	10.0	10.7	11.3	12.0	12.7	13.3
Beipiao	3.0	3.2	3.5	3.7	3.9	4.2	4.4	4.6
Yixian	1.6	1.7	1.8	2.0	2.1	2.2	2.3	2.5

Table 6 Output loss caused by climatic damages (in 100,00 yuan)

Names of Counties	Variable X_1	X_2	X_3	X_4	Y
Fuxin	-4	38	-11	-102.9	0.3594
Jianping	-4	40	5	-60.8	51.9917
Beipiao	-4	33	-3	-80.2	11.6105
Yixian	-4	26	-8	-128.1	-15.6981

Table 7 Observed meteorological elements and yield drops in 1983

Yield drop (kg/mu)	0-10	10-20	20-25	25-30	30-50
Class	1	2	3	4	5
Name	very light	light	middle	middle-heavy	heavy

Table 8 Classes of yield drop caused by climatic damages

example: Yang's curve was known as "long-term forecasting method of China" by Gils (a noted Soviet meteorologist) in 1960, it is clear that this method was very advanced in that period, but at present, because of the use of time series analysis and computer in meteorology, many quantitative techniques were used for forecasting (An, 1983); Applications of Chebyshev polynomials in agrometeorological yield forecast (Y_M, 1985) is also a development stage based on semi-qualitative method that had been used. In climatology, the extreme value that occurred once in many years was calculated in 1940'S with actual observed data, but in 1960'S it could be inferred theoretically (Jenkinson, 1955; Yao, 1963). This fact is another example which shows

the development from the subjective method to the objective one.In such studies as we have done, with the increase in data of subjective probabilities, still more objective results may be obtained with Monte Carlo Method for damage riskiness analysis.

Some errors of objective results frequencies. In order to minimize such errors, suitable step length ΔY has to be selected, which is relative to the exactness of climatic yields and requirement of problems. It is very difficult to test the forms of the distributions by means of histogram using general statistical test method, it is also true for those parameters in density functions and distribution functions. It is perhaps necessary to simulate their distributions for many points.

The forms of yield drops caused by climatic damages are very camplex. In our study, only four rough indice on damages were synthesized. In fact, the actual values of such indice should be continuous. It is possible to obtain a more satisfactory density functions and distribution functions, and those data can be treated by pattern recognition techniques in computer if expression such as $Y=f(X_1, X_2,...)$ is obtained by mathematical analysis method, and more values of independent variables are taken.

6. REFERENCES

An.H.Z., 1983: Time series analyses and their applications.Academic press, Beijing, 346pp.

Enslein.K.,et.al., 1977:Statistical methods for digital computers. Wiley & Sons, New York, 376pp.

Gold.H.J., 1977: Mathematical modeling of biological systems: an introductory guide book. Wiley and Sons, New York, 357pp.

Jenkinson.A.F., 1955: The frequency distributions of annual maximum (or minimum) values of meteorological elements, Quart. J. Roy. Meteorol. Soc., 81, 58-59.

Takahashi.A.,1983: Climatic change and food production, Meteorological Press. Beijing, 402pp.

Yao.Z.S., 1963:Climatic statistics, Academic Press, Beijing, 306pp.

Yu. X.M., 1985: Forecasting the distribution of meteorological yield using Chebyshev polynomial, Agricultural and Forest Meteorology, 34, 323-332.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

IMPROVING ESTIMATORS OF CROP YIELD DISTRIBUTIONS BY THE USE OF HISTORICAL RAINFALL DATA

Tom A.B. Snijders
Institute of Econometrics
University of Groningen, The Netherlands

INTRODUCTION AND SUMMARY

This paper was written in the context of a research project, called AGRISK, on risks in agriculture in the semi-arid zone of West Africa. This research project is concerned, among others, with the estimation of variability of crop yields. The paper presents a method for estimating mean and variance of crop yields, when in addition to crop yield data there is available historical rainfall data which can be transformed by a rainfall-yield model. The method is a generalization of sample augmentation by multiple regression, well known in hydrology (Moran, 1974). Nonparametric standard errors of the estimates are obtained by the bootstrap method (Efron, 1979). The method is illustrated for the case of sorghum yields in Burkina Faso.

SORGHUM YIELDS AND YIELD INDICES FOR BURKINA FASO

Burkina Faso (formerly called Upper Volta) is one of the so-called Sahel countries situated in sub-Saharan West Africa; most of its territory belongs to the semi-arid zone. The rainy season is practically completely confined to the months May-October. Sorghum and millet are the main food grains. They are cultivated mainly without irrigation in a semi-traditional way. Insufficient nutrient availability is an important yield limiting factor but, as this factor is not very variable from year to year, yield fluctuations for these grains are quite related to fluctuations in the yearly rainfall regime.

Let Y_j be sorghum yield in year j for Burkina Faso as a whole, and consider the estimation of its mean and standard deviation. It is assumed in this paper that this mean may depend on the year j while the standard deviation does not, leading to the notation

$$\mu_j = EY_j \qquad \sigma^2 = \mathrm{Var}\, Y_j .$$

Because of the relationship between rainfall and sorghum yields it can be useful to augment the data available on Y by data on a suitable rainfall-based yield index Z; this is meant here to be the result of a rainfall-based crop yield model, constructed so as to be fairly highly correlated with sorghum yields. For a certain set of years (called the primary sample), data (y_j, z_j) on both Y and Z are available; for a disjoint set of years (called the auxiliary sample), data z_j on Z only are available.

Sorghum yield data used in this paper are those presented in FAO production yearbooks. As statistical data collection in Burkina Faso is not highly developed, and as it is especially hard to collect reliable information on subsistence crops such as sorghum, this FAO data cannot be expected to be very reliable. Therefore the numerical outcomes presented in this paper are given only as an illustration.

For the yield index Z, daily rainfall data is used for 20 rainfall stations in the sorghum-growing zone in Burkina Faso, all stations situated between 11°10'N and 13°32'N. For every station i and every year j for which complete daily rainfall values for May-October were available for this station, an index

$$I_{ij} = E_{ij}/E^{o}_{ij}$$

was calculated, where

E_{ij} = actual evapotranspiration over the growing period

E^{o}_{ij} = potential evapotranspiration over the growing period.

The values E_{ij} and E^{o}_{ij} were calculated from a water balance model operating with daily time steps, using daily rainfall totals and climate and crop parameters relevant to each station; details are given in Snijders and Schweigman (1985). The ratio E_{ij}/E^{o}_{ij} indicates the extent to which rainfall was sufficient for optimal crop growth.

For every year j, the index values I_{ij} for available stations i must be combined into the single value z_j. It is argued in Snijders (1986), that, in view of the large number of missing data, one should not use for this purpose an average (weighted or unweighted) of the available I_{ij}, but that it is preferable to calculate z_j from an additive fit

$$I_{ij} = \beta_i + z_j + r_{ij}$$

where β_i are "station effects" and z_j "year effects" while r_{ij} are residuals. In view of the not uniformly high data quality, a robust method of fitting was used, namely a polish by 40% Winsorized means (Snijders, 1986). Results are given in Figure 1, while the relationship between y_j and z_j is shown in Figure 2.

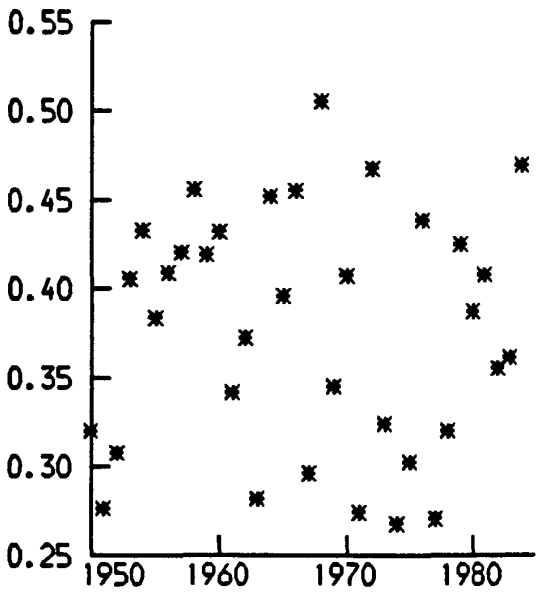


Figure 1. Rainfall-based yield index z_j for 1950-1984

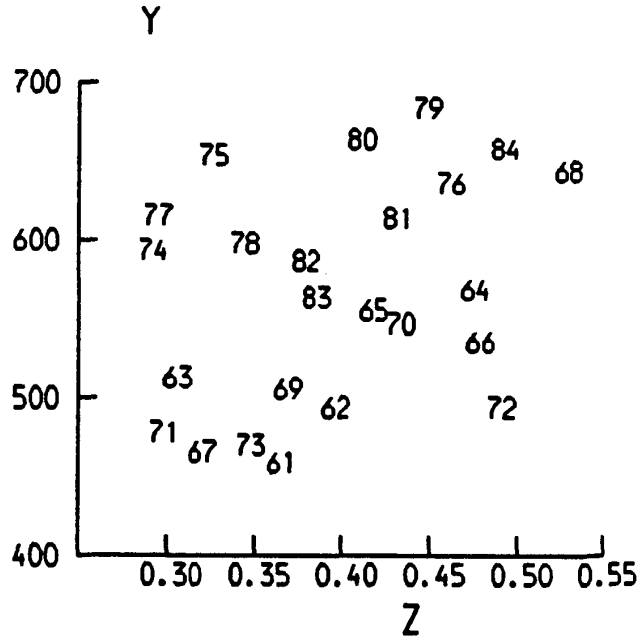


Figure 2. Relation between FAO yield data y_j and index values z_j for 1961-1984.

ESTIMATION WITH AUGMENTED DATA AND REGRESSION ON NON-RANDOM VARIABLES

Estimation with augmented data is a well-known technique; see Moran (1974). It is usually assumed for this method that the primary and the auxiliary samples are random. In the present application this is not a a realistic assumption because

a. agricultural methods have been advancing, so it is to be expected that the conditional distribution of crop yields, y_j, given the index z_j, is subject to an increasing trend as a function of the year date j;

b. the rainfall regime in the Sahel region has not been stationary, but is considered to be at a lower level since about 1970 than before 1970; see Lamb (1985) and Snijders (1986).

Point a is illustrated in Figure 2, where later years are seen to occupy higher positions than earlier years. Point b is illustrated to some extent in Figure 1. In order to incorporate these points in the model, it is assumed that the bivariate distribution of (Y_j,Z_j), of which (y_j,z_j) is the outcome, is subject to regression on the year date j; the regression function for both Y_j and Z_j is supposed to be a continuous piecewise linear function consisting of two linear parts with the bend in 1969. This amounts to linear regression on (x_{j1},x_{j2}) where $x_{j1} = j$ is the year date and $x_{j2} = \max(0,j-1969)$.

More generally, consider the model

$$\begin{aligned} Y_j &= \alpha_Y + x_j\beta_Y + U_j \\ Z_j &= \alpha_Z + x_j\beta_Z + V_j \end{aligned} \qquad (*)$$

where the $x_j = (x_{j1},\ldots,x_{jk})$ are given vectors of non-random explanatory variables, the numbers α_Y, α_Z and the vectors β_Y, β_Z are unknown parameters, and the (U_j,V_j) constitute a random sample from some bivariate probability distribution; for $j \in J_p$ (the primary sample) Y_j, Z_j and x_j are observed while for $j \in J_a$ (the auxiliary sample) only Z_j and x_j are observed. (It is assumed that $EU_j = EV_j = 0$ and that the column space of (x_{ji}) does not include the constant vector $(1,\ldots,1)'$.) Define

$$\sigma^2 = \operatorname{var} U_j, \qquad \tau^2 = \operatorname{var} V_j, \qquad \rho = \operatorname{corr}(U_j,V_j);$$

then

$$\mu_j = EY_j = \alpha_Y + x_j\beta_Y, \qquad \operatorname{var} Y_j = \sigma^2.$$

Let m and n, respectively, be the number of elements of J_p and J_a and let $N = m+n$. Denote by $\Sigma_{(p)}$ summation over the primary sample and by $\Sigma_{(t)}$ summation over the total sample, i.e. for $j \in J_p \cup J_a$. Define

$$x_\cdot = N^{-1}\Sigma_{(t)}x_j, \qquad x_\cdot^{(p)} = m^{-1}\Sigma_{(p)}x_j,$$

and $Y_\cdot^{(p)}$, $Z_\cdot$ and $Z_\cdot^{(p)}$ similarly; define

$$S_{xx} = N^{-1}\Sigma_{(t)}(x_j-x_\cdot)'(x_j-x_\cdot)$$

$$S_{xx}^{(p)} = m^{-1}\Sigma_{(p)}(x_j-x_\cdot^{(p)})'(x_j-x_\cdot^{(p)})$$

$$S_{xY}^{(p)} = m^{-1}\Sigma_{(p)}(Y_j-Y_\cdot^{(p)})(x_j-x_\cdot^{(p)})$$

and S_{xZ}, $S_{xZ}^{(p)}$, $S_{YY}^{(p)}$, $S_{YZ}^{(p)}$, S_{ZZ} and $S_{ZZ}^{(p)}$ similarly. The mean sums of squares and cross-products after regression on x are given by

$$\tilde{S}_{YY}^{(p)} = S_{YY}^{(p)} - S_{xY}^{(p)}(S_{xx}^{(p)})^{-1}(S_{xY}^{(p)})'$$

$$\tilde{S}_{YZ}^{(p)} = S_{YZ}^{(p)} - S_{xZ}^{(p)}(S_{xx}^{(p)})^{-1}(S_{xZ}^{(p)})'$$

and similarly for $\tilde{S}_{ZZ}^{(p)}$ and $\tilde{S}_{ZZ}$. Some computations show that ML estimators under the assumption of bivariate normality for (U_j,V_j) are given by

$$\hat{\beta}_Z = S_{xx}^{-1}S_{xZ}'$$

$$\hat{\beta}_Y = (S_{xx}^{(p)})^{-1}(S_{xY}^{(p)})' - \hat{\beta}((S_{xx}^{(p)})^{-1}(S_{xZ}^{(p)})'-\hat{\beta}_Z)$$

$$\hat{\alpha}_Z = Z_\cdot - x_\cdot \hat{\beta}_Z$$

$$\hat{\alpha}_Y = Y_\cdot^{(p)} - x_\cdot^{(p)}\hat{\beta}_Y - \hat{\beta}(Z_\cdot^{(p)} - x_\cdot^{(p)}\hat{\beta}_Z - \hat{\alpha}_Z)$$

$$\hat{\sigma}^2 = \tilde{S}_{YY}^{(p)} - \hat{\beta}^2(S_{ZZ}^{(p)} - \tilde{S}_{ZZ})$$

$$\hat{\rho} = \hat{\beta}\hat{\tau}/\hat{\sigma} \qquad \hat{\tau}^2 = \tilde{S}_{ZZ}$$

where

$$\hat{\beta} = \tilde{S}_{YZ}^{(p)}/\tilde{S}_{ZZ}^{(p)}.$$

The method was applied for the years 1961-1984 as the primary sample and 1950-1960 as the auxiliary sample. The estimated means for Y_j are given by

$$\hat{\mu}_j = 534.5 + 4.47(j-1969) + 2.45 \max(j-1969,0)$$

while $\hat{\sigma} = 52.3$, $\hat{\rho} = .38$ and $\hat{\beta} = 297$. (Units are kg/ha.) These numerical values are given for illustration purposes; they must be considered as referring to the probability distribution of the FAO estimates of Burkina Faso sorghum yield, not of the true sorghum yield.

COMPUTATION OF STANDARD ERRORS BY THE BOOTSTRAP

It is irrelevant to compute standard errors of these estimates under the assumption of normality, because it is known that distributions of crop yield are often quite skewed. The bootstrap method (Efron 1979, Efron and Tibshirani 1986) provides a distribution-free method for estimating standard errors. The principle of the bootstrap method is that, for random samples, the underlying distribution is estimated by the empirical distribution (the distribution of the sample) and accordingly, parameters of the underlying distribution such as standard errors of estimators can be estimated by resampling the observed sample, i.e., by computer simulation of random samples (with replacement) from the actually observed sample. In a regression situation, the bootstrap principle can be worked out by resampling observed residuals (Freedman, 1981).

In regression model (*), $\mu_j = EY_j$ and $\nu_j = EZ_j$ are estimated by

$$\hat{\mu}_j = \hat{\alpha}_Y + x_j\hat{\beta}_Y \qquad \hat{\nu}_j = \hat{\alpha}_Z + x_j\hat{\beta}_Z .$$

Define observed residuals by

$$\hat{U}_j = Y_j - Y_\cdot^{(p)} - (x_j - x_\cdot^{(p)})\hat{\beta}_Y \qquad j \in J_p$$

$$\hat{V}_j^{(p)} = Z_j - Z_\cdot^{(p)} - (x_j - x_\cdot^{(p)})\hat{\beta}_Z \qquad j \in J_p$$

$$\hat{V}_j = Z_j - \hat{\nu}_j \qquad j \in J_a.$$

These sets of residuals are centered in the sense that means are 0. See Freedman (1981) for the need to center residuals in bootstrapping regression models. The bootstrap now operates by independently generating a large number M of hypothetical data sets according to (*) with the original values for x_j, $j \in J_p$ and $j \in J_a$, using the estimated parameter values for α_Y, β_Y, and α_Z and β_Z, obtaining residuals for the primary sample by drawing random samples with replacement from

$$\{(\hat{U}_j, \hat{V}_j^{(p)}); j \in J_p\},$$

and obtaining residuals for the auxiliary sample by drawing from

$$\{\hat{V}_j; j \in J_p \cup J_a\}.$$

For every generated hypothetical data set the parameters are estimated, and the standard errors for the original parameter estimates are estimated by the standard deviations of these M hypothetical replicates. These standard errors are based on the assumption that the linear structure (*) is valid, but do not make use of any assumptions about the shape of the distribution of (U_j,V_j). This leads, using M = 1000, to the following estimated standard errors (note that $\hat{\mu}_{1986}$ = 652 kg/ha).

$$\text{s.e. } (\hat{\mu}_{1986}) = 28.9$$

$$\text{s.e. } (\hat{\sigma}) = 7.2 .$$

Similar computations were carried out for the usual estimators derived without data augmentation from only the primary sample. Denoting these estimators by $\tilde{\mu}_j$ and $\tilde{\sigma}$, this gives

$$\tilde{\mu}_{1986} = 654.6 \qquad \text{s.e. } (\tilde{\mu}_{1986}) = 30.2$$

$$\tilde{\sigma} = 52.7 \qquad \text{s.e. } (\tilde{\sigma}) = 7.9 ,$$

demonstrating a moderate gain in precision due to the data augmentation.

Acknowledgement. I thank the Office de Recherche Scientifique et Technique Outre-Mer and the Météorologie Nationale du Bourkina Faso for making available the rainfall data; and Dr. T.A. Buishand for calling my attention to Moran's paper.

REFERENCES

Efron, B., 1979: Bootstrap methods: another look at the jackknife. Ann. Statist., 7, 1-26.

Efron, B. and Tibshirani, R., 1986: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1, 54-77.

FAO, 1966-1985: Production yearbooks, vols. 19 (1965) - 38 (1984). FAO, Rome.

Freedman, D.A., 1981: Bootstrapping regression models. Ann. Statist., 9, 1218-1228.

Lamb, P.J., 1985: Rainfall in Subsaharan West Africa during 1941-1983. Zeitschr. für Gletscherkunde u. Glazialgeologie, 21, 131-139.

Moran, M.A., 1974: On estimators obtained from a sample augmented by multiple regression. Water Resources Res., 10, 81-85.

Snijders, T.A.B., 1986: Inter-station correlations and non-stationarity of Burkina Faso rainfall. J. Climate Appl. Meteor. 25, 524-531.

Snijders, T.A.B. and Schweigman, C., 1985: The stochastic nature of yields. Ch. 4 in C. Schweigman (ed.), Operations research problems in agriculture in developing countries, Tanzania Publ. House, Dar-es-Salaam, and Khartoum Univ. Press, Khartoum, 361 pp.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

SAMPLING THEORY OF OPTIMALLY WEIGHTED DATA TO DETECT CLIMATIC SIGNALS

Thomas L. Bell

Laboratory for Atmospheres
Goddard Space Flight Center
Greenbelt, MD 20771

A. Abdullah

Applied Research Corporation
Landover, MD 20785

1. INTRODUCTION

Detection of climatic change predicted by climate models is not easy because the changes are often small compared with the day-to-day variability of atmospheric variables. It is only by averaging data over time and space that the natural variability ("climatic noise") characterized by Leith (1973) can be reduced enough for there to be a chance of seeing the predicted changes ("climatic signal"). Madden and Ramanathan (1980) give an excellent discussion of this. But how should the averaging be done?

The most frequently used averaging scheme attempts to estimate what a hemispherically or globally averaged quantity such as surface air temperature would be, if we could obtain temperatures from stations uniformly distributed over the globe. Such an observing system does not exist, of course, but it is useful to try to estimate hemispherically averaged temperature, since it is easy to interpret physically. However, when a climate model prediction is available as a guide to what to look for in the data, a way of treating the data that is more sensitive to the predicted climatic change than a hemispheric average can be developed, at least in principle (Hasselmann, 1979; Bell, 1982). This consists in weighting data more from those geographical areas and seasons where the climate model predicts relatively large changes, but weighting less those areas and seasons where the climatic noise tends to obscure the signal.

As an example, suppose $T_z(t)$ denotes the zonally, annually averaged temperature for year t, where z labels zones. The data set developed by Angell and Korshover (1983), for which we present some results here, includes 4 zones covering the northern hemisphere, 90°N-60°N, 60°N-30°N, 30°N-10°N, and 10°N-10°S. For this case, z=1,2,3,4. If a climate model predicts the annually averaged temperature to change by an amount ΔT_z during one year, a weighted average of the 4 zonal temperatures T_z

$$A = \sum_z w_z T_z \qquad (1)$$

that has the best chance of differing significantly from the climatic mean should use weights (Bell, 1982)

$$w_z = \sum_{z'} (C^{-1})_{zz'} \Delta T_{z'} \qquad (2)$$

where C^{-1} is the inverse of the covariance matrix $C_{zz'}$ of the temperatures T_z, which can be estimated from

$$C_{zz'} \approx \overline{(T_z - \bar{T}_z)(T_{z'} - \bar{T}_{z'})} \quad , \qquad (3)$$

where the overbar indicates an average over data prior to the predicted change.

If we denote by ΔA the change in A predicted by the climate model,

$$\Delta A = \sum_z w_z \Delta T_z \ , \qquad (4)$$

and by σ_A^2 the variance of A estimated from prior data, then the signal-to-noise ratio

$$S/N = (\Delta A/\sigma_A)^2 \qquad (5)$$

is maximized by the choice Eq. (2) for the weights (Bell, 1982).

2. SAMPLING PROBLEMS

By using prior data to determine optimal weights via Eqs. (2) and (3), there is a risk of using up all the "information" (degrees of freedom) in the data set and having little left with which to estimate the variance of A. It can in fact be shown (Bell, 1986) that if weights are chosen according to Eq. (2) for p variables (p=4 in the above example), and if N=n+1 years of prior data are available to estimate the covariance as in Eq. (3), then the level for detecting a significant change in A must be increased by a factor n/(n-p) relative to what it would be if we had unlimited prior data. As a consequence of this, although the estimated signal-to-noise ratio may be increased by optimal weighting of data, if there is too little prior data this increase may be offset by the increased uncertainty in the weighted average.

As an illustration of this effect, consider the following example: suppose that the atmospheric variables are represented by amplitudes of the "true" empirical orthogonal functions (EOF's, or principal components) for the atmospheric system, standardized so that they have variance 1, but suppose that we have only N=26 years of data with which to estimate the covariance matrix of the variables, so that even though the true covariance matrix is diagonal, our sample estimate is probably not. Suppose that a model prediction for climatic change $\Delta\mu_i$ is available, with values $\Delta\mu_1 = 2$, $\Delta\mu_2 = \Delta\mu_3 = ... = 0.75$; in other words, the climate model predicts a 2σ shift in the first EOF and a 0.75σ shift in all other EOF's. If we form an optimally weighted average as in Eq. (1) of the first p variables, using the 26 year estimate of the covariance matrix, we find that as we increase the number p of variables included, the probability of detecting the climate change in the data increases for a while, and then begins to decrease because the confidence limit for detecting a significant change in A has increased too much. The probability of detecting climatic change in this example is shown in Fig. 1 as the solid curve.

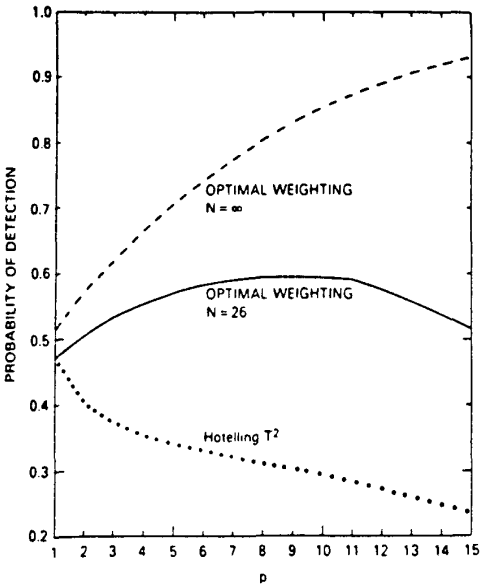


Fig. 1. Probability of detecting climatic change at the 2.5% significance level for p uncorrelated gaussian random variables with unit variance, and a climatic shift in the means $\Delta\mu = (2, 0.75, 0.75, ...)$. The dashed curve shows the detection probability for the optimal weighting method given an infinite amount of prior data. The solid curve gives the detection probability for the optimal weighting method with N=26 samples of prior data. The values are from Monte Carlo calculations accurate to ±0.007 (one sigma). The dotted curve shows the detection probability that Hotelling's T^2 test would yield.

Keeping more than 9 variables is therefore counterproductive. The upper curve shows how the method would fare if we knew the true covariance matrix exactly. The lower curve shows for comparison how the standard multivariate T^2 test (see Anderson, 1958, for a description) performs for the same problem.

A more complete description of this example may be found in Bell (1986). It is shown there that the effective signal-to-noise ratio [Eq. (5)] is approximately given by

$$(S/N)_{eff} = (1 - p/n)\,(S/N)_{\infty} \qquad (6)$$

where $(S/N)_{\infty}$ is the signal-to-noise ratio [Eq. (5)] we could obtain if we had unlimited prior data (i.e., if we knew the covariance matrix for the variables T_z exactly). If the values assumed in the example above are substituted into Eq. (6), the effective signal-to-noise ratio is found to have a maximum at p = 9.4, which agrees with the Monte Carlo results shown in Fig. 1.

Thus, in using the average (1) to detect climatic change, it is best to use relatively few variables, and to choose them so that they contain as much climate signal as possible. For a method of determining the best number of variables to keep, see Bell (1986).

3. EL CHICHON

In order to study the effect of having limited prior data to detect climatic change, data generously supplied by Angell and Korshover (1983) for zonally averaged surface air temperature for 4 zones covering the northern hemisphere were used in a test case. The eruption of El Chichón in the spring of 1982 was predicted to produce a cooling of the earth's surface, all other things being equal. We therefore set out to see how data in 1983 should be weighted in order to detect such a cooling.

The climate model prediction (Chou, 1984) was based on the nine-layer zonal-average energy-balance model of Peng et al. (1982) with the El Chichón aerosol radiative properties described by King et al. (1984). The delaying effects of a 68m deep mixed-layer ocean were included. A cooling of surface temperatures of the order of 0.3-0.5°C in the tropics and 0.7-1.2°C in polar latitudes was predicted for 1983 (December, 1982 to November, 1983). Polar latitude predictions tend to be more sensitive to uncertainties in the formulation of the model and to asumptions about the aerosol distribution. We have used intermediate values in the range of climate model predictions in our study.

Surface temperature data from 1958 to 1981 were used to develop a statistical model for the annually averaged data, designed to reproduce the covariance among zones and the 1-year-lag correlations of the data. Monte Carlo experiments based on the statistical model were then used to estimate the loss due to the small sample (24 years) in the power of the optimal weighting scheme to detect a cooling in the data if the cooling were to occur according to the climate model predictions.

It was found that even with the loss due to the small-sample effects, optimal weighting of the data gave the highest probability of detecting the cooling when all 4 zones were independently weighted. If the significance level for detecting cooling is set at 2.5% (1-sided test), which corresponds to about a "2-sigma" cooling, then the probability of actually seeing such a cooling is estimated to increase from 68% using ordinary hemispherically averaged temperatures, to 76% using optimal weighting of the 4 zones. Further improvement might be possible if the seasons are weighted differently.

All of this, of course, assumes that the climate model is "right", and that nothing else has occurred in 1983 to affect the climate that has not occurred often enough in the previous 24 years to be included in our estimate of climatic noise. In fact, analysis of the actual 1983 data shows over a 2-sigma warming during that year, relative to the previous 24-year average, presumably due to the "El Niño/Southern Oscillation event" that began at almost the same time as the volcanic eruption.

4. CONCLUSIONS

Finite sample effects can reduce the power of the optimal weighting of data to detect climatic change, and some thought should be given to choosing the variables that are used in forming the weighted average. In the El Chichón study, optimal weighting of the data increased the theoretical detectability of the climate signal in spite of the small sample used in determining the covariance matrix of the data.

5. REFERENCES

Anderson, T.W., 1958: An Introduction to Multivariate Statistical Analysis, John Wiley, New York, pp. 101-118.

Angell, J. K. and J. Korshover, 1983: Global temperature variations in the troposphere and stratosphere, 1958-1982. Mon. Wea. Rev., 111, 901-921.

Bell, T. L., 1982: Optimal weighting of data to detect climatic change: Application to the carbon dioxide problem. J. Geophys. Res., 87, 11,161-11,170.

Bell, T. L., 1986: Theory of optimal weighting of data to detect climatic change. J. Atmos. Sci. (in press).

Chou, M.-D., 1984: (private communication).

Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies, in Meteorology over the Tropical Oceans, edited by D. B. Shaw, Royal Meteorological Society, Bracknell, Berkshire, England, pp. 251-259.

King, M. D., Harshvardhan, and A. Arking, 1984: A model of the El Chichon stratospheric aerosol layer. J. Clim. Appl. Meteor., 23, 1121-1137.

Leith, C. E., 1973: The standard error of time-average estimates of climatic means, J. Appl. Meteorol., 12, 1066-1069.

Madden, R. A. and V. Ramanathan, 1980: Detecting climate change due to increasing carbon dioxide. Science, 209, 763-768.

Peng, L., M.-D. Chou, and A. Arking, 1982: Climate studies with a multi-layer energy balance model. Part 1: Model description and sensitivity to the solar constant. J. Atmos. Sci., 39, 2639-2656.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

QUALITY/VALUE RELATIONSHIPS FOR FORECASTS IN THE DYNAMIC COST-LOSS RATIO MODEL: INFINITE HORIZON, DISCOUNTED CASE*

Richard W. Katz

Environmental and Societal Impacts Group
National Center for Atmospheric Research†
Boulder, CO 80307 U.S.A.

Allan H. Murphy

Department of Atmospheric Sciences
Oregon State University
Corvallis, OR 97331 U.S.A.

INTRODUCTION AND SUMMARY

The purpose of this paper is to present some theoretical results concerning the nature of the relationship between the quality and economic value of imperfect forecasts of future weather or climate. The so-called dynamic cost-loss ratio decision-making model is considered. In this model, only two states of weather or climate are allowed (say, adverse or not adverse), with the weather or climate being viewed as a random variable O and p_0 denoting the climatological probability of the adverse weather or climate event. Moreover, only two actions are permitted (say, take protective action or do not take protective action). A cost C is incurred on each occasion that protective action is taken. If protective action is not taken and adverse weather occurs, then a loss L is incurred. To make the model dynamic, it is assumed that this loss L can be incurred, at most, once. Imperfect forecasts of future weather or climate are in the form of two possible conditional probabilities of adverse weather or climate, p_1 or p_0 say, with $p_0 \leq p_0 \leq p_1$.

The goal of the decision maker is to minimize the expected discounted expense totaled over an infinite horizon (i.e., infinitely many occasions), with future expenses being weighted by a discount factor α, $0 < \alpha < 1$. In previous work, an analytical quality/value curve has been derived for a static version of this optimization problem (i.e., only one occasion) (Katz and Murphy, 1986), and empirical quality/value curves have been obtained for a finite horizon, undiscounted case by means of stochastic dynamic programming (Murphy et al., 1985). In the present paper, an analytical quality/value curve is derived for this infinite horizon, discounted case. Like the static case, value is zero below a certain quality threshold. However, value rises nonlinearly above this quality threshold, unlike the linear increase in the static case. These results reinforce the need to explicitly consider the economic value of weather and climate forecasts, rather than relying on quality as a surrogate.

QUALITY AND VALUE OF INFORMATION

For convenience, we take the conditional probabilities, p_0 and p_1, to be related by

$$p_0 = (1 - p_1)p_0/(1 - p_0) \quad .$$

In this case, the single parameter p_1 represents "quality" of the weather or climate forecasts, with climatological and perfect information being special, limiting cases in which $p_1 = p_0$ and $p_1 = 1$, respectively. It is sometimes convenient to consider a linear transformation of p_1; namely,

$$q = (p_1 - p_0)/(1 - p_0) \quad .$$

†The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Note that $0 \leq q \leq 1$ with $q = 0$ for climatological information and $q = 1$ for perfect information. Hence, we regard q as a measure of the "relative quality" of the imperfect forecasts.

The economic value of imperfect weather or climate forecasts, and perfect information as well, is measured relative to the situation in which the decision maker has only climatological information available (e.g., Winkler and Murphy, 1985). Let E_C, E_F, and E_P denote the minimal expected discounted expenses totaled over an infinite horizon given climatological information, imperfect forecasts, and perfect information, respectively. Then the value of imperfect forecasts, say V_F, and of perfect information, say V_P, are given by

$$V_F = E_C - E_F \quad , \qquad V_P = E_C - E_P \quad .$$

For this infinite horizon, discounted version of the dynamic cost-loss ratio situation, we will derive an analytical expression for the value V_F as a function of the quality parameter p_1 (or q).

EXPENSE WITH CLIMATOLOGICAL INFORMATION

The expense E_C associated with climatological information satisfies

$$E_C = \min\{(1 - p_0)\alpha E_C + p_0 L,\ C + \alpha E_C\} \quad ,$$

where the first term (second term) within the brackets represents the total expected discounted expense when protective action is not taken (is taken) on the initial occasion. If the policy of always protecting were optimal, then

$$E_C(\text{protect}) = C/(1 - \alpha) \quad .$$

If the policy of never protecting were optimal, then

$$E_C(\text{not protect}) = p_0 L/[1 - \alpha(1 - p_0)] \quad .$$

Comparing these two expressions, protection is optimal only if both

$$0 < \alpha < 1 - C/L$$

and

$$p_0 > (C/L)/\{1 - [\alpha/(1 - \alpha)](C/L)\} \quad .$$

VALUE OF PERFECT INFORMATION

The expense E_P associated with perfect information satisfies

$$E_P = (1 - p_0)\alpha E_P + p_0 \min\{L,\ C + \alpha E_P\} \quad ,$$

where the first term (second term) within the brackets represents the total expected discounted expense when protective action is not taken (is taken), given that the adverse weather or climate event will occur on the initial occasion. If the policy of always protecting (whenever it is known that the adverse weather or climate event will occur) were optimal, then

$$E_P(\text{protect}) = p_0 C/(1 - \alpha) \quad .$$

If the policy of never protecting were optimal, then

$$E_P(\text{not protect}) = p_0 L/[1 - \alpha(1 - p_0)] \ .$$

Comparing these two expressions, protection is optimal (given the adverse weather or climate event) if

$$0 < \alpha < 1 - C/L$$

or if

$$1 - C/L < \alpha < 1$$

and

$$p_0 < [(1 - \alpha)/\alpha]\,[(1 - C/L)/(C/L)] \ .$$

Taking the difference in expected expenses between climatological and perfect information, the value of perfect information V_P is given by:

(i) $0 < \alpha < 1 - C/L$.

$$V_P = \begin{cases} p_0 \left[\dfrac{L}{1 - \alpha(1 - p_0)} - \dfrac{C}{1 - \alpha}\right] & \text{if } 0 < p_0 < p_0^* \ , \\ (1 - p_0)C/(1 - \alpha) & \text{if } p_0^* < p_0 < 1 \ , \end{cases}$$

where

$$p_0^* = (C/L)/\{1 - [\alpha/(1 - \alpha)](C/L)\} \ .$$

(ii) $1 - C/L < \alpha < 1$.

$$V_P = \begin{cases} p_0 \left[\dfrac{L}{1 - \alpha(1 - p_0)} - \dfrac{C}{1 - \alpha}\right] & \text{if } 0 < p_0 < p_0^{**} \ , \\ 0 & \text{if } p_0^{**} < p_0 < 1 \ , \end{cases}$$

where

$$p_0^{**} = [(1 - \alpha)/\alpha]\,[(1 - C/L)/(C/L)] \ .$$

VALUE OF IMPERFECT FORECASTS

The expense E_F associated with imperfect weather or climate forecasts satisfies

$$E_F = (1 - p_0)\min\{(1 - p_0)\alpha E_F + p_0 L \ , \ C + \alpha E_F\} + p_0 \min\{(1 - p_1)\alpha E_F + p_1 L \ , \ C + \alpha E_F\} \ ,$$

where the first term (second term) within the first set of brackets represents the total expected discounted expense when protective action is not taken (is taken) given the forecast probability of no adverse weather or climate is p_0, and the first term (second term) within the second set of brackets represents the total expected discounted expense

when protective action is not taken (is taken) given the forecast probability of adverse weather or climate is p_1. If the policy of always protecting were optimal, then

$$E_F(\text{protect}) = C/(1 - \alpha) \quad .$$

If the policy of never protecting were optimal, then

$$E_F(\text{not protect}) = p_0 L/[1 - \alpha(1 - p_0)] \quad .$$

Finally, if the policy of protecting only when given the forecast probability of adverse weather or climate is p_1 were optimal, then (making use of the assumed relationship between p_0 and p_1)

$$E_F(\text{protect given } p_1) = \frac{p_0[(1 - p_1)L + C]}{1 - \alpha[1 - p_0(1 - p_1)]} \quad .$$

Comparing these three expressions, the optimal policy is of the following form:

(i) $0 < \alpha < 1 - C/L$ and $p_0 > p_0^*$.

Always protect if

$$p_0 < p_1 < 1 - [(1 - p_0)/p_0]p_0^* \quad .$$

Protect only when given forecast probability p_1 if

$$1 - [(1 - p_0)/p_0]p_0^* < p_1 < 1 \quad .$$

(ii) $0 < \alpha < 1 - C/L$ and $p_0 < p_0^*$ or $1 - C/L < \alpha < 1$ and $p_0 < p_0^{**}$.

Never protect if

$$p_0 < p_1 < (C/L)\left[\frac{1 - \alpha(1 - p_0)}{1 - \alpha}\right] \quad .$$

Protect only when given forecast probability p_1 if

$$(C/L)\left[\frac{1 - \alpha(1 - p_0)}{1 - \alpha}\right] < p_1 < 1 \quad .$$

Taking the difference in expected expenses between climatological information and imperfect forecasts, the value of imperfect forecasts V_F is given by:

(i) $0 < \alpha < 1 - C/L$ and $p_0 > p_0^*$.

$$V_F = \begin{cases} 0 & \text{if } p_0 < p_1 < p_1^* \quad , \\ C/(1 - \alpha) - \dfrac{p_0[(1 - p_1)L + C]}{1 - \alpha[1 - p_0(1 - p_1)]} & \text{if } p_1^* < p_1 < 1 \quad , \end{cases}$$

where

$$p_1^* = 1 - [(1 - p_0)/p_0]p_0^* .$$

(ii) $0 < \alpha < 1 - C/L$ and $p_0 < p_0^*$ or $1 - C/L < \alpha < 1$ and $p_0 < p_0^{**}$.

$$V_F = \begin{cases} 0 & \text{if } p_0 < p_1 < p_1^{**} , \\ p_0 \left\{ \dfrac{L}{1 - \alpha(1 - p_0)} - \dfrac{(1 - p_1)L + C}{1 - \alpha[1 - p_0(1 - p_1)]} \right\} & \text{if } p_1^{**} < p_1 < 1 , \end{cases}$$

where

$$p_1^{**} = (C/L)[1 - \alpha(1 - p_0)]/(1 - \alpha) .$$

QUALITY/VALUE RELATIONSHIPS

For the infinite horizon, discounted version of the dynamic cost-loss ratio situation, examples of the nature of the quality/value relationships are shown in Figure 1. Like the quality/value curve for the static case, there is a threshold below which the value of imperfect weather or climate forecasts is zero. Above this quality threshold, economic value follows a continuous curve whose derivative increases as quality increases; that is,

$$\frac{\partial}{\partial q} V_F \text{ is a strictly increasing function of } q.$$

Hence, the sensitivity of economic value to changes in the quality of forecasts increases as perfect information is approached.

ACKNOWLEDGMENT

This research was supported in part by the U.S. National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8507495.

REFERENCES

Katz, R.W., and A.H. Murphy, 1986: Quality/value relationship for imperfect information in the umbrella problem. Submitted for publication.

Murphy, A.H., R.W. Katz, R.L. Winkler, and W.-R. Hsu, 1985: Repetitive decision making and the value of forecasts in the cost-loss ratio situation: A dynamic model. Monthly Weather Review, 113, 801-813.

Winkler, R.L., and A.H. Murphy, 1985: Decision analysis. Probability, Statistics, and Decision Making in the Atmospheric Sciences. A.H. Murphy and R.W. Katz, eds., Westview Press, Boulder, CO, pp. 493-524.

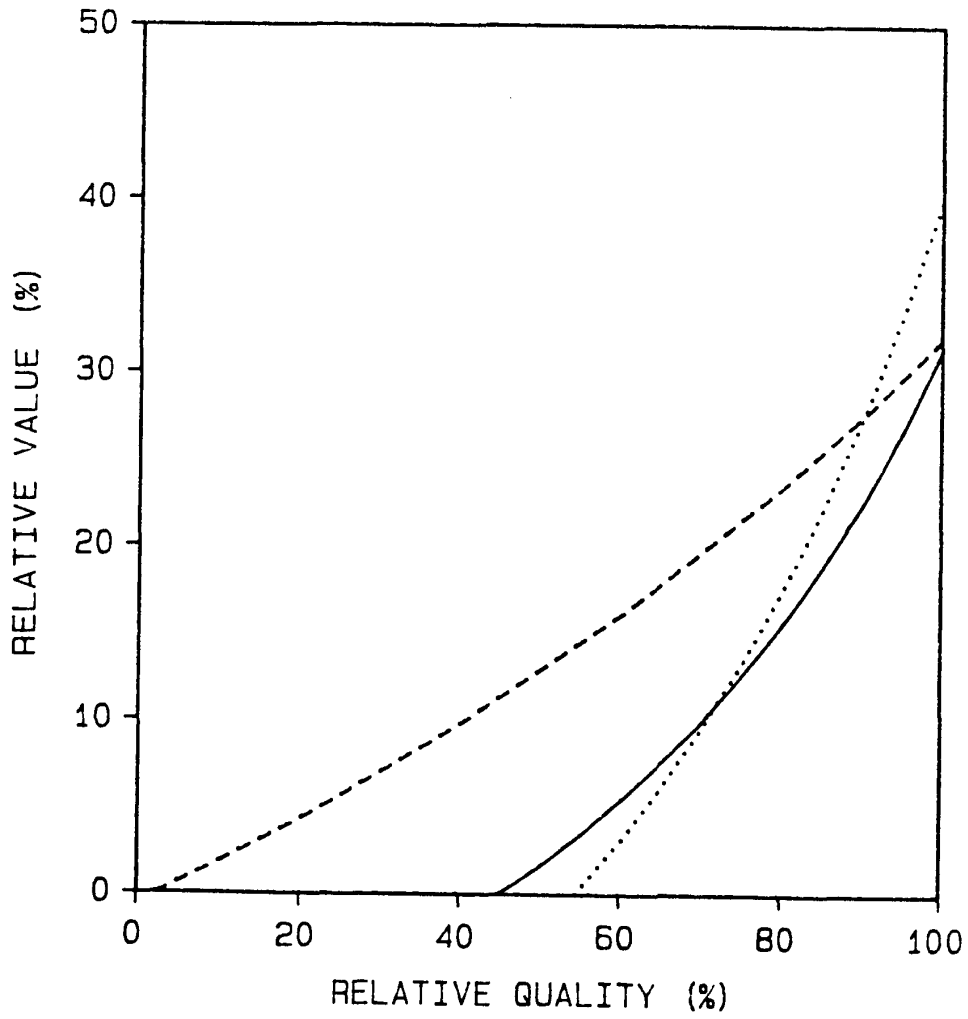


Figure 1 Relative economic value (V_F/L) as a function of relative quality (q) of imperfect weather or climate forecasts for discount factor $\alpha = 0.9$: (i) $C/L = 0.05$, $p_O = 0.05$ (dashed line); (ii) $C/L = 0.2$, $p_O = 0.2$ (solid line); and (iii) $C/L = 0.05$, $p_O = 0.2$ (dotted line).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE IMPACT OF CLIMATE FLUCTUATIONS ON AIRLINE SCHEDULES

Peter J. Robinson

University of North Carolina
Chapel Hill, NC 27514, USA

INTRODUCTION

The major impact that climate fluctuations have on aircraft operations is through the creation of variable flight times. A large fluctuation can lead to flight delays which inconvenience passengers and cause additional expense for the airline. The size of the fluctuation needed before delays occur depends on the flight schedule adopted by the airline. The schedule can incorporate extra time to provide a safety margin for the flights and for ground connections but operational factors dictate that these safety margins be kept to a minimum. Hence some climate fluctuations will lead to aircraft delays, but there has been no investigation of the magnitude needed to cause significant delay. This paper provides a preliminary estimate of the magnitude for two types of events: fluctuations in long-term average windspeed, where an airline has ample opportunity to adjust schedules; and thunderstorms at airports, short-term events which cannot be incorporated into schedules.

THE DELAY MODEL

An airline network consists of many routes and many interconnecting aircraft. A delay on any one flight can lead to delays in many subsequent flights. To estimate the climatic impacts on delay, a model was developed linking the amount of delay for a single incoming flight with the resultant departure delays for all connecting flights at that particular airport.

An arrival delay (DA) is defined for a particular airport, n, as the difference between the actual (A) and the scheduled (S) arrival time. It can be positive or negative. The difference can result from departure delays at the previous airport and from en route delays. Hence:

$$DA_n = DD_{n-1} + (A - S)_{n-1,n}, \qquad (1)$$

where single subscripts denote conditions at airports, double ones conditions during the flight. With the same notation, the departure delay DD is given by:

$$DD_n = DA_n + (A - S)_n. \qquad (2)$$

Delays can be positive or negative, but departures cannot occur before the advertised time. Hence Eq. (2) operates under the constraint:

$$\begin{array}{ll} DD_n = 0, & [DA_n + (A - S)_n] \leq 0 \\ DD_n = DA_n + (A - S)_n & [DA_n + (A - S)_n] > 0. \end{array} \qquad (3)$$

Assuming that connections must be made with a series of vehicles i (i = 1,2....,m) at airport n, then, introducing a superscript to denote individual aircraft, for vehicle j:

$$DD_n^j = \max_{i=1}^{m} | DA_n^i + (A - S)_n^i |. \quad (4)$$

The arrival delay at the next airport is then obtained using Eq. (1). The resultant delay, the the sum of all departure delays, can be determined as:

$$RD_n = \sum_{i=1}^{m} DD_n^i. \quad (5)$$

Equations (1) through (5) constitute the general model, which includes both in-flight and airport related delays, but does not consider the causes of the delay. The impact of thunderstorms at airports can be directly assessed, however, by adopting a value for DD_{n-1} consistent with the duration of a typical thunderstorm at the departure airport.

The influence of windspeed fluctuations in the creation of in-flight delays can be incorporated using (Robinson, 1986):

$$(A - S)_{n-1,n} = 0.017 \, [I \, (W'-W)]_{n-1,n}, \quad (6)$$

where I is the interairport distance and W and W' are the long-term average and short-term route equivalent windspeeds respectively. The route equivalent windspeed (Sawyer, 1950) is the average throughout the route of the wind vector directly affecting aircraft airspeed. Eq. (6) was developed from a correlation between long-term average windspeeds (Boeing, 1975) and scheduled flight times for numerous routes (Official Airlines Guide, 1985). Substitution of Eq. (6) into Eq. (1) yields the model for the impact of windspeed fluctuations on airline operations.

The above model considers delays over a single segment of an airline network. In this form, it allows a direct estimate of the magnitude of the atmospheric impacts for that segment, using a case study approach. These results can then be used, with the model expanded to provide a sample for the whole network, to assess the total impacts in the usual situation where there is spatial and temporal autocorrelation in climatic conditions.

DATA

The schedule data needed for the model, S_n and $S_{n-1,n}$, are available in the Official Airlines Guide (1985). $S_{n-1,n}$ was obtained directly as the difference between the published departure and arrival times for a flight. S_n is the difference between the scheduled arrival and departure times for a particular aircraft at an airport. The minimum turnaround time, represented by A_n in a delayed situation, is approximately 5 minutes less than the "minimum connecting times" given in the Official Airlines Guide

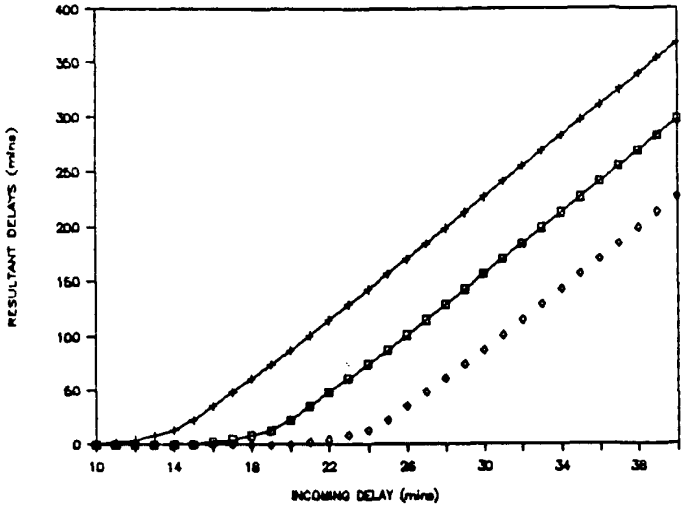


Figure 1. Resultant delays (minutes) at the Charlotte hub for morning flights, arising from incoming delays (minutes) in flights from Tampa - St. Petersburg (squares) and Greensboro (+) assuming a 15 minute aircraft turnaround time, and from Tampa - St. Petersburg assuming a 10 minute turnaround time (diamonds).

for a particular airport and airline (R. Powell, Vice-President, Piedmont Airlines, personal communication).

The data used for the case study were the Piedmont Airlines schedules for February 1985 at its Charlotte, NC. hub. This airline uses the increasingly popular "hub and spoke" strategy, bringing several flights into the hub at the same time, shuffling passengers, and then having flights leave together. Although operationally efficient, this strategy can lead to the development of downstream delays. Two of the four daily interchange periods were analyzed, the morning one, when most arriving aircraft had just completed their first flight, and an evening one, where most aircraft were about to start their last leg.

DELAYS AT CHARLOTTE, NC

The impact of a delay, of whatever cause, in a single incoming flight can be specified using data for the morning interchange period at Charlotte. Between 8:25 and 8:30 a.m. 14 flights are scheduled to arrive. The average time before these same aircraft depart is 31 minutes. In a situation with delays, the turnaround can be reduced to 15 minutes. Then an incoming delay in a flight arriving at 8:25 a.m., such as that from Tampa-St. Petersburg, FL, would have to be greater than 16 minutes before a downstream effect was felt (Fig. 1). Shorter flights, such as that from Greensboro, NC, are usually scheduled to arrive 5 minutes later, so that a delay of only 11 minutes would have a downstream effect. In addition, each minutes reduction in

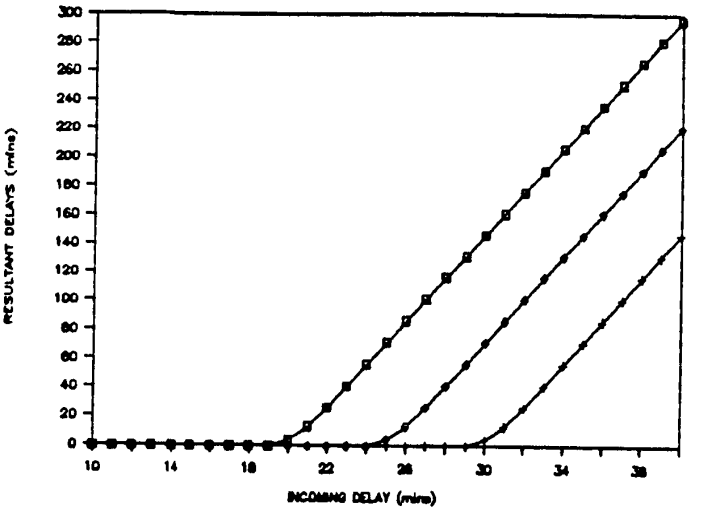


Figure 2. Resultant delays (minutes) at the Charlotte hub for evening flights, arising from incoming delays (minutes) in flights from Detroit (squares), Los Angeles (+) and San Francisco (diamonds), assuming a 15 minute aircraft turnaround time.

---

turnaround provides an extra minute before downstream delays start. Initially delays will affect only the early departures, but after only a few extra minutes the delay will affect all departures. Thereafter each additional minute of inbound delay leads to an additional 14 minutes in the total resultant outbound delay.

The major differences between the morning and evening interchanges are a longer interchange time, 41 minutes as opposed to 31 minutes, and longer incoming flights, which average 92 minutes as against 48 minutes in the morning. Two flights, 1 from Los Angeles, and 1 from San Francisco, have flight times of over 250 minutes. The former is the first flight to arrive, while the latter is in the middle of the pack. The latest arrival is the relatively short, 80 minute, flight from Detroit. Although the pattern of resultant delays (Fig. 2) is similar to that for the morning, the longer interconnection time allows a greater inbound delay before the effects are felt.

The potential impacts of thunderstorms at airports can be estimated using the Detroit data. The average duration of a thunderstorm at Detroit in late winter is around 45 minutes. This will create a departure delay from Detroit of about the same length. If the flight itself is unaffected, about 5 minutes of lost time can be made up because of the built in safety margin (Karam, 1978), and an arrival delay at Charlotte of 40 minutes is likely. This implies a resultant downstream delay from Charlotte of approximately 300 minutes, which is an average delay of 20 minutes for each outbound flight.

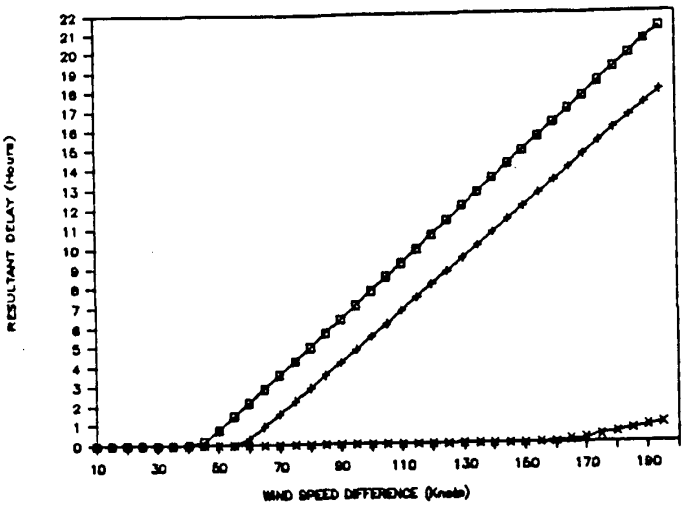


Figure 3. Resultant delays (minutes) at the Charlotte hub for evening flights, arising from route windspeed changes (knots) for incoming flights from San Francisco (squares), Los Angeles (+) and Detroit (x) assuming a 15 minute aircraft turnaround time.

Delays created specifically by windspeed fluctuations depend on the flight length (Eq. 5). For short incoming flights, including most of the morning ones, and the evening flight from Detroit, extremely large changes in windspeed would be necessary to create an inbound delay of even a few minutes (Fig. 3). However, the two long evening flights are more susceptible to windspeed effects. Both require windspeed decreases since they are affected by tailwinds in the region of the Westerlies. A windspeed change of 45 knots will lead to downstream delays for the flight from San Francisco. Because the Los Angeles flight is scheduled to arrive somewhat earlier, a windspeed change of 60 knots is needed. Although such anomalies are possible, they are unlikely to occur frequently, and development of a schedule to incorporate them would not be economically advantageous. However, if the airline had scheduled the same interchange time in the evening as in the morning, the windspeed anomaly needed to create delays would have been 40 and 30 knots for the Los Angeles and San Francisco flights respectively. This is more within the realm of day to day windspeed fluctuations.

CONCLUSIONS AND IMPLICATIONS

The results of the analysis of delays at the Charlotte hub suggest that the impact of windspeed fluctuations is sufficiently small that it can be incorporated into the schedule through the length of the interchange time allowed. The relatively long evening interchange period allows for the effects of windspeed fluc-

tuations on the long incoming flights. The morning period, with mainly short incoming flights, has a shorter interchange time. This simple response relationship suggests that these climatic effects are already incorporated into airline schedules. This conclusion, however, must be tested at different hubs, for different schedule periods and for different airlines.

The thunderstorm results indicate that major delays are most likely to be caused by individual weather events. The results suggest the magnitude of the event needed to create a delay, and allow specification of a climatology of events likely to create delays. The single hub results, however, provide only a partial analysis of the climatic impact. Rarely will an event be completely isolated, and delays at several airports, and on associated flight legs, are likely to be created simultaneously. This may lead to an amplification of the delays, in terms of both time and number of flights affected. Hence extension of the model is needed to provide this more realistic analysis.

REFERENCES

Boeing, 1975: "Winds on United States Domestic Routes", Document W3410, Boeing Commercial Airplane Company, Seattle, WA. U.S.A. 144pp.

Karam, G.L., 1978: "Night coach fares and airport delays", Transpn. J., 17, 4, 33.

Official Airlines Guide, 1985: "Official Airline Guide, North American Edition", Official Airlines Guides, Oak Brook, IL, U.S.A. (published semimonthly).

Robinson, P.J., 1986: "The impact of windspeed on airline schedules", in preparation.

Sawyer, J.S., 1950: "Equivalent headwinds: Application of upper wind statistics to air route planning", Meteorol. Rept. 2, 6, M.O. 535a, Her Majesty's Stationery Office, London, U.K.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

SOME STATISTICAL METHODS FOR THE APPLICATION OF CLIMATOLOGY

Jiuen Shi

Beijing Institute of Meteorology
Beijing, China

1. INTRODUCTION

Methodology drawn from the fields probability and statistics plays an important role in climatology, both in basic and applied research and in experimental and operational studies.

In climatology the uses of statistics may fall into four categories:

(1) To analyze the climatological records and reveal the behavior of the atmosphere pertaining to a specific geographical area. Essentially, this is a study of the atmospheric boundary profile and hence, edaphic data should be included in addition to geographic variates.

(2) To test the significance and validity of the climatic model through field and/or laboratory experiments.

(3) To make climatic/synoptic forecasts by designing forecasting procedures and methods as well as to evaluate the efficiency of forecasters.

(4) To assist engineering planning and design for nearly all human activities from the standpoint of climatological statistics with special emphasis on climatic control.

2. DISTRIBUTION OF CLIMATOLOGICAL VARIABLES

2.1 Type of frequency curves

Although many shapes of curves may be derived from many types of frequency distributions, three major shapes are generally noted: the normal, skewed, and bimodal. These shapes and their derivatives are shown in Fig. 1. A normal distribution curve takes the shape of a symmetrical (bell-shaped), smooth and monomodal form. It represents a continuous function when the expansion of the binomial $(p+q)^n$ approaches infinity, in which p is the probability of success in a single event and q, the probability of failure. The binomial distribution is one which is used to determine probabilities when events can be classified into only two categories, such as success-failure, cloudy-clear, or rain-no rain and when the trial are independent and the probability remains constant from trial to trial.

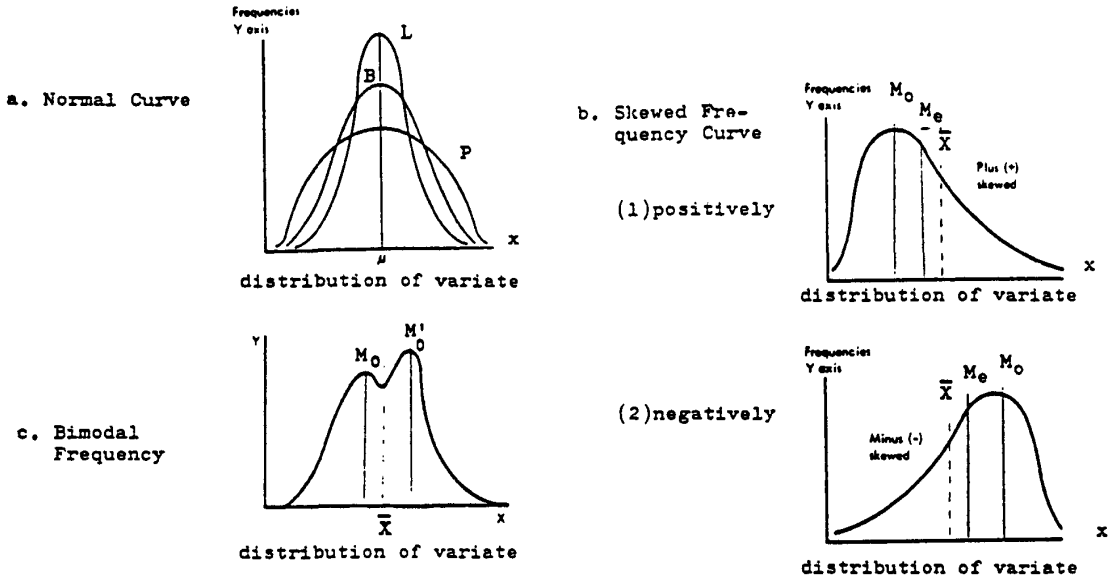


FIG. 1A (Wang) Three major types of frequency distribution

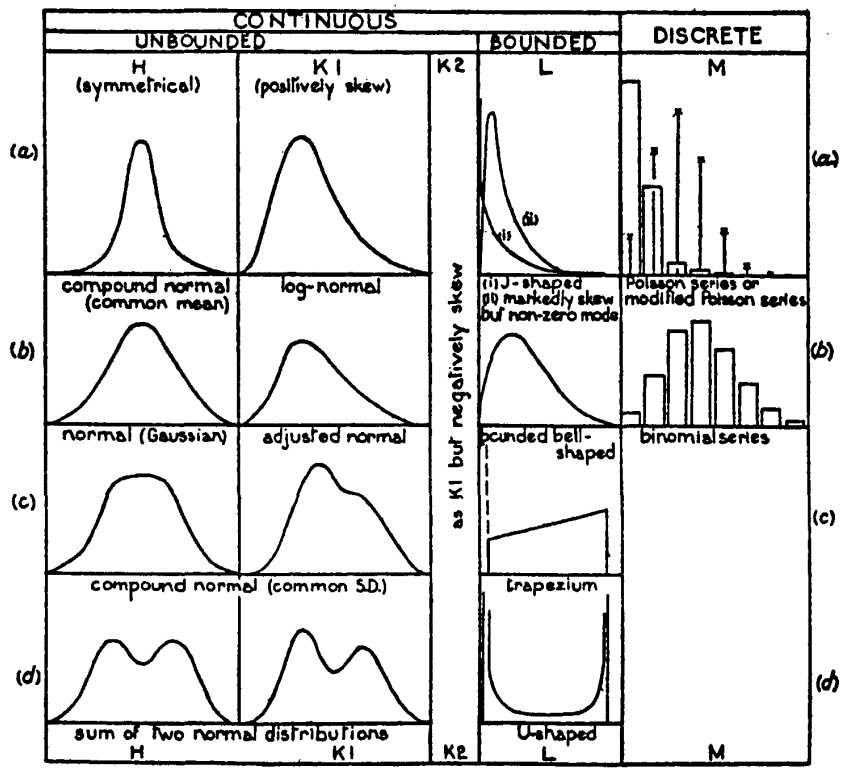


FIG. 1B (Brooks et al.) Classification of frequency distributions of common occurrence in meteorology

(a) Leptokurtic, (b) Nearly isokurtic, (c) and (d) Platykurtic, (d) Bimodal.

In climatological statistics, true normal distributions and multimodal distributions are rather rare. However, skewed, monomodal and bimodal distributions occur frequently. Although many such curvilinear frequency curves as J-curves, U-curves, L-curves, and growth and decay curves have been designated by some climatologists, the actual climatological statistics can only be approximated by these curves.

2.2 Angular data and frequency distributions

Angular observations arise from random experiments in various different ways. They may be direct measurements such as wind directions or cyclone movement. This kind of data is called as directional data. The directional data are considered as the points on two-dimensional cycle or higher dimensional sphere. Because of its own characteristical feature, the statistics of directional data much differs from common statistics in expression form with special treatment method. Most of the basic models on the circle have been derived either from transformations of the standard univariate (or bivariate) models or as circular analogues of important univariate characterizations. A distribution was introduced by von Mises to study the deviations of measured atomic weights from integral values. It plays a prominent role in statistical inference on the circle and its importance there is almost the same as that of the normal distribution on the line.

A circular random variable θ is said to have a von Mises distribution if its p.d.f. is given by

$$g(\theta;\mu_0,k)=\frac{1}{2\pi I_0(k)}e^{k\cos(\theta-\mu_0)},\quad 0<\theta\leq 2\pi,\ k>0,\ 0\leq\mu_0<2\pi, \tag{2.1}$$

where $I_0(k)$ is the modified Bessel function of the first kind and order zero, i.e.

$$I_0(k)=\sum_{r=0}^{\infty}\frac{1}{r!^2}\left(\frac{1}{2}k\right)^{2r}. \tag{2.2}$$

The parameter μ_0 is the mean direction while the parameter k is described as the concentration parameter. If a random variable θ has its p.d.f. of the form (2.1), we will say that it is distributed as $M(\mu_0,k)$. Fig.2 gives a polar representation of the density for $\mu_0=0$ and k=1.

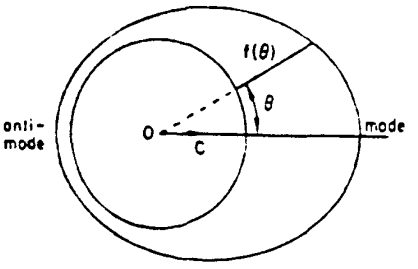


FIG. 2 Polar diagram of the density of the von Mises distribution for $\mu_0 = 0°$ and $\kappa = 1$.

3. MULTIVARIATE ANALYSIS

Multivariate analysis in climatology is an essential tool due to the fact that many applications can by make by it utilization. The characteristic feature of multivariate analysis is the consideration of a set of n objects, on each of which are observed the values of p variables. The set of objects may be complete or it may be a sample from a larger set. The variables may be continuous or discontinuous, and themselves may be a subset of a large group. Fig.3 may help to exhibit the coverage of the subject and the relationships among its component parts.

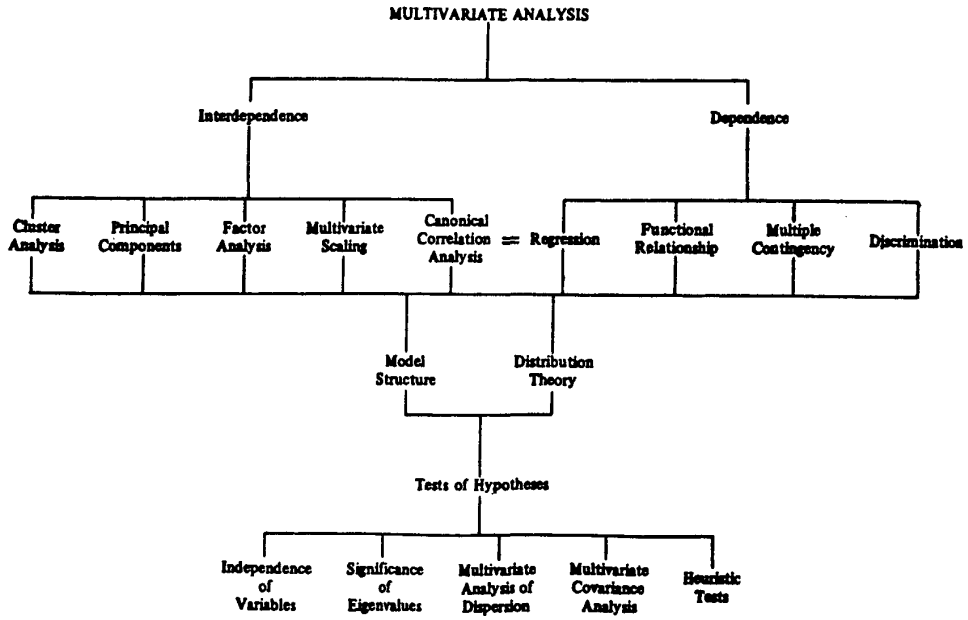


FIG.3(Kendall) The schematic of multivariate analysis

Some multivariate techniques in common use as follow:

(1) Principal component analysis

Given a vector of r random variables, $\underline{z} = (z_1 \cdots z_r)'$, we look for a linear combination of elements of $\underline{z}$, $\underline{\alpha}_1'\underline{z} = \sum_{i=1}^{r} \alpha_{1i} z_i$, which has maximum variance, then a second linear combination $\underline{\alpha}_2'\underline{z}$ which has maximum variance subject to being uncorrelated with $\underline{\alpha}_1'\underline{z}$, then $\underline{\alpha}_3'\underline{z}$ which has maximum variance subject to being uncorrelated with $\underline{\alpha}_1'\underline{z}$ and $\underline{\alpha}_2'\underline{z}$, and so forth. The idea is therefore to represent most of the r-dimensional variation of $\underline{z}$ in as few dimensions as possible. The new dimensions are defined by the new

variables $\underline{\alpha}_1'\underline{z}$, $\underline{\alpha}_2'\underline{z}$, $\underline{\alpha}_3'\underline{z}$,... which are the principal components. These principal components successively account for as much of the remaining variation as possible, and often the vast majority of the variation can be expressed in far fewer than r dimensions.

(2) Factor analysis

Factor analysis and principal component analysis are often confused because they frequently give similar results. In fact the two techniques take very different standpoint. Suppose we have a set of r variables $\underline{z}$; then factor analysis postulates that, apart from error terms, the r variables can be expressed as linear combinations of ($m \ll r$) hypothetical factors $\underline{f}$, so we have the model

$$\underline{z} = \underline{f} + \underline{e} , \qquad (3.1)$$

where Λ is (p x q) matrix of coefficients, and $\underline{e}$ a vector of errors. Various assumptions can be made regarding Λ , $\underline{f}$ and $\underline{e}$, leading to different estimates of the factors.

Principal component analysis imposes no such structure on the variables and so it does not fit into the framework of factor analysis. However, estimation in factor analysis has two stages, initial estimation followed by possible rotation, and the initial estimation is often based on principal components. Furthermore, the form of the estimated factors before rotation is often very similar to some of the first few principal components regardless of the method of estimation.

Factor analysis is rather appealing; it is nice to think that a set of climatic variables could be expressed in terms of a much smaller set of underlying factors. It remains to be seen whether this can be done, but even if it can, the factors which are found may not be very useful unless they can easily be interpreted. The possible uses of the factors once they have been found are similar to those for the principal components.

(3) Canonical analysis

We suppose that we have p climatic variables, $\underline{y} = (y_1 \ldots y_p)'$ whose values we wish to predict at some point in the future, and that predictions are to be based on the known value of q variables, $\underline{x} = (x_1 \ldots x_q)'$. $\underline{x}$ and $\underline{y}$ may, but need not, be the same set of variables measured at different point of time.

If separate principal component analyses are done on $\underline{x}$ and $\underline{y}$ there is no guarantee of strong relationships between the two sets of components. The idea behind canonical analysis is to look for linear combinations $\underline{\alpha}_1'\underline{x}$, $\underline{\beta}_1'\underline{y}$ of the elements of $\underline{x}$ and $\underline{y}$ which are as highly correlated as possible, then $\underline{\alpha}_2'\underline{x}$, $\underline{\beta}_2'\underline{y}$ which have maximum

correlation subject to being uncorrelated with $\underline{\alpha}_1'\underline{x}$, $\underline{\beta}_1'\underline{y}$, and so forth. These pairs of new variates are the canonical variates and the correlations between them are canonical correlations. Thus, $\underline{\beta}_1'\underline{y}$ is the linear function of $\underline{y}$ which is best predicted by a linear function of $\underline{x}$, and the appropriate linear function is $\underline{\alpha}_1'\underline{x}$, $\underline{\alpha}_2'\underline{x}$, $\underline{\beta}_2'\underline{y}$ have similar interpretations subject to being uncorrelated with $\underline{\alpha}_1'\underline{x}$, $\underline{\beta}_1'\underline{y}$, and so on.

It is clear that canonical analysis has appeal as a method for climatic analysis/forecasting, but will much again depend on the interpretability, and on the stability of the canonical variates.

(4) Multivariate regression

Multiple regression analysis is the statistical technique used most widely in data analysis and the development of empirical models. But in its usual form it is not strictly a multivariate technique since the independent variables are treated as having fixed values. An extension, multivariate regression, considers more than one dependent variable at a time, i.e. to fit linear relations to predict several dependent variables $\underline{y}$ from a common set of $\underline{x}$.

This type of analysis will obviously be easier to interpret than canonical analysis, although the predictive power of the equation for an element of $\underline{y}$ may be less than that for a canonical variate.

(5) Discriminant analysis

In discriminant analysis it is assumed that observations have come from a number (say k) of populations, whose general characteristics are known, and that a rule is required for assigning observations to one of the populations in an optimum way. Suppose that a typical observation is a r-variate random vector, $\underline{z}$. There are various possible approaches, of which two are well known. These two methods both rely on the assumption of a common covariance matrix Σ for all of the populations, and both base their assignment rules on the values of linear combinations of the elements of $\underline{z}$. For $k = 2$ the approaches are equivalent, but for $k > 2$ they are different.

In the climatology, discriminent analysis has obvious uses. If months (or seasons, years, etc.) can be divided into k categories, then we can set up rules which use past data $\underline{x}$ to predict the category of future months.

(6) Cluster analysis

Cluster analysis is a global term comprising techniques

which divide a set of observations into groups when there is litter or no prior knowledge of the form or the number of groups. There are many different methods of cluster analysis, and if there is no clear group structure in the data different methods may give very different results. There are at least two possible uses of cluster analysis in weather forecasting and climatic dignosis; the first is to use it on $\underline{y}$ so as to define the groups on which discriminant is to be used.

A second application is to divide the observations into groups using $\underline{x}$, and then to use different equations for predicting $\underline{y}$ from $\underline{x}$, depending on the group membership of $\underline{x}$.

4. TIME SERIES ANALYSIS

Most climatological data which are observed or recorded at successive time intervals, equal or unequal, form what is commonly known as a time series. The time series analysis in climatology may be identified as cyclic or non-cyclic, linear or non-linear, and short term (diurnal or seasonal) or long term (i.e., secular variation). It is beyond doubt that there is a multiplicative relationship among them. To separate the overlapping effects it is essential to do time series analysis. Quantification of each of the above types of time series would be useful in climate prediction/control.

The time series models widely used are models in the frequency domain. A lot of standard computer programs (e.g. power spectral analysis, cross spectral analysis, zero phase shift band pass filters for digital computers, etc) were written and included into the software packages of each computer firm, which especially after the invention of the fast Fourier analysis made the way open for the use of the variance spectrum analysis in climatology. Most of the applications were made in the field of climatic change in order to detect rhythms in the climatological time series. The advent of the maximum entropy method for the calculation of variance spectra was again a step forward in the application of the analysis of variance spectra. This method helps efficiently to find unknown periodicities in data series and permits also to make considerations about the significance of the spectral peaks hinting at the rhythms looked for.

However, the frequency domain models do not always give the same clear insight into the underlying statistical processes as the time domain models. The theory of these models in quite advanced and it is used in the practical work in the form of the so-called Box-Jenkins analysis, which has found further development by various authors recently. The Box-Jenkins approach to modelling is not fully objective but, nevertheless, it yields very good informations on the statistical process behind the observations. The method shows even the way for making forecasts, which of course not always may meet the final purpose of the forecaster, so one must be careful especially in this part of the work.

The nonlinear method of time series analysis---threshold autoregressive (TAR) model in discrete time is developed and used to climate analysis and forecasting. The basic idea of the TAR model is to deal with nonlinear systems by using the piece-wise linearization. Owing to the threshold control, the recursion stability is ensured, and limit cycle concept is introduced into the nonlinear random systems. The TAR model can effectively describe nonlinear fluctuation, explain various kinds of the stable periodical cycles in nature, and reflect the catastrophe similar to jump resonance. The TAR procedure consists of four parts: model building, statistical identification, parameter estimation and forecasting.

Previous investigations of applications of Chebyshev polynomials in meteorology dealt with only the expansion in equidistant grids. By introducing a nonlinear transformation, Chebyshev polynomials were gemeralized into irregular grids. It is worthwhile to point out that Chebyshev polynomials are suitable for interpolation of climatological data and predicting horizontal distributions of climatic elements.

5. PROBABILISTIC MODELS

Meteorological quantities exhibit variations both in time and space. To represent such meteorological quantities in a realistic manner, it is necessary to consider probabilistic models that allow for both variability and dependence. Such models are useful for characterizing climatological processes in terms of a few meaningful parameters, and they are also necessary to make valid statistical inferences about climatological data.

More recently, a two-state first-order Markov chain to fit daily precipitation occurrences was proposed. Since Gabriel and Neumann's work, more complex models have been suggested for fitting meteorological observations. Higher than first-order Markov chain, for instance, have been employed to fit daily precipitation occurrences. The Markov chain model for daily precipitation occurrence also has been generalized to bivariate processes that represent both daily precipitation occurrences and amounts. Yao (1981)applied Markov chain theory by using pseudovariables in the study of historical climate changes. Futher, point processes have been employed to model precipitation simultaneously over both time and space.

The distribution of extremes (e.g., maximum or minimum) is sometime required in climatological applications of probabilistic models. The asymptotic theory for the distribution of extreme values of a stochastic process is well established. In particular, under very general conditions the maximum of a sequence of random variables, suitably normalized, has one of the three so-called "extreme value distributions" as its limiting distribution. As with the central limit theorem, this extreme value theorem holds for many dependent processes.

6. CONCLUDING REMARKS

No single paper could be expected to cover completely the many probabilistic and statistical topics and applications of interest and importance to climatology, and this paper in no exception. It has been the intention here to place primary emphasis on two types of approaches:

(1) Surveys of relatively new developments in probability and statistics of potential interest to climatology;

(2) Reviews of selected area within climatology, in which probability and statistics have played and continue to play impotant roles.

There are many kinds of variates. They may be classified as vector or scalar, continuous or discrete and independent or dependent. Sometimes they may be further differentiated according to their temporal and spatial scales, cyclic or non-cyclic, as well as measured or estimated. Their identifications are the first step in the treatment of climatic data and are very important in statistical applications to climatology.

Although the fundamental principles underlying the various methods are identical regardless of the field of application, different statistical techniques are called for by different disciplines due to their varying inherited backgrounds.

When statistical methods are applied to climatology one must be highly selective. Obviously some methods are not applicable at all and others must be modified. In some cases new statistical methods and procedures are developed for certain types of climatological studies.

More new statistical methods are constantly being developed since the introduction of electronic computers to the atmospheric sciences. One is therefore advised to choose, modify or develop the proper method for solving his particular climatological problem. However, he would be familiar with the mechanisms of climatology in general, and that of his special interest in particular.

7. REFERENCES

- Brooks, C.E.P. and Carruthers, N., 1953: Handbook of statistical methods in meteorology. Her majesty's stationery office, London.
- Box, C.E.P. and Jenkins, G.M., 1970: Time series analysis, forecasting and control. Holden-Day, San Francisco.
- Cehak, K., 1983: A Box-Jenkins analysis of series of winter-strength in the Danube region. 2nd international meeting on statistical climatology (IMSC), Lisboa.

- Hobbs, J.E.,1980: Applied climatology. Dawson Westview Press.
- Jimenez, J.I. and Ballester, G., 1985: Multivariate analysis of local climatological series. 9th conference on probability and statistics in atmospheric sciences, Virginia.
- Jolliffe, I.T., 1980: Multivariate analysis and weather forecasting. The collection of papers presented at the WMO symposium on probabilistic and statistical methods in weather forecasting, Nice.
- Mardia, K.V., 1972: Statistics of directional data. Academic Press.
- Murphy, A.H. and Katz, R.W. edited, 1985: Probability, statistics, and decision making in the atmospheric sciences. Westview Press.
- Oke, T.R., 1982: Note on approaches toward modelling of the climatic effects of urbanization. Report of the meeting of experts on urban and building climatology, WMO WCP-37.
- Robertson, G.W. edited, 1983: Guidelines on crop-weather models. WMO WCP-50.
- Shi, J.E. et al, 1986: An application of the threshold autoregression procedure to climate analysis and forecasting. Advances in atmospheric sciences, Vol.3,No.1,134-138.
- -- 1986: Using directional data method for analysis on typhoon movement and its probability distribution features. 3rd international conference on statistical climatology,Vienna.
- Tong, H. and Lim, K.S., 1980: Threshold autoregression, limit cycle and cyclical data. Journal of the royal statistical society, series B, Vol. 42, No. 3, 245-292.
- Wang J. Y., 1982: Statistics in applied climatology. Milieu Information Service, Inc.
- Yao, C.S., 1983: Application of a regular Markov chain to computations of runs of historical flood and drought years. 2nd IMSC, Lisboa.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

OPTIMAL DESIGN OF THE REMOTE SOUNDING EXPERIMENTS WITH FINITE OBSERVATIONAL "WINDOWS"

V. Fedorov
International Institute for Applied Systems Analysis,
Laxenburg, Austria

1. INTRODUCTION

A number of publications concerning optimal design of experiments when controls belong to some functional space were published in the late 1970's. Now it is evident that the basic ideas behind these theoretical approaches are the same as in traditional experimental design theory (e.g., Fedorov, Uspensky 1977; Mehra 1974; Kozlov 1981; Pazman 1986). The differences become tangible in the application of general theoretical results to specific experimental problems.

In this paper these differences will be traced for experiments with spatially-averaged observations.

The simplest statistical model describing at least a part of the above mentioned experiments is the following one:

$$y_i = \vartheta^\tau x_i + \varepsilon_i \ , \quad i = \overline{1,N} \tag{1}$$

where $\vartheta \in R^m$ is a vector of unknown parameters; ε_i are independent random values with zero means and finite variances σ_i^2 (a more detailed assumption will be formulated later); τ stands for transposing. Variables x_i are defined by the integral

$$x_i = \int_V f(v) h_i(v) dv \ , \tag{2}$$

$f(v)$ is a vector of given basic functions; $h(v)$ are some functions which can be chosen (controlled) by an experimenter, $0 \le h \le 1$. In some cases integral (2) must be a Lebesque one. Function $h(v)$ describes the physical nature of an experiment and most typical examples will be given below in Section 2. If the least squares estimators

$$\hat{\vartheta} = Arg \min_{\vartheta} \sum_{i=1}^{N} \sigma_i^{-2} \{y_i - \vartheta^\tau x_i\}^2$$

are used to analyze experiments described by (1), then the quality of these estimators is defined by their dispersion (variance-covariance) matrices $D = E\{(\hat{\vartheta} - \vartheta_t)(\hat{\vartheta} - \vartheta_t)^\tau\}$, where the subscript τ stands for true values. It is well known (see Fedorov 1972) that in regular cases

$$D^{-1} = \bar{M} = \sum_{i=1}^{N} \sigma_i^{-2} x(h_i) x^\tau(h_i) \tag{3}$$

where σ^2 could depend upon $h(t)$ also. Matrix M is usually called "information matrix."

The objective of optimal experimental design is the search for controls $h_i^*(v)$ providing better dispersion matrices or (more accurately) some functions of them (see Sections 3, 4).

2. EXPERIMENTAL PROBLEMS

Example 1. Spectrometric Experiments

In these experiments the measurement tools (e.g., a spectrometer, a radiometer, etc.) have a finite spectral resolution and can measure only a spectrum (of absorbtion or radiation) averaged over some frequency interval which is defined by the so-called "slit" function. Formally the model of a measurement process can be described in the simplest case (when Beer's law is still working and the spectrum is linearly dependent upon concentrations of components) by the following model:

- The total spectrum intensity linearly depends upon the spectrum intensities of any component

$$\eta(v,\vartheta) = \vartheta^T f(v)$$

where v is frequency; $\vartheta^T = (\vartheta_1, \ldots, \vartheta_m)$ is the vector of concentrations.

- The observed signal y_i is a linear functional of the spectrum under analysis [compare with (2)]

$$y_i = \int_V \eta(v,\vartheta) h_i(v)dv + \varepsilon_i \quad , \tag{4}$$

where $h_i(v)$ is a slit function (or resolution function) of the i-th observation; V is the frequency interval available for observations; ε_i is the error of observation.

Usually a slit function can be satisfactorily approximated by

$$h_i = \begin{cases} 1, & v_{ij} \le v \le v_{ij} + \Delta_j \\ 0, & v < v_{ij}, v \ge v_{ij} + \Delta_j \end{cases}, \quad j = \overline{1,k} \quad . \tag{5}$$

- Errors ε_i are assumed to be random, independent for different i and their variance σ_i^2 can depend upon $h_i(v)$ and upon the time t_i spent on the i-th observation. In practice one can face several possibilities. For instance,

$$(a)\ \sigma_i^2 = \sigma^2 \quad (b)\ \sigma_i^2 = \sigma^2 [\int_V h_i(v)dv]$$

$$(c)\ \sigma_i^2 = \sigma^2 t_i \quad (d)\ \sigma_i^2 = \sigma^2 t_i [\int_V h_i(v)dv]. \tag{6}$$

The latter two cases correspond usually to the situation when an experimenter can choose the duration of an observation. If instead of y_i (these values can be called "total radiation") one considers average values ("average radiation"):

$$\bar{y}_i = y_i \{\int_V h_i(v)dv\}^{-1}$$

or

$$\bar{y}_i = y_i \{t_i \int_V h_i(v)dv\}^{-1} \tag{7}$$

then their variances will be correspondingly equal:

$$(a)\ \sigma_i^2 = \sigma^2 \quad (b)\ \sigma_i^2 = \sigma^2 \{\int_V h_i(v)dv\}^{-1}$$

or

$$(c)\ \sigma_i^2 = \sigma^2 t_i^{-1} \quad (d)\ \sigma_i^2 = \sigma^2 t_i^{-1} \{\int_V h_i(v)dv\}^{-1} \tag{8}$$

and variables x_i will be defined by the following formulae:

$$x_i = \int_V f(v)h_i(v)dv \Big\{ \int_V h_i(v)dv \Big\}^{-1}$$

or

$$x_i = \int_V f(v)h_i(v)dv \Big\{ t_i \int_V h_i(v)dv \Big\}^{-1}$$

In what follows below, the notation y_i will be used everywhere and the structure of variances σ_i^2 and values x_i will define the case.

Example 2. Remote Sounding of Atmosphere

The satellite measurements of the outgoing radiance in the infrared spectrum band become routine for distant retrieval of different physical parameters of the atmosphere (for instance, the vertical profile of temperature, humidity, ozone concentration, etc.; see Condratjev and Timofeev 1970, Strand and Westwater 1968, Twomey 1966). In a simplified form, the measured radiation $u(v)$ depends upon the vertical temperature profile $T(z)$, where z is an altitude, and the transmittance function of the atmosphere $\rho(v,z)$:

$$u(v) = B\{v,T(0)\}\rho(v,0) + \int_0^Z B\{v,T(z)\}\frac{\delta\rho(v,z)}{\delta z}dz$$

where Z corresponds to the altitude of a measurement tool; $B(v,T)$ is the Planck's function. Both $\rho(v,z)$ and $B(v,T)$ are assumed to be known.

The most crucial assumption is the possibility to approximate the function $T(z)$ by some parametric function $T(z,\vartheta)$ with subsequent linearization of the integral equation in the vicinity of an initial estimate of the temperature profile $\bar{T}(z)$ so that

$$u(v) \approx \vartheta^T f(v) = \vartheta^T\{\psi(v,0)\rho(v,0) + \int_0^Z \psi(v,z)\frac{\delta\rho(v,z)}{\delta z}\delta z\}$$

and integration with a weight function $h(v)$ over feasible frequency band V leads to the model described by (1), (2), (6) or (7) and (8).

3. OPTIMIZATION PROBLEMS IN EXPERIMENTAL DESIGN PROBLEM

Following Section 1, let us try to formalize a design problem for experiments from Section 2. As with traditional design theory (see above cited publications), the set of values

$$\xi_x = \{p_i, x_i\}_1^n \ , \quad \Sigma p_i = 1 \ , \qquad (9)$$

where the weights p_i are the shares n_i / N of total number of measurements (or the shares t_i / T of total time available), which have to be done under the conditions x_i (or at the supporting points x_i), will be called a design (of an experiment). In the traditional case, the set $X \in R^m$ of feasible controls (operability region) is explicitly given. In the considered case, X is the mapping of a feasible set H (in some functional space) of controls $h_i(v)$, and usually the construction of X (say, its boundaries) is a problem of great difficulty. Therefore, it could be useful to consider designs in the original space also:

$$\xi_h = \{p_i, h_i(v)\} \qquad (10)$$

From (3) it is evident that for model (1), (2) the information matrix (and subsequently dispersion matrix) depends upon the location of the supporting points x_i, and if at some points several measurements n_i are done, then on functions h_i also, but does not depend upon the results of measurements. In other words, the information matrix M depends upon a design ξ_x (or ξ_h).

Due to this fact, the design problem can be formulated as the following minimization problem

$$\xi^* = Arg \min_{\xi} \Phi\{\bar{M}(\xi)\} \tag{11}$$

where ξ can have both possible subscripts. The function Φ (optimality criteria) describes the objectives of an experimenter.

If the whole set of parameters present some interest, then it is reasonable to minimize the volume of the ellipsoid of concentration which is proportional to $|\bar{M}(\xi)|^{-1/2}$ and one can use $\Phi[\bar{M}] = |\bar{M}(\xi)|^{-1/2}$(everywhere $|A|$ means the determinant of matrix A).

If it is necessary to know a behaviour of some linear function of ϑ: $\psi^T(z)\vartheta$, then it is reasonable to minimize the average variance of this function over some region of interest. In this case

$$\Phi[M] = \int_Z \psi^T(z)\bar{M}^{-1}(\xi)\psi(z)dz$$

Rather detailed lists of the most popular optimality criteria can be found in Fedorov 1972; Silvey 1980.

In the future it is reasonable to distinguish between two types of designs: continuous and discrete ones.

In the first case, weights p_i can vary continuously between 0 and 1. This takes place when the weight is proportional to the time of measurement. We can go further and assume that any probabilistic measure $\xi_x = \xi(dx)$ or $\xi_H = \xi(dh)$ describes some design. In experimental practice, it could be impossible to realize, for instance, continuous measures. But fortunately for any design with continuous measure, it is possible to find the design with the same information matrix, but with measure concentrated in the final number of supporting points. It is easy to see that in the considered situation

$$\bar{M}(\xi_x) = N\int_X xx^T\xi(dx) = NM(\xi_x)$$

or

$$\bar{M}(\xi_H) = N\int_H x(h)x^T(h)\xi(dh) = NM(\xi_H)$$

In what follows below, the subscript x or h will be omitted without any comments if it will not lead to confusion.

Assuming that $\Phi(NM) = a(N)\Psi(M)$ (and it is true for the majority of optimality criteria used in practice) minimization problem (11) can be replaced by

$$\xi^* = Arg \min_{\xi} \Psi\{M(\xi)\} \tag{12}$$

where no values depend upon the total time or the total number of available measurements. This means that a continuous optimal design does not depend upon them also.

This useful property is not valid in the discrete case.

4. CONTINUOUS OPTIMAL DESIGNS

For the sake of simplicity in this section and all subsequent sections only the case when

$$\Psi\{M\} = |M^{-1}| \quad \text{and} \quad \xi^* = Arg \max_{\xi} |M|$$

will be considered. Other criteria can be handled in a similar way (see Fedorov 1980; Silvey 1980). It is most convenient to start with continuous version of designs ξ_x defined by (9). Then the celebrated Kiefer-Wolfovitz equivalence theorem (see Fedorov 1972) can be used and only one assumption is necessary for its fulfillment:

(a) Operability region X is compact.

Theorem 1

(1) There exists an optimal design ξ_x^* containing no more than $m(m+1)/2$ supporting points.

(2) The following problems are equivalent:

- maximization of $|M(\xi_x)|$,
- minimization of $\max_{x \in X} \lambda(x) d(x, \xi_x)$,
- $\max_{x \in X} \lambda(x) d(x, \xi_x) = m$,

  where $\lambda(x_i) = \sigma_i^{-2}$ and $d(x, \xi_x) = x^T M^{-1}(\xi_x) x$.

(3) At the supporting points of an optimal design ξ_x^* the function $d(x, \xi_x)$ approaches its maximum.

(4) The set of optimal designs is convex.

In a number of comparatively simple situations Theorem 1 gives a chance to construct optimal design analytically. For more complicated models it helps to develop numerical procedures and to understand some general features of optimal designs.

For instance, if one manages to prove that operability region X is compact, then he can be sure that (see point (3) of the theorem) all supporting points of an optimal design ξ_x^* are some boundary points of X.

Example 1

Let

$$f^T(v) = (1, v), \quad |v| \le 1, \quad \Psi(M) = |M|^{-1/2},$$

$$\gamma(x) = const, \quad x_1 = \int_{-1}^{1} h(v) dv, \quad x_2 = \int_{-1}^{1} v h(v) dv$$

The set X can be easily constructed because of the simplicity of integral (2) and its boundary is described by the following curves:

$$x_2 = \pm\{\frac{1}{2} - (x_1 - 1)^2/2\}, \quad 0 \le x_1 \le 2$$

From point 2 of Theorem 1 it follows that the supporting points of an optimal design must coincide with the points where the ellipse $x^T M^{-1}(\xi_x^*)x$ is tangent to X. This ellipse must have the least area (or $|M|$) between all ellipses containing X. The simple calculations show that the points with coordinates (2;0), $(\sqrt{5}-1,\ 2\sqrt{5}-4)$, $(\sqrt{5}-1,\ 4-2\sqrt{5})$ could be supporting ones.

From symmetry of X and point 4 of Theorem 1 it follows that the weights of the two last points must be equal. Straightforward maximization of the determinant $|M|$ gives

$$p_2 = p_3 = 0.5(\sqrt{5}-1)^{-1},\ p_1 = 1-2p_2$$

Finally, from the simple integral equations defined at the beginning of the example it is easy to find that

$$\xi_h^* = \begin{Bmatrix} h_1^*(v)\equiv 1 & h_2^*(v)=\begin{cases}1, & v\le\sqrt{5}-2\\ 0, & v>\sqrt{5}-2\end{cases} & h_3^*(v)=h_2(-v) \\ 0.19 & 0.405 & 0.405 \end{Bmatrix}, \quad |v|\le 1$$

with the information matrix

$$M = \begin{Vmatrix} 2 & 0 \\ 0 & 0.18 \end{Vmatrix}$$

For comparison of the traditional design with two Δ-windows at the points $v=\pm 1$, the same matrix equals

$$M = \begin{Vmatrix} \Delta & 0 \\ 0 & \Delta^2 \end{Vmatrix}.$$

This means the ratio of standard errors will be 0.5Δ (for $\hat{\vartheta}_1$) and $5.5\Delta^2$ (for $\hat{\vartheta}_2$). So the optimal design is essentially more effective than the traditional approach, especially for small Δ.

Example 2.

The characteristics of optimal designs (for instance, the location of supporting points) essentially depend upon the chosen basic function. To illuminate this, let us consider the regression problem formulated in the previous example with a new basic function $f^T(v) = (\sin\pi v,\ \cos\pi v)$. It can be proved that for any slit function $h(v)$ the vector (Krein, Nudelman 1973, VII:3)

$$x^T = \{\int_{-\pi}^{\pi} h(v)\sin\pi v dv\ ,\ \int_{-\pi}^{\pi} h(v)\cos\pi v dv\}$$

must belong to the circle $\{X: x_1^2 + x_2^2 \le 2\}$.

The optimal designs for this operability region and response function $v^T x = \vartheta_1 x_1 + v_2 x_2$ can be easily constructed. For instance, optimal design can consist of the supporting points coinciding with all vertexes of any regular polygon refined to the circle X and their weights must be equal. One of the simplest optimal designs is

$$\xi_x^* = \begin{Vmatrix} (0;2) & (2;0) \\ 0.5 & 0.5 \end{Vmatrix}$$

To find out the corresponding optimal design in slit function space the integral equations

$$x_{1i}^{*} = \int_{-\pi}^{\pi} h_i(v)\sin\pi v dv \quad \text{and} \quad x_{2i}^{*} = \int_{-\pi}^{\pi} h_i(v)\cos\pi v dv$$

must be solved. One of the solutions is

$$h_1^{*}(v) = \begin{cases} 1, & -0.5 \le v \le 0.5, \\ 0, & v < -0.5, \ v > 0.5, \end{cases}$$

and

$$h_2^{*} = \begin{cases} 0, & -0.5 \le v \le 0.5, \\ 1, & v < -0.5, \ v > 0.5. \end{cases}$$

It is worthwhile to note that the widths of slit function "windows" have the same order as intervals of typical variations of basic functions.

5. NUMERICAL METHODS

If assumption (a) holds and there is a way to find X then the following numerical can be used for optimal design construction (Fedorov 1972):

$$\xi_{s+1} = (1-\alpha_s)\xi_s + \alpha_s \xi(x_s) \tag{13}$$

where $\xi(x_s)$ is the design concentrated at the single point

$$x_s = Arg \max_{x \in X} \lambda(x) d(x, \xi_s) \tag{14}$$

where

$$d(x, \xi_s) = x^T M^{-1}(\xi_s) x$$

$$M(\xi_{s+1}) = (1-\alpha_s)^{-1} \left[1 - \frac{\lambda(x_s)\alpha_s M^{-1}(\xi_s) x_s x_s^T M^{-1}(\xi_s)}{1-\alpha_s + \lambda(x_s)\alpha_s d(x_s, \xi_s)} \right] M^{-1}(\xi_s)$$

The sequence $\{\alpha_s\}$ can be, for instance:

$$(a) \quad \sum_{s=0}^{s=\infty} \alpha_s = \infty , \quad \lim_{s \to \infty} \alpha_s = 0 \qquad (b) \quad \alpha_s = \frac{\lambda(x_s) d(x_s, \xi_s) - m}{\{\lambda(x_s) d(x_s, \xi_s) - 1\} m}$$

Both of them guarantee that

$$\lim_{s \to \infty} |M(\xi)| = \min_{\xi} |M(\xi)| \tag{15}$$

More sophisticated versions of this method are discussed in Ermakov 1983; Fedorov and Uspensky 1975.

In spite of the formal simplicity of iterative procedure (13), (14), its practical usefuliness is rather restricted: one must find out the way to construct X before using this procedure.

Procedure (13), (14) can be replaced by equivalent procedures in the space of functions $h(v)$. For that one has to replace the vector $\lambda^{1/2}(x)x$ with the vector $\int_V h(v) f(v) dv$ (for cases (a),(c) from (6) and (8)), and with the vector $\int_V h(v) f(v) dv \Big/ \sqrt{\int_V h(v) dv}$ (for cases (b),(d) from (6) and (8)).

For the sake of simplicity, the iterative procedure will be considered for the first cases when it can be presented in the following form:

$$\xi_{s+1} = (1-\alpha_s)\xi_s + \alpha_s \xi(h_s), \tag{16}$$

$$h_s = Arg \max_{h \in H} d(h, \xi_s), \tag{17}$$

where

$$d(h, \xi_s) = \int_V \int_V h(v)h(v')f^T(v)M^{-1}(\xi_s)f(v')dvdv'.$$

Unfortunately, maximization problem (17) is more complicated than (14). One of the simplest could be the following one:

- discretisize the set V (and therefore $V \times V'$ also), say with interval Δ;
- collect all points v_j on the corresponding grid which positively contribute to the sum

$$d(\xi_s) = \sum_{j,j'} f(v_j)M^{-1}(\xi_s)f(v_{j'}) \tag{18}$$

- put $h_s(v_j) = 1$ if v_j was chosen on the previous stage; otherwise $h_s(v_j) = 0$
- the fulfillment of the inequality (which is corollary to Theorem 1) $d(\xi_s h_s) \geq m$ tests that h_s can be used in (16).

This procedure is admissible for applications, for instance, when V is one-dimensional and that is the case for a number of applied problems (see Section 2).

Note 1.

Iterative procedure (16) will converge in sense (15) (for both original and discrete versions) if instead of (16) only the inequality

$$d(h_s, \xi_s) > m$$

will take place on every step.

Note 2.

The sequences $\{j\}$ and $\{j'\}$ must be identical. Therefore if $j = j_0$ is included in sum (18), then $j' = j_0$ must be included also.

Note 3.

When the slit function $h(v)$ can change its value at the modes of Δ-grid, then one can tell about Δ-optimal designs (ξ_Δ^*) which can be considered as some approximation of optimal designs defined by (12). The iterative procedure (16), (17) and (18) guarantees that

$$\lim_{s \to \infty} |M(\xi_s)| = M(\xi_\Delta^*).$$

The idea of Δ-optimal designs can be advanced further. As it was observed in Example 2, the widths of slit function windows were related to the intervals of variation of basic functions. Therefore, it is reasonable to decompose the set V into a comparatively moderate number of subsets Δ_j $(j = \overline{1,k})$, for instance, coinciding with the most typical fluctuations of basic functions. Assume that integrals

$$F_j = \int_{\Delta_j} f(v)dv \tag{19}$$

can be calculated (numerically or analytically). Then the operability region X can be approximated by X_Δ with elements

$$x = Fu ,$$

where $F=(F_1,\ldots,F_k)$, $u^\tau = (u_1,\ldots,u_k)\in U_\Delta$ and $u_j = 1$, if $h(v)=1$, $v\in\Delta_j$, and $u_j=0$; otherwise, $j=\overline{1,k}$. Observing that $\vartheta^\tau x=\vartheta^\tau Fu=\omega_\tau u$ and the information matrix equals

$$M = F\int_{U_\Delta} uu^\tau \xi(du)F^\tau = FM_\omega(\xi_u)F^\tau$$

where $\xi(du)$ describes a design ξ_u with supporting points in U_Δ, one can conclude that rank F must be equal to m (number of estimated parameters). Therefore, the decomposition of V should contain at least m subsets Δ_j ($k\geq m$).

When $k=m$ and of course $|F|\neq 0$, then $|M|=|F|^2|M_\omega|$ and the design problem is reduced to the maximization of $|M_\omega|$. The latter problem coincides with the routine problem of "optimal weighting." (See Ermakov 1983)

If $k>m$ then iterative procedure (13),(14) can be used with the replacement of the vector $\lambda^{1/2}(x)x$ with the vector Fu [or $Fu/\sum_{j=1}^{k} u_j$; compare with comments to (16),(17)]:

$$u_s = Arg\max_u uF^\tau M^{-1}(\xi_s)Fu. \qquad (20)$$

The maximization problem (20) is a discrete one and at every s-th stage it demands no more than 2^k calculations of $uF^\tau M^{-1}(\xi_s)Fu$.

6. STRUCTURE OF SLIT FUNCTION

In the previous sections, it was assumed that the slit function can equal 1 or 0. Some "physical" arguments were behind this assumption. The compactness of operability region X (see Section 4) was also an essential assumption which was done to slimplify all final results. If one refuses this assumption, then instead of optimal designs, so-called optimal sequences (see, for example, Ermakov 1983) must be considered and that leads to some technical difficulties. The following results (which are straightforward corollaries of well-known results from classical approximation theory; see, for example, Karlin and Studden 1966, Chapter VIII) illuminate that both above mentioned assumptions are not very restrictive. For the sake of simplicity, we consider a one-dimensional case ($V\in R^1$):

Assume now that:

(a) $0\leq h(v)\leq 1$, for any $v\in(a,b)$

(b) Functions $f(v)$ constitute a Tchebysheff system on the open interval (a,b), where a and b are possibly infinite. This assumption requires that the functions $f(v)$ be continuous on (a,b) and the determinants

$$\begin{vmatrix} f_1(t_1) & f_1(t_2) & \ldots & f_1(t_m) \\ f_2(t_1) & f_2(t_2) & \ldots & f_2(t_m) \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ f_m(t_1) & f_m(t_2) & \ldots & f_m(t_m) \end{vmatrix}$$

$$t_i\in(a,b),\quad i=\overline{1,m}$$

are positive.

Theorem 2.

The operability region

$$X = \{x = \int_a^b f(v)h(v)dv:\ 0 \le h(v) \le 1\}$$

is a compact convex set in R^m.

From Theorem 1 it is clear that all supporting points of any optimal design must be boundary points of X. Therefore, only these points had to be considered in the previous sections, and for them the following result takes place:

Theorem 3.

The necessary and sufficient condition for x to be a boundary point of X is the fulfillment of the condition

$$h(v)\{1-h(v)\} = 0 \tag{21}$$

almost everywhere in (a,b).

Let $h(v)$ be a function satisfying (21) and let $I(x)$ be the number of separate nondegenerate intervals (windows of a slit function) where $h(v)=1$ with the special convention that an interval whose closure contains point a or b, is counted as 1/2. For any point $x \in X$, $I^*(x)$ stands for the least possible $I(x)$ (it could be several different functions $h(v)$ giving the same x).

Theorem 4.

A necessary and sufficient condition that x belongs to the boundary of X is that $I^*(x) \le (m-1)/2$. Moreover, every boundary point corresponds to a unique $h(v)$ with $I(x)=I^*(x)$.

Theorems 3 and 4 allow for the development of a comparatively simple algorithm of optimal design construction.

Let $\bar{v}=(a,v_1,\ldots,v_{m-1},b)$, where $a \le v_1 \le \cdots \le v_{m-1} \le b$. According to Theorem 4, there exist optimal designs with all supporting points (in the operability region H) which have the following structures:

$$\bar{h}(v) = \{1, v \in (a,v_1);\ 0,\ v \in (v_1,v_2);\ 1,\ v \in (v_2,v_3);\ \cdots\}$$

and $\underline{h}(v) = 1-\bar{h}(v)$.

That allows for the modification of the iterative procedure (16),(17) to the procedure with maximization in space which dimension is less or equal $(m-1)$, where m is a number of basic functions:

$$\xi_{s+1} = (1-\alpha_s)\xi_s + \alpha_s \xi(h_s), \tag{22}$$

$$h_s = Arg \max_{\gamma,x} d(x_\gamma, \xi_s),\quad \gamma = 1,2, \tag{23}$$

where $a \le v_2 \le \cdots \le v_{m-1} \le b$,

$$x_1 = \int_a^b f(v)\bar{h}(v)dv$$

$$x_2 = \int_a^b f(v)\underline{h}(v)dv.$$

Procedure (22),(23) in a computational sense coincides with iterative procedures used for traditional design problems and can be handled with software developed for the latter one (see Fedorov, Uspensky 1975).

REFERENCES

Condratjev, K.J. and Timofeev (1970) The Thermal Remote Sounding of Atmosphere from Satellites. Leningrad, 410.

Ermakov, S.M., editor (1983) Mathematical Theory of Experimental Design. Moscow: Nauka, 391 (In Russian).

Fedorov, V.V. (1972) Optimal Design of Experiments. New York: Academic Press, 292.

Fedorov, V.V. and A.B. Uspensky (1975) Numerical Aspects of the Least Squares Methods. Moscow: Moscow State University, 168.

Fedorov, V.V. and A.B. Uspensky (1977) On the Optimal Condition Choice of Spectroscopic Measurements. Moscow: Proc. of State Scient. and Research Center for the Study of Earth's Environment and Natural Resources, 4:42-53 (In Russian).

Fedorov, V.V. (1980) Convex Design Theory. Math. Oper. Stat., 11:403-413.

Karlin, S. and W.J. Studden (1966) Tchebycheff Systems: With Applications in Analysis and Statistics. New York: Wiley & Sons, 586.

Kozlov, V.P. (1981) Design of Regression Experiments in Functional Spaces, in "Mathematical Methods in the Design of Experiments", ed. Penenko, V.V. Moscow: Nauka, Novosibirsk, 74-101 (In Russian).

Krein, M.G. and A.A. Nudelman (1973) Markov Moment Problem and Extremal Problems. Moscow: Nauka, 552.

Mehra, R.K. (1976) Optimal Input Signals for Parameter Estimation in Dynamic System. IEEE Trans. on Autom. Control AC-19:753-768.

Pazman, A. (1986) Foundations of Optimum Experimental Design. Dordrecht: D. Reidel Publishing Company, 228.

Silvey, S.D. (1980) Optimal Design. London: Chapman and Hall, 86.

Strand, O.N. and E.R. Westwater (1968) Minimum RMS-Estimation of the Numerical Solution of a Fredholm Integral Equation of the First Kind. SIAM J. Numer. Anal. 5:287-295.

Twomey, S. (1966) Indirect Measurements of Atmospheric Temperature Profiles from Satellites: Mathematical Aspects of the Inversion Problem. Monthly Weather Review, 99:363-366.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE USE OF CLIMATOLOGICAL STATISTICS IN THE PREDICTION OF EXTREME WIND SPEEDS IN TROPICAL CYCLONE REGIONS

Peter N. Georgiou

Boundary Layer Wind Tunnel Laboratory
The University of Western Ontario
London, Ontario, Canada N6A 5B9

1. INTRODUCTION

Tropical cyclones, or hurricanes and typhoons as they are known in some parts of the world, are the most devastating of all natural phenomena. Measured by their violence, duration and size of area affected, they are without equal in terms of the amount of destruction they can cause. The primary hazards are threefold: high winds causing the destruction of buildings, trees, crops and power and communication lines, intense rainfall resulting in river and local flooding, and storm surge, the rapid increase in sea-level along the coastline as water is driven ashore by the wind. The design of buildings, especially emergency services and hurricane shelters, to withstand tropical cyclone winds clearly occupies an important position in the disaster preparedness of all vulnerable countries. In many smaller nations, for example in the Caribbean, this becomes particularly acute as recovery from a hurricane disaster may depend upon the survival of a single hospital, power plant or communication centre.

Basic to the mitigation of the tropical cyclone hazards - through the assessment of the areas likely to be hardest hit and the design of structures to resist the hurricane-force winds - is reliable information on what wind speeds might occur and their relative risk. In many tropical cyclone-prone areas of the world, meeting this need has been made difficult by the lack of suitable long-term wind records. With a frequency of occurrence of wind speeds of at least tropical storm intensity at any particular locality of only once per annum or less tropical cyclones constitute a relatively rare event. This is aggravated in the case of surface records by the vulnerability of surface anemometers to damage; frequently the crucial observations of wind speed during a severe storm are lost. Furthermore, the routine relocation of anemometers and changes to the urban environment affect the uniformity of the exposure and continuity of the records. Thus the wind data required to define the tropical cyclone wind climate and hence estimate extreme wind speeds and the associated risk rarely, if ever, exist.

In this paper, these obstacles are avoided through the use of a tropical cyclone simulation procedure, developed during the past decade and described in several technical papers (Georgiou et al., 1983; Georgiou, 1984, 1985). A key ingredient in the simulation approach is the statistical representation of tropical cyclone climatological characteristics in the area of interest, i.e. local occurrence rates and parameters defining the local storm intensity, size, speed and track characteristics. Examples of these will be given for the North Atlantic region, indicating how this climatological information is incorporated in the simulation method to predict extreme wind speeds. The regional tropical cyclone wind risk maps which can then be developed by repeated application of the method will be shown. The potential for identifying areas prone to very high winds and the consequent benefit to the future siting of sensitive facilities such as communication towers will be indicated for example locations in the Caribbean.

2. MONTE-CARLO SIMULATION OF TROPICAL CYCLONE WIND SPEEDS

The Monte-Carlo simulation method for estimating tropical cyclone wind speeds has been so-named because of the statistical nature of the computations involved. The method relies on the following two observations. Firstly, the windfield of a tropical cyclone, i.e. the wind speeds and wind directions, can be calculated with sufficient accuracy knowing only certain key characteristic

parameters of the storm. These include the central pressure difference, Δp, the radius of maximum winds, RMAX, the storm translation speed, VT, the translation direction of its track, Θ, and lastly the position of the point of interest relative to the storm centre. Secondly, long-term historical records of these same parameters exist, in the case of the North Atlantic, extending back to 1886, a record length of 100 years. Similar records of varying length documenting the tracks and intensities of these storms exist for the other tropical cyclone regions of the world. A major advantage of the simulation approach in estimating tropical cyclone winds is that the meteorological records of storm characteristics used are not prone to the same sampling and interpretation problems which affect wind speed and direction records.

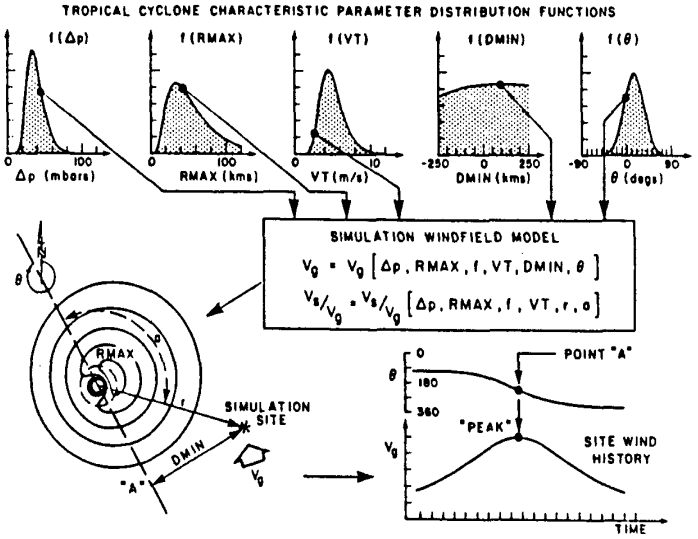


Figure 1

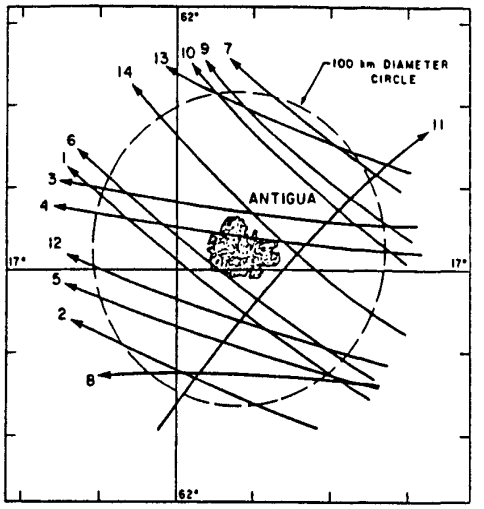
Monte-Carlo simulation procedure flow chart

These two notions form the basis of the simulation method, as indicated in Figure 1. **Step 1:** At any particular location, the historical data records are analyzed to produce a statistical representation of the tropical cyclone parameters (Δp, RMAX, VT, Θ etc.) required by the simulation windfield model. All recorded tropical cyclones within a given specified distance are included in the analysis, typically 250 km for North Atlantic storms. **Step 2:** These distributions are then sampled sequentially at random to generate sets of tropical cyclone characteristic parameters. From each set a new tropical cyclone is defined and its windfield computed. **Step 3:** The windfield is moved past the site of interest and the wind speeds and wind directions at the site recorded. A separate record is kept of all wind speeds and their respective directions, the "parent" distribution, as well as the individual peak wind speeds recorded for each tropical cyclone generated, the "extreme" distribution. The parent distribution yields information regarding the directional characteristics of the tropical cyclone wind climate while the extremes define wind speed levels suitable for design purposes. **Step 4:** A sufficient number of storms are run so that the results, e.g. the set of extreme wind speeds, stabilize statistically. The storms created during the simulation bear the same climatological characteristics as those that were determined at the site from the historical records, as well, it is assumed, as to the future. The simulation however can be run over virtually any length of time (i.e. over thousands of years).

The windfield model employed in the simulation procedure utilized in this paper has undergone considerable refinement in recent years and is described in detail in Georgiou (1985). The model predicts wind speeds and directions at three levels, (i) at the 700 mb height, where gradient balance is assumed to apply, (ii) at the 500m height, where surface friction and translation of the storm system itself combine to modify the circulation in a manner which is computed using a recently developed numerical hurricane boundary layer model, and (iii) at the surface or 10m height. The simulation windfield model is capable of modifying the winds at each level once a storm makes landfall and can adjust surface winds to account for a uniform change in terrain roughness. The accuracy of the windfield model has been assessed by comparing actual wind records obtained in tropical cyclones with those predicted by the windfield model using the known storm values of Δp, RMAX, VT and track parameters. This has been done for tropical cyclones in the North Atlantic, Western North Pacific and Australian regions. The conclusion from these comparisons is that, although variations can be found from storm to storm, the errors from a statistical viewpoint are comparatively small.

Figure 2

Tropical cyclones reaching at least tropical storm intensity, passing within 50 km of Antigua: 1886-1985



1.	SEPT.	3	1889	8.	AUG.	22	1950 "BAKER"
2.	SEPT.	21	1898	9	SEPT.	1	1950 "DOG"
3.	AUG.	3	1900	10.	SEPT.	11	1955 "HILDA"
4.	SEPT.	6	1910	11.	OCT.	28	1963 "HELENA"
5.	SEPT.	25	1910	12.	SEPT.	3	1973 "CHRISTINE"
6.	AUG.	28	1924	13.	SEPT.	3	1979 "FREDERIC"
7.	AUG	13	1943	14.	SEPT.	4	1981 "FLOYD"

The statistics of the tropical cyclone characteristic parameters required by the simulation windfield model are found for any particular geographic region from records of past storms. For the North Atlantic these have been conveniently compiled by the National Hurricane Centre, Coral Gables, Miami, Florida, on the so-called "HURDAT" tape. This tape documents the tracks and intensity of all known tropical cyclones reaching at least tropical storm intensity for the period 1886 to 1985 inclusive. An example of the historical data compiled on the HURDAT tape is given in Figure 2, showing the 14 tropical cyclones passing within 50 km of Antigua in the period 1886-1985. Much effort has gone into determining which probability distributions are suitable for representing the various parameters. In the North Atlantic, an appropriate choice of distributions (Georgiou, 1985) is: the Poisson distribution for the occurrence rate; the Weibull distribution for the central pressure difference, Δp; the lognormal distribution (conditionally dependent upon Δp) for the radius of maximum winds, RMAX; the lognormal distribution for the translation speed, VT; the Von Mises distribution for the storm direction angle, Θ; and a cubic polynomial distribution for the minimum approach distance of the storms to the site of interest. These distributions do not universally apply to all tropical cyclone areas. For example, Davenport et al. (1984) found that the lognormal distribution was better suited to representing the distribution central pressure difference in the Western North Pacific region.

3. EXAMPLE SIMULATION RUN - ANTIGUA

The simulation procedure is illustrated with a single simulation run for the Caribbean island of Antigua. The first step involves determining the characteristic parameter distribution functions. A simulation circle is defined around the site of interest. In the simulation of Antigua's tropical cyclone wind climate, as for all North Atlantic locations, all documented storms that passed within 250 km of the island in the period 1886-1985 were used to determine the statistics of storm occurrence rate, central pressure difference, speed, direction etc. Data obtained from the HURDAT tape showed that the average number of storm occurrences per year for this size circle is 1.07. Figure 3a shows the random nature of the occurrence of these tropical storms and hurricanes. For this reason tropical cyclone occurrence is modelled by the Poisson distribution using the measured annual occurrence rate to define the Poisson parameter of the process. Figure 3b shows the distribution of translation directions for all storms passing within 250 km of Antigua, indicating a mean direction of 291° (69° west of North). The distribution chosen to model the translation direction of tropical cyclones, the Von Mises distribution, is the counterpart on the circle to the Gaussian distribution on the straight line. The distribution of translation speeds for tropical cyclones in the Antigua simulation data circle is shown in Figure 3c. The distribution exhibits a skewness which lends itself to be modelled by the lognormal distribution. The mean value indicated in the area is 6.3 m/sec.

The results of the Antigua simulation using 2500 generated tropical cyclones are shown in Figure 4. The 100-year recurrence interval extreme wind speed estimates are 51 and 42 m/sec for the 500m and surface heights respectively. These values are readily converted to extreme wind pressures, a direct measure of the structural loads imposed on buildings. It should be noted that the surface wind speeds refer to flat, open-water exposure terrain, and would have to be corrected to account for topographic and local surface features at individual sites on the island.

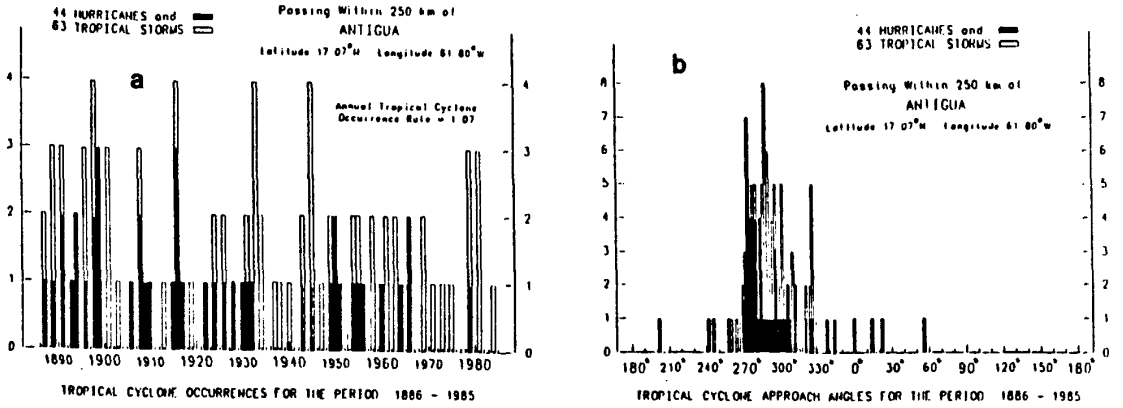
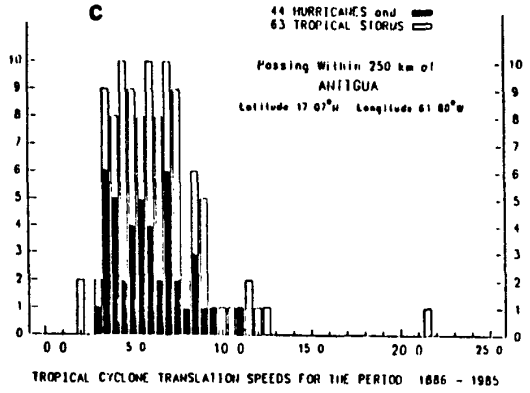


Figure 3 Tropical cyclone characteristics used in ANTIGUA simulation, for 500 km diameter data circles



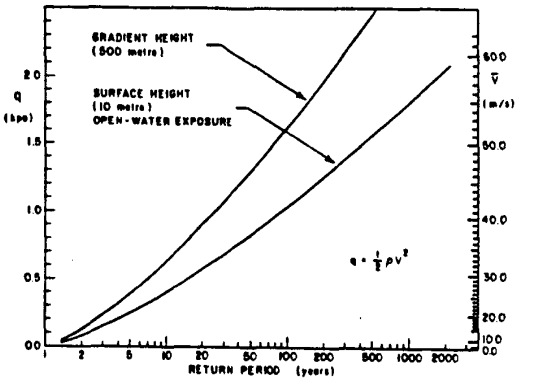


Figure 4 ANTIGUA simulation extreme mean wind speeds and wind pressures

4. TROPICAL CYCLONE REGIONAL WIND RISK MAPS

4.1 Tropical cyclone wind risk in the Lesser Antilles region

The utility of the simulation approach can be demonstrated by its repeated application over a given region to develop regional tropical cyclone wind risk maps. Figure 5 shows the variation of mean annual occurrence rate, mean central pressure difference and mean translation speed for the Lesser Antilles region of the Caribbean based on 250 km radius simulation data circles. The mean annual occurrence rate varies from less than 0.2 in the south to a local maximum of 1.1 just to the west of Guadeloupe. This latter area is a convergence zone for storm paths in the region. The variation of mean central pressure difference shows a strong gradient in a north and west direction. This agrees well with the observation that the southeast area of the region, i.e. east of Trinidad and Barbados, is a formation location for the North Atlantic with storms then typically moving in a northwesterly direction. Thus storms would tend to intensify and their central pressure difference values would increase as they moved into the northern and western areas of the region. There appears an interesting plateau of tropical cyclone intensity surrounding the islands from Dominica north to Antigua, where the regional increase in storm intensity holds constant. In this area, the relatively dense network of island masses has often been observed to cause temporary filling of traversing tropical cyclones. The mean translation speed of tropical cyclones in the region shows a gradual decrease from south to north, reflecting the "steering" influence of the trade winds in the area, which also decrease in strength towards the north.

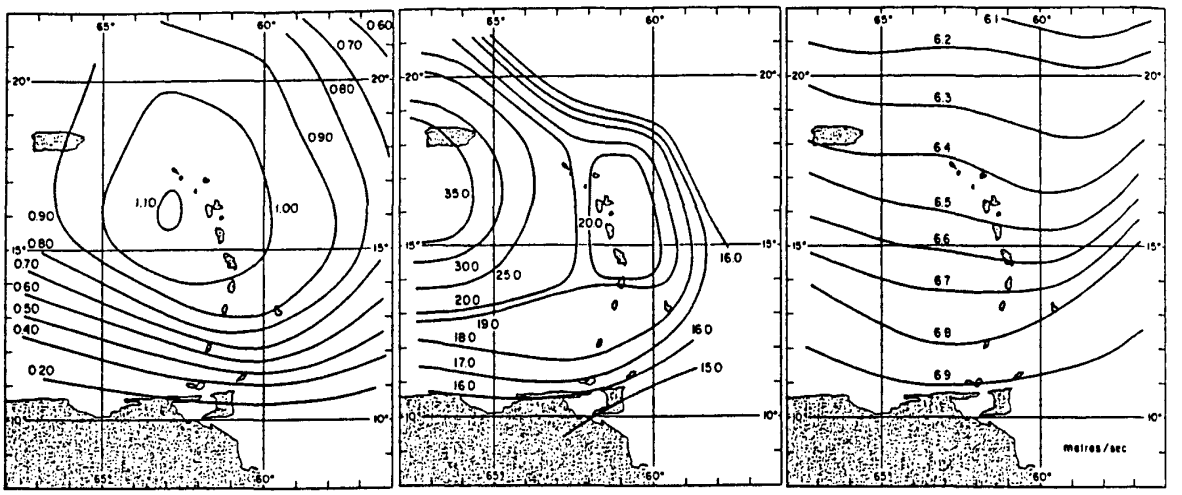


Figure 5

Variation of tropical cyclone characteristic parameters in LESSER ANTILLES, for 500 km diameter data circles

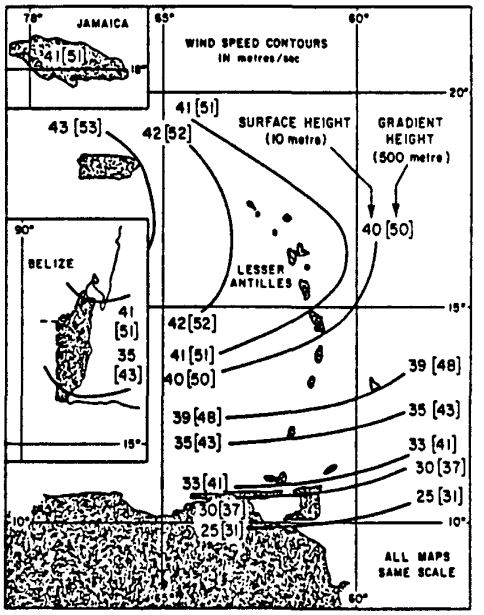


Figure 6 Simulation 100-year extreme mean wind speeds for LESSER ANTILLES

The results of repeatedly applying the simulation procedure throughout this region are shown in Figure 6, giving the 100-year recurrence interval mean (10-20 minute average) wind speeds at the surface and 500m height. Tropical cyclone simulations were run based on the generation of 2500 storms. At each site, the actual simulation time length varied according to the site's local occurrence rate. It can be seen that the wind risk in the region generally increases from south to north reflecting both the increased frequency of occurrence of tropical cyclones and the general increase in storm intensity in this direction.

4.2 Tropical cyclone wind risk along the U.S.A. coastline

The statistics of tropical cyclone characteristic parameters along the Gulf and Atlantic seabords of the United States has been studied by Georgiou (1985). The area experiencing the highest tropical cyclone frequency of occurrence is Cape Hatteras, related to its geographically-exposed location on the Atlantic coast. On the other hand, the area experiencing the maximum tropical cyclone intensity, reflected in the variation of central pressure difference value shown in Figure 7, is the southern Florida peninsula region. The translation speed of tropical cyclones increases monotonically with latitude. The average storm speed ranges from under 5 m/sec on the western Gulf Coast to over 15.0 m/sec on the Atlantic Coast north of Boston. Figure 7 also indicates the variation of direction of motion exhibited by tropical cyclones around the U.S. coastline. At the western end of the Gulf of Mexico storms are typically moving west-northwest and northwestward. From New Orleans till the Florida west coast the mean value of translation direction is almost exactly zero (due North). North of the 30° latitude tropical cyclones typically veer northeastward away from the coast.

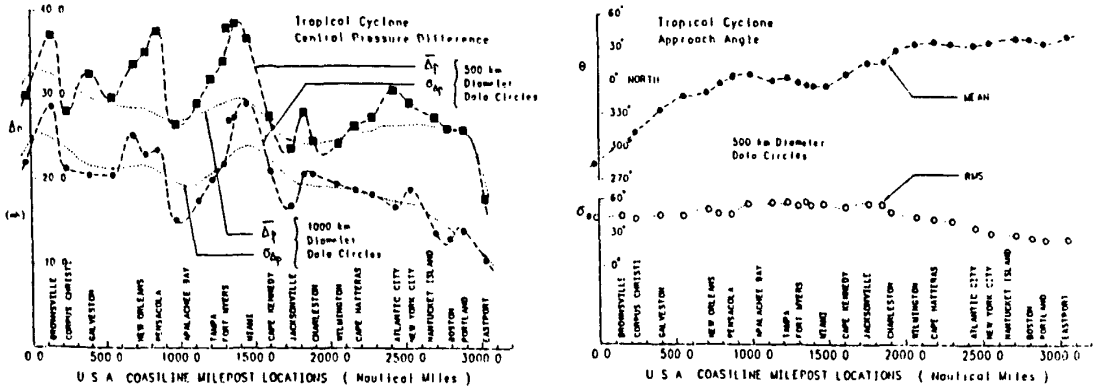


Figure 7 Variation of tropical cyclone intensity and translation direction along U.S.A. coastline, for 500 km diameter data circles

The simulation procedure was applied to the Gulf and Atlantic Coasts of the United States at 250 km intervals around the coastline. The resulting variation in extreme wind speeds is shown in Figure 8. The highest extremes are found, not unexpectedly, in the southern Florida region. Secondary local maxima are found at the U.S.-Mexican border, the Mississippi Delta-Pensacola area, the Charleston-Cape Hatteras area and Cape Cod. Differences between neighbouring sections of coastline can be large, e.g. the 100-year surface wind speed estimate drops from 47.8 to 37.3 m/sec in the 500 km distance from Miami to Jacksonville. This translates into a 50% drop in the extreme estimate of wind pressure, with consequent effects on risk to buildings etc. Figure 9 shows a surface risk map for U.S. coastal regions produced by additional application of the simulation procedure to both offshore and inland locations. The potential to identify areas where the risk is high is of obvious value to planners, engineers and other interested groups (e.g. the insurance industry).

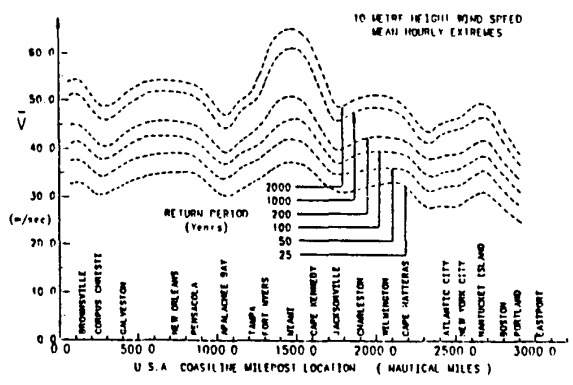


Figure 8 Simulation extreme wind speeds for U.S. coastline

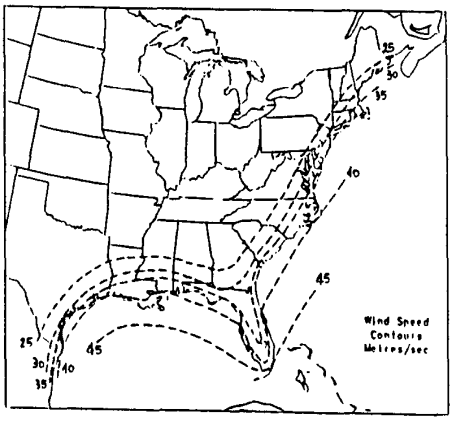


Figure 9 100-year simulation extreme 10-metre mean wind speeds

5. THE INFLUENCE OF TOPOGRAPHY ON TROPICAL CYCLONE WIND RISK

The surface wind speed predicted by the simulation model is a 10m height wind speed and applies to open-water exposure conditions. In the case of many of the typically rugged and mountainous Caribbean island nations however, wind speeds may be substantially lower in sheltered areas on the leeward coast or may be subject to significant speed-up in exposed bluff locations. The influence of topography has been investigated by Davenport et al. (1985) using the Caribbean island of Nevis as a pilot study. A 1:3000 scale model of the island was constructed and tested in the low-speed section of The University of Western Ontario's Boundary Layer Wind Tunnel Laboratory. Wind speed measurements were made at a variety of locations on the island. The results for two sites are shown in Figure 10, giving the speed-up of local winds at the surface compared to the undisturbed gradient height wind speed upstsream of the island. For the Charlestown Hospital site, where the topography is gently rolling and not extreme, winds are relatively insensitive to direction, with some minor shielding when the wind is coming over 1000m high Nevis peak. For the second site at Rawlins Estate, located closer to the peak and on higher ground, a dramatic speed-up is evident in the surface wind when the oncoming wind is from the east-northeast or west-southwest.

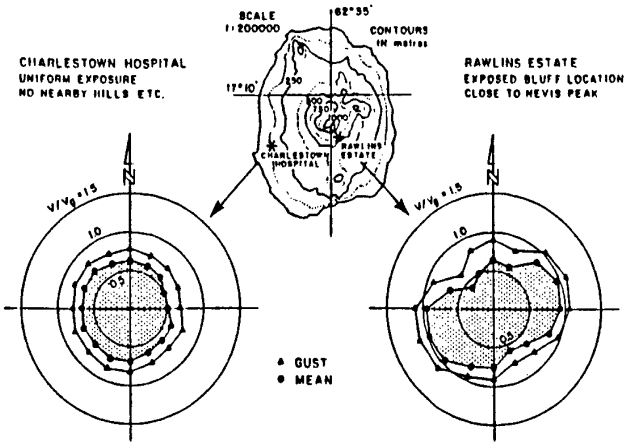


Figure 10 NEVIS surface to gradient wind speed ratios

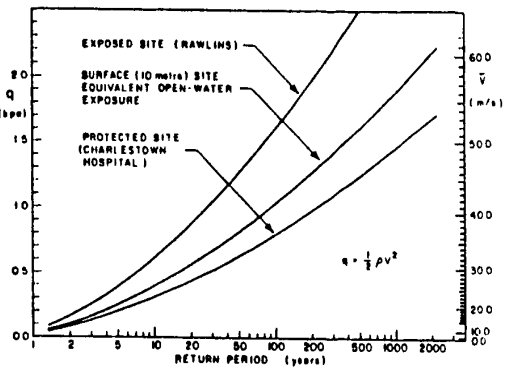


Figure 11 NEVIS extreme mean surface speeds for different topographies

The simulation procedure may be used to establish the local wind risk if such correction factors accounting for topography are available. As the simulation procedure is being executed, the hour by hour wind speeds and directions being generated at the site of interest can be multiplied by the relevent topographic correction factors and a corrected set of extremes be compiled and analyzed. This was done for the two Nevis locations discussed above. The results are shown in Figure 11, indicating the 100-year recurrence interval wind speed at Rawlins is about 25% higher than over open-water, and at the Charlestown Hospital site about 10% less. These correspond to wind pressures (proportional to the square of the wind speed) which are, respectively, approximately 50% more and 20% less than the open-water exposure. This single comparison indicates differences of the order of 100% and greater in estimated wind pressures between sheltered and exposed sites. Needless to say, this information would be of great value in the decision-making process involved in say the siting of a future communication tower, whose integrity would have to be guaranteed in the case of a severe tropical cyclone event.

ACKNOWLEDGEMENTS

Much of the Caribbean regional data discussed in this paper is taken from the study of hurricane wind risk in the region by The University of Western Ontario's Boundary Layer Wind Tunnel Laboratory for the Pan-Caribbean Disaster Prevention and Preparedness Project (PDCPPP). The project was carried out under the auspices of the United Nations Disaster Relief Organization (UNDRO) with support from the Canadian International Development Agency (CIDA). Adaption of the simulation procedure to the United States coastline and Eastern Caribbean region was facilitated by assistance from the National Hurricane Centre, Miami, Florida, and the Caribbean Meteorological Institute, Barbados, respectively.

REFERENCES

Davenport, A.G., Georgiou, P.N., Mikitiuk, M., Surry, D. and Lythe, G., 1984: Design wind speeds in the Hong Kong region. 3rd Intl. Conf. on Tall Buildings, December 10-15, Hong Kong and Guangzhou.

Davenport, A.G., Georgiou, P.N. and Surry, D., 1985: A hurricane wind risk study for the Eastern Caribbean, Jamaica and Belize with special consideration to the influence of topography. Report to UNDRO-PCDPPP, BLWTL Rep. SS31, The University of Western Ontario, Canada.

Georgiou, P.N., Davenport, A.G. and Vickery, B.J., 1983: Design wind speeds in regions dominated by tropical cyclones. Proc. 6th Intl. Conf. on Wind Eng., Gold Coast, Australia; Auckland, New Zealand.

Georgiou, P.N., 1984: Simulation of hurricane wind speeds. ASCE Specialty Conf. "Hurr. Alicia - One Year Later", August 16-17, Galveston, Texas, U.S.A.

Georgiou, P.N., 1985: Design wind speeds in tropical cyclone-prone regions. Ph.D. Thesis, BLWTL, The University of Western Ontario, Canada.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

AN ANALYSIS OF RAIN-PRODUCING SYSTEMS ON COASTAL GUYANA

Anthony B. Shaw
Department of Geography
Brock University
St. Catharines, Ontario
Canada

1. INTRODUCTION

Rainfall variability constitutes the dominant characteristic of the climate of Guyana. Problems to farmers arise from variability in onset dates of the two main rainfall seasons and from occurrences of dry spells interspersed with periods of excessive rainfall amounts. Adequate precipitation does not always coincide with crop growth stages due to annual variability in the onset and advance of the rainfall seasons. The annual rainfall distribution at coastal stations shows typically a bi-modal pattern with major peaks in June and December when this region is under the direct influence of the Inter-Tropical Convergence Zone. The two rainy seasons have been usually defined by the length of period with rainy days: (1) a long wet season which extends from mid-April to the middle of August and (2) a short wet season which extends from mid-November to the middle of February. The intervening periods are referred to as the long dry season and short dry season. However, the seasonal variability is such that distinct alternating wet and dry seasons may not materialize in a particular year and consequently crop yields may fluctuate. In fact, this highly generalized description of the rainfall regime purported to be true of numerous coastal stations is misleading.

Guyana's coastal rainfall is typically equatorial and is produced primarily by mesoscale synoptic systems throughout the year as a result of strong daily surface heating. This rainfall mechanism, which is associated with individual convection columns accounts for the large variations of daily and monthly rainfall over short distances. While one station might be receiving well above average monthly rainfall, another station a few miles away might not receive even a normal amount. These differences of rainfall over short distances occur prior to the arrival of the ITCZ, i.e. during the two intervening periods of low precipitation (the dry seasons).

Mesoscale synoptic systems over the coastal region are in turn intensified by large-scale systems of the Inter-Tropical Convergence Zone. The ITCZ is a narrow east-west band of vigorous cumulonibus convection and heavy precipitation which forms along the Equatorward boundary of the trade wind regimes. Within this zone precipitation exceeds evaporation by a factor of two or more, the excess moisture being supplied by the moist, converging low-level airflows (Holton and Wallace, 1970). The ITCZ over land, in general, seems to follow the seasonal march of the sun. But the zone of heavy precipitation may lag behind in the sun's path by as much as one month over land and two months over sea (Riehl, 1979). On the Guyana coast northward movement of the ITCZ brings heavy and prolonged summer rains with peak rainfall in June. A second major rainfall peak occurs in December when the coastal region is once again under its direct influence.

Temporal and spatial variations of precipitation is most likely linked to seasonal shifts in the average position of this equatorial trough as it migrates north and south of the equator. In some years peak precipitation has occured as late as July for the summer rainy season and as late as January for the winter rainy season. Satellite observations have also revealed a year to year variability in the intensity of weather associated with the ITCZ. Major shifts, especially of latitudinal alignment, seem to be related to the shifting zone of maximum insolation to which the intertropical cloud is closely tied (Gruber, 1972).

The cloud band itself may mitigate against its own survival in that position by shielding the surface beneath from direct insolation and thereby reduce surface heating that is so necessary to maintain the trough (Barret, 1974). The cloud and rain areas associated with the trough are often irregular and the clouds and rain appear to jump between locations without affecting the intervening areas and are often counter to seasonal trends (Jackson, 1977). This sequence of events could account for the observations which shows that the trough appears to move discontinuously and is made up of "cloud clusters", separated by relatively large expanse of clear skies (Chang, 1970).

Meridional shifts of the ITCZ are also linked to the passage of easterly waves which act as trigger mechanisms creating areas of low convergence and ascent. Moreover, high latitude circulation systems also exert their forces through the strength of their air streams which influence the location of the ITCZ and the intensity of weather. Such forces operating in each hemisphere are by nature variable, hence equatorial areas such as the Guyana coastlands experience a variety of weather even though its climatic pattern suggests year-long monotony. The climatological significance of the ITCZ to agriculture, therefore, rests on careful observation of its movements and associated weather systems as it can provide a time frame for assessing the intensity and spread of seasonal rainfall.

An attempt is made in this paper to determine the extent to which coastal stations are synchronously affected by the same major rain-producing systems and the extent to which the occurrence of rainfall is related to atmospheric processes operating on a small scale (Figure 1). Rain-producing systems can be broadly described as (1) stationary with limited spatial extent and the tendency for rainfall to develop in situ, and (2) non-stationary, which produces widespread precipitation that is spatially highly variable. The latter is usually associated with the ITCZ which slowly traverses the coastal region seasonally bringing widespread and heavy precipitation. Hence, its latitudinal location and spatial dimension determines what stations will be synchronously affected.

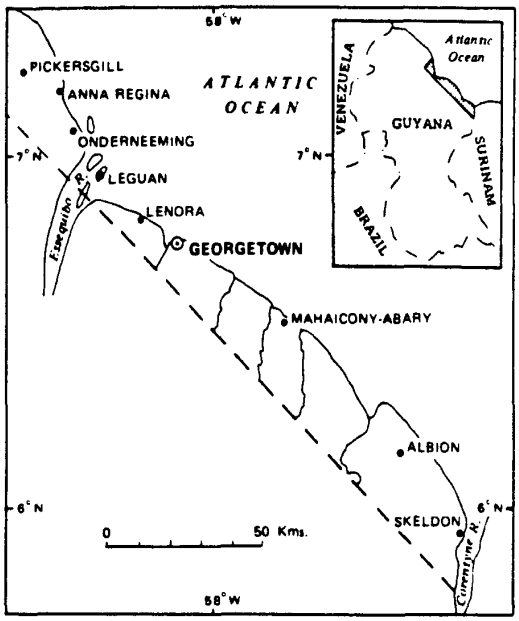


FIGURE 1. Coastal region of Guyana and location of stations.

An appreciation of the size and direction of movement of rain-producing systems can be obtained with the use of correlation analysis. The correlation coefficient r_{ij} between any two stations i and j in a network can be expressed as a function of the distance d_{ij} between two respective points

$$r_{ij} = r(d_{ij})$$

The rate of decrease of the resulting correlation function has been shown to be directly related to the size of rain-producing cells (Sharon, 1972; Longley, 1974). The larger systems should result in greater spatial coherence of seasonal rainfall regime and high interstation correlations of seasonal rainfall. If the seasonal rainfall is highly localized, then marked spatial variations should be evident and should be reflected in low interstation correlations.

Interstation correlations for yearly and seasonal rainfall totals tend to be low in the tropics and subtropics as documented in several studies (Jackson, 1977; Snijders, 1986). Typically, interstation correlations with distance tend to show a decreasing trend, suggesting the localized nature akin to tropical rain. In contrast, trends for mid-latitude regions,

show a less rapid decrease of correlations with distance especially in winter under the influence of large-scale cyclonic systems.

2. DATA AND APPROACH

All analyses are based on daily rainfall data for eight coastal stations. All rainfall stations are located within major cultivated areas and have thirty or more years of continuous rainfall data. Only the more recent years (1947-1977) were utilized in this study in order to maintain a consistent time period of analysis for all stations. These stations were also selected to provide a good geographic representation of the climatic regime of the coastal region. The annual rainfall is distributed mainly over two seasons: (1) a summer rainfall season beginning from April and through to the end of September, and (2) a winter rainfall season beginning from October through to the end of March of the subsequent year. The latter is usually referred to as the short rainy season for the obvious reason that it contains a lesser number of rainy days. Mean annual rainfall totals range from 2622mm in the northwest to 1912mm in the east and is characterized by a high degree of variability at individual stations (Table 1).

TABLE 1 : Mean Total Rainfall (mm) and Coefficient of Variation

	Summer	%	Winter	%	Annual	%
Pickersgill	1612	25	1020	35	2622	24
Anna Regina	1265	30	839	42	2104	28
Onderneeming	1447	27	872	43	2319	26
Leguan	1381	34	932	48	2313	33
Lenora	1490	26	904	42	2394	25
Mahaicony-Abary	1278	20	776	43	2054	23
Albion	1189	21	723	45	1912	27
Skeldon	1320	19	729	33	2049	18

Throughout its length, the coastal plain rises less than ten centimeters per kilometer from the low water mark on the coastline to approximately eighteen kilometers inland. This region has a flat topography with an average elevation of less than two meters above sea level. It is ideally suited for the production of field crops such as rice and sugar, which are extensively cultivated throughout the coastal plain. However, its low-lying nature makes it highly susceptible to flood waters from the Atlantic Ocean and from excessive precipitation.

The following approaches are used to determine the extent to which coastal stations are synchronously affected by the same rain-producing systems and whether the spatial variations in coastal rainfall can be considered typical of the tropics.

(1) interstation correlations of seasonal rainfall and annual rainfall totals. Here the objective is to assess the similarity of rainfall pattern in terms of the magnitude of precipitation associated with rain-producing systems (see Table 1).

(2) interstation correlations of the seasonal and annual totals of days with precipitation. This approach attempts to measure the similarity of rainfall in terms of the dimensions of rain-producing systems (see Table 2).

(3) interstation correlations of the onset dates of heavy rainfall associated with the movement of the ITCZ. This approach attempts to determine the extent to which large-scale weather systems associated with this zone of convergence affects the onset of heavy persistent rainfall synchronously at all stations in its seasonal path.

TABLE 2 : Mean Total Number of Days with Precipitation and Coefficient of Variation

	Summer	%	Winter	%	Annual	%
Pickersgill	128	18	93	27	221	61
Anna Regina	101	19	68	25	169	48
Onderneeming	94	28	59	30	152	50
Leguan	102	31	63	36	165	49
Lenora	120	15	80	20	200	48
Mahaicony-Abary	112	19	74	27	186	47
Albion	98	13	63	30	161	40
Skeldon	120	11	75	18	195	20

A day with precipitation in this study is defined as any day with 0.25mm or more of rainfall. This is the lowest quantifiable value recorded at any of the stations used in this study. The daily precipitation range from low of 0.25mm to maximum of 130mm. Onset of the rainy season is defined here as the date when accumulated rainfall from the beginning

of each season, equals or exceeds 10 percent of the mean seasonal total, provided there are no dry spells exceeding 7 days after this date. A prolonged dry spell following the date of the accumulated threshold value could lead to a 'false start' of the rainfall season. When this occurs, accumulation begins again at the end of the dry spell.

Spatial coherence and the localness of rainfall was investigated by the coefficient of correlations between pairs of stations and tested for their statistical significance using the Student's t test. The results are presented in the form of graphical plots showing the relationships between the correlation coefficients and distance between stations and in correlation matrices for each parameter.

3. RESULTS AND ANALYSES

(a) Interstation Correlations of Rainfall Totals

The spatial coherence of rainfall in terms of the magnitude of precipitation associated with rain-producing systems is shown in Figure 2. A clear negative trend is evident in all three situations but it appears to be less pronounced for the winter (short) rainfall season. As the distance between stations increases the coefficients decrease and become negative for pairs of stations that are widely separated. At a distance of about 80 km the coefficients are positive and statistically significant while at a distance of about 160 km and greater, they become zero and negative, and are not statistically significant.

Tables 3(a), 3(b) and 3(c) show interstation correlation coefficients that are statistically significant. What emerges is a group of six stations with high positive interstation correlations and that are statistically significant at 0.05 level, while two stations, Albion and Sheldon, appear to be in a distinct group of their own showing low negative interstation correlations. These two sub-groups of stations within the coastal region clearly reflect a variation in the amounts of precipitation which appears to be directly related to the intensity of the rain-producing systems. The bulk of the precipitation which is normally associated with large-scale

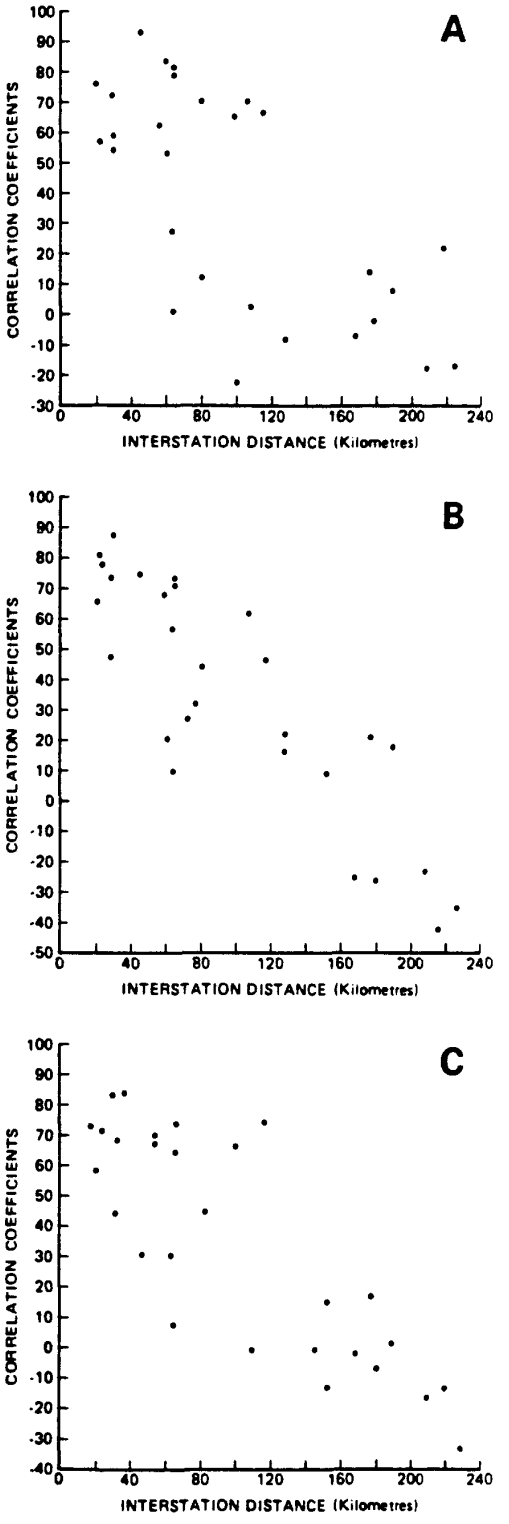


FIGURE 2. Plot of interstation correlation coefficients against distance for winter (A), summer (B) and annual rainfall totals.

systems of the ITCZ, is not uniformally distributed. Evidently the magnitude of this rain-producing system is not uniformly felt over this region.

TABLE 3a : Interstation Correlations of Seasonal Rainfall Totals (Winter)

	1	2	3	4	5	6	7	8
1. Pickersgill	1.0							
2. Anna Regina	.56*	1.0						
3. Onderneeming	.75*	.57*	1.0					
4. Leguan	.84*	.40*	.60*	1.0				
5. Lenora	.80*	.62*	.92*	.76*	1.0			
6. Mahaicony-Abary	.74*	.67*	.71*	.70*	.81*	1.0		
7. Albion	.15	.07	-.22	.14	-.08	.01	1.0	
8. Skeldon	-.18	.22	-.19	-.03	-.06	.03	.55*	1.0

TABLE 3b : Interstation Correlations of Seasonal Rainfall Totals (Summer)

	1	2	3	4	5	6	7	8
1. Pickersgill	1.0							
2. Anna Regina	.81*	1.0						
3. Onderneeming	.87*	.79*	1.0					
4. Leguan	.68*	.46*	.47*	1.0				
5. Lenora	.74*	.68*	.76*	.67*	1.0			
6. Mahaicony-Abary	.57*	.47*	.61*	.44*	.71*	1.0		
7. Albion	.21	.19	.10	.18	.22	.10	1.0	
8. Skeldon	-.34	-.43*	-.22	-.26	-.24	-.06	.20	1.0

TABLE 3c : Interstation Correlations of Annual Rainfall Totals

	1	2	3	4	5	6	7	8
1. Pickersgill	1.0							
2. Anna Regina	.72*	1.0						
3. Onderneeming	.84*	.72*	1.0					
4. Leguan	.61*	.30	.47*	1.0				
5. Lenora	.72*	.68*	.85*	.60*	1.0			
6. Mahaicony-Abary	.66*	.75*	.67*	.46*	.75*	1.0		
7. Albion	.18	.03	-.12	.15	-.07	.08	1.0	
8. Skeldon	-.34*	-.12	-.16	-.05	-.02	.01	.31	1.0

* Coefficients are statistically significant at 0.05 level using the Student's t test.

(b) Interstation Correlations of Days with Rainfall

An overall negative trend is also observed (see Figure 3). Again, the short wet season does not show a distinct trend as the long wet season and only a few correlations are statistically significant (Table 4). The annual totals are all positively correlated and statistically significant, suggesting a moderate overall degree of spatial coherence.

However, spatial coherence varies with season, being much higher during the long wet season as evidenced by the larger number of positive correlations. Correlation coefficients decline to zero at a distance of about 200 km for the summer rainy season and at about 160 km for the winter rainy season.

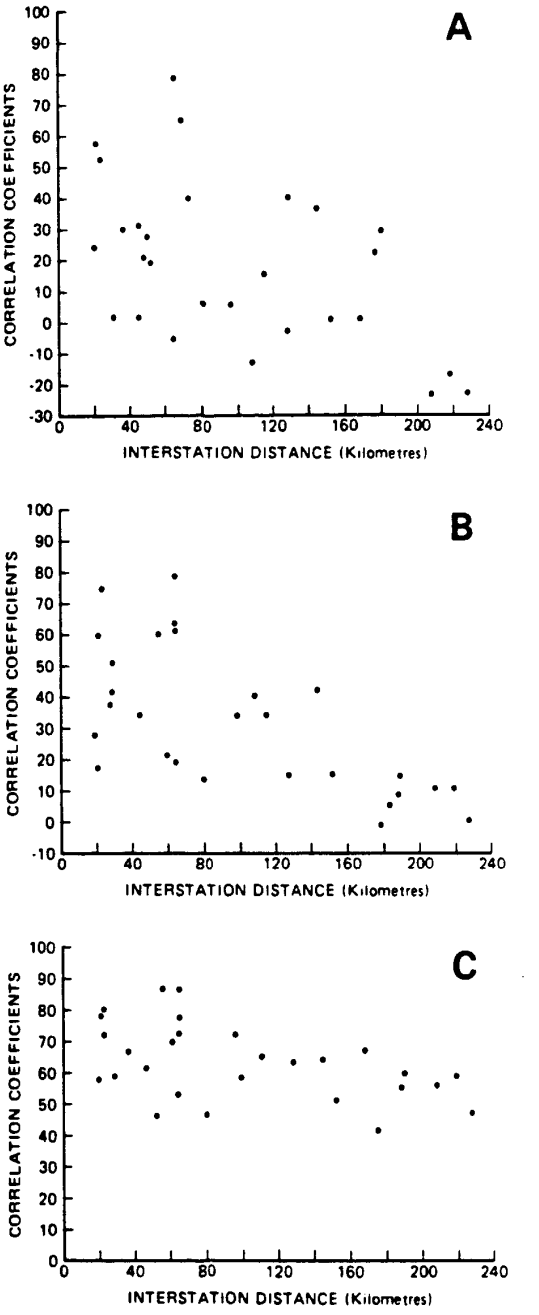


FIGURE 3. Plot of interstation correlation coefficients against distance for the total number of rainy days in winter (A), summer (B) and year (C).

Higher numbers of days with precipitation in the summer season than those in the winter season might be associated with more widespread convective activities suggesting a non-uniform spatial pattern of the ITCZ in its seasonal migration.

TABLE 4a : Interstation Correlations of Days with Precipitation (Winter)

	1	2	3	4	5	6	7	8
1. Pickersgill	1.0							
2. Anna Regina	.53*	1.0						
3. Onderneeming	.25	.54*	1.0					
4. Leguan	.20	.32	.02	1.0				
5. Lenora	.79*	.66*	.30	.21	1.0			
6. Mahaicony-Abary	.40*	.16	.06	.06	.45*	1.0		
7. Albion	.11	.24	.01	.37*	-.03	-.06	1.0	
8. Skeldon	-.19	-.14	-.17	.31	.02	-.13	.20	1.0

TABLE 4b : Interstation Correlations of Days with Precipitation (Summer)

	1	2	3	4	5	6	7	8
1. Pickersgill	1.0							
2. Anna Regina	.51*	1.0						
3. Onderneeming	.38*	.74*	1.0					
4. Leguan	.18	.37*	.42*	1.0				
5. Lenora	.79*	.60*	.37*	.28	1.0			
6. Mahaicony-Abary	.64*	.33	.35*	.13	.63*	1.0		
7. Albion	.07	.14	.15	.42*	.15	.19	1.0	
8. Skeldon	.02	.11	.11	-.02	.06	.40*	.22	1.0

TABLE 4c : Interstation Correlations of Days with Precipitation (Year)

	1	2	3	4	5	6	7	8
1. Pickersgill	1.0							
2. Anna Regina	.72	1.0						
3. Onderneeming	.57	.81	1.0					
4. Leguan	.46	.60	.56	1.0				
5. Lenora	.89	.81	.66	.58	1.0			
6. Mahaicony-Abary	.71	.60	.57	.46	.78	1.0		
7. Albion	.41	.59	.51	.64	.58	.52	1.0	
8. Skeldon	.48	.58	.55	.55	.67	.64	.71	1.0

* Coefficients are statistically significant at 0.05 level using the Student's t test. All coefficients are significant in Table 4c.

(c) Interstation Correlation of Onset Dates

Heavy and persistent rainfall is initiated by the arrival of the ITCZ over the coastal region. With respect to summer rains onset dates, the interstation correlations are all positive and with the exception of a few, are statistically significant (Table 5). Moderately low interstation correlations are obtained for the winter rain onset dates and less than half of them are statistically significant. These results support earlier evidence that the rain-producing mechanisms appear to be more spatially uniform in the summer season than in the winter season. Although a few stations exhibit high interstation correlations, the results suggest that the onset dates are not synchronous for the coastal region. Correlation coefficients remain well above zero up to a distance of 220 km for the summer rains while in the winter season they decline to zero at a distance of 140 km. and begin to increase from this point on (Figure 4). The latter case is attributed to positive interstation correlations which exist between

TABLE 5a : Interstation Correlations of Winter Rains Onset

	1	2	3	4	5	6	7	8
1. Pickersgill	1.00							
2. Anna Regina	.35	1.00						
3. Onderneeming	.38	.60*	1.00					
4. Leguan	.65*	.61*	.61*	1.00				
5. Lenora	.29	.49*	.68*	.73*	1.00			
6. Mahaicony-Abary	.06	.21	.46*	.39	.75	1.00		
7. Albion	-.23	.02	.32	-.03	.27	.43*	1.00	
8. Skeldon	.35	.80*	.60*	.61*	.49	.20	.02	1.00

TABLE 5b : Interstation Correlations of Summer Rains Onset

	1	2	3	4	5	6	7	8
1. Pickersgill	1.00							
2. Anna Regina	.83*	1.00						
3. Onderneeming	.74*	.76*	1.00					
4. Leguan	.75*	.69*	.68*	1.00				
5. Lenora	.66*	.60*	.48*	.73*	1.00			
6. Mahaicony-Abary	.57*	.64*	.58*	.57*	.06	1.00		
7. Albion	.65*	.42*	.31	.26	.47*	.03	1.00	
8. Skeldon	.80*	.56*	.54*	.85*	.65*	.35	.43*	1.00

* Coefficients are statistically significant at 0.05 level using the Student's t test.

stations located in the east and west and suggests a secondary effect might be responsible for initiating the heavy rains over either of these areas.

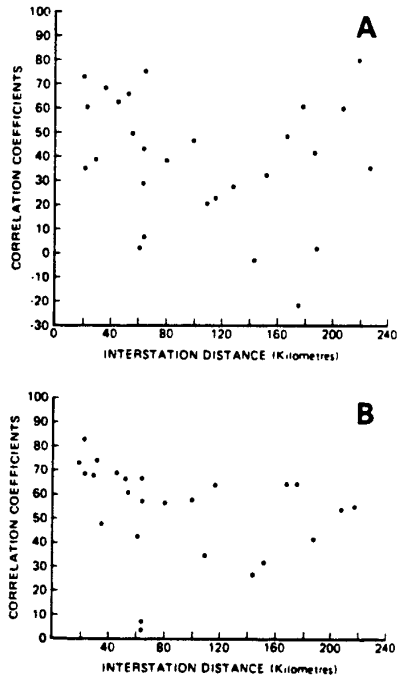


FIGURE 4. Plot of interstation correlation coefficients against distance for the onset dates of winter (A) and summer (B) rainfall.

CONCLUSIONS

Rainfall on the Guyana coastal region is produced by mesoscale synoptic systems occurring throughout the year and by large-scale systems associated with the seasonal movement of the ITCZ. Consequently, the rainfall is highly variable in time and space. The spatial variability has been analyzed using correlation analysis to determine the degree to which rainfall can be considered localized and whether the onset of the heavy rains occur simultaneously at all stations throughout the coast. The results suggest the tendency for rainfall to be more spatially coherent in the summer months as indicated by higher interstation correlations and lower rate of decay with distance. Rainfall is primarily produced by large-scale systems and onset dates of heavy persistent rainfall vary only moderately from one station to the next.

Spatial and temporal variability are more pronounced in the winter season. Interstation correlations are low and in some cases negative and decay at a greater rate with distance. Onset dates of heavy persistent precipitation show high interstation variations. Rainfall therefore tends to be more localized and produced by fewer large-scale weather systems.

The ITCZ which is primarily responsible for most of the rainfall is not spatially coherent in terms of the magnitude of rainfall. Rainfall totals and onset dates were good indicators of this, especially during the winter rainfall season, while uniformity of pattern was better reflected in the annual totals of rainy days.

Spatial variability is partly a function of topographic influence in adjacent regions, the latitudinal location of stations and their orientation with respect to the seasonal changes in the magnitude and direction of prevailing winds.

REFERENCES

Barrett, E. C., 1974, Climatology from Satellites, Methuen & Co., London.

Chang, C. P., 1970, Westward propagating cloud patterns in the Tropical Pacific as seen from time-composite satellite photographs, Journal of Atmospheric Sciences, 27, 132.

Gruber, A., 1972, Fluctuations in the position of the ITCZ in the Atlantic and Pacific Oceans, Journal of Atmospheric Sciences, 9, 194.

Holton, J. R., Wallace, J. M. and Young, J. A., 1970, On Boundary Layer Dynamics and the ITCZ, Journal of Atmospheric Sciences, 28, 275.

Jackson, I. J., 1977, Climate and Weather in the Tropics, Academic Press.

Longley, R. W., 1974, Spatial Variations of Precipitation over the Canadian Prairies, Monthly Weather Review, 102, 307-312.

Riehl, H., 1977, Climate and Weather in the Tropics, Academic Press, London, 16.

Sharon, D., 1972, The Spottiness of Rainfall in a Desert Area, Journal of Hydrology, 17, 161-175.

Snijders, T. A.., 1986, Interstation Correlations and Nonstationarity of Burkina Faso Rainfall, Journal of Applied Meteorology, 25, 524-531.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

REGRESSION PREDICTION OF MONTHLY PRECIPITATION IN SUMMER

Hongxing Cao

Academy of Meteorological Science, State Meteorological Administration, 46 Baishiqiaolu, Beijing, China

Shanxian Yu

Department of Computer, Zhejiang Meteorological Administration, Genshanmen, Hangzhou, Zhejiang Province, China

1. INTRODUCTION

A monthly and seasonal prediction of precipitation in summer is very useful for water resource and agriculture. The regression is an essential method to predict the precipitation quantitively. In order to improve quality of the precipitation prediction, in this paper some types of regression model, such as stepwise regression, optimal subsets regression, robust regression and ridge regression are employed to build the statistical models for monthly precipitation. The feature of different regression models are studied numerically. The comparative study shows that the ridge regression has higher prediction skill.

2. SCREENING OF PREDICTOR

To obtain some dozen of efficient predictors for the monthly prediction from a number of potential predictors, the preliminary screening procedure is carried out by calculating correlation coefficients between the monthly precipitation at a location and the monthly 500 hPa height fields over the Northern Hemisphere. The areas which cover the grid points with great value of correlation coefficient whose confidence level is significant are considered as key areas, the mean of height at the grid points over every key area is taken as remote predictors for the monthly precipitation prediction at a location. Taking account of some experience of local weatherman and weather proverbs, the meteorological elements at the same location, whose correlation is significant

with the monthly precipitation, are selected as local predictors.

For the precipitation in August at Nanjing, China, denoted by Y, a preliminary screening is carried out. As a result, 12 predictors are taken for building the regression equation in which x_1, x_2, ..., x_8 are the mean height of 500 hPa level over key area respectively. x_9 is the mean temperature during the period lasting from December 22 through 30 called Toujiu in Chinese calender, x_{10} the date of first frost, x_{11} the mean temperature during the third decade of April, x_{12} total precipitation during the period lasting from December 7 through 22. The data of the pedictand Y and 12 predictors with 1952-1984 records are taken to study feature of 4 types of regression and their prediction skill.

3. ANALYSIS OF RIDGE TRACE IN REGRESSION MODEL

The least square estimation of regression coefficient is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where $\hat{\beta}$ is an estimation of the coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, X the matrix composed of P predictor with n individual, T shows transpose of matrix X, $(X^T X)^{-1}$ inverse of covariance matrix $S = X^T X$, Y is the predictand vector composed of n individual. It is proved that when S becomes "ill", a collinear relation should exist among predictors. Therefore, least square estimation of β is not accuratd and instable. As an alternative, the ridge estimation

$$\hat{\beta}(k) = (\hat{\beta}_1(k), \hat{\beta}_2(k), \ldots, \hat{\beta}_p(k))^T = (S+kI)^{-1} X^T Y \qquad k > 0$$

is suggested (Horel and Kennard, 1970). Let

$$U = \begin{bmatrix} Y^T Y & Y^T X \\ X^T Y & X^T X \end{bmatrix}$$

and

$$A U A^T = \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_{p+1}) = \Lambda$$

where A is an orthogonal square matrix. Then the $\hat{\beta}_j(k)$ can be calculated (Chatterjee S. and B. Price, 1977) by

$$\hat{\beta}_{j-1}(K) = -\left[\sum_{i=1}^{p+1} \frac{a_{i1}a_{ij}}{\lambda_i + K}\right] \Big/ \left[\sum_{i=1}^{p+1} \frac{a_{i1}^2}{\lambda_i + K}\right] \qquad j=2,3,\ldots,p+1$$

Taking k=0, 0.1, 0.2, ..., 1.0, the $\hat{\beta}(k)$ are calculated respectively for 12 predictors. Consequently, a diagram of the ridge trace is drawn. In Figure 1, the ridge traces of x_1, x_4, x_8 and x_{11} are shown exemplarily. Although the coefficient of x_8 is large when k=0, it becomes small when k> 0.1. According to the view of ridge estimation, the x_8 should be rejected, although x_8 is selected in the regression equation with 5 predictors, which is built by an optimal subsets regression. Having rejected the predictor x_8, the prediction skill of ridge regression is improved.

Further, an examination for the diagram of ridge trace of all predictors are made. The screening criteria for the predictors are that

i. the predictor has stable coefficients values with increase of k and the absolute value of $\hat{\beta}(k)$ is not small,

ii. the accuracy of predictors for the independent sample composed of 1982-1984 year's data is higher.

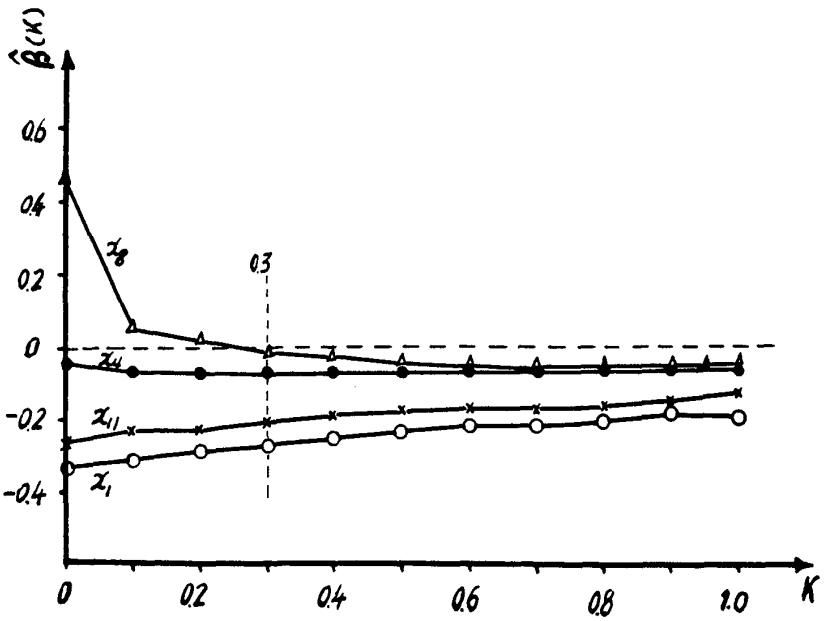


Fig.1 The ridge trace of predictors x_1, x_4, x_8, x_{11}.

4. FITTING AND PREDICTION

Taking k=0.3, as the accuracy of the 1982-1984 year's prediction is higher, consequently, the ridge regression is

$$\hat{Y}_k = 13.1906 - 0.146286x_1 - 0.0585130x_2 - 0.00396506x_4 + 0.0647038x_7 - 0.121590x_{11} .$$

The equation of stepwise regression, which is equivalent to k=0 in the ridge regression, is

$$\hat{Y}_s = 14.5863 - 0.190084x_1 - 0.0759031x_2 - 0.00425660x_4 + 0.0740485x_7 - 0.113643x_{11} .$$

The optimal subsets regression and the robust regression are also built based on the screening criterion ii. in Section 3. The optimal subsets regression equations is

$$\hat{Y}_o = 13.6453 - 0.199838x_1 - 0.0769246x_2 + 0.0656489x_7 - 0.147237x_{11}$$

and the robust regression equation is

$$\hat{Y}_r = 14.4522 - 0.187093x_1 - 0.0916675x_2 - 0.00493794x_4 + 0.0824509x_7 - 0.0734975x_{11} .$$

The predictions for the independent sample composed of the 1982-1984 year's data using above 4 types of regression equation are given in Table 1. It is shown that the mean absolute error of ridge regression with k=0.3 is smallest. This implies when we carefully select predictor whose regression coefficient is quite large and stable with the

increase of ridge parameter k, the prediction skill of ridge regression can be improved. The prediction skill of the different type of regression is certainly different. For special problem,it is required to select an appropriate regression type to get higher prediction accuracy.

Tab.1 Verification of the independent sample for 4 regressions

estimation / year	Y (observations)	$\hat{Y}_{k=0.3}$	$\hat{Y}_s$	$\hat{Y}_o$	$\hat{Y}_r$
1982	68.0	68.4	62.5	62.4	58.7
1983	99.7	125.7	136.1	128.1	135.1
1984	83.3	155.8	180.6	206.4	181.4
$\bar{E}$		32.7	46.4	52.4	47.6

5. REFERENCE

Chatterjee, S. and B. Price, 1977 : Regression Analysis, John Wiley & Sons, New York, 181-192.

Horel, A.E. and R.W. Kennard, 1970 : Ridge regression : Biased estimation for nonorthogonal Problems, Technometrics, 12, 69-82.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

A NEW METHOD FOR INTERPOLATION OF CLIMATOLOGICAL DATA

Zhou Jiabin

Institute of Atmospheric Physics

Academia Sinica

Beijing, China

Some climatological data can not be obtained because of various kinds of difficulty and this will have influence upon the research and application of climatology. In this paper a new method for interpolation of climatological data is proposed.

1. INTERPOLATION FORMULAE FOR ONE-DIMENSSIONAL DATA

Let us discuss the interpolation for one-dimenssional data first.

Set on x axis Io grids (x=1,2,... Io) and assign Z(1), Z(2),...Z(Io) to values of a meteorological element at these grids respectively in which Z(io) is unknown (1≤io≤Io). A scheme for interpolating the Z(io) is shown as follows.

The function Z(i) at Io grids are expanded in term of Chebyshev polynomials, that is (Bagrov, 1958),

$$\tilde{Z}(i)=\sum_{k=0}^{K_o} A_k \varphi_k(i) \qquad (1)$$

and

$$\hat{Z}(i_o)=\sum_{k=0}^{K_o} A_k \varphi_k(i_o) \qquad (2)$$

In above expressions Z(i) is the fitting value of Z(i) at grid i (i=1,2,...,io,...Io), $\varphi_k(i)$, the value of normalized Chebyshev polynomial with order k at grid i, Ko, the truncation order in Exps.(1) and (2), Ak, Chebyshev coeffient which can be obtained from the following(Bagrov, 1958),

$$A_k=\sum_{i=1}^{I_o} Z(i)\,\varphi_k(i) \qquad (3)$$

This expression can be rewritten as that

$$A_k=\sum_{\substack{i=1\\ i\neq i_o}}^{I_o} Z(i)\,\varphi_k(i)+Z(i_o)\varphi_k(i_o) \qquad (4)$$

which indicutes that Ak deponds two terms, the first of which is known and the second is unkown because it relates to the value of Z(io).

Letting the initial value of Z(io) be $Z^{(0)}(io)$ and replacing Z(io) in Exp.(4) with it, we have the first approximation of Ak, that

$$A_K^{(1)}=\sum_{\substack{i=1\\ i\neq i_o}}^{I_o} Z(i)\,\varphi_k(i)+Z^{(0)}(i_o)\varphi_k(i_o) \qquad (5)$$

Substituting $A_k^{(1)}$ in Exp.(2) we obtain the first approximation of Z(io),

$$Z^{(1)}(i_o)=\sum_{K=0}^{K_o} A_K^{(1)}\varphi_K(i_o) \qquad (6)$$

Likewise we have the second approximation of Ak by replacing Z(io) in Exp.(4) with $Z^{(1)}(io)$,

$$A_K^{(2)}=\sum_{\substack{i=1\\ i\neq i_o}}^{I_o} Z(i)\varphi_K(i)+Z^{(1)}(i_o)\varphi_K(i_o) \qquad (7)$$

Then the second approximation of Z(io) can be found from Exp.(2),

$$Z^{(2)}(i_o)=\sum_{K=0}^{K_o} A_K^{(2)}\varphi_K(i_o) \qquad (8)$$

Thus we can have the general forms,

$$A_K^{(\nu)}=\sum_{\substack{i=1\\ i\neq i_o}}^{I_o} Z(i)\varphi_K(i)+Z^{(\nu-1)}(i_o)\varphi_K(i_o) \qquad (9)$$

$$Z^{(\nu)}(i_o)=\sum_{K=0}^{K_o} A_K^{(\nu)}\varphi_K(i_o) \qquad (10)$$

It is easy to see that Exps.(9) and (10) are just the predicted formulae for Z(Io) when the value of Z(1), Z(2),..., Z(Io-1) are known. Therefore, they are the generalization of prediction formulae for time series proposed by the auther (Zhou, 1985a) previously.

2. ERROR ESITMATES

As shown in Zhou (1985a), to make the treatment easier some equivalents and definitions are given,

$$Z = Z(i_o), \tag{11}$$

$$\hat{Z} = \hat{Z}(i_o), \tag{12}$$

$$Z^{(\nu)} = Z^{(\nu)}(i_o) \quad (\nu = 0, 1, 2, \cdots), \tag{13}$$

$$\varepsilon^{(\nu)} = Z^{(\nu)} - Z, \tag{14}$$

$$\tilde{\varepsilon} = \hat{Z} - Z, \tag{15}$$

$$\sigma = \sum_{k=0}^{K_0} \phi_k^2(i_o). \tag{16}$$

From Exps.(14),(6),(2),(5),(3), (15) and (16), we get,

$$\varepsilon^{(1)} = \sigma \varepsilon^{(0)} + \tilde{\varepsilon}. \tag{17}$$

This expression is the same with Exp. (17) in Zhou (1985a) formally and we can have the similar expression as follows,

$$\varepsilon^{(\nu)} = \varepsilon_1^{(\nu)} + \varepsilon_2^{(\nu)} = \sigma^{\nu} \varepsilon^{(0)} + \sum_{\tau=0}^{\nu-1} \sigma^{\tau} \tilde{\varepsilon}. \tag{18}$$

It can be proved that (Zhou 1985b), $0 < \sigma < 1$.

Thus we have

$$\lim_{\nu \to \infty} \varepsilon_1^{(\nu)} = 0, \tag{19}$$

$$\lim_{\nu \to \infty} \varepsilon_2^{(\nu)} = \frac{\tilde{\varepsilon}}{1-\sigma}. \tag{20}$$

$$\lim_{\nu \to \infty} \varepsilon^{(\nu)} = \frac{\tilde{\varepsilon}}{1-\sigma}. \tag{21}$$

Since Exps.(18)-(21) are the same with the corresponding expressions in Zhou(1985a, 1986) the relationship between the error of interpolation and the initial value may be discussed in a similar way.

Here two kinds of initial value are most important, that is

(1) Ideal initial value (IIV)

It satisfies the following conditions,

A. The error of initial value $\varepsilon^{(0)}$ and the fitting error $\tilde{\varepsilon}$ have the opposite sign.

B. $|\varepsilon^{(0)}| > \frac{|\tilde{\varepsilon}|}{\sigma}$.

In this case, $\varepsilon_1^{(\nu)}$ is opposite to $\varepsilon_2^{(\nu)}$ in sign and the two errors offset each other in part. $\varepsilon^{(\nu)} \approx 0$ and a smaller error may be got after some times of iteration, ν_c, called critical time of iteration.

(2) Non-ideal initial value (NIIV)

It satisfies,

A. $\varepsilon^{(0)}$ and $\tilde{\varepsilon}$ have the same sign.

B. $|\varepsilon^{(0)}| > \frac{|\tilde{\varepsilon}|}{1-\sigma}$.

In this case, $\varepsilon_1^{(\nu)}$ is the same as $\varepsilon_2^{(\nu)}$ in sign and the two errors are found to superimpose each other and we get the largest interpolation error in all kinds of initial value.

A similar discussion may be done for the relationship between $\varepsilon^{(\nu)}$, $\varepsilon^{(\nu)}$, $\varepsilon^{(0)}$ and Io(See Zhou 1985a,b, 1986).

3. THE INTERPOLATION BASED ON THE FINAL VALUE OF ITERATION

From (14) and (21) we have,

$$Z^{(\infty)}(i_o) = Z(i_o) + \frac{\tilde{\varepsilon}}{1-\sigma}. \tag{22}$$

We can give the following interpolation formulae based on the final value of iteration,

$$Z_\infty(i_o) = Z^{(\infty)}(i_o) - \frac{\bar{\tilde{\varepsilon}}_+}{1-\sigma} \quad (\text{for } \tilde{\varepsilon} > 0), \tag{23}$$

$$Z_\infty(i_o) = Z^{(\infty)}(i_o) - \frac{\bar{\tilde{\varepsilon}}_-}{1-\sigma} \quad (\text{for } \tilde{\varepsilon} < 0), \tag{24}$$

where Z_∞(io) is the interpolation value based on the final value of iteration, $\bar{\tilde{\varepsilon}}_+$ and $\bar{\tilde{\varepsilon}}_-$ are the mean of fitting error $\tilde{\varepsilon}$ for $\tilde{\varepsilon} > 0$ and $\tilde{\varepsilon} < 0$ respectively.

4. OPERATIONAL PROCEDURES

The procedures are similar with that shown by Zhou(1986) and are indicated briefly as follows.

4.1 Treatments of Dependent Data

(1) Date pretreatment.

(2) Determination of the grid number Io.

(3) Fitting and seeking the condition to judge the sign of fitting error $\hat{\varepsilon}$.

A. Fitting and evaluating $\tilde{\varepsilon}, \tilde{\hat{\varepsilon}}, \tilde{\bar{\varepsilon}}$.

B. Evaluating the final value of iteration $Z^{(\infty)}(io)$.

C. A figure of distribution for final value of iteration is drawn and a condition judging the sign of $\tilde{\varepsilon}$ may be found.

(4) Evaluating the interpolation value by using the final value of iteration.

(5) Determing the interpolation by using IIV.

In each case the critical iteration time ν_c may be determined and the frequent $\nu_c(\nu_c^*)$ be obtained. Thus $Z_i^{(\nu_c^*)}(io)$ is got. For NIIV the interpolation value $Z_{ni}^{(\nu_c^*)}(io)$ can also be obtained after iterating ν_c^* times. For a definite initial value a figure of distribution of $Z_i^{(\nu_c^*)}(io)$ and $Z_{ni}^{(\nu_c^*)}(io)$ can be drawn.

(6) Intergrating different results of interpolation.

4.2 Application to the Independent Data

(1) Iterating and determing the final value. Using the conjuctive condition obtained from dependent data, the sign of $\tilde{\varepsilon}$ is determined.

(2) Evaluating the interpolation value based on the final value of iteration.

(3) Determing the interpolation by IIV. The results of iteration by a larger positive initial value and a negtive initial value with a larger absolute value are selected to determined which of the two results can be taken as the interpolation value by IIV using the distribution of $Z_i^{(\nu_c^*)}(io)$ and $Z_{ni}^{(\nu_c^*)}(io)$ in the dependent data.

(4) Intergrating the different results of interpolation.

5. THE INTERPOLATION OF DATA IN TWO-DIMENSSIONAL RECTANGULAR GRIDS

In the case of two-dimenssinal rectangular grids we have the following formulae.

$$Z^{(\nu)}(i_o, j_o) = \sum_{K=0}^{K_o} \sum_{s=0}^{S_o} A_{Ks}^{(\nu)} \phi_K(i_o) \psi_s(j_o), \quad (25)$$

$$A_{Ks}^{(\nu)} = \sum_{j=1}^{J_o} \sum_{i=1}^{I_o} Z(i,j) \phi_K(i) \psi_s(j) + Z^{(\nu-1)}(i_o, j_o) \phi_K(i_o) \psi_s(j_o). \quad (26)$$

$(i,j) \neq (i_o, j_o)$

Here (i, j) represent the grid and (i_o, j_o) is the grid at which the value of function Z is unknown.

As treated in the previous sections, the following equivalents and definitions are given.

$$Z = Z(i_o, j_o), \quad (27)$$

$$\tilde{Z} = \hat{Z}(i_o, j_o) = \sum_{K=0}^{K_o} \sum_{s=0}^{S_o} A_{Ks} \phi_K(i_o) \psi_s(j_o), \quad (28)$$

$$Z^{(\nu)} = Z^{(\nu)}(i_o, j_o), \quad (29)$$

$$\hat{\varepsilon} = \hat{Z} - Z, \quad (30)$$

$$\varepsilon^{(\nu)} = Z^{(\nu)} - Z, \quad (31)$$

$$\sigma = \sum_{K=0}^{K_o} \sum_{s=0}^{S_o} \phi_K^2(i_o) \psi_s^2(j_o). \quad (32)$$

Based on the Exps, (30), (25), (29), (26), (31), (32) and

$$A_{Ks} = \sum_{j=1}^{J_o} \sum_{i=1}^{I_o} Z(i,j) \phi_K(i) \psi_s(j), \quad (33)$$

we obtain,

$$\varepsilon^{(1)} = \sigma \varepsilon^{(0)} + \hat{\varepsilon}. \quad (34)$$

This expression in the same formally with Exp.(17) in section 2 and the expressions similar with Exps.(18)-(21) may be got. The relationship between the initial value and errors of interpolation could be discussed in a similar way. Thus the operational procedures are also similar with that indicated above.

1) See Bagrov (1958)

6. THE INTERPOLATION OF DATA IN TWO-DIMENSSIONAL NON-RECTANGULAL GRIDS

In this case the distribution of grid can be represented as follows,

$j=1,2,\ldots, Jo;$

$i=I_1(j), I_1(j)+1, I_1(j)+2, \ldots I_2(j)$

The corresponding interpolation formulae are

$$Z^{(\nu)}(i_o,j_o)=\sum_{k=0}^{K_o}\sum_{s=0}^{S_o} A_{ks}^{(\nu)}\varphi_{k,j}(i_o)\psi_s(j_o), \quad (35)$$

$$A_{ks}^{(\nu)}=\sum_{j=1}^{J_o}\sum_{i=I_1(j)}^{I_2(j)} Z(i,j)\varphi_{k,j}(i)\psi_s(j) + Z^{(\nu-1)}(i_o,j_o)\varphi_{k,j}(i_o)\psi_s(j_o). \quad (36)$$

Here Chebyshev polynomials $\varphi_{kj}(i)$ have two subscripts, k and j.k denotes the order of polynomial and j is the number of row. $\psi_s(j)$ indicates the value of Chebyshev polynomial with order s at grid j. The above expressions are convient for use of the data where appropoiate and can be derived based on the Chebyshev expansion at non-rectangular grids (Zhou and Li, 1985). Similar treatments could be done for the error of interpolation.

7. THE INTERPOLATION OF DATA AT IRREGULAR GRIDS

Chebyshev expansion is generalized into irregular grids by Zhou (1982) and Zhou and Li (1985). Therefore, Exps.(9),(10),(25),(26), (35) and (36) can be used at irregular grids if the variables i and j in them are explained as ordinals.

In consider of the dependence of interpolation exactness on the grid length, it is reasonable the variation of distances between grids are smaller.

8.CALCULATED CASES

We have treated the following data.

(1) Beijing monthly rainfall and total precipitation from January to Septmber over the Xin'anjiang Valley.

(2) Distribution of monthly mean rainfall in June over the middle and lower reaches of the Changjiang River and of mean temperature in June over the Northeastern China. Two kinds of grid, rectangular and irregular, are used.

As an example Fig.1 shows the results for monthly rainfall in April in Beijing. Comparison between interpolation (solid line) and observation (dashed line) indicates that the scheme proposed in this paper is avaliable.

9. SOME REMARKS

The scheme shown here can also be used in the vertification of data. In this satuation the interpolation value may be considered as a standard in vertification of data.

10. REFERENCES

Bagrov, N.A., 1958, Analiticheskoe predstavlenie polei, Trugy TsIP, 64: 3-25.

Zhou Jiabin, 1982, On the Expansion of Chebyshev Polynomials in Irregular Grids, Kexue Tongbao, 27:516-518.

——,1985a, A New Method of Time Series, Scientia Atmospherica Sinica, 9:27-35.

——,1985b, A New Type of Time-Series-Forecasting Method, Advances in Atmospheric Sciences, 2:385-401.

——,1986, Some Problems on the Time-Series-Forecast by Chebyshev Polynomials, Kexue Tongbao, 31:101-105.

—— and Li Huang, 1985, Futher Studies on the Expansion of Chebyshev Polynomials in Irregular Grids, Scientia Sinica, B, 28:199-201.

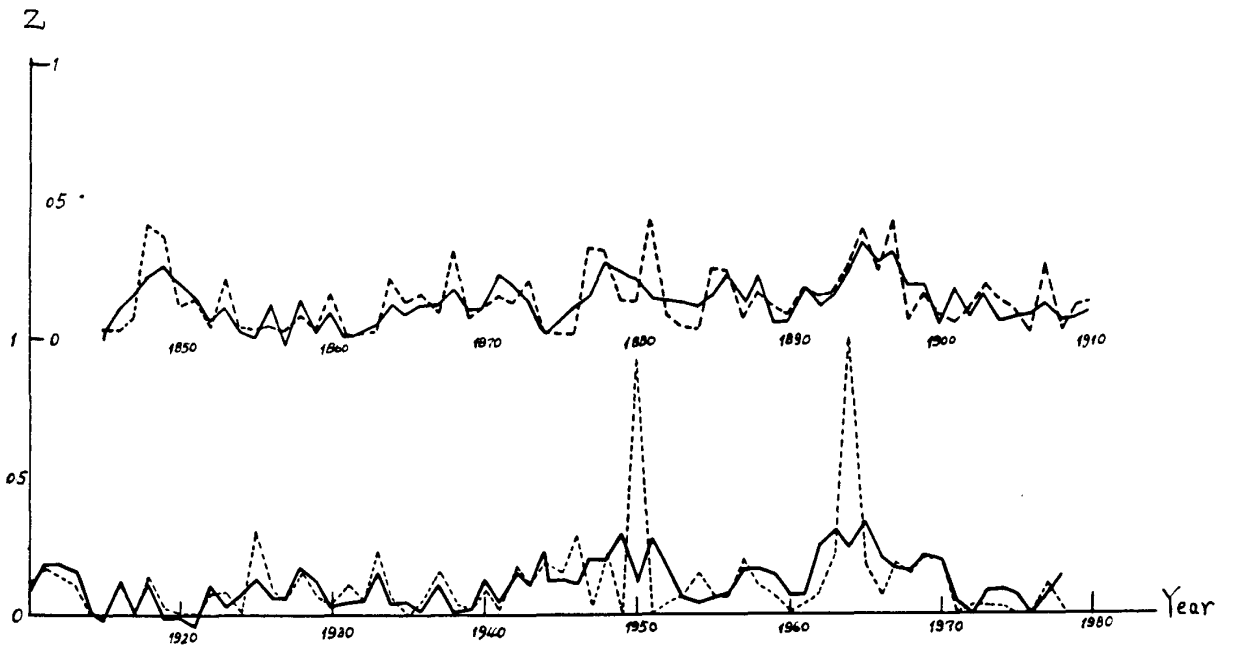


Fig.1 Mothly Rainfall in April in Beijing. Figures in ordinate denote the rainfall which has been pretreated. Abscissa represent the year.

Third International Conference on Statistical Climatology
June 23-27, 1986, Veinna, Austria

AN INDIRECT MODEL FOR PROBABILITY DISTRIBUTION OF PRECIPITATION

Ding Yuguo

Department of Meteorology, Nanjing Institute of
Meteorology, People's Republic of China

1. INTRODUCTION

The theoretic distribution of monthly (or yearly) precipitation has been verified by many researchers (Ison, N.T. et al, 1971, Mooley, D.A., 1973, and Ozturk, A., 1981). It has been shown that the monthly rainfall follows the Gamma distribution but the yearly rainfall only partly follows the normal distribution. However, the yearly rainfall shows some skewness at quite a number of stations. Tests by the author also show that the monthly rainfall follows the Gamma distribution in vast areas of China.

It is well-known that the only way to improve the precision of the parameters of the estimation model is to increase the sample information. Thus, a perfect estimation of the rainfall probability distribution in a place depends on the utmost use of the information from a large-sample station (long-term records). China is a country with vast territory and various climates. The theorectic distribution model for a meteorological element varies with the season and climatic region, but on the other hand the local difference in climate between neighbouring stations within one climatic region is generally stable, i.e., there exists some similarity. In China many stations have weather records of less than 50 years. It is, however, fortunate that we have some stations with records of 50–100 years and these stations are surrounded by those with short-term records. Having taken the above-mentioned fact into account, we present in this paper a theoretic model which, based on the indirect estimation of the distribution of rainfall probability for a small-sample station using the parameters from a neighbouring station having a large-sample record. In this sense, it is a correction model for the Gamma distribution between neighbouring stations.

2. AN INDIRECT MODEL FOR THE GAMMA DISTRIBUTION BETWEEN NEIGHBOURING STATIONS

It is generally believed that the ratio of the yearly or monthly rainfall between two immediately neighbouring stations is quite steady, in other words, there exists an approximate formula

$$y = bx \tag{1}$$

In general, the approximate relationship of rainfall between adjacent stations

is given by a linear regression equation

$$y = a + bx \tag{2}$$

where y denotes the rainfall at station Y and x the rainfall at station X, and a, b are the regression coefficients respectively.

Let the rainfall (monthly or yearly) at station X follows the Gamma distribution, and its density function be expressed as

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{3}$$

From the density formula for the linear function of a random variable, an indirect probability distribution model for station Y can be obtained from (2)

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{(y-a)^{\alpha-1}}{b} e^{-\frac{\beta(y-a)}{b}} \frac{1}{|b|} \tag{4}$$

If $b>0$, then

$$f(y) = \frac{\beta'^{\alpha}}{\Gamma(\alpha)} (y-a)^{\alpha-1} e^{-\beta'(y-a)} \tag{5}$$

where $\beta' = \beta/b$ is the scale parameter, which is 1/b of the scale parameter for station X, and α denotes the shape parameter and a is the location parameter.

From the above mentioned, when the rainfall between two adjacent stations follows a linear regression equation, the rainfall probability distribution for station Y is characterized by the parameter for station X (the Gammadistribution). In other words, when there exists (1) or (2), the Gamma distribution model for station X can be transposed to station Y by (5). In reality, this is an indirect estimation.

It seems possible theoretically that two neighbouring stations are under the same circulation system. As a result, the rainfall records are usually synchronous so that they have much the same shape parameter α. Nevertheless, their scale parameter β may differ systematically because of the difference in topograph or geographic location. This is easily seen from the following formula

$$\sigma^2 = \frac{\alpha}{\beta^2} \tag{6}$$

As the variances for different stations are not the same, the scale parameters β may differ greatly although there is a tendency for the shapes of the probability distribution to be similar.

In a rigorous sense, Eqs.(1) and (2) are used as an approximate function in the above indirect model without involving the linear regression theory itself. Flueck and Mielke (1975), having derived the two-dimensional Gamma distribution

model, pointed out that the condition distribution is in fact the Gamma distribution and the regression function corresponding to the condition distribution is the linear regression. However, the correlation coefficient is only $\rho > 0$. This agrees with the fact given in this paper, that is, the the same climatic elements at two neighbouring stations have positive correlation in most cases.

3. EFFICIENCY AND PRECISION OF THE MODEL

The indirect model for precipitation between two adjacent stations aims at i) higher efficiency and precision for the theoretic distribution parameter obtained by the indirect model than by the direct model, and ii) simplification of the calculating process.

From the estimation theory (moment estimation, for example), the efficiency of an estimator is given as follows

$$E = \frac{\mathrm{Var}\ \hat{\theta}}{\mathrm{Var}\ \hat{\theta}_m} \tag{7}$$

where E is the efficiency statistics, θ the efficiency estimator, and θ_m the amount to be estimated. In the same way, the moment estimator for the parameters α, β in the indirect model has the following formula

$$\mathrm{Var}\ (\hat{\alpha}) = \frac{2\alpha}{n}(\alpha+1)$$
$$\mathrm{Var}\ (\hat{\beta}) = \frac{\beta^2}{n\alpha}(2\alpha+3) \tag{8}$$

Suppose $\hat{\alpha}_{yN}$ and $\hat{\beta}_{yN}$ are the parameters estimated from the indirect model and $\hat{\alpha}_{yn}$ and $\hat{\beta}_{yn}$ are the parameters estimated from the direct model. Thus, in order to make the indirect model better than the direct model, it is necessary to have the following relationship

$$\mathrm{Var}\,\hat{\alpha}_{yN} < \mathrm{Var}\,\hat{\alpha}_{yn}$$
$$\mathrm{Var}\,\hat{\beta}_{yN} < \mathrm{Var}\,\hat{\beta}_{yn} \tag{9}$$

where the subscript N represents the parameter estimated for the large-sample station in N years and the subscript n the parameter estimated for the small-sample station in n years. Therefore, $\mathrm{Var}\,\hat{\alpha}_{yN}$ and $\mathrm{Var}\,\hat{\beta}_{yN}$ denote respectively the variances for the shape and scale parameter estimation based on the large-sample records from station X, and $\mathrm{Var}\,\hat{\alpha}_{yn}$ and $\mathrm{Var}\,\hat{\beta}_{yn}$ the variances for the shape and scale parameter estimations based on the small-sample records from station Y.

From (4) or (5), we have

$$\alpha_y = \alpha_x \tag{10}$$

$$\beta_y = \beta_x / b \tag{11}$$

Combining (9) and (8), we may write

$$\frac{2\alpha_x}{N}(\alpha_x + 1) < \frac{2\alpha_x}{n}(\alpha_x + 1) \tag{12}$$

$$\frac{\beta_x^2}{N\alpha_x b^2}(2\alpha_x + 3) < \frac{\beta_y^2}{n\alpha_x}(2\alpha_x + 3) \tag{13}$$

Obviously, in (12) the inequality depends on the size of the sample N and n, i.e.,

$$n/N < 1 \tag{14}$$

The above formula is always established so far as $N > n$. For this reason, N should be greater than n. On the other hand, we know

$$b = \rho \frac{\sigma_y}{\sigma_x} \tag{15}$$

where ρ is the population ccorelation coefficient for precipitation between station X and station Y, and σ_x and σ_y are the variances corresponding to X and Y. From (6) and (13), we have

$$\frac{n}{N\rho^2} < 1 \tag{16}$$

Obviously, the inequality depends on the ratio of N to n and the size of the coorelation coefficient ρ. It is seen from (16) that the indirect model is always efficient so long as n/N is small enough and ρ^2 large enough.

For evaluating the precision of the model, the following index is suggested

$$E = 1 - \frac{\mathrm{Var}\beta_{yN}}{\mathrm{Var}\beta_{yN}} \tag{17}$$

where E denotes the efficiency of the model. Evidently, if the variance of β_{yN} is not improved, then $E \longrightarrow 0$, i.e., the model is not efficient. On the contrary, when $\mathrm{Var}\beta_{yN}$ is much smaller than $\mathrm{Var}\beta_{yn}$, then $E \longrightarrow 1$, i.e., the indirect model is more efficient. Generally, we have

$$0 < E < 1 \tag{18}$$

And from (16) and (17), we have

$$E = 1 - \frac{n}{N\rho^2} \qquad (19)$$

Tables 1-3 (omitted) show the fewest numbers of N under different n, and E. And the coorelation coefficient critical values under different E and N are given in Table 4.

Table 4. Correlation coefficient critical values when N = 100.

n \ E	0.5	0.6	0.7	0.8	0.9
10	0.45	0.50	0.58	0.71	1.00
20	0.63	0.71	0.82	1.00	
30	0.75	0.87	1.00		
40	0.89	1.00			

It is seen from Table 4 that in order for the small-sample records in different years to have a Gamma distribution indirect model of the same large sample (for example, N = 100), it is necessary to make the correlation coefficient between station X and station Y reach a certain critical value so that the parameters for the model can be estimated.

4. CASE STUDIES

For the purpose of evaluating the fitting effect of the indirect model, two pairs of representative stations, i.e., Shanghai-Hanzhou and Beijing-Tianjin, are chosen as case studies. The theoretic distribution of January monthly rainfall for Hangzhou and Tianjin are fitted based on the large-sample records of Shanghai (N = 110) and of Beijing (N = 109), respectively.

First, the linear regression equations for Hangzhou-Shanghai and Tianjin-Beijing are calculated as follows

$$Y_H = 2.4731 + 1.3609\ X_S$$
$$Y_T = 0.4761 + 0.8929\ X_B$$

Here, the size of the sample is n = 32. From the indirect model (5) and the direct model (3), the values of some related parameters have been calculated as shown in Table 5.

For convenience, the moment estimator is used in all the cases. In fact, both Hangzhou and Tianjin have long-term records and we have taken only n = 32 for assessing the fitting effect. This suggests that the theoretic distribution for the small-sample station is fitted by means of the indirect model based on the large-sample station.

Table 5. Parameters of indirect model (I.M.) and direct model (D.M.)

Model	Hangzhou I.M.	Hangzhou D.M.	Tianjin I.M.	Tianjin D.M.
C_s	1.4431	1.6197	3.7216	3.4282
S	33.7270	44.0590	6.7000	4.8940
α	1.9207	1.5247	0.2888	0.3404
β	0.0302	0.0280	0.0989	0 1192
γ	0.8524		0.8825	

Tables 6 and 7 show (Table 7, omitted) the theoretic distribution of January monthly rainfall by using model (5) and model (3) for Hangzhou and Tianjin respectively.

Table 6. Comparison of fitting effect between I.M. and D.M. at Hangzhou.

Class Interval	Midpoint	np_i D.M.	np_i I.M.	P_i D.M.	P_i I.M.	V_i	$\frac{(V_i-np_i)^2}{np_i}$ D.M.	$\frac{(V_i-np_i)^2}{np_i}$ I.M.
0.0–30.0	15.0	26.6	18.5	0.3917	0.2729	16	4.224	0.338
30.0–60.0	45.0	20.4	21.9	0.3009	0.3221	16	0.949	1.589
60.0–90.0	75.0	11.4	13.9	0.1680	0.2044	19	5.067	1.871
90.0–120.0	105.0	5.9	7.6	0.0874	0.1119	8	0.747	0.021
120.0–150.0	135.0	2.9	3.9	0.0431	0.0572	7		
150.0–180.0	165.0	1.39	1.9	0.0205	0.0282	1	3.431	0.790
180.0–210.0	195.0	0.65	0.9	0.0096	0.0136	1		

According to the criterion for fitting goodness test, the statistics (both theoretic frequency and actual frequency) for Hangzhou are as follows

$$\text{direct model} \quad X^2 = \sum_{i=1}^{5} \frac{(V_i-np_i)^2}{np_i} = 14.418$$

$$\text{indirect model} \quad X^2 = \sum_{i=1}^{5} \frac{(V_i-np_i)^2}{np_i} = 4.609$$

where V_i denotes the actual frequency in the ith group and np_i the theoretic frequency in the ith group.

In the same way, we have the statistics for Tianjin as follows

direct model $X^2 = 5.463$

indirect model $X^2 = 2.781$

Given $\alpha = 0.05$, we have $X^2 = 5.99$ (the degree of freedom $f = 5-2-1$) for Hangzhou and $X^2 = 9.488$ ($f = 7-2-1$) for Tianjin, and thus both models are in agreement with the actual distribution, the indirect model always having a better effect than the direct model as shown above.

From (19), the efficiency of the indirect model for the above stations can be obtained

$E_H = 0.5996$

$E_T = 0.6627$

The results are in agreement with the above X^2 test, i.e., X_T^2 is better than X_H^2 and E_T is higher than E_H. All these are fit for the figures shown in Tables 1-4.

5. CONCLUSION

i) According to the approximate linear relationship for the rainfall between adjacent stations, the indirect Gamma distribution model can established based on the large-sample estimated parameters and can be used to estimate the precipitation probability distribution for the surrounding stations where information is insufficient. This idea may be extended to some other weather elements or distributions, which will be discussed in another article written by the author.

ii) Analyses and case studies show that this indirect model is more efficient and simplified for estimating the theoretic distribution than the direct model based on the small-sample records of its own. That is to say, it has the fitting effect equivalent to extending the sample records, thus improving the reliability of the probability and cumulative probability calculation. The model can also be used to unify the length of the sample for simulating the probability distribution for many other stations.

REFERENCES

1. Ison, N.T., Feyerherm, A.M., and Bark, L.D., Wet Period Precipitation and Gamma Distribution, J. Appl. Met., 10, 658-665, 1971.

2. Mooley, D.A., Gamma Distribution Probability Model for Asian Summer Monsoon Monthly Rainfall, Mon. Wea. Rev., 101, 2, 160-175, 1973.

3. Oztürk, A., J. Appl. Met., 20, 1499-1505, 1981.

4. Yao, C.S.,Climatic Statistics, Science Publ., Beijing, (in Chinese), 1963.

5. Yao, C.S., Fundamentals of Climatic Statistics, Science Publ., Beijing (in Chinese), 1984.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

VARIATION OF SPATIAL CORRELATION PATTERNS IN DENMARK

Peter Allerup and Henning Madsen
Meteorological Institute
Copenhagen Denmark

1. INTRODUCTION

The series of investigations of which the present one is the last was originally inspired by a WMO draft concerned with climatic atlasses and standard normals 1931-60. A comprehensive statistical model for recorded variations in monthly precipitation values was undertaken (Allerup, 1981) and this model is then applied to interpolations (Allerup and Madsen, 1983).

Very often in statistical analyses of precipitation data interpolated values are needed, e.g. in case of quality control of precipitaion series, completion of series input to hydrological models, calculation of areal precipitation in non- homogeneous regions etc. It then becomes urgent to know the spatial correlation structure in the region of interest.

In Denmark a complete series of monthly precipitation values from 148 stations for 1931-60 exist, and it is possible, then to interpolate new values of sites having incomplete data.

The paper demonstrates how the correlation structure is assesed, based on residuals from a log-additive model with spatially intercorrelated observations.

The annual amount of precipitation in Denmark is about 650 mm, varying from 800 mm in the interior of Jutland to 500 mm at the coasts. The wettest month is August, 80 mm and the driest March, 33 mm. Cyclonic activity is dominating in autumn and winter. In summer convective precipitation often occurs.

2. DATA

Fig. 1 shows the distribution of the 148 base stations with complete data. The selection of these stations have been carried out in such a way that those stations situated at cliffs or other places, where the measurements have been influenced by strong aerodynamic effects, have been disregarded together with stations, where displacements have occured.

For this investigation data from stations inside the two encircled areas (see Fig. 1) has been used, only assuming that a homogeneous precipitation regime is prevailing in each of the two areas. The one represents coastal conditions and the other inland conditions. A full analysis of all sub-regions in Denmark will be published elsewhere (Allerup and Madsen, 1987).

To study the influence on interpolations arising from the type of precipitation, monthly values for the summer: June, July and August and autumn: September, October and November representing the convective and cyclonic type of precipitation respectively, have been selected.

3. STATISTICAL ASSUMPTIONS

The mathematical background for the interpolation procedures is given in Allerup, 1981 in which monthly precipitation values have been analysed. The main conclusions from these analyses were that logarithmic monthly totals can be estimated by means of a site component, α_i and a temporal component, β_t (called the 'loading') with a residual error e_{it}, which is fairly well described by a normal distribution. Allerup, 1981, Allerup et al., 1982, Allerup, 1982 and Allerup and Madsen, 1983 discuss the problems connected with a proper statistical formulation of the correlation structure among the e_{it}'s which is at the same time estimable for practical use.

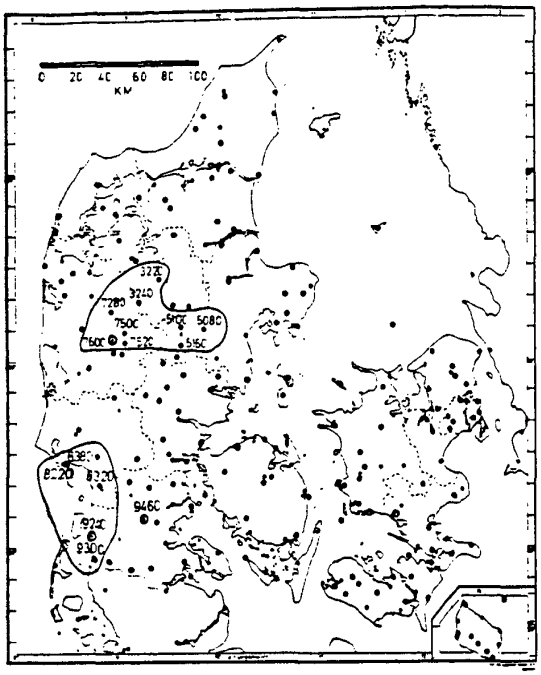


Fig.1. 148 precipitation stations in Denmark. Two subsets of stations for interpolation are encircled.
●: Site to be interpolated.

The most efficient version of our total statistical model for analyzing monthly precipitation values is collected in the following model assumptions

$$E(x_{it}) = \xi + \alpha_i + \beta_t \qquad (1)$$

α_i is a site specific parameter

β_t is a temporal specific parameter (loading)

$$\mathrm{cov}(x_{it_1}, x_{it_2}) = 0, \; t_1 \neq t_2 \qquad (2)$$

$$\rho(x_{it}, x_{jt}) = 1 - \gamma(\beta_t)\cdot[D(i,j)]^{\tau} \qquad (3)$$

ρ is yhe correlation between x_{it} and x_{jt}

$D(i,j)$ is the distance between site i and j

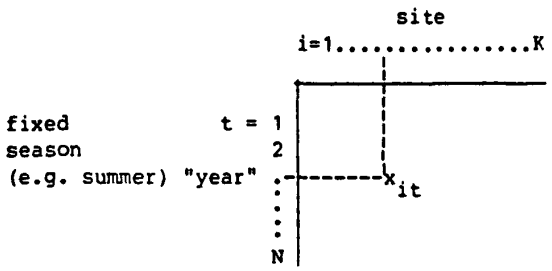
$$x_{it} \approx E(x_{it}) + e_{it}, \; e_{it} \sim Nf(0, \sigma_i^2) \qquad (4)$$

The assumptions state, that the expected logarithmic monthly precipitation can be calculated as a sum of the site specific component α_i and the temporal specific component β_t - a year specific 'season' component, cf. the description of data. The model design is therefore a two way analysis of variance with one observation per cell and with 'column-wise' intercorrelated residuals, according to (2) and (3).

We have earlier (Allerup and Madsen, 1983) applied these model assumptions to interpolation problems, but the method was at that occasion restricted to few sites, deliberately selected homogeneously and inhomogeneously in two parts of Denmark; further, the correlation structure (3) was estimated without free variation of $\gamma(\beta_t)$ (in fact the calculations were pooled over β_t-groups in the data). The present calculations are based upon an a priori sub-devision of Denmark into regions which from a meteorological point of view are expected to be homogeneous and we will study, specifically the multifactorial structure of the spatial correlation function (3) given by β_t-levels.

In summary the analyses presented below follow this scheme of calculations:

We start with a given set of observations (x_{it}) - fixed season (e.g. summer) over the years 1931-60.



With x_{it} = ln(monthly precipitation) at site No. i, year No. t, we first estimate $\alpha_1,\ldots,\alpha_K$, specific for the sites, and the β_t's, the loadings (interpreted as the log-average precipitation over the region with the measurements i = 1,...,K). Next, the residuals, e_{it} from the additive structure (1) are calculated and after grouping of these residual values according to β_t (i.e. residuals arising from months with low versus high average precipitation) we can estimate the multifactor correlation function (3).

The interpolation 'climate' for a region within a given season is then assessed as the degree of accuracy by which we can interpolate a random spot inside the region. This accuracy is,

of course dependent upon the number of interpolating base-stations used and how much we can profit from the stochastic correlation structure in the region (given the season) in comparison with results obtained as simple neighbour averages. The mathematics behind the interpolation of a site 'q' by means of INT_q, is for short given as the conditional mean value $E(x_{i_q} | x_{i_1}, \ldots, x_{i_p}) = INT_q$

$$INT_q = \underbrace{\frac{\alpha_{i_1} + \cdots\cdots + \alpha_{i_p}}{p} + \beta_t}_{\text{structural part}} + \underbrace{V_{21} V_{11}^{-1} (\underline{x}^{(p)} - m_1)}_{\text{stochastic part}} \qquad (5)$$

(see Allerup, 1981, and Allerup et al., 1982 for details).

The structural part represents the interpolation value one would get if simple neighbour values $\alpha_{i_1}, \ldots, \alpha_{i_p}$ and the loading, β_t of the year in question, are used. (This is much like simple day-to-day average over neighbouring stations, although refined according to the additive structure (1)). The last term the stochastic part is a 'correction' term, in which the spatial correlation function is used to modify the simple structural 'guess' according to knowledge about the interpolating base-observations being above or below their individual typical values (V is covariance matrix of interpolating base stations, m_1 the vector of expected precipitation values for the interpolating base stations and V_{21} a vector containing information concerning correlation between the interpolated site 'q' and the base stations).

4. STATISTICAL ANALYSES

The inland region (see Fig. 1) contains 9 stations and the coastal region contains 5 stations they are comparable in size and in order to compare the spatial correlation structure between these two regions we first estimate the multifactorial correlation function. This is based upon a grouping of monthly precipitation (by means of ß) into three groups: low, medium and high. Since we have chosen two seasons: summer and autumn we therefor end up with 2(regions) x 2(seasons) x 3(precipitation grouping)= 12 spatial correlation functions, i.e. 12 estimates of $\gamma(\beta_t)$ and τ, which determine the function, cf. (3).

To see if the mathematical structure (3) is correct and make comparisons easy, we re-write (3) in the form

$$\ln(1 - \rho_{ij}) = \alpha(\beta) + \tau(\beta) \ln D(i,j) \qquad (6)$$

with $\alpha(\beta)$ and $\tau(\beta)$ referring to three ß-levels. The straight line structure (6) can be checked, and least square estimates of intercepts (α) and slopes (τ) can then be calculated. The results from estimation of α, τ are given in Table 1.

Two examples on the fit of correlation function (6) are given in Fig. 2 and 3. These graphs show some few 'outliers' (marked by arrows) where a distinct station enters into station combinations, which are mainly outside the linear structure (6). In view of the specific correlation function (3) this procedure is a kind of second-order quality control of the precipitation measurements.

Inspection of test graphs like Fig. 2,3 reveals a general impression, that the linear structure is more distinct in inland regions that in coastal regions. It seems, that the mathematical structure of (3) forms a workable hypothesis, which is a conclusion arrived at earlier, even for daily precipitation (Allerup and Madsen, 1981, 1982 and 1983, and Allerup, 1982).

Further, analysis of how sensible the estimates in Table 1 are towards different ß-groupings have been undertaken and the results showed rather robust intercept and slope values.

The four correlation functions upon pooled ß-groups are shown in Fig. 4.

If we compare the functions in Fig. 4 we see that they are 'equal' across type of region and show a distinct pattern of 'difference' across season. This is not the impression one gets from Table 1 when reading the slope estimates, but they must, in fact be combined with the intercept values before 'equalities' or 'differences' can be properly understood.

The difference pattern concerns the fact, that correlations are systematically lower in summer than in autumn, i.e. the correlation function drops quickly towards zero in summer. This is in accordance with meteorological expectations since shower dominated weather situations in summer are likely to produce less correlations than the frontal precipitation occuring in the autumn.

For the statistical model it means that it is 'worth-while' collecting information at greater distances in autumn than in summer, if we want to improve simple interpolation estimates.

	Coast				Inland			
	summer		autumn		summer		autumn	
	intercept $\alpha(\beta)$	slope τ	intercept $\alpha(\beta)$	slope τ	intercept $\alpha(\beta)$	slope τ	intercept $\alpha(\beta)$	slope τ
low	-2.16	0.24	-2.65	0.34	-3.17	0.56	-3.59	0.62
medium	-1.27	0.32	-1.32	0.28	-2.02	0.51	-2.26	0.52
heigh	-3.89	0.93	-3.25	0.57	-3.11	0.70	-3.79	0.62
pooled	-2.39	0.42	-2.60	0.40	-2.84	0.56	-3.30	0.59

Table 1. Estimates of coefficients α,τ of correlation functions, specific for precipitation levels: 'low', 'medium' and 'high'.

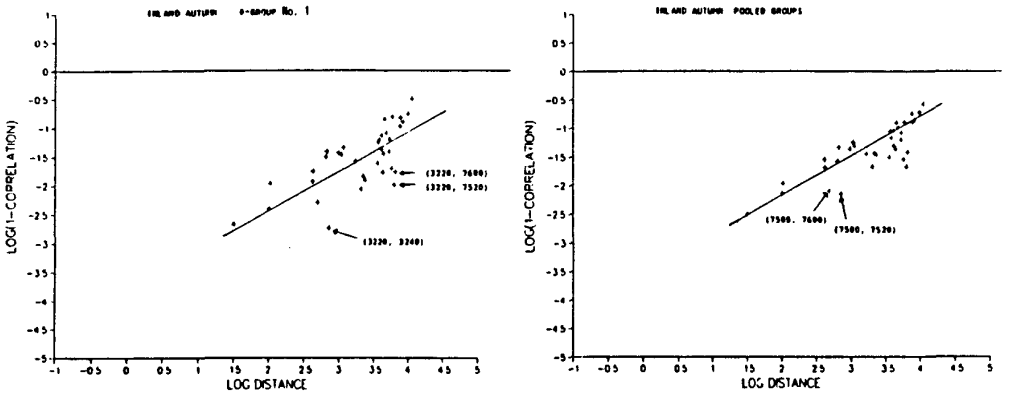


Fig. 2-3. Relations between $\ln(1 - \rho_{ij})$ and logarithmic distance ln D(i,j). (Indland, lowest ß-group and inland, pooled ß-group)

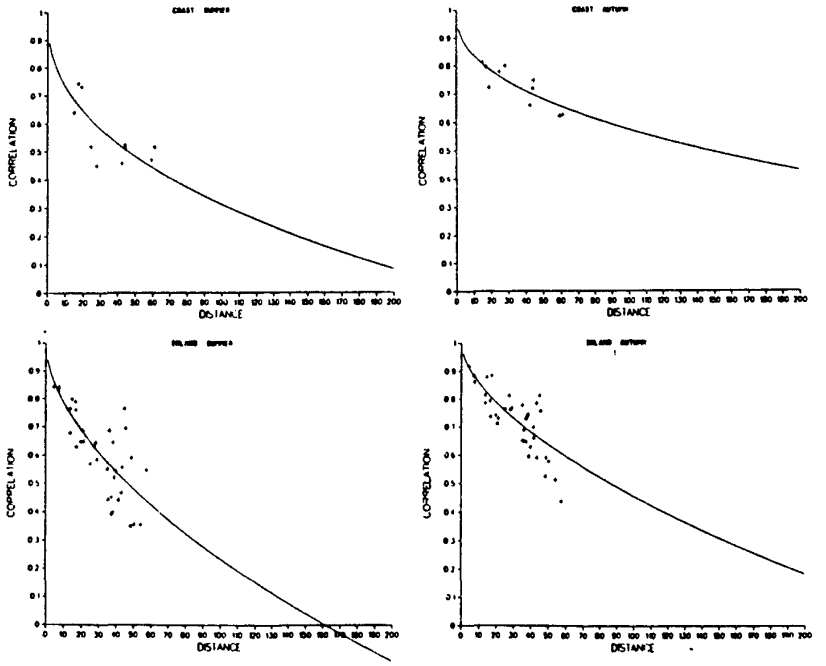


Fig. 4 Estimated spatial correlation functions for coastal- and inland region in two seasons.

Or we may put it in this way: at a given distance one expect greater benefits from referring to how neighbours co-variate in autumn than in summer.

It is furthermore seen from Table 1 that heavy precipitation ('high' ß-groups) is characterized by greater slope values than found for small monthly precipitation values (low ß-groups). It means that great slope values indicate relatively little distant-dependence in contrast to small slope values. The equal slope values for inland, autumn might be due to the general frontal-type precipitation dominating this season.

Weak structures (i.e. not distinct) in Fig. 2,3 - analyses and/or low slope (τ) values point to the fact, that very little can be 'learned' about the stochastic co-variation between points in the neighbourhood. It is therefor to be expected that e.g. in the coastal region in summer the level of interpolation precision will not be increased effectively by 'adding' stochastic knowledge to the primitive structural guess formed by structural part of formula (5). It is to be emphasized that this is independent of the fact, that neighbouring sites may have α_i-values close to the unknown α of the interpolation site. The α's are estimated by temporal averages (for each site) and the β_t is estimated from the spatial average over all stations within a region. The difference between a simple neighbour-average, $\bar{x}$ and the structural part of (5), $(\alpha_1+\cdots+\alpha_p)/p+\beta_t$ is therefor that (i): long term temporal information and (ii): long term spatial information is included in the structural part, while $\bar{x}$ is dealing with model-free references to few neighbours. It is of course interesting to study at which limit, i.e. how close interpolation neighbours should be placed to the interpolation site, before the model-free $\bar{x}$'s will be closer to the 'thruth' (true precipitation) than practically every model based estimate. these last estimates are produced by a model, which is known not to be a complete picture of reality - like all statistical models.

The next step of analysis is the demonstration of interpolations within each of the regions and to show interpolation of a known site outside the two regions. The interpolation scheme is as follows (see the map):

	Interpolate site No.	by means of site No.
coastal region	9240	8320 9300
inland region	7600	7520 7500 7520 7500 7280 7520 7500 7280 3240
outside regions	9460	8280 8320 9240 (coast) 5160 7520 7600 (inland)

In both regions all base stations (5 resp. 9) have been used to assess the correlation function (3), specific for the region and for the season. The inland interpolations of 7600 are executed by referring to 2,3 and 4 neighbour base stations in order to see the possible advantage by increasing the number of neighbour points.

The strategy for the comparisons is that we compare the true values known at 9240, 7600 and 9460 by (i): the primitive, structural guess i.e. the structural part, cf (5) obtained from the model without referring to spatial correlations, (ii): the complete model guess obtained by (5) in which we add spatial, stochastical knowledge about co-variation and finally (iii): a simple local average over the same neighbour stations as used in the structural part of the model. The last procedure is the 'usual' one applied at many met-offices.

The complete model guess is calculated from one correlation function, obtained by pooling the β-group specific correlations to one function. Probably the effectiveness of model interpolation should be increased if one include knowledge concerning the expected level (low, medium , high) of the unknown interpolated site, since we can then refer to one of the sub functions of the multifactorial correlation function.

One could chose to evaluate interpolation efficiency by means of ordinary correlation coefficients but these measures are not very informative. Non-normalized calculations upon differences between observed and interpolated values are dependent on site specific vatiances σ_i, cf the specification of the model.

In consequence, we simply ask whether the stochastic part of (5) effectively lead to a closer guess of the interpolated site (true value) in comparison with the guess based upon the structural part only. The results are presented in the table below, in which the frequency of positive (i.e. effective in the sense of closer fit) is noted, i.e. the frequency of 'good' modifications of the structural guess by adding the stochastic part.

It is in good accordance with the results obtained below concerning the spatial correlations, that we find 'no' effect of including spatial, stochastic knowledge in the coastal region. In fact, we are not better off than tossing a coin. This is true for the attempt of interpolating the outside station 9460 as well, but it is seen that in spite of longer distance between 9460 and the inland region it seems that we gain precision here by adding the stochastic part (62% is significantly deviating from 50%).

Region	Season summer	autumn	Interpolated site	
coast	47%	47%	9240	
inland	57%	68%	7600	2 neighbours
	63%	62%	7600	3 "
	63%	63%	7600	4 "
outside regions	48%	47%	9460	from coast
	56%	62%	9460	from inland

In inland interpolations only summer with 2 neighbours are not significantly pointing to a positive effect of adding the stochastic part. It is furthermore noteworthy, that so little effect of increasing the number of neighbour stations can be seen.

We conclude therefor that within the statistical model we profit from adding spatial knowledge concerning co-variation to the primitive structural part in case of inland regions.

For coastal regions the results point to the fact, that it is not woth-while calculating complex correlation quantities in order to improve the simple structural estimate. Of course, these results are so far obtained as 'examples' gathered in two specific regions in Denmark. Further analyses will include all regions in the country (Allerup and Madsen, 1987).

Since these arguments are carried out by means of a model, which principially could be 'wrong', it remains to see what happens if no statistical model is applied at all for the interpolations. In other words we compare, rather roughly which of the two values (i): the complete interpolated value from the model and (ii): a simple average over the (same) neighbour points are closest to the true value.

The result of these comparisons is that the complete interpolated value is closer to the true value than the 'primitive' average in more than 50% of all month-by-month interpolations in the scheme (minimum 55%, maximum 70%). This means, that a simple choise between interpolations carried out by an average value over neighbours versus using the model comes out in favour of the model. It is however interesting to see, that the 'odds' for the model in this game varies un-systematically over regions and over the number of neighbour stations.

Fig. 5 shows two examples of the relation between true precipitation values and interpolated values calculated by the model. The points are all evenly distributed around the identity line. We summarize the spread around the identity line by calculating mean square deviations (square root) between observed and model-expected values. In terms of relative errors for the original (not-transformed) observations we find 12-13% (inland , autumn), 20-22% (inland, summer and coast, autumn) and 25% (coast, summer). These values are in agreement with earlier experience (Allerup and Madsen, 1983). It seems to be lower than predicted by the model, referring to its internal variance, defined as

$$V_q = \mathrm{Var}(x_{qt} | x_{i_1 t}, \ldots x_{i_p t})$$

which is the variance of the conditional distribution of x_{qt} = interpolated value given the p neighbours $x_{i_1}, \ldots, x_{i_p}$.

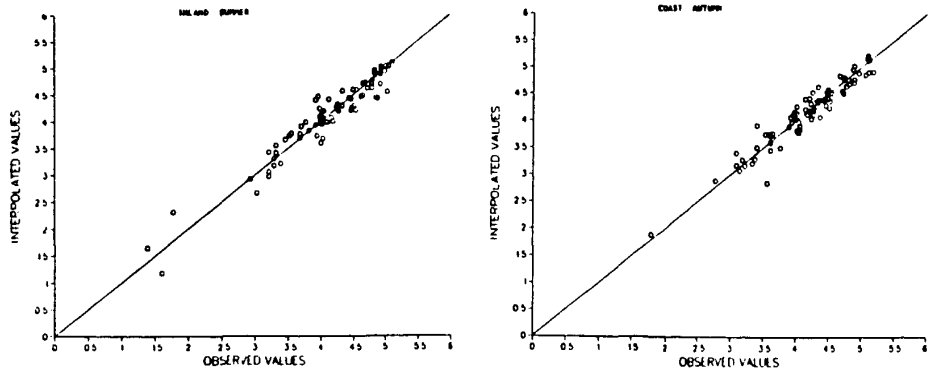


Fig. 5 Relation between observed (log) and predicted (log) values by the model.

5. CONCLUSIONS

We have applied a complex statistical model to two regions in Denmark with different meteorological characteristics, in order to study the conditions for accomplishment of interpolations. The examples given in this paper are the beginning of a comprehensive study of interpolation and quality control among the approx. 700 precipitation stations in Denmark. To day this job is done manually, with little use of automatic interpolation procedures but the installment of a large computer at the Meteorological Institute where all precipitation data are now part of a data base has raised the need of automatic quality/interpolation procedures. It is the aim that the total set of stations can be sub-grouped into 'homogeneous' regions so that interpolations within each region can be carried out with maximum statistical precision. The further analyses will show whether this will be done allways by referring to the complex model with 2,3,4,... neighbour stations or - in some cases - it might be better to use 'no' model, e.g. the primitive average value of the neighbouring stations.

6. REFERENCES

Allerup, P., 1981: A statistical analysis of monthly precipitation in Denmark. Climatological Papers No. 7. Met. Inst.

Allerup, P., et. al., 1981: Nedbør. Suså undersøgelsen rapport No. 1. Dansk Komite for Hydrologi.

Allerup, P., 1982: TTSI method for calculating areal precipitation. Proc. Int. Symp. on Hydrol. Bern.

Allerup, P., et. al., 1982: Methods for calculating areal precipitation. Nordic Hydrol, 13.

Allerup, P., and Madsen, H., 1983: A statistical model for precipitation data - with application to interpolation problems. II international Meeting on statistical Climatologi.Lisabon.

Allerup, P., and Madsen, H., 1987: Analysis of precipitation regions in Denmark (to appear).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

CARTOGRAPHY OF STATISTICAL PLUVIOMETRIC FIELDS WITH AN AUTOMATIC ALLOWANCE FOR REGIONAL TOPOGRAPHY

Patrick BENICHOU
Direction de la Météorologie Nationale
2 Avenue Rapp - 75340 PARIS Cédex 07 - FRANCE

ABSTRACT

Herein is developed the notion of "landscape" encompassing a geographical point, stemming from a principal component analysis concerning the whole of the French relief (kilometric altitude file). 27,000 French sites situated on a 5km -mesh grid have been codified, along with 9,000 raingauge sites. This concept of landscape is fed into the regional analysis of statistical pluviometric fields which through multiple regression makes the most of "landscape" information contained in the measured value, while an optimum interpolation method is used only for a minimal percentage of this value. Not only does this analytic method prove interesting for the user - as it supplies a set of realistic charts - but it can also help climatologists in so far as they can thereby assess the "response" of the various physiographic features in a given season or weather type. The method is now currently in use and helps practise field analysis in any region in France. Various applications have been realized on a mountainous area of central France.

I - INTRODUCTION

To reconstitute a precipitation field from punctual values recorded at a number of pluviometric sites is no easy matter, generations of climatologists have gone through this ordeal and are still doing so. The problems we are faced with are nevertheless well known (spatial and temporal variability of rain, problem of rainfall gauge measurement...) without the solutions being really satisfactory.

The needs for a pluviometric cartography (normal values, rare frequency rainfalls, rainfall amounts on watersheds...) are more and more pressing and for the time being the climatologist has two tools at his disposal in order to meet these needs. On the one hand he can rely on the knowledge he has of the area for which the analysis is asked for : he draws the charts by hand, integrating topography or wind effects. On the other hand, he can use interpolation methods (optimum interpolation, kriging) : the drawing is then made automatically and objectively (Creutin et alii, 1980). The former tool (laborious manual drawing) lacks objectivity and lies greatly on human experience : this is its greatest defect. As for the latter, despite the sophistication of analysis techniques being more and more advanced, it falls short of integrating the reduced scale effects due to topography. Now, a rainfall cartography, whatever it may be, cannot be carried out in a credible way without allowing for parameters likely to be linked with significantly. Hence the idea of setting up an analysis system involving, through linear regression, the local topography coded in vector form. Once the regression has been carried out, it only remains, in principle, to analyze "mathematically" (kriging) the regression residuals whose variance is hoped to be much lower than that of the basic field. The "mathematical" portion of interpolation is then greatly completed with a "physical" portion by making allowance for topography parameters.

All this can be formalized by simply expanding the value P observed at site S_i whose geographical coordinates are (x_i, y_i) and topographic environment is T_i :

$$P(S_i) = P(x_i, y_i, T_i) = \underbrace{f(T_i)}_{\text{regression}} + \underbrace{\varepsilon(x_i, y_i)}_{\text{residual}} \qquad (1)$$

Assessment of rain at any given point M(x,y,T) is made as follows :

$$P^*(G) = P^*(x, y, T) = \underbrace{f(T)}_{\substack{\text{Application of} \\ \text{regression equation}}} + \underbrace{\varepsilon^*(x, y)}_{\substack{\text{interpolated value of} \\ \text{regression residual}}} \qquad (2)$$

The idea of such a method is not new since it dates back to the sixties (Peck and Brown, 1962) where it went by the name of anomaly procedure. At the time the handling was time consuming for it was entirely manual.

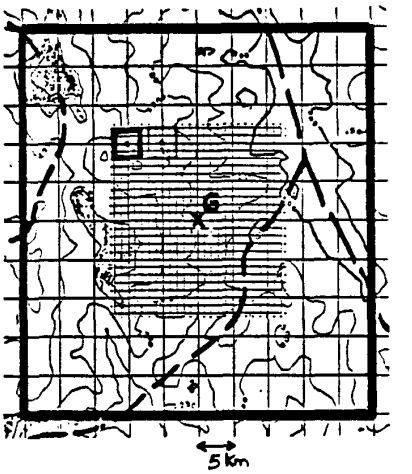
The progress made ever since in the data processing fields make it possible to imagine topography parameters elaborated automatically on the whole country on the one hand and to utilize sophisticated interpolation schemes on the other hand.

II - CODING THE TOPOGRAPHY SURROUNDING A POINT

The expanding relation (1) for the rainfall observed at a site implies that physiographic parameters (symbolized by vector Ti) connected with the measurement site Si are known. In the same way the reconstitution relation (2) for rainfall at any point G in France (grid point for example) is of interest only if it is possible to know automatically and without interpolation the physiographic parameters (vector T) of the site surrounding the point G. Hence the need in the first place for a digitized topography of France : for this purpose, the "Direction de la Météorologie Nationale" has acquired a file, from the "Institut Géographique National", providing an "estimated" elevation (more or less erroneus) at any point of a kilometric grid. On this basis we have made several trials for coding sites through a system of physiographic parameters : distance to the sea, aspect, exploration of a site through concentric rings, evaluation of the solid angle under which the horizon can be seen from a given point... (Laborde, 1984), (Whitmore, 1972), (Solomon, 1968), (Storr and Fergusson, 1972). These coding systems, although interesting, have proved difficult to use in view of their lack of robustness when regression is being made (see following paragraph), and the results are often awkward to interpret physically. We have then tried to achieve an approach both more global and robust to characterize the topographic environment of a site.

II.1 Landscape surrounding a point

In a first stage we carried out a smoothing of France topography by calculating, at each point of a 5km-mesh grid, an average altitude from the 16 surrounding points of the kilometric grid. This operation enables to gain in simplicity and the loss of information which ensues is not very important in so far as it is believed (Laborde, 1984) that topography disturbances less than 5 km do not appear to intervene in a significant way in the spatial distribution of precipitation. The "landscape" surrounding a grid-point G is determined by considering a square matrix (11 x 11) of mean altitudes surrounding the point G. The site G is thus characterized fy 121 values. We have assessed these 121 values for about 27,000 sites of the 5 km-side square grid over France (Figure 1).



II.2 Principal components analysis (P.C.A) over french landscapes

It is obvious that 121 altitudes are redundant when characterizing a site. Hence the necessity to condense information to a maximum in order to obtain a reduced system of parameters in losing a minimum of information : the Principal Components Analysis (P.C.A) lends itself to this operation perfectly : it was applied to a "relative" landscape, obtained from the basic landscape by taking out the mean elevation of the central point from each of the 121 altitudes. This "centering" operation makes it possible, on the one hand, to isolate as a first predictor the average altitude of the central point, to avoid the occurence of a first main component containing an abusive share of variance, on the second hand. The P.C.A is then carried out on a consistent sample (27,000 sites). The results are comparatively satisfactory from two viewpoints :

1° The first fifteen principal components suffice to account for about 90 % of the field variance.

2° The eigenvectors associated with the first eigenvalues of the covariance matrix of the relative landscapes have an interesting physical significance.

II.3 Physical interpretation of the eigenvectors

After the P.C.A, the relative landscape in each point G is given with a satisfactory accuracy through the linear combination of 15 eigenvectors V1, - , V15, each Vj being weighted by a principal component Cj. The relative landscape in G is :

$$\sum_{j=1}^{15} C_j(G) \cdot V_j$$

In view of this, each eigenvector Vj appears as a typical landscape. Each "landscape" in France is a certain combination of these typical landscapes. The eigenvectors or typical landscapes are then interesting to visualize. This is what Figure 2 suggests.

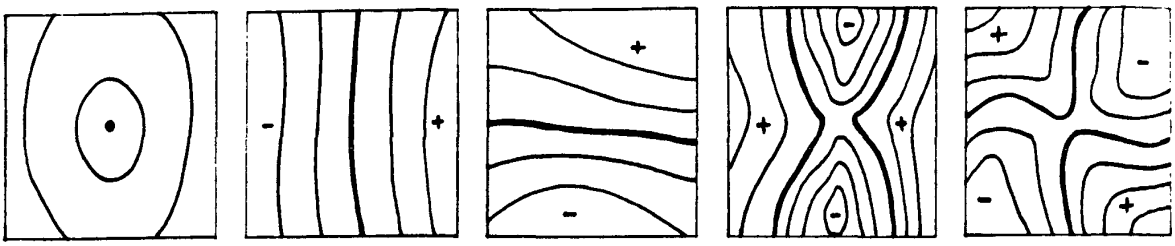


Figure 2 : first five eigenvectors or typical landscapes

The first five typical landscapes are characterized by :
V1 : global embankment effect (39% of variance accounted for)
V2 : presence of a West/East slope (24% of variance accounted for)
V3 : presence of a North/South slope (9% of variance)
V4 : North/South saddle effect (3% of variance)
V5 : Nord-East/South-West saddle effect (3% of variance).

The other eigenvectors reveal finer spatial stuctures but the essential (78%) of a relative landscape is then accounted for by giving its first five principal components.

To sum up, each site G is now coded by a vector T comprising 16 predictors :
- the average altitude of the central point : $\bar{Z}$
- the first 15 principal components C1, - , C15 of the relative landscape.

Finally, the coding shown above suggests as independant predictors to a regression scalar variables which have the advantage to have an important spatial range, each principal component being associated with a typical landscape over a zone 50 km square. Once the topography coding has been carried out, at he 5x5 grid points as well as at the rainfall reporting stations Si, it is suggested to quantify the fraction accounted for by the landscape in a precipitation field, given that the more the accounted for fraction will be important the more the fraction to interpolate will be reduced.

III - REGRESSION AS AGAINST TOPOGRAPHY. EXAMPLE : THE MASSIF CENTRAL

III.1 Field of study - an example : The Massif Central.

We limit our study to statistical pluviometric fields which are less fluctuating than daily precipitation : monthly or seasonal normals, reckoned for the "Mémorial 1951-1980". The field of study must be vast enough for the sample of measuring points to be sufficient, but it must not be too vast either, for the regression equation to make sense, which brought us to divide France into ten regions, each of them representing a geographical entity : Southern Alps, Pyrenees... and in particular the Massif Central (250 km x 370 km) : whithin this field, 270 pluviometric sites show computed normal values. These sites are quite evenly distributed, which is not generally the case for all the regions in France.

III.2 Selection of predictors

From this sample of 270 values, we select the significant predictors (described above) - through an ascending progressive selection method - according to a Fisher test on the increments of the multiple correlation coefficient.

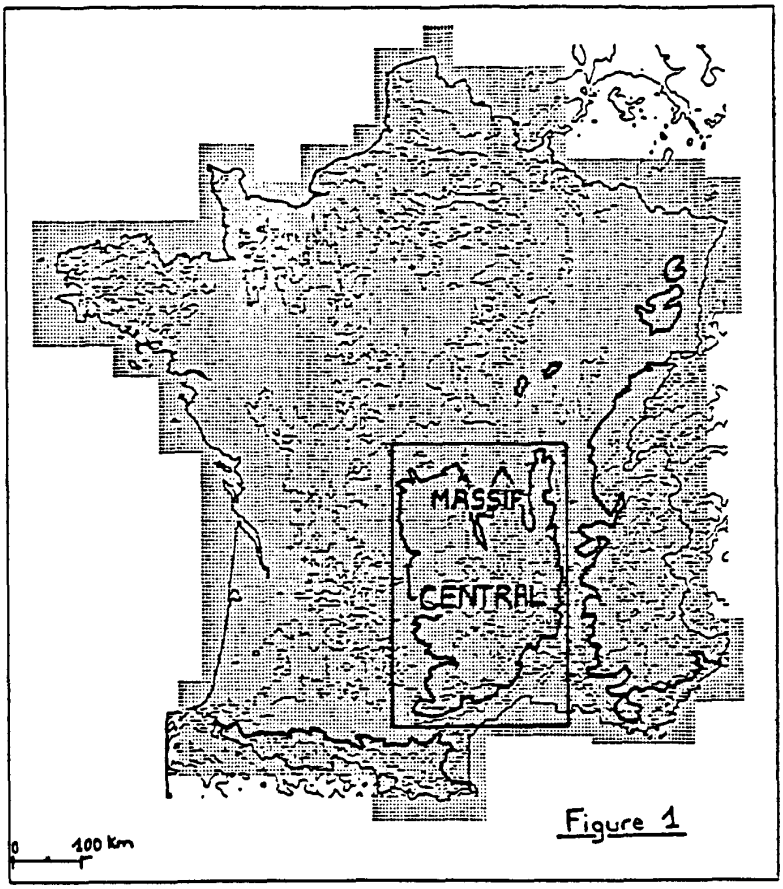


Figure 1 : 5 km-mesh grid over France

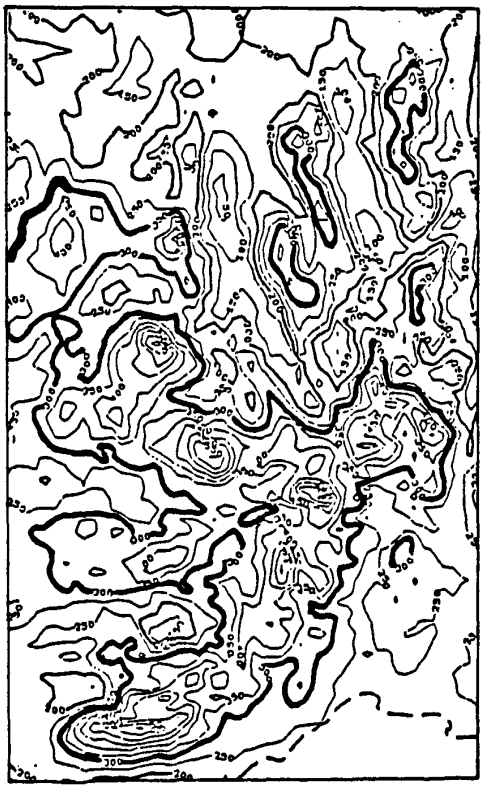


Figure 3 : application of the regression equation

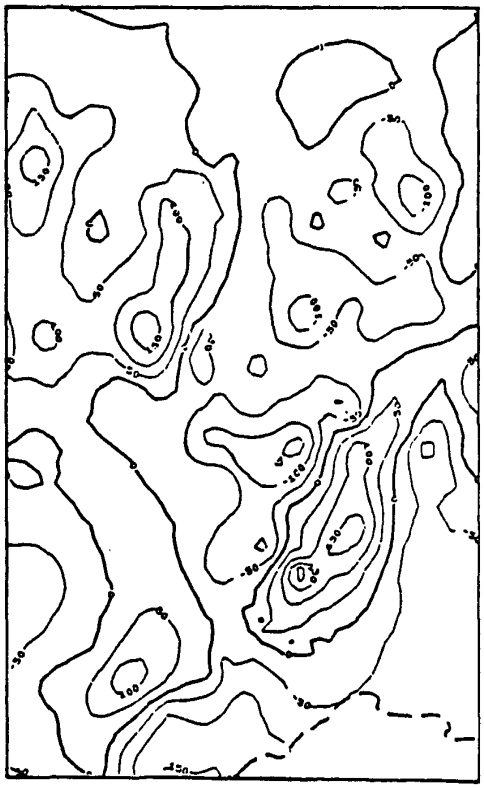


Figure 4 : kriging of the regression residual

0 25 50 Km

In our example of winter normals (December + January + February) over the Massif Central, 4 predictors have thus been selected :
- third component C3 (R^2 = 0.308)
- first component C1 (R^2 = 0.525)
- mean elevation Z (R^2 = 0.561)
- second component C2 with a final multiple correlation coefficient (R^2 = 0.591)

III.3 Regression equation and interpretation

We can then compute the coefficients of the regression equation valid at the sites

$$f(T) = 2184 + 7.2\ C3 - 2.2\ C1 + 0.7\ Z + 1.4\ C2 \qquad (3)$$

It is established that it is not the mean altitude Z which is selected in the first place contrary to received ideas but C3 (North-South exposure). The sign of coefficients in the regression equation f(R) is important : that of C3 is positive, which means that the precipitation seem more abundant - and then linked with a mediterranean regime - on the South facing slopes (example : the Cevennes Barrier). The coefficient of C1 is negative, which shows that it rains more on the ridges than in the valleys. The average altitude Z comes only third with a positive coefficient. The component C2 is assigned a positive coefficient which causes the westerly exposed slopes to be the wettest (example : Auvergne volcanoes), contrary to the east facing slopes, like the oriental edge of the Massif Central. Regions like "La Montagne Noire" can be observed where mediterranean ($C3 > 0$) and oceanic ($C2 < 0$) influences are in opposition to each other.

III.4 Cartography of regression

The regression equation **(3)** is then applied at any point G in the studied field (5km mesh). However, it is necessary to ensure that both samples of predictors concerning the sites on the one hand and the grid points on the other hand are extracted from the same population. A test is currently being prepared. After taking this precaution we can map f(T) on to the Massif Central (Figure 3). The relief having explained about 60 % of the variance, the residual $\varepsilon(x,y)$ represents only 40% of it and it is this residual that we are going to assess through a kriging method.

IV - KRIGING THE FIELD OF RESIDUALS

The kriging method is a geostatistical method which amounts to a probabilistic approach to interpolation. It is beyond the scope of this paper to discuss the theory of kriging (Delfiner, 1973). It is comparatively clumsy and implies some hypotheses on the field to analyze in particular it only applies to a regionalized variable. It has been seen that it is not the case for precipitation since the latter depends on the "topography" but it is reasonable to think that the residual:

$\varepsilon(x, y) = P(x, y, T) - f(T)$ draws near to it. This advantage linked with a comparatively low spatial variability makes out of this residual a variable easier to krige. The kriging program makes an automatic structural identification out of the $\varepsilon(x, y)$ field. We have chosen a model which respects the measured values (model without any nugget effect). The assessment $\varepsilon^*(x, y)$ at the grid points yields a chart of the field of residuals (Figure 4)

V - RECONSTITUTION OF THE PRECIPITATION FIELD - DISCUSSION

V.1 Reconstitution

At each point G of the grid covering the Massif Central we know the value f(T) given by the regression equation, together with the assessed residual $\varepsilon^*(x, y)$. We infer :

$$P^*(x, y, T) = f(T) + \varepsilon^*(x, y)$$

and the final chart of winter normals assessed over the domain (Figure 5)

In order to facilitate comparison we have mapped (Figure 6) the field of normals without taking account of topography (direct kriging on the normal values). This chart seems little realistic on the one hand and clumsily linked with the position of pluviometric sites on the other hand. It will also be noticed that the application ot the regression equation alone (Figure 3) provides a good approach to the final chart, as the kriged field of residual (Figure 4) intervenes as a complement only. The gain in quality brought by taking topography into account appears to be consistent.

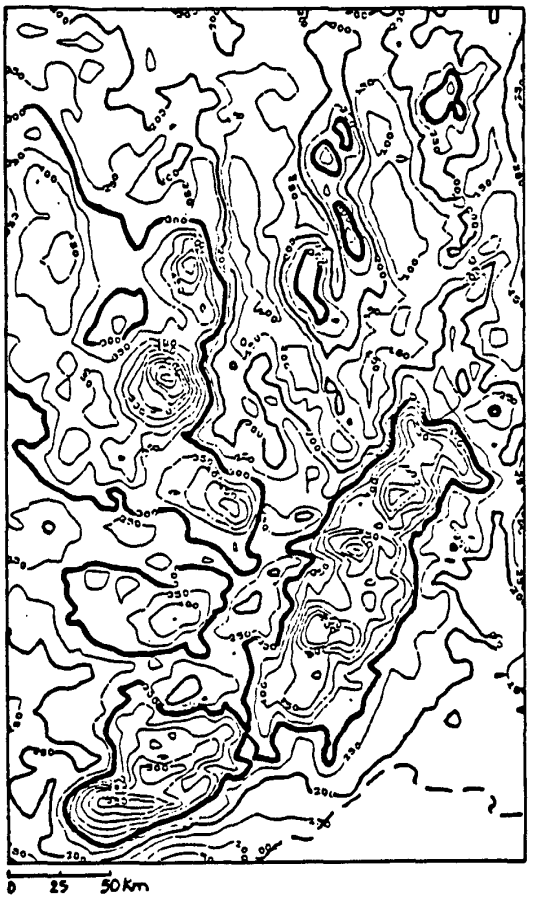


Figure 5 : final winter normal chart



Figure 6 : direct kriging of winter normals

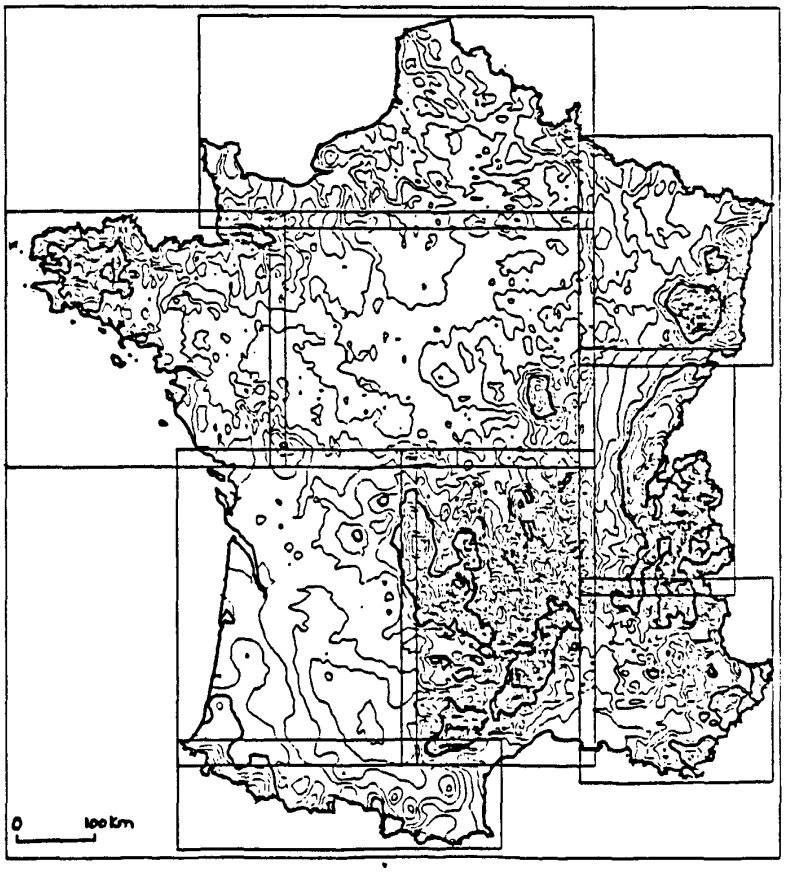


Figure 7 : analysis of february normal precipitation over continental France (nine zones)

V.2 Test of the method under less favourable conditions

It has been seen (paragraph II) that the network available over the Massif Central was remarkably dense. We have then tested the analysis method over a more and more reduced fraction of sites (from 75 to 25 %). We notice an extreme stability of Figure 3 (regression equation) and therefore of the final chart. It is not all the case for the chart obtained by direct kriging (Figure 6) which can only gets worse rapidly.

VI - OPERATIONAL USE AND PROSPECTS

The analysis method explained here has been termed AURELHY : Analysis Using the RELief for HYdroclimatology. It is now operational in the hydroclimatology department of the "Direction de la Météorologie Nationale". It seems to be in agreement with the recommendations of the WMO (Nordenson, 1971). The corresponding software has been elaborated and is run on a routine basis on a Control Data Computer. It comprises several alternative plans (regression on logarithms, forcing of predictors...) and is run over any French region with an accuracy (mesh) of 5 km. The method is being used to provide - in addition to statistical field charts (monthly normals, quintiles) - reconstituted values for sites with no sufficiently long chronological series (automatic stations for example). Average annual rainfall charts are then inferred by adding the twelve monthly charts, which allows to take into account the various monthly pluviometry/topography relations. In the same way, the national rainfall charts are obtained by putting the ten regional charts (see paragraph III.1) together (Figure 7). In a remote future, the method will have to be used provided certain arrangements are made (allowance for field drift) for the analysis of individual events of short duration (dekad, day). Radar information together with meteorological information (weather types, Benichou, 1985) could then constitute valuable complementary tools.

REFERENCES

BENICHOU P. (1985) : Classification automatique de configurations météorologiques sur l'Europe Occidentale. Note SCEM n° 18. Météorologie Nationale.

BLEASDALE A. and CHAN Y.K. (1972) : Geilo (Norway) symposium. WMO/OMM n° 326 orographic influence on the distrubution of precipitation.

CREUTIN D. et alii (1980) : Analyse spatiale et temporelle des épisodes pluvieux cévenols. La Météorologie. VI n° 20 Mars Juin 1980.

LABORDE J.P. (1984) : analyse des données et cartographie automatique en hydrologie : éléments d'hydrologie lorraine - thèse de doctorat - Institut National Polytechnique de Lorraine.

DELFINER P. (1973) : Analyse objective du géopotentiel et du vent geostrophique par krigeage universel. La Météorologie. V n° 25.

NORDENSON T.J. (1971) : Rapport n° 6 sur les projets OMM/DHI. Préparation de cartes coordonnées des précipitations. de l'écoulement et de l'évaporation.

PECK E.L., BROWN M.J. (1962) : An approach to the development of Isohyetal Maps for Mountainous Areas. Journal of Geophysical Research Vol. 67.

SALTER P.M. (1972) : Areal rainfall analysis by computer. Geilo symposium. WMO/OMM n° 326.

SCHULZE R.E. (1977) : Practical considerations in the application of surface fitting techniques for estimations of areal rainfall. 2nd Conference on hydrometeorology. Toronto - Canada.

SOLOMON S.I. et alii (1968) : The use of a square grid system for computer estimation of precipitation, temperature and run-off. Water Resource Research 1968 pp 919-929.

STORR D. and FERGUSSON H.L. (1972) : The distribution of precipitation in some mountainous canadian watersheds. Geilo symposium WMO/OMM n 326.

WITHMORE J.S. (1972) : The variation of mean annual rainfall with altitude and locality in South Africa. Geilo symposium WMO 'OMM n° 326.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

EXTREME WIND SPEEDS DUE TO TROPICAL CYCLONES IN FRENCH DEPARTMENTS AND TERRITORIES OVERSEAS

D. DELAUNAY
Centre Scientifique et Technique du Bâtiment Nantes (France)

INTRODUCTION

Construction dimensions, among other things, must take account of the loads due to extreme winds liable to be sustained during the life of the buildings. This question is particularly important in regions which are prone to tropical cyclones.

Methods based on statistical analysis of the maximum annual measured wind speeds are not adapted to these regions. Indeed, part of the maximum annual winds are produced by statistically different populations, depending on whether they are or are not due to the passage of a cyclone ; in addition, extreme winds due to tropical cyclones are relatively rare for a given site (approximately 1 to 3 occurences in ten years). What is more, the strongest winds are not generally measured.

For these areas, another approach must be used. For a few years now, a number of studies have been carried out (GEORGIOU et al., 1983, BATTS et al., 1980 and GOMES and VICKERY, 1976...) for which extreme wind speed distributions are obtained by simulation (Monte-Carlo method) on the basis of the climatological characteristics of tropical cyclones of a region and modeling of the wind field as a function of the pressure at the center and the speed of translation.

This method was applied in the French Overseas Departments and Territories (DOM-TOM) : Guadeloupe, Martinique, Reunion, New Caledonia and Tahiti.

1. WIND IN TROPICAL CYCLONES

The fundamental cyclone dynamic mechanism is the driving of hot, damp air from the ocean surface towards the upper layers of the atmosphere. A depression is formed, to which the air from neighboring regions moves, in a spiralling trajectory, produced by the Coriolis force and the centrifugal inertia force. Therefore, at the center, there is a region within which the wind is relatively low : this is the cyclone's "eye".

In addition, the translation of the cyclone results in assymetry in the structure of the wind field : depending on the position with respect to the trajectory, the speed of translation increases or decreases the speed of the purely cyclonic wind. This phenomenon is well known to sea men, who make a distinction between the "dangerous half-circle" and the "manoeuvrable half-circle".

On the basis of the equilibrium of force equation, and accepting an exponential decrease in atmospheric pressure as the distance to the center changes, MYERS and MAKLIN (1961) demonstrated that the mean gradient wind speed (V_g) can be approximated by the formula :

$$V_g = -X + \left[X^2 + \frac{\Delta P}{\rho} \frac{R}{r} \exp\left(-\frac{R}{r}\right) \right]^{1/2} \quad \text{with } X = (fr - T \sin\theta)/2$$

where r is the distance from the center of the cyclone, ρ the air density, ΔP the pressure difference between center and periphery of cyclone, f the Coriolis parameter, T the speed of translation of the cyclone, θ the azimuth of the point considered with respect to the direction of the trajectory, and R the radius of maximum wind speeds.

The ratio between the mean wind speed 10 m above the sea surface and the mean gradient wind speed varies according to SCHWERDT et al. (1979) between 0.75 and 1.0. The value C = 0.865 is generally used and has been applied to this study.

2. CLIMATOLOGY OF CYCLONES IN THE DOM-TOM REGIONS

The method of simulation used requires a knowledge, for all cyclones relevant to each site studied, of both the distributions of parameters ΔP, R, T and the distribution of the cyclone center position with respect to the site, when it moves closest to it. This position is characterized by the minimum distance d between the site and the center, and is given a positive sign when the site is in the "dangerous half-circle" and a negative sign when it is in the "manoeuvrable-circle". The distribution of N_c is also determined, i.e. the number of cyclones affecting the site during a cyclonic season (one cyclone season per year).

The distributions of these different parameters are linked with a number of choices which set the studies phenomenon precisely :

a/ Nature of the phenomenon : according to documentary sources, it is possible to limit the statistics merely to those phenomena of high intensity during passage closest to the site (maximum 10 min. mean speed greater than 33 m/s, classification "tropical cyclone"), or it is necessary to include in them tropical storms (maximum speed between 17 and 33 m/s) when the intensity of the phenomenon is not determined, as in the case of the South Pacific.

b/ Neighborhood of site : the statistics apply to phenomena which have evolved within a 1000 km diameter circle (or 1200 km for Tahiti) whose center is the site of interest. Beyond these distances, the influence of a cyclone on the site is negligible.

c/ Reference period : precise statistical analysis of the data has led us to choose the following reference periods : in the Antilles 1915-1985 ; Reunion 1950-1985 ; South Pacific 1948-1983.

The distribution characteristics for parameters ΔP, T, d, N_c, R are obtained from a systematic survey of these phenomena.

The assumption of a log-normal distribution, generally used for ΔP appears to be justified at Reunion and in the South Pacific. Conversely, a Weibull distribution is better suited to the Antilles cyclones, where the log-normal law tends to overestimate the largest values of ΔP.

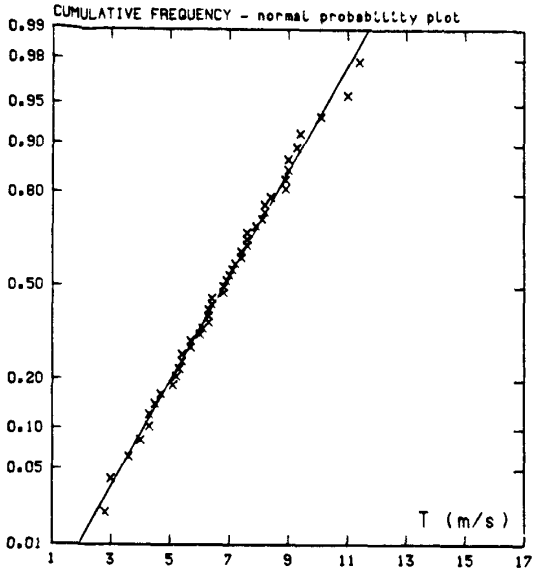
In the Antilles, the speed of translation T is well-represented by a normal distribution at Reunion and in the South Pacific, a log-normal distribution is more appropriate, as the values of T may be relatively high.

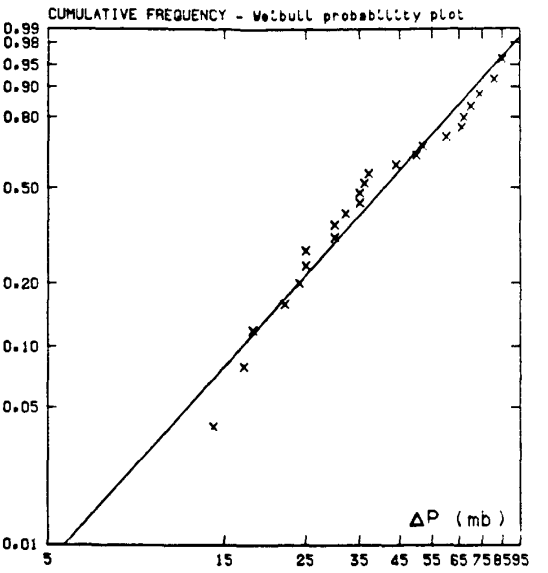
The position parameter d appears to have a uniform probability at Reunion and in New Caledonia. For the Antilles, as the trajectory directions are principally East-West, and as the number of cyclones increases towards the North, the probability of d decreases with its algebraic value. A linear probability density law, limited to the boundaries of the area in question, has therefore been fitted ; the same applies to Tahiti.

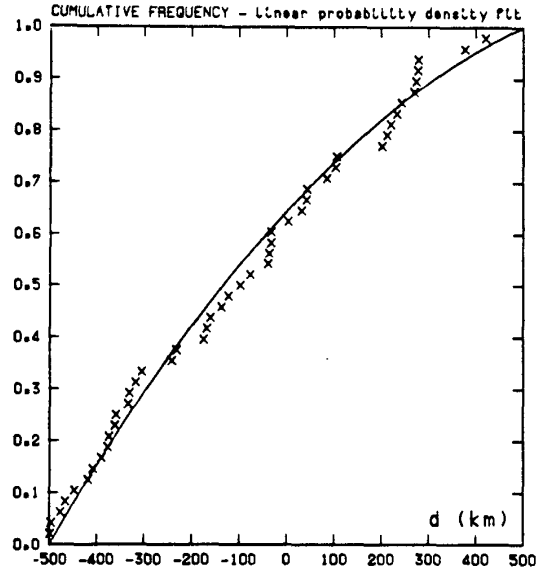
The number of cyclones per cyclonic season is generally well adjusted by a Poisson's law. However, for Tahiti, a negative binomial law gives better account of the risk of sustaining several cyclones in the same season.

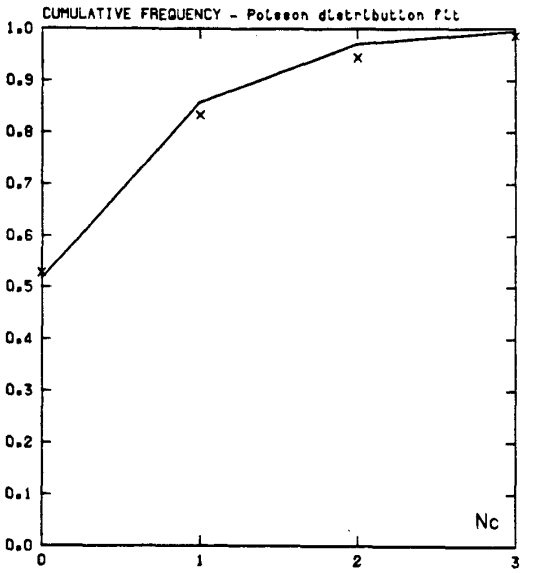
The determination of distributions of maximum velocity radius R poses a particular problem because it is more difficult to access and, in addition, it has been demontrated for the North Atlantic that it depends upon the intensity of the phenomenon. According to the data available to us for the regions in question, R-distributions can be fitted by a log-normal law and $\ln(R)$ appears to be negatively correlated with $\ln(\Delta P)$ in the Antilles and the Reunion.

Figure 1 shows the distributions of the ΔP, T, d, N_c and R parameters for Guadeloupe.









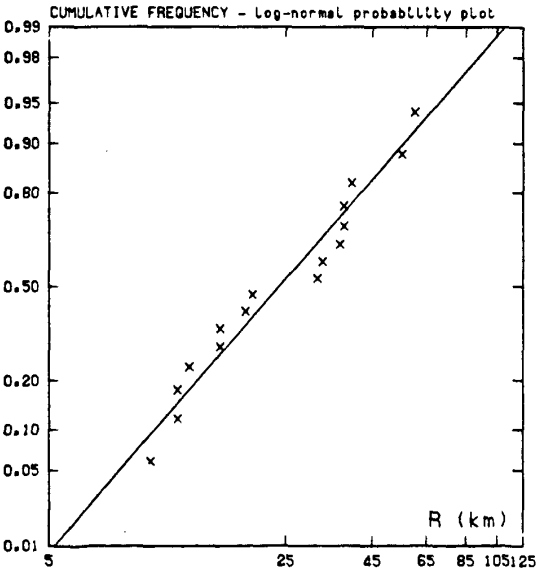


Fig. 1 : Characteristics of tropical cyclones in GUADELOUPE (1915-1985)

3. DISTRIBUTIONS OF EXTREME CYCLONIC WINDS

3.1. Procedure - results

The problem posed is to determine, during a cyclonic season, the probability of attaining a given wind speed at a set location. The Monte-Carlo method, on the basis of the climatology of phenomena passing near the site in question, serves to calculate the distributions of the maximum annual speed due to tropical cyclones : for a given year, on the basis of the distribution of N_c, a random choice is made of the number of phenomena (n) passing in the neighborhood of the site. Then, on the basis of distributions and any correlations observed n cyclones are choosen at random with characteristics ΔP_i, R_i, T_i, the site being in position d_i during closest passage. For each of them, the physical model makes it possible to calculate the wind speed V_i induced on the site. The maximum speed of the season is the greatest of speeds V_i (it is zero when n = 0).

The distribution of the maximum annual speeds due to tropical cyclones is obtained by renewing the operation on a great number of cyclonic seasons, in this case, 20 000. The wind speeds liable to be exceeded during a 50-year period are deduced from this, with the associated return periods (table 1).

Return periods (years)	25	50	100	225	500	1000
Probability of exceedance in 50 years	87 %	64 %	40 %	20 %	10 %	5 %
SAINT BARTHELEMY	32	37	41	45	48	51
GUADELOUPE	31	36	40	44	47	50
MARTINIQUE	29	35	39	44	47	50
REUNION	38	40	43	46	48	50
NEW CALEDONIA	38	40	43	45	48	49
TAHITI	30	34	39	43	(47)*	(52)*

* estimate unreliable

Table 1 : 10-min wind speeds (m/s), 10 m above sea, for different return periods

It will be noted that these distributions may be fitted by a Weibull law over all the values with frequencies less than 10 % (return period greater than 10 years) for Reunion and South Pacific, and for values with frequencies less than 2 % (return period greater than 50 years) for the Antilles. The Gumbel distribution often used for extreme wind estimations, is unsuitable here (figure 2).

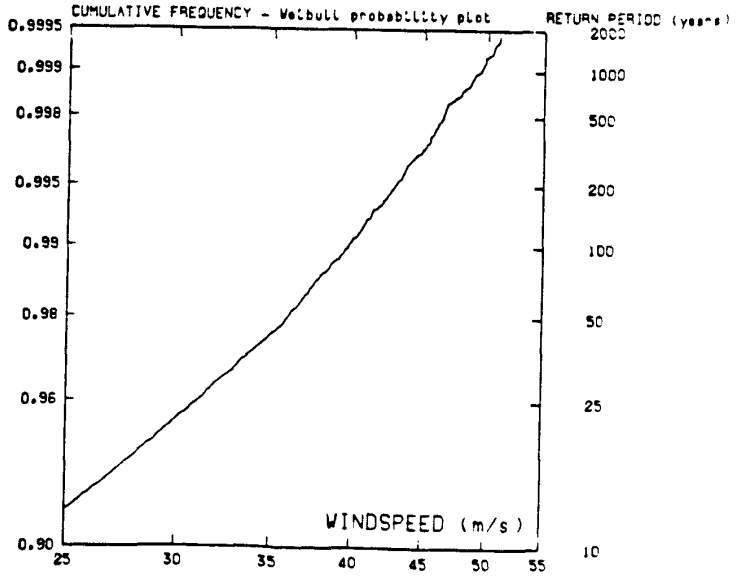


Fig. 2 : Distribution of annual maximum 10-min windspeed in GUADELOUPE (10 m above sea)

It should be noted that the 1000 year return period wind can result from a cyclone with intensity ΔP = 85 mb and speed of translation T = 8 m/s, which values have been observed on a few of surveyed cyclones. The rareness of such winds at a site is due to the fact that the speed is reached in a relatively small zone of the cyclone. Thus, it appears that this method of simulation can be used for estimation high-return-period speeds, without the extrapolation of the observed parameter distributions, and therefore with a sampling error which can be assumed to be modest.

3.2. Estimation errors

3.2.1. Sampling errors

The distributions of the characteristic parameters of the cyclones, ΔP, R, T are determined on the basis of samples of limited size which can give only an estimate of the characteristics of all the cyclones in this area, similar limitations apply for the distribution of the number of cyclones per season, determined on the basis of a limited number of years.

The method proposed by BATTS et al. (1980 b) is used here : the distribution characteristics of a parameter are estimated on the basis of a sample of n observed cyclones (or n cyclonic seasons). However, it would have been possible to observe n other cyclones (for cyclonic seasons) which would have given another distribution, and therefore other wind speed values.

Therefore on the basis of the original distribution of each parameter, assumed to be exact, a random choice of 1000 samples can be used to estimate 1000 distributions of this parameter, whose dispersion is due to the sampling error. This operation is performed for parameters T, ΔP, R, N_c to obtain 1000 "sets" of distributions and correlations between R and ΔP enabling, by a method of simulation, the determination of 1000 values of V_{50} and V_{1000} - i.e the 10-minute average winds at 10 m above sea level, with 50 and 1000 year return periods. The characteristics of the sampling error in V_{50} and V_{1000} (table 2) are then deduced from these values.

	ANTILLES	REUNION	NEW CALEDONIA	TAHITI
V_{50}	2.5	2.0	1.7	3.9
V_{1000}	4.9	4.0	2.7	12.3

Table 2 : Standard deviation of sampling error in V_{50} and V_{1000} (m/s)

It will be observed that the sampling errors are high at Tahiti because of the small number of data available.

3.2.2. Observation errors

It is reasonable to expect that errors of observation for ΔP, R and T have a random character, and therefore that these observations errors are included within the sampling errors.

However, to test the effects of a systematic error, the values of ΔP and T are increased by 10 %, and those of R by 20 %. Similarly, it may be feared that all the cyclones which have crossed the area in question were not listed. Simulation was therefore carried out with an average value of N_c increased by 10 %.

It appears that these modifications result in an increase of the values of V_{50} and V_{1000} not exceeding 1.5 m/s except for ΔP, for which a systematic over-evaluation of 10 % leads to an increase of V_{50} and V_{1000} between 2 and 2.5 m/s.

3.2.3. Errors due to physical model

The ratio C between wind at 10 m above sea level and wind gradient has been chosen as constant and equal to 0.865. However, it may be presumed that this ratio is a random variable. The hypothesis of normal distribution of C, average 0.865, standard deviation 0.09, used by GOMES and VICKERY (1976) has been tested. It would appear that the influence of this choice on V_{50} is small but that it has resulted in an increase of V_{1000} of 2 to 3 m/s.

This hypothesis is equivalent to stating that the physical model is valid on average for a high number of cyclones, and that the standard error of the model is 10 %, which value is obtained by comparison between the wind values measured in the DOM-TOM stations and the model-predicted values (DELAUNAY, 1985).

CONCLUSION

A statistical study of cyclone characteristics in the DOM-TOM's, associated with a method of simulation, has made it possible to determine, for each region, the extreme cyclonic wind speed distributions. These speeds are reference values for an average wind over 10 min. at 10 m height above sea level.

Estimation errors are essentielly due to sampling errors, for which the standard deviation is generally less than 10 % of the estimate, except at Tahiti, for winds with high return periods.

In coastal areas, the sea wind values give an excellent estimate of the cyclonic risk. Indeed, during the passage of a cyclone, the mean wind direction sweeps over a wide sector, and the probability of the wind coming from the sea at a given moment is high. For interior regions, thé wind speed reduction coefficients have been calculated by SIMIU et al. (1976). The case of mountainous areas does, however, pose a major problem : experience reveals that terrain-induced speed-up may be substantial. In this case, it would appear necessary to consider numerical modelling.

BIBLIOGRAPHY

BATTS M.E., RUSSEL L.R., SIMIU E., 1980 a "Hurricane wind speeds in the United States", J. Struct. Div., ASCE, vol. 106, pp 2001-2016

BATTS M.E., CORDES M.R., SIMIU E., 1980 b, "Sampling Errors in Estimation of Extreme Hurricane Winds", J. Struct. Div., ASCE, vol. 106, pp 2111-2115

DELAUNAY D., 1986 "Vents extrêmes dus aux cyclones tropicaux dans les DOM-TOM" Cahiers du C.S.T.B., livraison 269, mai 1986

GEORGIOU P.N., DAVENPORT A.G., VICKERY B.J., 1983, "Design Wind Speeds in Regions Dominated by Tropical Cyclones", J. Wind Eng. Ind. Aerodyn., vol 13, pp 139-152

GOMES L., VICKERY B.J., 1976 "On the Prediction of Tropical Cyclones Gust Speeds along the Northern Australia Coast", Res. R. 278, School of Civil Engineering University of Sydney

MYERS V., MALKIN W., 1961, "Some Properties of Hurricane Wind Fields as Deduced form Trajectoires", National Hurricane Research Project, Rep. n° 49, nov. 1961

SCHWERDT R.W., HO F.P., WATKINS R.R., 1979, "Meteorological Criteria for Standard Project Hurricane and Probable Maximum Hurricane Wind Fields, Gulf and East Coasts of the United States", NOAA Tech. Rep. NWS 23, Washington D.C.

SIMIU E., PATEL V., NASH J.F., 1976, "Mean Speed Profiles of Hurricane Winds", Journal of the Engineering Mechanics Division, ASCE, vol 102, pp 265-273.

COMPARISON OF METHODS TO ESTIMATE MISSING DATA AND TO BUILD LONG CLIMATOLOGICAL SERIES

by JOEL GUIOT, Laboratoire de Botanique Historique & Palynologie, U.A. 1152 du CNRS, rue Poincaré, F-13397 Marseille, France.

1. INTRODUCTION

Various methods usually based on standard statistics are used to estimate the missing data of climatological series. More complex statistics must be used to extrapolate climatological series, frequently short, from long proxy series such as historical archives series or tree-ring series.

From the first point of view, Tabony (1983) has compared several techniques: the total or partial correlation, the linear or optimum interpolation, the principal components and a particular method of the U.K. meteorological office. Goodin et al. (1979) has compared several interpolation methods for sparse data (it is the most frequent case in our application domain). Although they conclude that the least-squares polynomial interpolation is a good compromise between accuracy and computational costs, and that optimum interpolation is often complex, mainly when the data are numerous and diversified (as in this paper), it seems that weighted interpolation is easy to implement.

Concerning the transfer function problem, which is the usual name of the methods allowing to extrapolate proxy series, the review made by Webb and Clark (1977) shows that methods such as multiple regression, stepwise regression, regression after extracting principal components and canonical analysis are slightly different solutions of the same linear model. The difference between the results are mainly function of the data set, the importance of noise and/or the lack of analogues in the calibration data. A partial solution of these problems is to experiment various methods.

2. THE METHODS

The aim is to estimate the missing data of the matrix X(n,m), where n is the number of temporal observations (the raw of the matrix) and m is the number of variables (the columns of the matrix). Two periods are distinguished within the observations period [1,n]: [1,nc] the calibration interval on which the estimations are based, and [nc+1,n], the independent verification interval on which the results are checked.

The mean and the variance of the variables are given by:

$$x_{.J} = 1/n_J \sum_{i=1}^{n_J} x_{iJ} \qquad (1)$$

$$s_J^2 = 1/n_J \sum_{i=1}^{n_J} (x_{iJ} - x_{.J})^2$$

where n_j is the number of not missing observations in the calibration interval [1,nc] for the variable j. The correlation between the variables j and k is given by

$$r_{Jk} = 1/n_{Jk} \sum_{i=1}^{n_{Jk}} (x_{iJ}-x_{.J})(x_{ik}-x_{.k})/(s_J s_k) \qquad (2)$$

where n_{jk} is the number of observations not missing for both variables j and k within interval [1,nc].

2.1. The simple regression

The estimation of the elements x_{ij} of the matrix X is based on a particular regression. Two input parameters are necessary: nr, the maximum number of regressors and rc, the critical correlation below which the regressor is rejected.

For each jth series (j=1,..,m):

The vector $u_j = (|r_{j1}|n_{j1}^{\frac{1}{2}}, \ldots, |r_{jm}|n_{jm}^{\frac{1}{2}})$ is used to determine which variable is the best correlated with the jth one. The correlation is multiplied by the number of observations really used to take account of the number of degrees of freedom. The nr maxima of the vector u_j are found. If a corresponding correlation $|r_{jk}|$ is below the critical value rc, the variable k is removed from the regressors set.

For each ith observation (i=1,...,n): after exclusion of the weakly correlated and/or missing regressors, p_{ij} regressors are available for the observation i

$$\hat{x}_{ij} = 1/p_{ij} \sum_{k=1}^{p_{ij}} (s_j/s_k)(x_{ik}-x_{.k}) + x_{.j} \qquad (3)$$

Then the final estimate of x_{ij} is given by the average of p_{ij} estimates from the best correlated variable. It must be pointed out that the coefficient of variable k is the ratio of the concerned standard deviations and not this ratio multiplied by the corresponding correlation coefficient so that the variance of the estimates is preserved.

The ratio of the standard deviations may be replaced by the ratio of the averages. It is better adapted for data which are not normally distributed and must be greater than zero, e.g. the case for the precipitation data. This "normals ratio" method is proposed by Paulhus and Kohler (1952) to estimate the missing precipitation records. Equation (3) then becomes:

$$\hat{x}_{ij} = 1/p_{ij} \sum_{k=1}^{p_{ij}} (x_{.j}/x_{.k})\, x_{ik} \qquad (3')$$

The method, which is based on a comparison of the column of X, is able to extrapolate beyond the limits of the actual data, but, in this case, the prevision errors can be much more important.

2.2. The best analogues method

Now the variables are no longer compared, but the observations are. So an observation vector i is compared to another observation vector k, using this weighted euclidian distance:

$$d^2_{ik} = \sum_{j=1}^{m_{ik}} w_j^2 \, (e_{ij}-e_{kj})^2 \qquad (4)$$

where m_{ik} is the number of observations available for both records i and k, and where w_j is a loading used to weight the influence of the variable j in the distance: these w_j may be either equal to 1 (unweighted case: UNW) or not (weighted case: W);

$e_{ij} = x_{ij}$ (unstandardized case: UNST)

or $e_{ij} = (x_{ij} - x_{.j}) / s_j$ (standardization by the standard deviation: SD)

or $e_{ij} = \log x_{ij}$ (natural logarithm transform: LOG)

or $e_{ij} = x_{ij} / x_{.j}$ (standardization by the mean: MEAN)

Other cases are possible. But these four ones are largely sufficient in most of the analyses. The first one is well adapted when the variables are really homogeneous, and the second one is good in most cases. When the variables are necessarily positive, the third case may be efficient to attenuate the high variations of some variables and the fourth case corresponds better to their probability distribution and is equivalent to the normal ratio method of section 2.1.

A threshold dc must be choosen for this distance. If a tolerance of one standard deviation is acceptable for the variables, we have:

$$dc = \sum_{j=1}^{m} w_j^2 \, s_j^2 \qquad (5)$$

with $S_j^2 = 1$ in the case of the standardization by the standard deviation and is equal to 2 in the logarithm case. This last value is justified by the fact that the variance of the ratio of the logarithm of two random variables tends to 2 when the sample size increases. In the case of the standardization by the mean, the threshold becomes:

$$dc = \sum_{j=1}^{m} w_j^2 \, (S_j / x_{.j})^2 \qquad (6)$$

All the p_i observations analogues to the ith observation, i.e. such as d_{il}^2 is less than dc, are selected in the period [1,nc]. The estimate of x_{ij} is then given by

$$\hat{x}_{ij} = \sum_{k=1}^{p_i} (x_{kj}/d_{ik}^2) \; / \; \sum_{k=1}^{p_i} d_{ik}^{-2} \qquad (7)$$

This method is not an extrapolative method, as it is not possible to obtain estimates beyond the limits of the actual values.

2.3. The multiple regression

The multiple regression may not be applied straightforwardly because the number of regressors is constant neither on the total interval [1,n], nor on the calibration interval [1,nc]. m_j is the number of regressors available for observation i:

$$\hat{x}_{ij} = a_{oj} + \sum_{k=1}^{m_i} a_{kj} x_{ik} \qquad (8)$$

This particular equation is solved iteratively by the partial quadratisation technique (Legras, 1980). The next quantity L must be minimized:

$$L = \sum_{i=1}^{n_j} (x_{ij} - \hat{x}_{ij})^2 \qquad (9)$$

where n_j is the number of not missing observations of the jth variable. It is clear that if the method may be good when a maximum of regressors are available in [1,n], it is efficient when the regressors set is too different from the set used in [1,nc].

2.4. The weighted interpolation after reduction to the sea-level

This method is useful only if the columns of X correspond to a geographical position. Then, the following vectors are associated to X:

$(u_1, \ldots, u_m)$ the vector of the longitudes
$(v_1, \ldots, v_m)$ the vector of the latitudes
$(h_1, \ldots, h_m)$ the vector of the elevations

For each ith observation and for each jth variable: the distances between each column k of X and the jth column are given by:

$$d_{jk}^2 = (u_j - u_k)^2 + (v_j - v_k)^2 \qquad (10)$$

m_i is the number of variables available for observation i and such as

$$d_{jk}^2 < dc \qquad (11)$$

where dc is a radius defining the working zone centered on the point (u_j, v_j).

A regressive equation is computed between x_{ij} and the elevation h_j on the m_i observations. If the degree 1 is selected, the linear gradient g_1 is computed by minimization of L:

$$L = \sum_{k=1}^{m_i} (x_{ik} - g_1 h_k)^2 \qquad (12)$$

If degree 2 is selected, equation (12) is modified to take into account the quadratic gradient:

$$L = \sum_{k=1}^{m_i} (x_{ik} - g_1 h_k - g_2 h_k^2)^2 \qquad (13)$$

If degree 2 is selected and if g_2 does not explain more than 10% of the variance of x_{ik}, g_2 is set to 0. If degree 1 is selected and if g_1 explain less tha 10% of the same variance, g_1 is also set to 0. So the m_i elements x_{ik} are reduced to the sea-level:

$$e_{ik} = x_{ik} - g_1 h_k - g_2 h_k^2 \qquad (14)$$

The element x_{ij} is estimated by this equation:

$$\hat{x}_{ij} = g_1 h_k + g_2 h_k^2 + \hat{e}_{ij} \qquad (15)$$

$$\text{with } \hat{e}_{ij} = \sum_{k=1}^{m_i} (e_{ik}/d_{jk}^2) \; / \; (\sum_{k=1}^{m_i} d_{jk}^{-2})$$

This method is different from the others, because the comparison of climatic variables is based on external variables (coordinates of the variable). It is not an extrapolative method, such as the best analogues method.

3.THE DATA

Four data sets with different characteristics are used for an exhaustive comparison of the methods. The first set is composed by 12 precipitation stations from a homogeneous network. It is supposed to represent the most standard case in the estimation of the missing meteorological data. The second set is composed by 14 tree-ring series which are also relatively homogeneous but their geographical position cannot be translated into correlation terms because other characteristics (e.g. the species) can be more important. The third set is proposed to represent the extrapolation of short climatic series from much longer proxy series. Here important problems may arise from no-analogues situations encountered when the calibration set does not contain all the situations existing in the extrapolation period. The last set is used to represent a very defavourable case: a full disconnection between the calibration space-domain and the extrapolation time-domain. The present time is subject to various space disturbances which did not exist in the past and also to a great risk of no-analogues situation. In each of these typical cases, the best method will be indicated.

Data set 1: The monthly total precipitation in 12 stations of southeastern France between 1871 and 1980 (1320 observations): (1) Aix-en-Provence, 1882-1980, 5°28E, 43°31N, 215 m; - (2) Gardanne, 1892-1980, 5°28E, 43°27N, 283 m; - (3) Salon-de-Provence, 1884-1980, 5°06E, 43°37N, 59 m; - (4) Avignon, 1871-1979, 4°49E, 43°58N, 25 m; - (5) Aubagne, 1880-1980, 5°36E, 44°00N, 127 m; - (6) Orange, 1871-1980, 4°51E, 44°08N, 53 m; - (7) Brignoles, 1900-1980, 6°04E, 43°25N, 215 m; - (8) Trets, 1881-1980, 5°48E, 43°27N, 249 m; - (9) Montelimar, 1921-1980, 4°44E, 44°35N, 73 m; - (10) Nice, 1884-1979, 7°13E, 43°41N, 2 m; - (11) Marseille, 1871-1979, 5°24E, 43°18N, 30 m; - (12) Gap, 1871-1980, 6°05E, 44°36N, 735 m.

Data set 2: The 14 tree ring-width series in Europe back to 1000 A.D.: (1) pines in Calabre, southern Italy (Serre, 1985), 1148-1974; - (2) larches in Les Merveilles, southern Alps (Serre, 1978), 1100-1974; - (3) firs on Mont Ventoux, southeastern France (Serre, unp.), 1660-1975; - (4) larches in Northern Italy (Bebber and Fasani, unp.), 1000-1984; - (5) larches in l'Orgere, western Alps (Tessier, 1981), 1353-1974; - (6) northern Finland (Siren from Lamb, 1977), 1180-1960; (7) oaks in West Rhine, Germany (Hollstein from Lamb, 1977), 1000-1964; - (8) oaks in Spessart forest (Huber and Giertz-Siebenlist from Lamb, 1977), 1000-1949; - (9) oaks in Belfast, North Ireland (Baillie, 1977a), 1001-1970; - (10) oaks in Dublin 1, South Ireland (Baillie, 1977b), 1000-1306; - (11) oaks in Dublin 2, South Ireland (Baillie, 1977b), 1357-1556; - (12) oaks in southwestern Scotland (Baillie, 1977c), 1000-1975; - (13) oaks in Exeter, southwestern England (Hillam, 1980), 1000-1216; (14) oaks in Lockwood, Scotland (Pilcher and Baillie, 1980), 1571-1975.

Data set 3: 19 various proxy-series in Europe and 6 annual mean temperature series: (1-10) the ten tree-ring series of set 2 fully available between 1600 and 1958; - (11) number of weeks each year with drift ice at the coast of Ireland (Koch from Lamb, 1977), 1600-1975; - (12) dates of final opening of the port of Riga counted from January, 1, on the Baltic sea (Betin and Preobazensky from Lamb, 1977), 1530-1958; - (13-16) the thermal indices in Switzerland for the four seasons (Pfister, 1981), 1550-1829; - (17) the dates of wine harvests in northern France (Le Roy-Ladurie and Baulant, 1981), 1484-1879; - (18) the summer temperature index for northern France and Germany (Bray, 1982), 1453-1973; - (19) the dates of wine harvests in Switzerland (Legrand, 1979), 1502-1979; - (20-25) the gridded annual temperature in the northern hemisphere at (0°W, 60°N), (0°W, 50°N), (0°W, 40°N), (20°W, 60°N), (20°W, 50°N), and (20°W, 40°N) derived by Jones et al. (1985), 1851-1984.

Data set 4: pollen and climate (Guiot and Pons, 1986): - (1-26) 26 pollen types observed on 182 modern sites in Europe and on 9 fossil sites in France (244 observations: maximum period = 0 to 18500 BP by steps of 500 years); - (27-34) 8 climate series observed on the same 182 modern sites (mean monthly temperature and total monthly precipitation of January, April, July and Octobre) and to be extrapolated on the 244 fossil time observations.

4. VERIFICATION OF THE ACCURACY

Thirty-eight series of 500 random observations have been taken from independent and normal distributions. Approximatively 50 % of the observations (randomly selected) have been considered as missing. Three methods have been applied to this matrix for esimating these gaps (the interpolation method is not relevant here).

In Table 1, the checking is based on various statistics: the mean residual variance (MSSR) calculated over the m series and the corresponding percentage related to the mean variance of the series (MSSD). These two statistics are biased, because they are calculated on the calibration interval [1, nc]. The other statistics are calculated on the interval [nc+1, n] and then allow a unbiased verification. DM allows to verify the ability of the method to reconstruct the true mean of the series:

$$DM = 100 \left(\sum_{j=1}^{m} (x_{.j} - \hat{x}_{.j})^2 \right) / MSSD \qquad (16)$$

DSD allows to verify the capacity of the method to reconstruct the true variability of the series:

$$DSD = 100 \left(\sum_{j=1}^{m} (s_j - \hat{s}_j)^2 \right) / MSSD \qquad (17)$$

The accuracy of the estimates is also verified by the correlation, for each series, between the actual observations and their estimates. These m correlations are summed up by the minimum (rmin), the maximum (rmax) and the mean correlation (rmean).

Table 1 shows that the three methods do not enable to estimate missing random data, which is fairly natural. Nevertheless some correlations may be relatively high but have no significance. The three methods reconstruct well the mean, but not the variations nor the variability, which is strongly underestimated mainly for the (unweighted and unstandardized) best analogues method. It is understandable since this last method is the only one without any extrapolation ability.

An other important criterion of selection of the method is the computation costs. As an indication, the run time are given for an IBM 3033. The interpolation program (even with a quadratic gradient) is the less time consuming. The multiple regression and the best analogues method are expensive even when, like here, only 10 variables are used as regressor or for defining an analogue.

method	calibration [1,250] MSSR	% MSSR	verification [251,500] DM	DSD	rmin	rmax	rmean	CPU
before	1.00	100.0	-	-	-	-	-	-
Simple regr.	1.09	109.0	4.1	42.0	-0.23	0.14	0.00	3
Multiple regr.	0.88	88.2	2.4	41.0	-0.10	0.52	0.06	20
Analogues (UNW/UNST)			2.0	92.0	-0.14	0.15	0.00	14
Interpolation (deg.2)			-	-	-	-	-	2

TABLE 1. 38 standardized variables taken from a gaussian distribution with 500 observations from which 50 % are randomly missing; see text (this section) for the meaning of the symbols. The CPU time costs are given in seconds for a IBM 3033 computer.

5. HOMOGENEOUS VARIABLES (FROM A NETWORK)

The data set 1 is a typical case of climatological data preparation. The simple regression and the normals ratio method, using 5 regressors and a critical correlation of 0.50, give similar results: a very good reconstruction of the means, the variabilities and the covariations. The interpolation methods reconstruct well the covariations, but not so well the means and the variabilities. The preliminary sea-level reduction (using a second or a first degree gradient) improves the results. The multiple regression is a better method on the calibration period, but the results are quite similar to those of the previous methods on the verification period. The unweighted best analogues method reconstructs at best the covariations but underestimates the variabilities.

Consequently, if the respective time costs of the methods are taken into account, it is better to use either the simple regression (or the normals ratio method for the precipitation) or the weighted interpolation after reduction to the sea-level. This case is also discussed in Guiot (1985a) where it is shown that sparse series with a strong seasonal cycle are better processed using a combination of canonical regression and digital filtering called "spectral canonical regression". A combination of the simple regression and the harmonic analysis is shown to be sufficient to analyse dense network of seasonal data.

method	calibration [1871-1960] MSSR	VR	verification [1961-1980] DM	DSD	rmin	rmax	rmean
before	3113.5	100.0	-	-	-	-	-
Simple regression	992.8	31.9	0.6	0.8	0.61	0.95	0.86
Normal ratio	1036.7	33.2	0.5	0.8	0.64	0.95	0.87
Multiple regression	622.6	20.0	3.4	1.1	0.82	0.95	0.91
Analogues (UNW/SD)	-	-	2.4	5.1	0.88	0.97	0.92
(UNW/UNST)	-	-	2.6	9.7	0.85	0.95	0.91
(UNW/MEAN)	-	-	1.8	5.9	0.80	0.96	0.90
(UNW/LOG)	-	-	0.6	17.4	0.78	0.92	0.86
Interpolation (deg.0)		-	3.3	3.8	0.59	0.95	0.87
Interpolation (deg.1)		-	3.2	3.5	0.78	0.95	0.89
Interpolation (deg.2)		-	3.2	3.6	0.79	0.95	0.89

TABLE 2. 12 precipitation stations in southeastern France (data set 1, section 2); see text (section 3) for explanation of the symbols.

6. EQUIVALENT HETEROGENEOUS SERIES

For the tree-ring series (data set 2), that play an equivalent role in the matrix, three methods only can be tested (Table 3): the simple regression is not well adapted because the individual correlations are not high enough. The multiple regression is better, but, here it is not efficient, because the number of regressors varies too much from an observation to another. Only the unweighted best analogues methods is adaped here, even if the variability is underestimated.

method	calibration [1750-1984]		verification [1000-1749]				
	MSSR	VR	DM	DSD	rmin	rmax	rmean
before	685.0	100.0	-	-	-	-	-
Simple regression	905.1	132.1	42.7	5.0	-0.13	0.84	0.27
Mult. regression	370.0	54.3	135.0	15.3	-0.13	0.91	0.25
Analogues (UNW/SD)	-	-	2.4	24.6	0.50	0.98	0.87

TABLE 3. 14 tree-ring series in Europe (data set 2, section 3); see text (section 4) for the explanation of the symbols.

7. TIME-EXTRAPOLATION OF SHORT TIME-SERIES

This analysis is often known as a tranfer function. Guiot (1985a) has also shown the efficiency of the spectral canonical regression and Guiot (1985b) has applied multiple regression combined to best analogue method to produce reliable reconstructions of some canadian climatic series from historical archives data and tree-ring data. These two methods are now compared on the same kind of data (set 3). Two subsets are selected: the first is composed of 15 long proxy series and the 6 annual gridded temperature series, and the second is composed by the same 15 long proxy series and the 4 thermal index series in Switzerland. Table 4 shows that the covariations are better reproduced with the multiple regression, but that the means are better approached with the best analogues method, which was already a conclusion drawn by Guiot (1985b).

A combination of both methods can be obtained by weighting the proxy series, in equations (5) and (6) with the regression coefficients of equation (8). The weights w_j must reflect the integrality of the variables to reconstruct. Here the regression coefficients for thermal indices of spring, summer and autumn are different from those of winter. Consequently, the weights w_j are set equal to the standardized regression coefficient between regressor j and summer index. In cases where the predictands should be more heterogeneous, a canonical analysis could be used. Table 4 shows that the means, the variabilities and the covariations are the best reconstructed with such a weighted analogues method.

An important characteristic of the best analogues method is that the reconstructed variability decreases with the number of analogues included in the estimates. A correction may be brought by decreasing the critical distance dc. It is really justified when the proxy series are very correlated.

method	calibration		verification				
	MSSR	VR	DM	DSD	rmin	rmax	rmean
6 annual temp.: calibration [1851-1904] - verification [1905-1958]							
before	39.9	100.0					
Multiple regression	17.9	45.0	41.0	6.0	0.07	0.67	0.35
Analogues (UNW/UNST)			28.0	70.0	0.02	0.33	0.22
Analogues (UNW/MEAN)			23.0	48.0	-0.09	0.34	0.22
Analogues (UNW/SD)			20.0	15.0	0.01	0.40	0.14
4 thermal ind.: calibration [1600-1751] - verification [1752-1829]							
before	2.8	100.0					
Multiple regression	1.9	67.6	6.0	64.3	0.01	0.73	0.33
Analogues (UNW/SD)			6.1	77.9	0.01	0.52	0.19
Analogues (W/SD)			1.8	29.6	0.00	0.70	0.37

TABLE 4. 29 long proxy series in Europe from which 6 short annual temperature series and 4 thermal indices of Pfister are to be extrapolated (data set 3, section 3); see text (section 4) for explanation of the symbols.

8. TIME-EXTRAPOLATION OF SPATIAL SERIES

The problem analysed now is the transfer function calibrated in the space domain (182 modern sites in Europe) and applied to time observations. This case is particularly difficult, because of the lack of homogeneity between these two domains. The data are defined and the final results are presented in Guiot and Pons (1986). The discontinuity between the two domains obliges to check also the results in the time domain. Since no true observations are available for that, only the coherency between the 9 reconstructions can be calculated. This coherency is given by the mean of the variance of the 8 climatic variables (calculated over the 244 fossil observations) divided by the corresponding space variance.

The first method is the multiple regression between the 8 climatic variables and the logarithm of the pollen series. This logarithm transform improves strongly the results in this pollen analysis. The second method is the best analogues method using logarithm transform and weights determined on the modern data. These weights are the multiple correlation between the corresponding pollen type and the climatic variables. This way to compute the weights does not allow to remove the heterogeneity between the two domains. So a third method uses weights computed only on the extrapolation time-domain. So these weights are given by the correlation, for the concerned pollen type, between two groups of contemporary observations from different sites. More explanations over these two kind of weights are given in Guiot and Pons (1986).

Table 5 shows that, in the space domain, the analogues method gives the best results. In all the cases, the variability is underestimated. The use of "modern weights" seems to improve the reconstruction in the space domain, but it is paid by a loss of coherency in the time domain. The use of "fossil weights" seems to optimize the time-coherency. Finally, in this type of transfer function, the best results are obtained by the best analogues method using a logarithm transform and weights based over extrapolation domain.

method	calibration [92,182]		verification: [1,91]					time
	MSSR (T)	MSSR (P)	DM	DSD	rmin	rmax	rmean	variance
before	2859.0	128147.0						
Mult. regr.	431.7	14993.2	20.3	6.0	-0.02	0.51	0.29	71 %
Analogues (modern weights/LOG)			10.5	11.2	0.11	0.62	0.33	75 %
Analogues (fossil weights/LOG)			22.5	20.5	-0.01	0.55	0.47	55 %

TABLE 5. 26 pollen series and 8 climate series: 182 modern observations in Europe and 244 fossil observations (data set 4, section 3); see text (section 4) for explanation of the symbols; the variance of the time-extrapolations is checked on the 244 fossil observations.

9. CONCLUSIONS

Four typical cases of missing data estimation have been analyzed. In the case of a climatological network, the less expensive methods (simple regression, normal ratio method, weighted interpolation after reduction to the sea-level) are largely sufficient. In the case where the series have an equivalent role, but not extracted from a network and not highly correlated, multiple regression is adequate. When too many data are missing, the best analogues method is more efficient. For extrapolating either time-series or space series in the time- domain, multiple regression reconstructs at best the high frequencies, and the best analogues method is more efficient in the low frequencies. A combination of both methods, i.e. by the use of weights based on a multiple regression to determine an analogue improves largely the results. In the case of a full discontinuity between the calibration domain and the extrapolation domain (e.g. for the reconstruction of the Quaternary climate), the use of weights based on the extrapolation domain improves largely the coherency between the reconstructions from various series.

BIBLIOGRAPHY

Baillie, M.G.L., 1977a: The Belfast oak chronology to A.D.1001. Tree-Ring Bulletin, 37, 1-12.

Baillie, M.G.L., 1977b: Dublin medieval dendrochronology. Tree-Ring Bulletin, 37, 13-30.

Baillie, M.G.L., 1977c: An oak chronology for South Central Scotland. Tree-Ring Bulletin, 37, 33-44.

Bray, J.R., 1982: Alpine glacial advance in relation to a proxy summer temperature index based mainly on wine harvest dates, A.D.1453-1973. Boreas, 11, 1-10.

Goodin, W.R., McRae, G.J., Seinfeld, J.H., 1979: A comparison of interpolation methods for sparse data: application to wind and concentration fields. Journal of Applied Meteorology, 18, 761-771.

Guiot, J., 1985a: The extrapolation of recent climatological series with spectral canonical regression. Journal of Climatology, 5, 325-335.

Guiot, J., 1985b: Reconstruction of seasonal temperature and sea-level pressure in the Hudson Bay area back to 1700. Climatological Bulletin, 19, 11-59.

Guiot, J. and Pons, A., 1986: Une méthode de reconstruction quantitative du climat à partir de chroniques pollenanalytiques: le climat de la France depuis 15000 ans. C.R. Acad. Sc. Paris, T.302, série II, 911-916.

Hillam, J., 1980: A medieval oak tree-ring chronology for southwest England. Tree-Ring Bulletin, 40, 13-22.

Jones, P.D., Raper, S.C.B., Santer, B., Cherry, B.S.G., Goodess, C., Kelly, P.M., Wigley,T.M.L., Bradley, R.S., Diaz, H.F., 1985: A grid point surface air temperature data set for the northern hemisphere. U.S. Dept. of Energy, Washington, TR022.

Lamb, H.H., 1977: Climate: present, past and future. Methuen, London, 825pp.

Legrand, J.P., 1979: Les variations climatiques en Europe Occidentale depuis le Moyen-Age. La Météorologie, 16, 167-182.

Legras, J., 1980: Algorithmes et programmes d'optimisation non linéaire avec contraintes. Masson, Paris, 188pp.

Le Roy Ladurie, E., Baulant, M., 1981: Grape harvests from the 15th through the 19th century. Sudies in Interdisciplinary History, R.I. Rotberg and T.K. Rabb (eds), Princeton University Press, 259-269.

Paulhus, J.L.H., Kohler, M.A., 1952: Interpolation of missing precipitation records. Monthly Weather Review, 80, 129-133.

Pfister, C., 1981: An analysis of the Little Ice Age climate in Switzerland and its consequences for agricultural production. Climate and History, Wigley et al. (eds), 214-248.

Pilcher, J.R., Baillie, M.G.L., 1980: Eight modern oak chronologies from England and Scotland. Tree-Ring Bulletin, 40, 23-34.

Serre, F., 1978: The dendroclimatological value of the european larch (Larix decidua Mill.) in the French Maritime Alps. Tree-Ring Bulletin, 38, 25-33.

Serre-Bachet, F., 1985: Une chronologie pluriséculaire du Sud de l'Italie. Dendrochronologia, 3, 45-66.

Tabony, R.C., 1983: The estimation of missing climatological data. Journal of Climatology, 3, 297-314.

Webb III, T., Clark, D.R., 1977: Calibrating micropaeontological data in climatic terms: a critical review. Annal of the New-York Academy of Sciences, 288, 93-118.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

EXTRACTING INFORMATION FROM RAINFALL DATA

K.J.A. Revfeim, Meteorological Service, Wellington, N.Z.

Abstract

As some time-space models of the rainfall process become more elaborate so the number of existing data sets to which these models can be applied become more restricted. The apparent tendency is for sophisticated models tailored to particular data sets obtained for calibration of those models.

On the other hand the simplest Compound Poisson process model has not been so widely applied that the component characteristics of recurrence rate and mean amount have been fully assessed. How well do these characteristics represent a rainfall regime and what can be deduced from the components that is not available from simple statistics of rainfall totals?

1. Introduction

Being the simplest of all environmental measurements rainfall is observed at more locations than other phenomena. Even so the areal sampling rate is very small considering the amount of information that is extracted from records, and may be restrictive for testing certain hypotheses.

A survey of member countries by WMO (Operation Hydrology Report No. 10, 1977) shows the number of manual raingauges as just over 120,000 for a land surface area of 106 million km^2 (about 2/3 of the total land surface area). Many countries not furnishing a return to WMO could be described as developing and are likely to have a less dense network of rainfall stations. With an intercepting area of between 0.025 m^2 (5 inch diameter) and 0.1 m^2 the areal sampling is of order 106 x $10^{12}/(120,000 \times 0.05)$ or less than 1 in 10^{10}.

It is a salutary exercise to consider that all the rain gauges on earth would barely cover a football ground. Yet this data source is only used to characterise rainfall over land or less than 30% of the earths total surface area.

For any one site we usually have a record of daily (00-24, 07-07, 09-09 hrs etc) and monthly totals (sometimes left blank in published records because of missing readings despite little loss of information from what could be a tight lower bound). The extensive literature on rainfall analyses is dominated by immediate recourse to statistical manipulations on raw data. Such statistical models reveal very little about the processes generating the daily rainfall amounts or monthly totals. An alternative model that displays some elements of the rainfall mechanism has parameters which can be related to an intuitive understanding of a rainfall climate (Buishand, 1976, Revfeim, 1982, Alexanderson, 1985). Additional parameters with physical

meaning can be added to the basic model, but not without some mathematical and computational difficulties.

Parameters of some long term series are investigated with possible uses for testing hypotheses of broad scale climatic progression. Mathematical problems associated with development of more realistic models are briefly sketched. The potential of complementary climate data to improve model fitting will also be considered.

2. Hypothesis testing

The following quotation conjectures a possible relationship the identification of which would enhance any climatologists reputation (however, the 'noise' of local variation in observations usually mask elusive global 'signals').

"If, as may be easily imagined, a definite relation exists between the seasons of the two hemispheres and different countries, their correlation must be sought for by inductive and deductive methods from both the records of meteorological phenomena and solar physics. The sun is the dominant factor in all climatic changes, and sun-spots show that it itself is subject to great variations. Other alterations may go on within the luminous veil which are invisible to us, but their effects may be radiated and possibly be traced in the vegetation, &c. When the sun crosses the Equator and leaves that side of the hemisphere where the land is more extensive it may carry forward conditions and continue effects which are modified and not so evident when the insolation spends itself on the wider oceans of the south." (Bates, 1905)

Given long term rainfall series at several sites on the globe what structural parameters should one use to test this hypothesis? A typical analysis would involve the plotting of n-year moving-averages and a search for elusive trends or cycles, before going on to spectral analysis of the smoothed series (or first or second differences). Believing that there must be information in seasonal patterns a second approach is to carry out a principal-component (P-C) analysis of the vectors of monthly totals. This P-C analysis might be tried on the same 'window' as used for smoothing the annual totals to give a series of eigen values or over the whole data set to give a series of 'scores' for the dominant principal component (Nicholls, 1985).

In contrast to these two empirical approaches the Compound Poisson (C-P) model can be fitted over the same data window to give several time-series of physically meaningful parameters. There seems to be a better prospect of indentifying a relationship in time at widely spaced locations between a series of parameters with the same physical meaning than between series of principal components whose interpretation may differ at each location.

For example consider the 207-year rainfall record for Milan (45°N, 9°E 1764-1970) available from World Weather Records (1927, 1934, 1947, 1959, 1965, 1979). C-P model analysis of the rainfall record in disjoint 30 year blocks gives fairly consistent parameter estimates as shown in Table 1. With no suggestion of gross inhomogeneity the C-P model was fitted over an 11-year moving window with 2 harmonics to represent seasonal variation (i.e. 10 parameters estimated from 11 x 12 = 132 data values).

Table 1. Parameters of first 2 harmonics of C-P model fitted to monthly rainfall data at various locations.

Location	Period	r	a_{r1}	p_{r1}	a_{r2}	p_{r2}	m	a_{m1}	p_{m1}	a_{m2}	p_{m2}
Milan	1764-1790	5.3	1.3	191	1.9	41	15.6	3.0	37	3.3	168
	1791-1820	5.4	0.6	163	1.8	43	16.0	2.6	64	1.8	158
	1821-1850	4.5	1.4	161	1.1	79	20.7	5.1	13	1.8	220
	1851-1880	4.5	1.2	185	1.8	56	20.1	5.7	48	4.9	173
	1881-1910	4.8	1.1	198	1.2	104	18.7	4.2	29	1.6	75
	1911-1940	4.9	0.9	204	0.7	87	16.9	3.2	31	2.5	105
	1941-1970	4.8	1.3	157	1.7	27	16.9	3.0	-23	3.5	160
	1764-1970	4.6	1.0	177	1.2	57	18.4	3.7	30	2.4	157
Oslo	1866-1970	5.6	0.9	147	0.6	-31	9.4	3.6	90	0.7	181
Tunis	1887-1960	3.3	2.0	-36	0.4	99	10.2	3.6	7	2.0	128
Calabar	1898-1960	10.7	5.3	137	5.3	112	24.1	7.6	145	7.7	-79
Dunedin	1852-1984	6.8	1.6	0	0.9	63	12.0	2.8	181	1.5	-54
Montreal	1874-1970	11.7	2.5	-28	2.6	-35	7.7	1.8	131	1.1	158

r = recurrence rate (events/month), m = mean amount (mm/event), 'a' amplitudes, 'p' phase angles

One or more of the 10 time series may show an identifiable pattern which cannot be detected in one series of smoothed totals. A pattern or significant spectral frequency may be detectable in the product rm, which represents the mean total with seasonal effects removed. Series for r, a_{r1}, a_{r2} and m are shown in Figure 1. Spectra of a_{r2} and a_{r4} showed significant peaks with a period of 30 and 20 years respectively. The mean rate r gave a non-significant peak for a 30 year period.

Are these local or global effects? The rainfall record for Oslo, with similar longitude but latitude almost 15° further north, is given in WWR for 1866 - 1970. Unlike the Milan rainfall regime the first harmonics of recurrence and amount for Oslo are nearly in phase as shown in Table 1. Seasonal model parameters were estimated over the same 11-year window and those identified as having a likely pattern in the Milan rainfall regime were compared. There was no apparent correlation between

the mean rates, r, and mean amounts m. There was negative correlation between the amplitudes of the first harmonics of -0.44 for a_{r1} and -0.22 for a_{m1}.

WWR data for Tunis (37°N) and Calabar (5°N) at similar longitude were hoped to be used for further comparisons but there were difficulties in obtaining parameter estimates with the 11-year window (Whole record parameter estimates are shown in Table 1). The nearest to an antipodean site Dunedin, N.Z. (46°S, 170°E) and a near 90° shift to Montreal (45°N, 74°W) were also investigated and the whole record rainfall parameters are shown in Table 1. The 11-year window provided smoothly varying parameters and for Milan and Montreal the amplitudes of the first harmonics were positively correlated; 0.42 for a_{r1} and 0.23 for a_{m1}.

3. Model extensions

The elementary C-P model representing rainfall as a point process is easy to fit to data using maximum likelihood procedures for parameter estimation. The iterative procedures benefit from the use of a reasonable approximation to the Bessel function component of the frequency function. This greatly reduces computation and similar properties are sought in the development of more realistic models.

A duration parameter has been added to the C-P model with possibly over lapping rectangular pulse (uniform intensity) storms (Rodriquez Iturbe et al, 1984). A model for non-over-lapping storms with independent occurrence, amount and duration is not difficult to specify as an explicit propability distribution for totals. However, the functional form needs to be expressed as a 2-parameter approximation for computationally simple iterative procedures. The frequency function of totals is:-

$$h(x) = x^{-1}\, e^{-\rho - x/\mu} \sum_{r=1}^{\infty} (\rho x/\delta\mu)^r / r!(\Gamma r)^2 \int_0^1 \theta^{r-1}\,(1-\theta)^r\, e^{(\rho - 1/\delta)\theta}\, d\theta$$

$$= x^{-1}\, e^{-\rho - x/\mu}\, A(y)$$

c.f. the C-P model where $A(y) = y\, I_1(y)$, $y = 2(\rho x/\mu)^{1/2}$

The assumption of an exponential distribution of rainfall is not realistic since it implies that the most frequent storm amount is infinitessimally small. As shown by Eagleson (1978) it is not analytically difficult to include more modal gamma distributions underlying totals leading to the frequency function

$$h(x) = x^{-1}\, e^{-\rho - \lambda x/\mu}\, A_\lambda(y)$$

where $A_\lambda(y) = \sum_{r=1}^{\infty} y^{\lambda r}/r!\Gamma\lambda r, \quad y = \rho^{1/\lambda}\lambda x/\mu$

Initial estimates can be obtained from $\bar{x} = \rho\lambda\mu$, variance $v = \rho\lambda(\lambda+1)\mu^2$ and third moment $\rho\lambda(\lambda+1)(\lambda+2)\mu^3$. However

for maximum likelihood estimation it is desirable to derive an approximation for $A'_\lambda(y)/A_\lambda(y)$ in the form of an inverse polynomial similar to $I_0(y)/I_1(y) \approx (1+1/2y)(1+1/2y^2)$ for $\lambda = 1$.

$A_\lambda(y)$ is not a standard transcendental function and it can be shown that A'_λ/A_λ tends to $(\lambda/y)^{1/(\lambda+1)}$ for large y. As should lie in a limited range (1 to 2+) an approximate form for A'_λ/A_λ might only involve simple linear expressions in λ.

One further matter that does not seem to have been addressed is data bounds. Parameter estimation procedures where some data are given as bounds are easily developed for extremes but are not so simple for totals. A solution for this problem would be useful as so many totals are recorded as missing wheras the information is almost complete (only a few days data missing). It would also increase the utilizability of short incomplete records as 5-10 years of monthly data can form a representative sample for a rainfall regime (Revfeim, 1985).

4. Complementary information

Perhaps the routine climate observation that is most closely related to rainfall, and for which integrated monthly values are given, is sunshine hours. Starting with the simplest specification of randomly distributed interceptions of direct sunlight by cloud during daylight hours a distribution of total sunshine is derivable.

Thus for about half the time that rainfall is possible there is a negatively associated variable which could be used in a joint model. At least conceptually the information in sunshine should lead to more precise estimates of the rainfall component parameters.

Development of the sunshine model above is not without mathematical and statistical difficulty. However, such a sunshine model may also lead to a better specification of the solar radiation process with improved characterisation of the quality or utilizability of radiation (Gordon and Zarmi, 1983).

Finally it is inevitable that satellite information on cloud cover may be used alone and in parallel with rainfall data for atmospheric process models when sufficient data have been obtained - about 50 years from now for long term patterns!

5. Summary

The simplest model of the rainfall process has physically meaningful parameters which have potential for observing patterns in rainfall climates, comparing different regions, and predicting or simulating extreme situations. Improvement of these models could be a field for fruitful research.

6. References.

Alexanderson, H., 1985: A simple stochastic model of the precipitation process. J. Climate Appl. Meteor, 24, 1285-1295.

Bates, D.C., 1905: Meteorological Report for October, New Zealand Gazette.

Brazdil, R., Samaj, F., and Valovic, S., 1985: Variation of spatial annual precipitation sums in Central Europe in the period 1881-1980. J. Climatology, 5, 612-631.

Buishand, T.A.,1977: Stochastic modelling of daily rainfall sequences. Meded. Landbouwhogesch. Wageningen, 77-3.

Eagleson, P.S., 1978: The distribution of annual precipitation derived from observed storm sequences. Water Resour. Res., 14, 713-721.

Gordon, J.M., and Zarmi, Y., 1983: The utilizability function_1: theoretical development of a new approach Solar Energy, 31, 529-536.

Nicholls, N., 1985: Towards the prediction of major Australian droughts. Aust. Met. Mag, 33, 161-166.

Revfeim, K.J.A., 1982: Comments 'On the study of a probability distribution for precipitation totals'. J. Appl. Meteor., 21, 1942-1945

_____1985: A comparison of theoretical and empirical quantiles for monthly rainfall totals. Atmos. - Ocean,23,

Rodriguez-Iturbe, I., Gupta, V.K., Waymire, E., 1984: Scale consideration in the modeling of temporal rainfall. Water Resour. Res., 20, 1611-1619.

WMO No. 464. 1977: Statistical information on activities in operational hydrology. Report No. 10, WMO, Geneva.

World Weather Records, Smithsonian Misc. Collns,
(Ed .H.H. Clayton)

_____ Vol. 79, -1920, Smithsn. Instn., Washington, 1927.
_____ Vol. 90, 1921-1930, " " " , 1934.
_____ Vol 105, 1931-1940, " " " , 1947.
_____ 1941-50, Weather Bureau.
U.S. Dept. of Commerce, Washington, 1959.
_____ 1951-1960, Environmental Data Service.
U.S. Dept. of Commerce, Washington, 1965.
_____ 1961-1970, National Oceanic and Atmospheric Admin.
U.S. Dept. of Commerce, Washington, 1979.

Richman, M.B., and Lamb, P.J. 1985: Climatic pattern analysis of 3- and 7- day summer rainfall in the central United States. J. Climate Appl. Meteor., 24, 1325-1343.

Guzman, A.G., and Torrez, W.C., 1985: Daily ranfall probabilities conditional upon prior occurrence and amount of rain. J. Climate Appl. Meteor., 24, 1009-1014.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

STATISTICS OF THE POLLUTANT CONCENTRATIONS DUE TO SINGLE SOURCES AND APPLICATIONS TO POLLUTANT DISPERSION NUMERICAL MODELLING

Sergiu Tumanov
Institute of Meteorology and Hydrology
Bucharest Romania

INTRODUCTION

One of the most important applications of the air pollution dispersion models lies in estimating, on account of the climatic features of an area, the statistical characteristics of the concentrations due to industrial facilities prior to their being in operation. The statistical characteristics thus assessed are compared to the air quality standards which helps anticipating the degree of alteration in the purity of atmosphere over the area being investigated. This application is supported by the acceptance of the hypothesis that dispersion models are suited to render, with sufficient accuracy, the required statistical characteristics. Still there are some doubts as far as the performances of the dispersion models in the range of upper quantiles are concerned. Some of these objections,based on the comparison of estimated to measured concentrations were gathered in a previous paper (Tumanov, 1983). Among other authors, Cats (1978) found out that the matching between measured and estimated values decreases towards the upper quantiles,i.e. in the very range of greatest application interest. The quoted author explained this limitation by that the Gaussian model is devised for stationary homogeneous conditions while the upper quantiles correspond to extreme weather conditions occurring only a few days in a year. His conclusion, as in fact of other authors too, is that the Gaussian model is suited only for the estimation of long-term average concentrations. Ludwick et al.(1980), as a result of a measuring programme performed in the surroundings of a coal-fired power plant in U.S.A., noticed that the dispersion models used at present underestimate the frequencies of the concentrations due to thermal power stations in points remote from the source in rough terrain.

Considering these findings, the idea seems reasonable that the use of dispersion models should be confined to the estimation of those statistical characteristics for which accurate estimation can be obtained, i.e. the arithmetic mean and the mean square deviation, other statistical characteristics, quantiles included, having to be computed by means of a suitable probability law (Tumanov, 1983).

THEORETICAL DISTRIBUTIONS APPLICABLE TO CONCENTRATIONS DUE TO SINGLE SOURCES

The idea put forward in the end of the previous section is not a new one. Venkatram (1979), for instance, quoted Tikvart and Freas who had suggested in 1977 the estimation of maximum and second highest concentrations making use of probability laws consistent with frequency distributions of the concentrations provided by dispersion models. In particular, the exponential distribution was found to describe with sufficient accuracy the statistical behaviour of the concentrations due to single sources.

A similar idea was forwarded by Cats(1978), who, without recommending a probability law proper, suggests estimating the o.98-quantile from the long-term arithmetic mean on account of the finding that the ratio between $C_{0.98}$ and the average is approximately constant.

Interest in this field was shown by Camps and Bonometti (1974)as well; they found that in points situated under the influence of single sources the beta distribution is applicable.

The use of experimental data provided by a diffusion experiment carried on in a thermal power station site in Romania (Moţiu et al., 198o) helped testing the applicability of the Eggenberger-Polya discrete distribution(Tumanov, 1979). This was found to approximate sufficiently well the empirical distribution of sulphur dioxide hourly concentrations. Denoting by C the pollutant concentration and assuming that it is an integral quantity (an assumption which may be accepted if the concentration is expressed in conveniently chosen metric units),the Eggenberger-Polya cumulative probability is given by the sum

$$P[C<c] = \sum_{C=o}^{c} p(C) \qquad (1)$$

where

$$p(C) = \frac{m(m+d)(m+2d)\ldots(m+[C-1]d)}{C!(1+d)^{m/d+C}}, \quad C = 1,2,\ldots$$

$$p(O) = \frac{1}{(1+d)^{m/d}} \qquad (2)$$

where $d = s^2/m-1$, m being the arithmetic mean and s^2 the variance of the concentrations (Brooks and Carruthers, 1953). The probability that a given concentration C should occur is equivalent to the occurrence of C units of the concentration, the latter not being independent events but affected by the persistence factor d. A simple recurrence formula between the probabilities p(C) makes the electronic computation for the sum (1) not too lengthy even for large values of the concentration.Mention should be made, though, that the Eggenberger-Polya probabilities are not absolutely invariant to the change of metric unit; it is, therefore, advisable to use $\mu g\ m^{-3}$, a unit for which the tests proved satisfactory.

Figures 1 and 2 show the results of a rank-statistics based test (Romanian National Water Council - Report, 1981) on the applicability of the Eggenberger-Polya probability law to the hourly SO_2 concentrations measured in different points located around the thermal power station mentioned above (Moţiu et al., 198o; Tumanov, 1979). The same test was applied with promising results to some short series of 3-hour average concentrations of

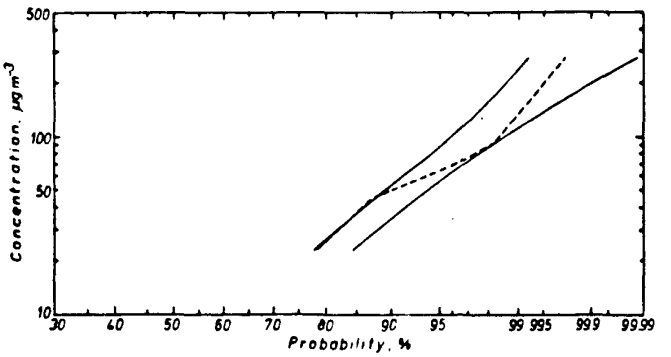


Fig.1 Confidence domain (5%-significance level) for the Eggenberger-Polya probability law (continuous lines) adjusted to the empirical distribution function (dotted line) of the hourly SO_2 concentrations (Point 9; 12.8 km from the source)

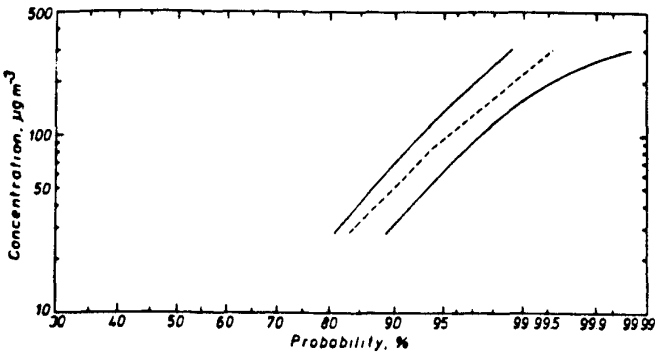


Fig.2 Confidence domain (5%-significance level) for the Eggenberger-Polya probability law (continuous lines) adjusted to the empirical distribution function (dotted line) of the hourly SO_2 concentrations (Point 1o; o.5 km from the source)

sulphur dioxide measured in the vicinity of another power plant as well as to a relatively long series of average concentrations over the same time interval in an urban area situated under the direct influence of a single sulphur dioxide source (RNWC - Report, 1981).

APPLICATIONS OF THE EGGENBERGER-POLYA DISTRIBUTION

To resume the idea of confining the application of a dispersion model (the Gaussian one in particular) to the assessment of statistical characteristics m and s, one can now specify that the Eggenberger-Polya probability could be an adequate tool in assessing the other statistical characteristics required in the case of single sources.

In principle, there should be restrictions on the procedure applicability according to the designed source location in either an environment with pre-existing pollution background (urban environment) or an unpolluted environment. In the former case, the use of an adequate statistical law such as the log-normal distribution should be desirable instead of the probability (1) while in the latter case the Eggenberger-Polya probability should be resorted to. On another hand, in many cases the influence of the designed source (or sources) upon larger areas is also of interest as they may cover urban -together with rural- type portions. Therefore the computation programme should switch over from one probability law to another which would result not only in operative difficulties, but also in undesirable discontinuities at the urban-rural boundary.

In an attempt to avoid this type of difficulties the Eggenberger-Polya distribution was tested in two points situated in an urban zone, where the log-normality of the measured SO_2 concentrations (average daily values) was proved. It was found that the Eggenberger-Polya distribution approximates with sufficient accuracy the empirical distributions in such cases, too. In other words, the distribution proved flexible equally fitting the frequency distributions with the mode situated at the end of the range (J-shaped) and the distributions with the mode within the range of the concentration (RNWC-Report,1983).

As there have been so far few tests in urban environment with the Eggenberger-Polya distribution they cannot yet represent the basis of a final conclusion. It seems reasonable though, that the procedure might possibly be applied without restrictions, i.e. in both cases of single and multiple source.

The results of an application of the suggested procedure are presented in Figures 3 to 8. The ammonia concentrations due to 13 sources of 4o to 7o m height with emission rates of 1 to 12 g s^{-1} placed within a designed chemical works site are analysed. The concentrations in 1 km- (o.5 km- within a smaller square centred on the works site)spaced points of a grid were calculated through the Gaussian plume model.

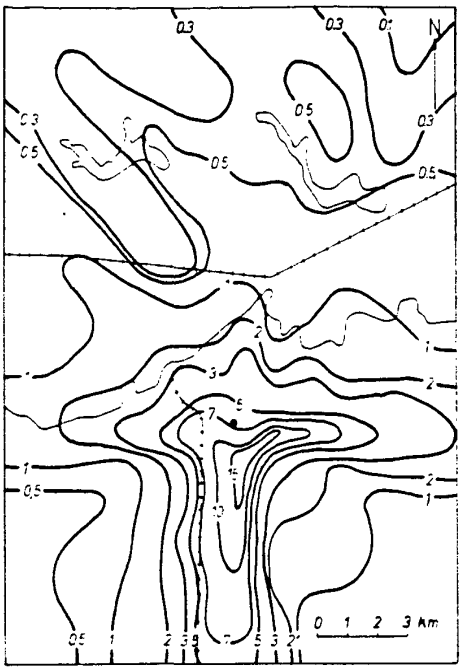


Fig.3 Arithmetic mean of the concentrations, $\mu g\ m^{-3}$

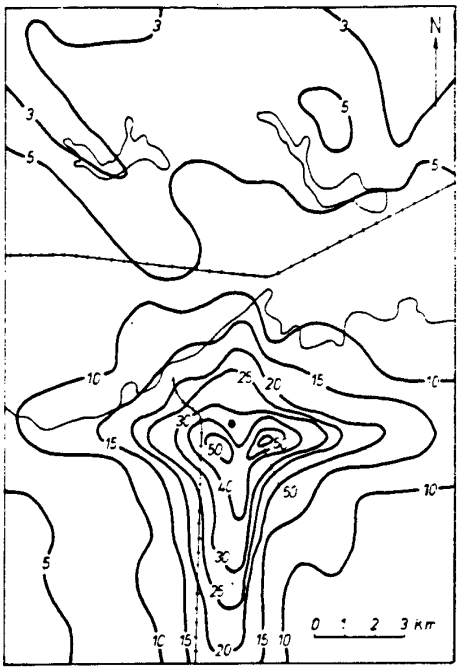


Fig.4 Mean square deviation of the concentrations, $\mu g\ m^{-3}$

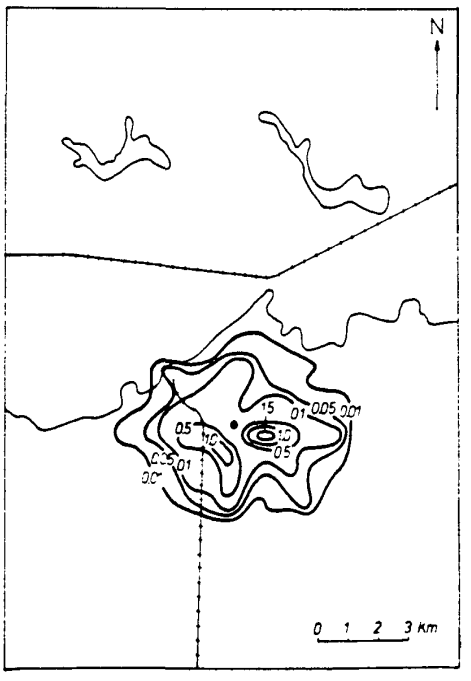


Fig.5 Frequency of equalling or exceeding the 300$\mu g\ m^{-3}$ concentration computed by means of the Gaussian model

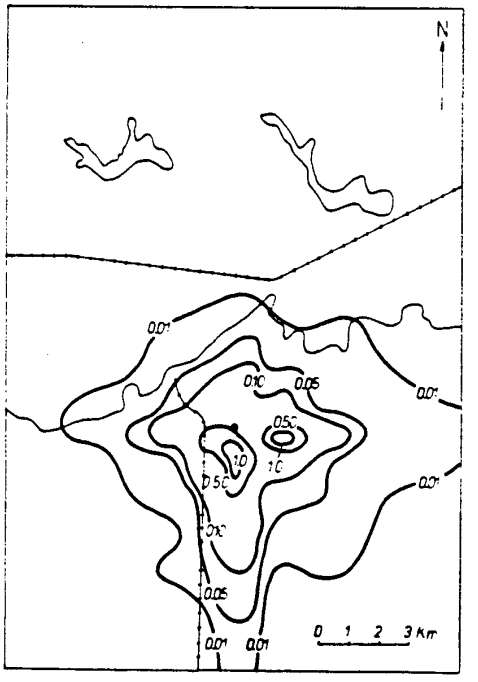


Fig.6 Frequency of equalling or exceeding the 300$\mu g\ m^{-3}$ concentration computed from the Eggenberger-Polya probability law

The long-term arithmetic mean and the mean square deviation are shown in Figures 3 and 4, respectively. The expected probabilities of equalling or exceeding 300 $\mu g\ m^{-3}$ concentration (the 30-minute air quality standard for ammonia in Romania) estimated through the Gaussian model and computed from the Eggenberger-Polya probability law are shown in Figures 5 and 6, respectively. It should be noted that the Eggenberger-Polya probabilities are greater then those estimated by means of the Gaussian model for locations remote from the sources while the probabilities for locations close to the sources are almost identical. This observation does not mean, of course, a validation of the procedure of the Eggenberger-Polya probability as there is not a comparison to experimental data but a possible index that the procedure suggested is appropriate to yield more realistic values, moreover if the observation forewarded by Ludwick et al., that dispersion models underestimate frequencies in points more distant from the source, were considered.

At last, two quantiles computed through the Eggenberger-Polya distribution, i.e. those of o.999 and o.999943 ranks are shown in Figures 7

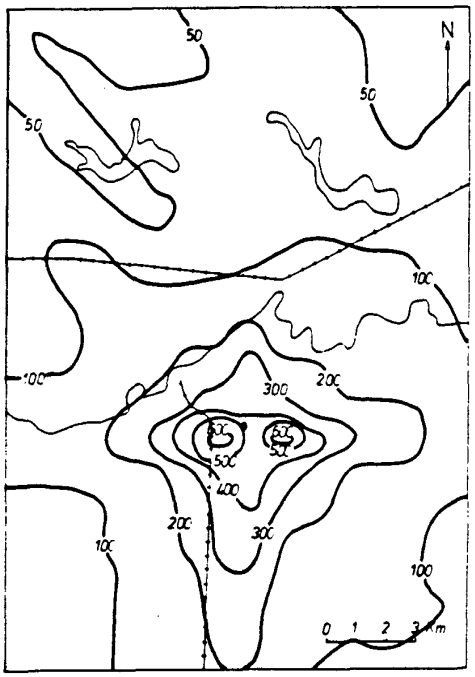


Fig.7 The o.999-quantile of the concentration distribution ($\mu g\ m^{-3}$), computed from the Eggenberger-Polya probability law

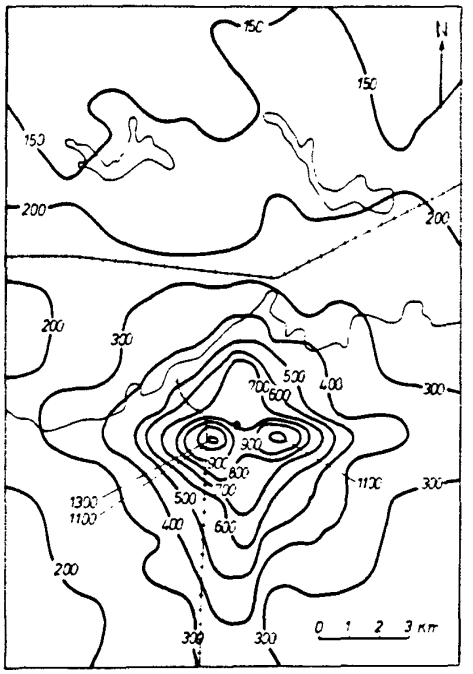


Fig.8 The o.999943-quantile of the concentration distribution ($\mu g\ m^{-3}$) (corresponding to the frequency of 1 occurrence a year), computed from the Eggenberger-Polya probability law

and 8. The latter rank was chosen according to the way of setting an air quality standard; a frequency of once a year is assigned to the 30-minute annual maximum which means a frequency of equalling or exceeding the maximum of 1/1752o=o.oooo57 (there are 1752o intervals of 30 minutes within a year), the complementary frequency being o.999943. Following the 300 $\mu g\ m^{-3}$ isopleth in Fig.8 one can easily notice that it has essentially the same configuration as the o.o1% isopleth in Fig.6 which comes out obviously as the o.o1% probability to equal or exceed the 300 $\mu g\ m^{-3}$ concentration is but the rounded value of the o.oooo57 frequency transformed into percents of one occurrence a year. The same figure shows two nuclei of maximum concentrations out of which one reaches 1300 $\mu g\ m^{-3}$ while the concentrations computed directly from the Gaussian model do not exceed 750 $\mu g\ m^{-3}$.

Fig.7 shows the area where the 300 $\mu g\ m^{-3}$ concentration is exceeded, i.e. with a probability of 1-0.999=0.001 or 0.1%(cf. the 0.1% isopleth in Fig.6).

As the computation programme supplies a set of 16 quantiles, there is a possibility to assess the areas where the critical concentration is expected to be equalled or exceeded with given probabilities.

CONCLUSIONS

As a result of the tests performed it may be stated that the Eggenberger-Polya probability law approximates with sufficient accuracy the empirical distributions of the concentrations due to air pollution single sources.

This law will help calculating some statistical characteristics of the concentrations after the long-term arithmetic mean and mean square deviation were estimated by using a dispersion model, e.g. the Gaussian plume model. The estimations thus obtained are plausible in magnitude and seem to be more realistic than those directly supplied by models.

The results presented in this paper cannot stand for a validation of the procedure as the experimental data available were not sufficient. However, it can be taken as being illustrative to a significant extent of the capabilities of the Eggenberger-Polya probability law as to the statistical treatment of the concentrations due to single sources in view of comparing the concentrations to the air quality standards.

References

Brooks, C.E.P. and Carruthers, N., 1953: Handbook of statistical methods in meteorology. Her Majesty's Stationary Office, London, 412 pp.

Camps, M. and Bonometti, G., 1974: Modèle pour la représentation de la pollution sur le site de LACQ. 11-ème Colloque international sur les atmosphères polluées, Paris, 13/1-10.

Cats, G.J., 1978: A simple method for quick estimates of frequency distribution of air pollution concentrations from long-term average concentrations. WMO Symposium on Boundary Layer Physics Applied to Specific Problems of Air Pollution, Norrköping, 19-23 June 1978, WMO No. 510, 157-162.

Ludwick, J.D., Weber, D.B., Olsen, K.B. and Garcia, S.R., 1980: Air quality measurements in the coal-fired power plant environment of Colstrip, Montana. Atm.Env., 14, 523-532.

Moţiu, C., Romanof, N. and Schmidt, R., 1980: Studiul dispersiei SO_2 în zona centralei termoelectrice Mintia. (Dispersion study of SO_2 in the area of Mintia thermal power plant. Romanian original). Energetica, 28, 493-495.

Tumanov, S., 1979: Frequency distributions of gas pollutant concentrations measured in the neighbourhood of a high single source. Meteorology and Hydrology, Bucharest, 2, 11-16.

Tumanov, S., 1983: Aplicaţii ale metodelor statistice la evaluarea gradului de poluare a aerului (Applications of statistical methods to the assessment of air pollution level. Presented in the annual scientific session of IMH) (Romanian manuscript).

Venkatram, A., 1979: Applications of pollutant frequency distributions. J.Air.Poll.Contr.Assoc., 29, 251-253.

Raport CNA, 1981: Cercetări asupra poluării aerului datorită surselor multiple (Study on air pollution due to multiple source. IMH, Bucharest, Chap.5)(Romanian manuscript).

Raport CNA, 1983: Cercetări asupra poluării aerului datorită surselor multiple (Study on air pollution due to multiple source. IMH, Bucharest, Chap.3)(Romanian manuscript).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

HYDROMETEOROLOGICAL DATA AND STATISTICAL ESTIMATION CHECKING UP BY MEANS OF STOCHASTIC MODELS

Ionel Zlate
Institute of Meteorology and Hydrology
Bucharest - Romania

The classical statistical computation and the extrapolation of the information from a sample outside the observation period is connected with the use of a risk criterion, such as the average recurrence time (Kite, 1976). In this respect, an event having a "T" average recurrence time, longer than the life time "L" of the economic object designed, may occur even during the "L" duration. A certain event, against which the dimensioning is achieved, must not occur during the "L" period or the risk of its occurrence should be quite reasonable - o.1 as a rule.

Information extrapolation based on a distribution curve supposes a previous analysis of the sample,its elements being independent and the series being representative, consistent with the conditions of homogeneousness, stationarity and consistency.

The study of stationarity, the second order stationarity, is generally sufficient.

As the statistical estimations through distribution curves involve a certain degree of error, and additional specifications starting from the same sample finally depend on the sample restricted volume, the problem arising is the necessity of expanding the information contained in the sample either by means of some different variables or by the wider use of a complete series. This paper considers extrapolating hydrometeorological time series of maximum annual values, up to T = looo years.

For the study of stationarity the trend, the correlogramme and the continuous spectrum of the variance are analysed in view of emphasizing the eventual variability of the first and second order statistics in the annual series and in the monthly values, respectively.

For the checking-up of classical statistical estimations the series with monthly values are analysed by means of the Fourier and auto-regressive analyses later on emphasizing the independent component which provides the basis for the Monte Carlo generation of the desired set of sequences with independent values; by supperposing the independent component over the dynamic one the possibility of some exceptional maximum annual values to occur in an "L" time interval is checked up.

ANALYSIS OF STATIONARITY

The study of stationarity enables spotlighting the effects of the inconsistencies and non-homogeneousness in the data series being analysed on the sample statistics. On the average, the trend achieves the clearest image on the sample statistics variability in time, the first order non-stationarity bringing about, naturally, the second order non-stationarity.

Under a polynomial form,the trend for average and variance is written as below (Croley, 1977):

$$(1) \qquad T_m(i) = \sum_{j=o}^{k} a_j t_i^j \quad ,$$

$$(2) \qquad T_s(i) = \sum_{j=o}^{k} b_j t_i^j \quad , \; t = 1, 2, \ldots , n \; ,$$

where n is the number of observation years.

observing the notations for average and variance at moment "i", a_j and b_j are the coefficients of the polynomial to be determined the solving of which was based in these examples, on the maximum annual values regression analysis. The variability of the two statistics is generally attached to the change of observed data obtaining or to changes at

the level of the surface where the phenomenon being investigated makes manifest; with water discharges the trend is justified by progressive dike building in the riverbed (fig. 1) while with precipitations it depends on the effect of urban zone extent around the raingauging station (fig. 2).

As the effects of the modifications are little felt in the series under analysis it is admitted that the condition of stationarity is met from the standpoint of trend in both cases.

The correlogramme is achieved in terms of maximum monthly values and allows of the second order stationarity study is the time range. In the time interval [1,m] the auto-correlation coefficient (ACF) is derived from the experimental data in terms of the equation:

$$(3) \quad r(k) = \mathrm{cov}(x_t, x_{t+k}) \,/\, \mathrm{var}\, x_t \,, \quad k = 1, 2, \dots, m \,,$$

where, $m \leq N/4$, $N = 12n$ and n = the number of years (Haan, 1977).

The results obtained are presented in figures 3 and 4 for discharges and precipitations, respectively. In the case of water discharges the correlogramme indicates the presence of a low frequency harmonics overlapping an annual periodic component. With precipitations, it indicates a quite high degree of independence.

The continuous spectrum of the variance is the most efficient method to trace persistencies in the frequency range. The version using the Fourier transform is given by equation:

$$(4) \quad S(f) = \int_{-\infty}^{\infty} \varrho(\tau) \cos(2\pi f \tau) d\tau$$

The numerical integration of equation 4 is achieved in terms of r(K) (Haan, 1977) in the case of a given sample:

$$(5) \quad S(f) = \Delta t \left[r(0) + 2 \sum_{k=1}^{m-1} r(k) \cos(2\pi k f \Delta t) + r(m) \cos(2\pi m f \Delta t) \right],$$

with m, a quantity established as for correlogramme $\Delta t = 1$, the Nyquist frequency is $f_N = 1/2$ and $f = {}^{k}f_N/m$.

Checking up the numerical integration for the experimental data in equation 4 is based on the equation below:

$$(6) \quad \int_{-\infty}^{\infty} S(f)\, df = 2 \int_{0}^{\infty} S(f)\, df = 1 \,,$$

numerically experimented as:

$$(7) \quad 2\Delta f \left\{ 0{,}5 \left[S(0) + S(f_N) \right] + \sum_{k=1}^{m-1} S(k f_N / m) \right\} \approx 1$$

In order to observe restriction 7, smoothing procedures are applied (Yevjevich, 1972), lying in the correction of the computed values for the ACF in the range [1,m], or in smoothing the final values by means of an average sliding over three scans.

The outcomes of the analysis performed on a set of samples given in figures 5 and 6. With certain water, discharges, the visual inspection of the spectrum reveals the presence of the low frequency harmonics, but it is quite difficult to specify them. Fourier components with 24-year periods were attempted. In the case of precipitation studies the spectrum clearly emphasizes the annual periodicity as against the correlograme but not other harmonics.

CHECKING UP THE STATISTICAL ESTIMATIONS

The improvement of information in the sample of annual values by means of expansion from the series of monthly values is achieved in a similar way as the analysis of stationarity when other periodicities were sought after the annual periodicity was emphasized (as it is knouwn that the series must indicate the annual periodicity). In the case under study here, the same annual periodicity allows of tracing a dynamic component and it is advantageous that the latter's weight were greater; the residual component brought to the independence condition, checked up on account of the An-

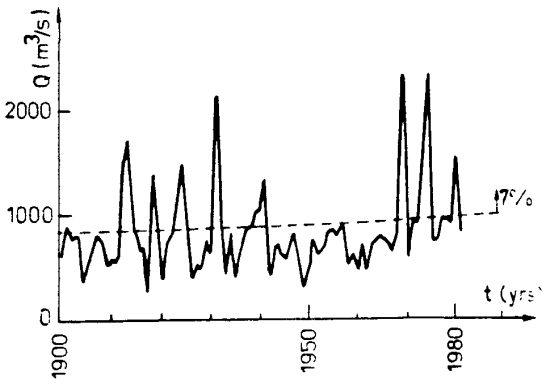


FIG.1 Trend average for the annual maximum discharge sample

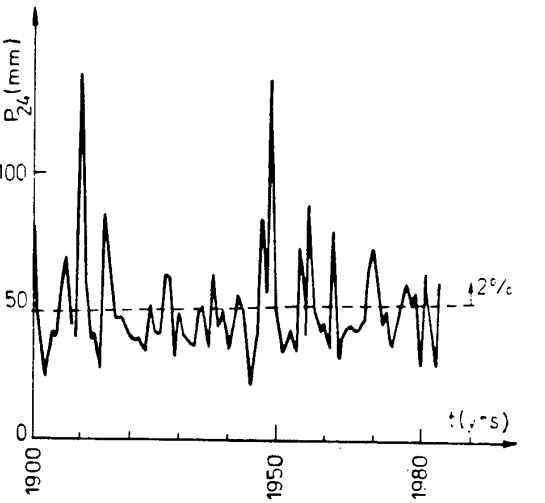


FIG.2 Trend average for maximum annual precipitation over 24 h

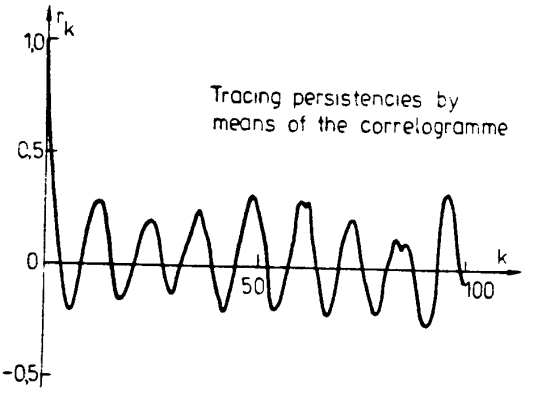


FIG.3 The correlogramme for the maximum monthly discharge sample

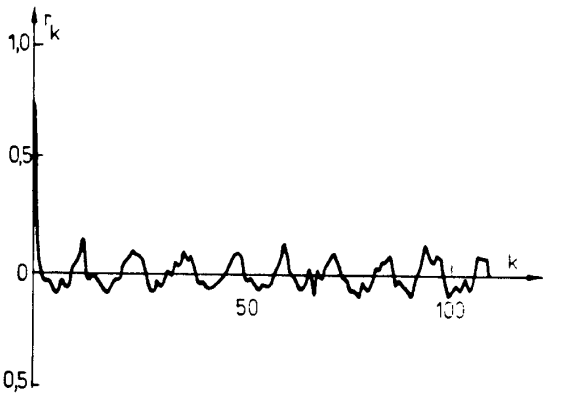


FIG.4 Correlogramme for the maximum monthly precipitation sample over 24 h

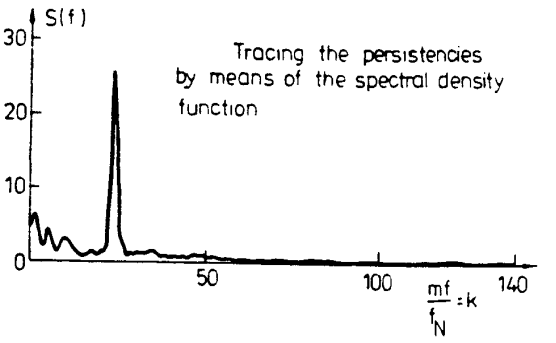


FIG.5 Spectral density function for the maximum monthly discharge sample

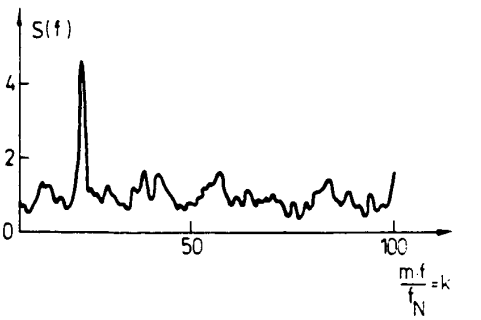


FIG.6 Spectral density function for the maximum monthly precipitation sample over 24 h

derson test, is a normal variable when there is a good agreement between the real series and the synthetic one; otherwise, the initial series is turned logarithmic, eventually translated, thus reaching the independent component as a normal variable. The parameters of normal distribution are computed for the volume of monthly values.

The independent component is a stationary variable with 12 × n realisations. By sliding, the same realisations overlapping the dynamic component allow of making up a series of annual maxima of a volume 12 × n. Since a good correlation between the real series and the synthetic one is sought it was admitted, in the examples here that the empirical distribution for the independent component is best represented by the normal distribution and therefore the Monte Carlo set of sequences covering looo years is generated further on transformed in variables to be analysed. A reframing of the independent component undergoes dynamic deformations.

This way, the history of the monthly values, whereof the annual values are derived, enables the verification of the recurrence capacity of the exceptional values the sample of annual values may provide by eytrapolation.

The dynamic components are determined by astronomical causes or by the phenomenon inertia.

A representation of the form below (Yevjevich, 1972) is attempted:

$$(8) \qquad X_{i,j} = m_j + s_j\left(\sum_{l=1}^{n_l} \alpha_{l,j}\ \varepsilon_{i,j-l} + \xi_{i,j}\right)$$

where, the average m_j, the variance s_j and the Markov $\alpha_{l,j}$ model parameters are periodical quantities mutually dependent.

Against this framework, Fourier representations are emphasized for means, variances and autocorrelation coefficients corresponding to each month in the observation series, according to the relation below (Croley, 1977).

$$(9) \qquad m_j = M_m + \sum_{k=1}^{n_m}\left[A_{m_k}\cos\left(\frac{2\pi kj}{\omega}\right) + B_{m_k}\sin\left(\frac{2\pi kj}{\omega}\right)\right]$$

where, m_j is the mean for each month, M_m is the general mean, n_m is the number of harmonics for the average, ω - the period, k=1, ... $\omega/2$ (ω is even) or k=1,2, ... , $(\omega-1)/2$ (ω is odd), $\frac{2\pi kj}{\omega}$ is the frequency corresponding to harmonics k while A_{m_k} and B_{m_k} are the Fourier coefficients determined on account of relations:

$$(10) \qquad A_{m_k} = \frac{2}{\omega}\sum_{j=1}^{\omega} m_j' \cos\left(2\pi kj/\omega\right)$$

$$(11) \qquad B_{m_k} = \frac{2}{\omega}\sum_{j=1}^{\omega} m_j' \sin\left(2\pi kj/\omega\right)$$

with m_j' the experimental mean corresponding to the j month.

Likewise, for the set of n_s harmonics the general variance M_s, A_{s_k} and B_{s_k}, the s_j variances are obtained for each month.

The initial variable is succesively normalized as against the m_j and s_j, then related to the general mean $\bar{Y}$ and the general variance S_Y:

$$Y_{i,j} = \left(X_{i,j} - m_j\right) / s_j$$

$$\varepsilon_{i,j} = \left(Y_{i,j} - \bar{Y}\right) / S_Y$$

The last variable obtained may still be dependent. To separde this dependence, a Markov model is used which establishes the auto - correlation between the succesive months for which the Fourier representation is also admitted, certain months being more closely correlated than others, function of the M_l - the first order coefficient average, n_l - number of harmonics A_{l_k} and B_{l_k} - Fourier coefficients.

Since a Markov representation is of the following form:

$$X_i = \sum_{l=1}^{n_l} \alpha_l \cdot X_{i-l} + W_i \ , \quad l = 1,2,3 \ ;$$

where α_l are dependent on r(k) and having a Fourier representation one has to emphasize the W_i component for which a normal distribution N is admitted $(\bar{W}, \sigma_W^2)$ starting from the real X variable or the logarithmic one ln X

Testing the W independent variable one takes into account the Anderson test for the confidence level of 95% (96% for the discharge series and 93% for the precipitations). Variable W is stationary or the initial series was deseasonalized (Croley, 1977).

A synthetic representation of the series under consideration was reached covering periodical components, autoregressive component and a normal independent component characterized by $\bar{W}$ and σ_W^2 which stands as the basis for the Monte Carlo generation. The assessment of the highest generated values is still to be solved.

As the positive asymmetry favours the occurrence of high values over a given time interval the probability of meeting such values against the period of L = looo years must be studied (Borgman, 1964).

The analysis of the probability of very rare values occurrence

It was assumed that the occurrence of such values must be studied around the probability E = o.5 which represents the average value and that the occurrence of the values of interest follows a normal distribution This assumption is quite cautious meaning, in the case of a normal variable, an equal chance to occur or not and, in the case of a positive asymmetry distribution, the greater chance for realisation than for the contrary.

The average being something in the range $[\bar{X} - \sigma, \ \bar{X} + \sigma]$, the average recurrence times for the generated values correspond to the probabilities in the range o.16 ; o.84 i.e. [5735, 546] years for L = looo years. The interval $\bar{X} - \sigma$, $\bar{X} + \sigma$ brings about the consideration of 6 classes in the time interval [o ; 1] and for the classes 2 - 5 the ends are considered thus obtaining 5 probabilities o.16; o.33; o.5; o.67; o.84 for which T takes the values: 5735, 2497, 1443, 9o2, 546.

The largest generated values are figured on the probability diagramme at T_i mentioned, then the correlated distribution curve is extrapolated onto the sample of annual values and the correlation, if any, is checked up for T_i between the real and generated values. The numerical examples in figures 7 and 8 showing discharge and precipitation extrapolations, the Kritzki-Menkel and log-Gumbel distributions were used respectively.

CONCLUSIONS

The trend analysis excepted, the other procedures are based on the series of monthly values, 12 x n in volume.

For this series the basic characteristics refering to non-stationarity are known and other persistencies were detected as side-effects, in fact periodical low frequency components compatible with the trend representation under polynomial form.

With water data the three procedures applied indicate non-stationarity of the series but from the standpoint of the trend this aspect may be overlooked. With precipitation data, the series under study is stationary.

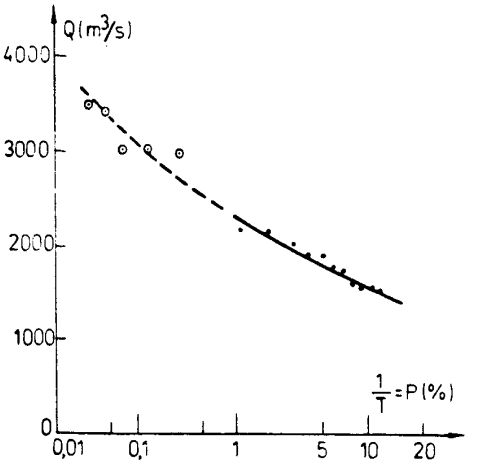


FIG.7 Checking up the extrapolation Kritzki-Menkel distribution curve for the maximum anual dischearge sample

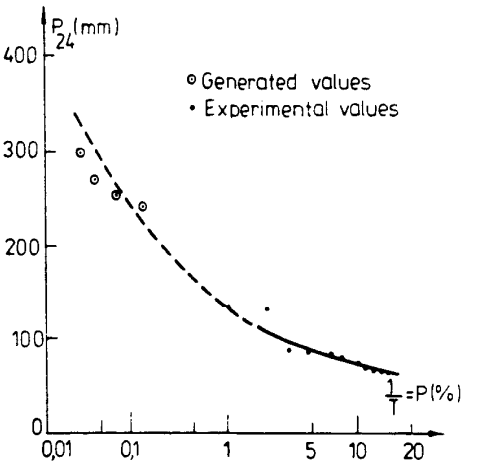


FIG.8 Chenking-op the extrapolation of the log-Gumbel distribution curve for the maximum annual precipitation sample over 24 h

In checking up the statistical estimations, the dynamic component is of a more significant weight for the discharge series, the Anderson test indicating 68% independence compared to 95% after its being withdrawn, whereas the independence was 89% on the initial precipitation series as against 93% after separating the dependent component. This is the reason why the results in the first numerical example are more consistent.

References

Borgman, E.L. 1964: Risk criteria. J.ww and Harbors Division, 3, 4-17, PASCE.

Croley, E.Th., 1977: A manual for Hydrologic time series deseasonalization and serial dependence reduction, IIHR Rep No. 199.

Haan, T.Ch., 1977: Statistical methods in Hydrology The Iowa State University Press, 279-286.

Kite, W.G., 1977: Frequency and risk analyses in Hydrology 2nd edition, Wat.Res.Publ., Fort Collins, Colorado.

Yevjevich, V., 1972: Stochastic processes in hydrology Wat.Res. Publ., Fort Collins, 81-82, 94-96.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE EFFECT OF VIEWING ASPECT ON CLIMATOLOGICAL CLOUD DISTRIBUTION

Donald D. Grantham and Albert R. Boehm
Air Force Geophysics Laboratory
Hanscom Air Force Base, Massachusetts, U.S.A., 01731-5000

Currently there are nonhomogenious cloud observations and data bases with many types of biases. These cause minor problems with weather forecasting but for climatology the error introduced can be major. The purpose of this paper is to present an overview of methods developed to use current data bases to estimate various climatic statistics about clouds.

OBSERVATION EFFECTS

To record sky cover, a surface based observer estimates the fraction of the sky dome that is cloudy. Practical problems such as surface obscurations and thin clouds are handled in a methodical if not completely rigorous fashion. The observed coverage is subjective and is affected by various physiological factors (Neuberger, 1965). On a weather chart, sky cover is plotted at the observation location, but some clouds, e.g. high level cirrus near the horizon, may be located more than a hundred kilometers away.

When weather satellite data first became available, it was hoped that many of these shortcomings would be resolved. It soon became apparent, however, that satellite data has some of the same old problems plus some new problems: viewing angle, asynopticity with lower orbits, pixel (footprint) size, degradation of sensors, etc.

Cloud measurement problems

Here is a brief list of the problems associated with measuring clouds:

Cloud or no cloud -- On a bright day with well-formed cumulus, it is relatively easy for a surface observer or instrument (ground or space based) to identify clouds from clear areas. But there are situations - thin cirrus, sun angles, snow backgrounds (viewing from space), dawn, dusk, night time, haze, etc. - where it is difficult to say with any certainty whether there is a cloud or not. Undoubtably these problems cause biases in cloud climatology. Sometimes the magnitude of the bias can be estimated or deduced from climatic analysis.

Viewing angle -- There appears to be more cloudiness near the horizon than when a scene is viewed straight up (or down).

Field of view -- The frequency distribution of cloud amount varies with the size of the area viewed or evaluated. A small area has a very U-shaped distribution with mostly clear or overcast; as the area becomes larger the distribution approximates a normal distribution.

Pixel or footprint size -- If one looks at a cloud through binoculars, one will see many smaller clouds near the larger clouds. The same thing happens when a higher resolution sensor is used on a weather satellite. If the cloud elements are smaller than the pixel size, they may be missed completely or the whole pixel may be counted as cloudy.

Recording errors

Weather observers from many nations for several generations have kept remarkably good records. We in the atmospheric sciences should take great pride in this extraordinary feat of record archiving. Nevertheless, many types of errors have crept into the records. Fortunately, most types of random errors cancel out in climatology, but some biases remain.

Reportable cloud cover values -- Sky cover has been recorded in hundredths (by satellite), tenths, octas, and four categories: CLR, SCT, BKN, OVC - clear, scattered, broken, and overcast. In addition, some data bases carry only the probability of below 2/8 and above 7/8. Amazingly, we have had to point out that directly calculating a mean with the courser type of data can lead to considerable error.

Interpolated cloud cover values -- If sky cover values are interpolated to a grid, they will contain too many intermediate values and the climatic distribution will be too low at the end points. Similarly, if climatic values are smoothed, the distribution will no longer be accurate. Specifically, many of the objective analysis schemes devised for use with numerical weather prediction produce very poor climatic distributions.

Asynoptic data -- Low orbiting satellites pass over at various times viewing one area at one time, another area at another time. Data is often stored at a nominal time that could be considerably different than the actual viewing time. Data stored this way causes error in the diurnal distribution, and spatial and temporal correlation. Also, it is often difficult to obtain the actual number of observations.

MODELING THE DISTRIBUTION OF SKY COVER

The distribution of sky cover (N) is bounded at 0 and 1. Although a small amount ($N < .05$) may be recorded as clear, completely clear or overcast does occur frequently. Karl Pearson (1898) used a Type I distribution (generalized Beta) to fit sky cover. Johnson (1949) fitted sky cover with his S_B distribution which uses the transformed variable $y=Ln[N/(1-N)]$, as a normally distributed variable similar to the more common log-normal distribution. Bean and Somerville (1981) computerized a worldwide (daytime) atlas of sky cover using the Beta distribution. Errors in these distributions are mostly in the clear and overcast categories because they do not represent truly clear and overcast probabilities. Greaves (1973) used the truncated normal distribution which has the probability for clear and overcast conditions. In addition, he described methods to account for variations in viewing area.

The Burger distribution

Burger (1985) described a new model for specifying sky cover distributions. Model distributions are shown in Figure 1 and the typical variation of its parameters in Figure 2. It is a numerical fit to the results of multiple simulations of a sawtooth simulation model (Burger and Gringorten, 1983).

The sawtooth simulation model -- This model has two parameters; the mean condition and scale distance (a parameter of spatial correlation). Briefly, the sawtooth wave model consists of sums of the heights of randomly oriented sawtooth wave formations. The height of a randomly oriented sawtooth is uniformly random, zero to one. Since the sum of uniform heights approximate a normal distribution, the summed heights approximate a multi-dimensional Gaussian process. Rather than some other waveform, e.g. cosines, sawtooths are used because: they are the fastest periodic function to calculate; the sum of their uniform heights approach normality with relatively few waves; but mostly because the resultant correlation is very similar to that found between observations of sky cover as shown in Figure 3.

The Burger Areal Algorithm, BAA — Burger's fit to the result of multiple simulations is somewhat complex although it is quite fast when programmed on computer. It is a general algorithm applicable to many weather elements. For sky cover, a small part of the algorithm is sufficient as will be shown.

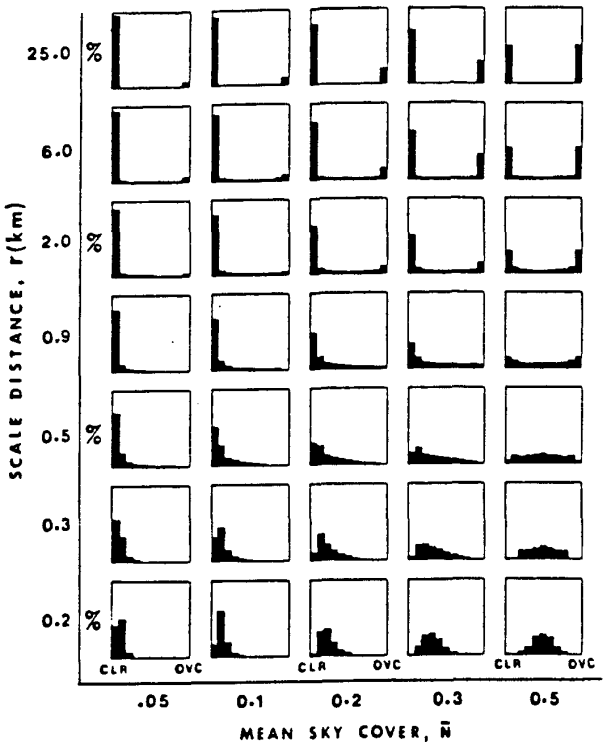


Fig. 1. Frequency distributions of sky cover categories (from clear to overcast) as produced by the BAA for various mean sky covers and scale distances.

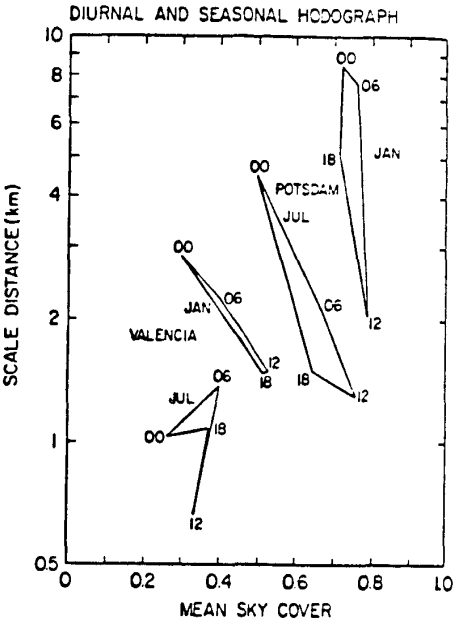


Fig. 2. Diurnal variation of mean and scale distance ofr Potsdam and Valencia in January and July. For most locations scale distances are shorter (i.e., more cumulus) in summer and afternoon.

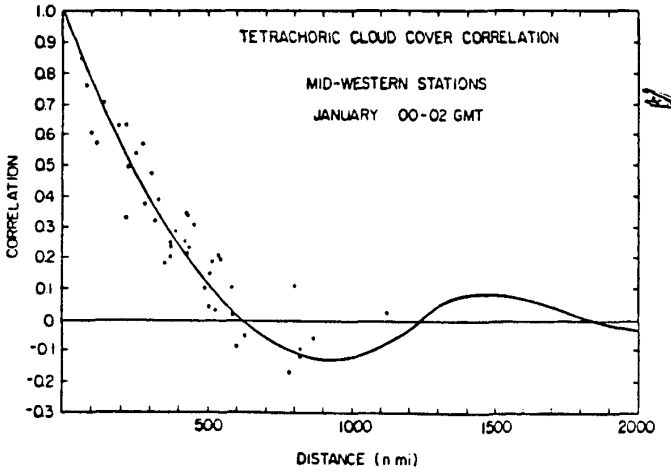


Fig. 3. Tetrachoric cloud cover correlation between pairs of 17 mid-western U.S. stations. Curve shows 3D sawtooth correlation.

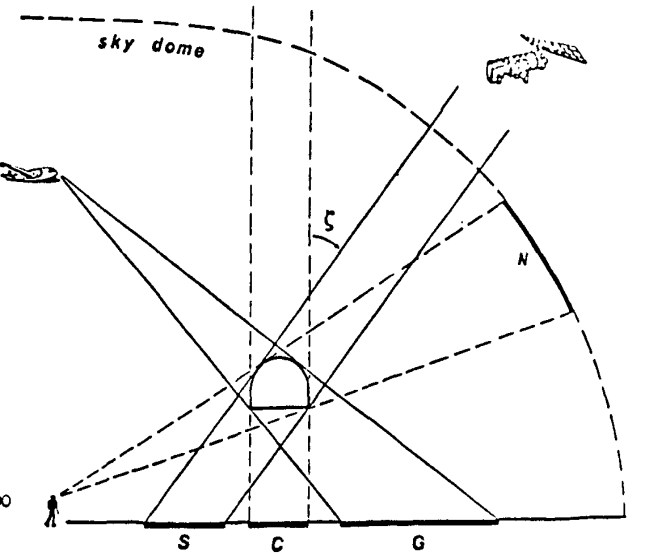


Fig. 4. Cloudiness quantifications: sky cover, N, from a ground observer; apparent cloud cover, S, from a satellite; vertical projection of cloud cover, C; and ground cover, G, from an aircraft. is view angle.

Allow the sky cover (tenths) to be N and the mean sky cover to have an equivalent normal deviate (END) of y_o. Then the END (y) of the cumulative probability for sky cover less than or equal to N is obtained as follows:

For $N \leq 1/2$, set $y_o' = y_o$, $N' = N$
For $N > 1/2$, set $y_o' = -y_o$, $N' = 1-N$

Find $i = 5 - INT(10N')$, where INT is the integer function, e.g. INT(4.2) = 4

Find $f = 10N' + i - 6$

$$\text{Find } y' = y_o'.G(6,z) + f.G(i,z) - \sum_{j=1}^{i-1} G(j,z) \qquad (1)$$

where z is a dimensionless function:

$$z = [\ln(\sqrt{A}/r)]/\ln 2$$

where A is the area of the sky dome which we set to 2424 km^2, that is, a radius of 15 nm, and where r is the sky-cover scale distance (km) which varies by station, by time of day, and by time of year.

$$G(k,z) = a_k + b_k \exp(c_k z); \text{ except } G(0,z) = 0 \qquad (1a)$$

k	a	b	c
1	-.0071	.0086	.53
2	-.0205	.0097	.52
3	.0229	.0045	.64
4	-.0517	.0224	.45
5	.0260	.0087	.71
6	.9981	.0011	.89

For $N \leq 1/2$, set $y = y'$
For $N > 1/2$, set $y = -y$.

The algorithm given in Eq. (1) is satisfactory for sky cover. Specifically, it is limited to $1<z<7$, .385 km$<r<$24.6 km, and a mean between .0062 and .9999. See Burger (1985) for the more general algorithm.

Obtaining the mean and scale distance from data -- Scale distance was originally defined as the separation distance at which correlation is 0.99. As new models were developed, we needed to use scale distances found with older models, but it was important that they resembled each other over the entire range of correlation. We chose Gringorten's (1979) Model B as the standard. Thus, in the 3D sawtooth the scale distance has a correlation of 0.988, but closely resembles Model B over the range of correlation 1. to 0.1.

Burger developed an algorithm to derive scale distance given the distribution of sky cover:

$$r(km) = \sqrt{2424}/2^z, \text{ where } z = a + b \ln(\Delta y + c) \qquad (2)$$

choose if:	Δy	a	b	c
	FOR DATA IN OCTAS			
CLR & OVC frequent	$y_{7/8} - y_{0/8}$	4.44437	1.78645	0.75
Prob. OVC near 0	$y_{4/8} - y_{0/8}$	5.49416	1.80228	0.43
Prob. CLR near 0	$y_{7/8} - y_{4/8}$	5.88697	1.78077	0.34
	FOR DATA IN AIRWAYS - CLR, SCT, BKN, OVC			
CLR & BKN frequent	$y_{BKN} - y_{CLR}$	4.35802	1.75995	0.67
Prob. OVC near 0	$y_{SCT} - y_{CLR}$	5.57793	1.75995	.335
Prob. CLR near 0	$y_{BKN} - y_{SCT}$	5.57793	1.75995	.335

y_q is the equivalent normal deviate of the cumulative probability up to and including q.

There is considerable error when the mean sky cover is calculated with the standard formula using the broad categories of CLR, SCT, BKN, OVC. There is measurable error even for the narrower categories of octas. Burger developed the following equation to minimize error. Given z from the algorithm above:

$$y_o = (y_{CLR} + y_{BKN})/(1.9962 + .0022e^{.89z}) \quad (3)$$

The Burger distribution as a function of area size

Above we used the Burger distribution for sky cover of the sky dome which we took to be 2424 km^2. However, the distribution is more general. In the equation for z, we can change the size of the area by replacing the value 2424 by the desired area in km^2. As the area becomes small, the distribution becomes very U shaped and in the limit becomes only probabilities at clear and overcast. As the area becomes very large, the distribution tends toward a normal distribution centered at mean cloud cover. The figure showing the Burger distribution for sky cover assumes a 2424 km^2 area, but this figure could be used for other sizes by use of a false scale distance (r'):

$$r' = r(2424/A)^{1/2} \qquad \text{where A is the area in } km^2 \quad (4)$$

The Burger Areal Algorithm can also give directly the distribution for a different size area. The limited version given above can be used as long as the z which is a function of area remains within stated limits.

Cloud coverage in layers given ceiling distributions

Gringorten (1982) developed some useful procedures for estimating cloud coverage in layers above the surface. The development of the new algorithms by Burger makes these procedure very attractive. Here we limit our discussion to one straightfoward application.

A ceilng is the lowest cloud deck that covers 6/10 or more of the sky. Climatological summaries of the frequency of ceilings as a function of height above the surface have been made for many locations. Rearranging the Burger algorithm given above, the equivalent normal deviate (y_o) of the mean sky cover from the surface to the level H is:

$$y_o(H) = [y(H) + G(1,z)]/G(6,z) \quad (5)$$

where y(H) is the equivalent normal deviate of the cumulative probability of ceilings up to height H. The function G and z are calculated as above.

The mean sky cover is associated with one or more clouds (possibly overlapping) between the surface and the height H. To obtain the probability of clouds in a layer of thickness h the correlation between levels must be taken into account. This was modeled (Gringorten, 1982) with the following algorithm:

$$\overline{P}_h = (\overline{P}_{H+h} - \overline{P}_H)/[(1 - \overline{P}_H)\cdot(1 - e^{h/(1.69 - 1.94\sqrt{H} - 0.475h)})] \quad (6)$$

where $\overline{P}_h$ is the mean cloud cover in the layer with bottom at H km and top at H+h km. P_H and P_{H+h} are the mean sky cover up to H and H+h respectively and are given by the cumulative probability associated with the equivalent normal deviates $Y_o(H)$ and $y_o(H+h)$. The root mean square error of the above estimate of the mean cloud cover in a layer using the systematic aircraft data collected in Germany (deBary and Moller, 1963) was 0.032 (3.2%).

CLOUD COVER AS A FUNCTION OF VIEWING ANGLE

Figure 4 depicts various methods used in assessing cloud cover from three platforms. Cloud cover, C, is defined as the vertical projection of clouds onto the earth. Cloud amount quantification from a ground observer, aircraft, and satellite are shown as N, G, and S, respectively.

PCFLOS as a function of zenith angle

A major application of cloud climatologies is to specify the probability of clouds along a line-of-sight (PCFLOS). Based on whole sky cloud photos at Columbia, Missouri, Lund and Shanklin (1973) developed a matrix model to estimate PCFLOS as a function of sky cover and elevation angle. Malick, Allen, and Zakanycz (1979) improved this model by ensuring that the PCFLOS values integrated over the total sky dome equalled one minus the total sky cover. The curves produced by this model (called the SRI model since it was developed at Stanford Research Institute) can be easily calculated:

$$PCFLOS = [1 - N\,(1 + 3N)/4][1 + (.55 - N/2)\tan Z] \qquad (7)$$

where N is sky cover as observed from the ground and Z is the zenith angle.

Lund had realized that the matrix model overestimated PCFLOS and stated to the authors that the SRI adjustment was an improvement (Lund, 1985).

In addition to the Columbia whole sky photos, we compared the SRI model to three additional sets of data: (1) data collected in the AFGL CFLOS program - a large data base consisting of aircraft observation of CFLOS (Bertoni, 1967, 1977), (2) whole-sky photography data in Estonia and the Canary Islands that were published, along with a theory, in Feigelson (1984), (3) the SRI model using several series of high resolution space shuttle photos looking at the same location from various angles (Snow et al., 1985 & 1986). We concluded that for elevation angles above 30° there was very good agreement - differences being only a few percent.

For angles within about 10° above the horizon, we incorporated the influence of earth curvature and cloud height by calculating the angle (α_c) that the line-of-sight intercepts the cloud:

$$\cos(\alpha_c = \cos(\alpha_g)\, R/(R + H) \qquad (8)$$

where α is the elevation angle measured at ground level

R is the radius of the Earth
H is the height of the cloud above the ground.

The climatological probability of CFLOS

Given the probability of a cloud-free line-of-sight (CFLOS) for a given zenith angle (Z) and for a given sky cover, the climatological probability of CFLOS can be calculated by summing this probability over each sky cover category weighted by its climatological frequency of occurrence. This probability is still a function of zenith angle. The calculation of the climatological probability of CFLOS for non-geostationary satellites involves many zenith angles and can become time consuming. Burger found the climatological probability nearly independent of scale distance and so developed a more direct method of calculating the climatological probability of cloud-free line-of-sight (CPCFLOS):

$$CPCFLOS(Z) = 1 - aP_o + (a - 1)P_o^2 \qquad (0 < Z < 80°) \qquad (9)$$

where P_o = PCFLOS as given in Eq. 7 and $a = (c_1 + c_2\, r + c_3\, r^2) + (c_4 + c_5\, r + c_6\, r^2)Z + (c_7 + c_8\, r + c_9\, r^2)Z^2$.

n	c_n for $Z > 30°$	c_n for $Z < 30°$
1	.9137	1.2958
2	.06032	-.2339
3	-.0191	.1001
4	4.7423×10^{4}	-1.2255×20^{-2}
5	-3.2087×10^{-4}	9.4255×10^{-3}
6	9.0008×10^{-5}	-3.812×10^{-3}
7	8.2725×10^{-6}	1.1425×10^{-4}
8	-5.8336×10^{-6}	-8.6538×10^{-5}
9	1.8496×10^{-6}	3.3768×10^{-5}

Cloud coverage as seen by satellite

The detection of clouds by satellite sensors is plagued by the overestimation of cloud amount at viewing angles other than zero. Figure 5 illustrates the problem. The fraction of the scene assessed as cloudy, increases as the viewing angle increases because the sides as well as the tops are detected. This problem is greatest for geostationary satellites and viewing high latitudes and for orbiting satellites scanning limb angles. The dashed curve in Figure 6 shows this effect for 10 passes of a polar orbiter over the Atlantic Ocean. The bottom four curves in Figure 6 presents the result of four sequences of space shuttle photos as described and modeled by Snow, et al, (1985 & 1986). The function SCC (ζ) is seen to be smooth and remarkably repeatable.

ACKNOWLEDGMENTS

We thank Irving Gringorten for Figure 3, for checking many of the formulas, and for originating many of the ideas here presented. We present here some of Charles Burger's remarkable unpublished algorithms since he is no longer at the A.F. Geophyics Laboratory. We thank J. William Snow for Figures 4, 5 and 6.

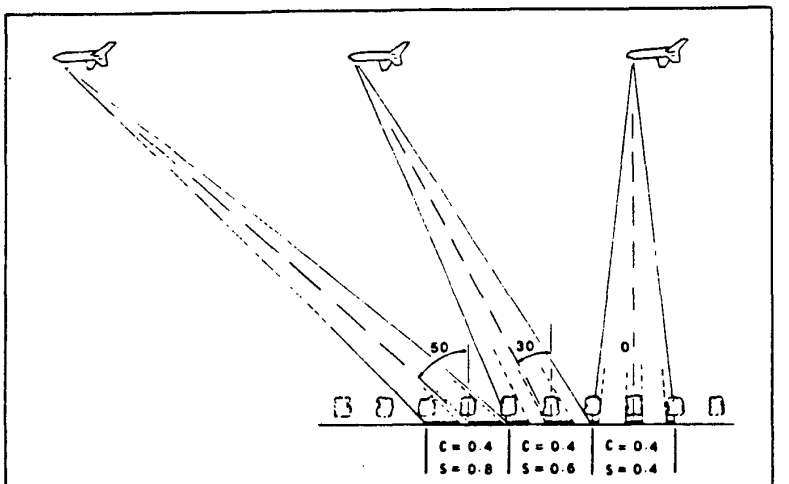


Fig. 5. Cloud cover C and satellite cloud cover S as seen from the space shuttle orbiter at viewing angles of 50°, 30°, and 0°. Surface area is displaced for clarity. The cloud field is invariant in the third dimension, that is, each cloud indicates an entire cloud row. Heavy lines represent the portion of area assessed from the satellite as cloudy. S is a function of cloud depth, viewing angle, and C. The illustration is for a cloud depth to width ratio of 0.866.

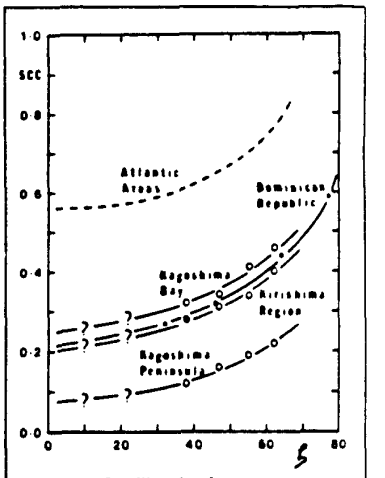


Fig. 6. Satellite cloud cover versus viewing angle ζ. Open circles indicate Kyushu Island data; solid circles show Dominican Republic data. The dashed curve is Average oceanic cloud cover as measured by meteorological satellite sensors operating in the visible wavelengths.

REFERENCES

Bean, J.S., and Somerville, P.N., 1981: Some new worldwide cloud-cover models, J. Appl. Meteor., 20, 223-228.

Bertoni, E.A., 1967: Clear line-of-sight from aircraft, AFSG 196, AFCRL-TR-67-0435, 186 pp, AD657801.

Bertoni, E.A., 1977: Clear and cloud-free lines-of-sight from aircraft, AFCRL-TR-77-0141, 21 pp, ADA046352.

Burger, C., 1985: A new model for specifying cloud distributions for climatologies, Preprints: 9th Conf. on Prob. and Stat. in Atmo. Sci., AMS, Boston, MA., 459-462.

Burger, C., and Gringorten, I., 1983: Lineal and areal probabilities of weather conditions, Preprints: 8th Conf. on Prob. and Stat. in Atmo. Sci., AMS, Boston, MA., 39-44.

deBary, E., and Moller, F., 1963: The vertical distribution of clouds, J. Appl. Meteor. 2, 806-808.

Feigelson, E.M. ed., 1984: Radiation in a cloudy atmosphere, D. Reidel Publishing Co., Dordrecht/Boston/Lancaster, 32-35.

Greaves, J.R., 1973: Development of a global cloud model for simulating earth-viewing space missions, J. Appl. Meteor., 12, 12-22.

Gringorten, I.I., 1979: Probability models of weather conditions occupying a line or an area, J. Appl. Meteor. 18, 957-977.

Gringorten, I.I., 1982: Climatic Probabilities of the Vertical Distribution of Cloud Cover, AFGL-TR-82-0078, pp 48, ADA 118753.

Johnson, N.L., 1949: Systems of frequency curves generated by methods of translation, Biometrika, 36, 149-176.

Lund, I.A., 1985: Private communication.

Lund, I.A., and Shanklin, M.D., 1973: Universal methods for estimating probabilities of cloud-free line-of-sight through the atmosphere, J. Appl. Meteor., 12, 28-35.

Malick, J.D., Allen, J.H. and Zakanycz, S., 1979: Calibrated analytical modeling of cloud-free intervals, SPIE, 195, 142-147.

Neuberger, H., 1965: Introduction to Physical Meteorology, Penn. State U., University Park, PA.

Pearson, K., 1898: Cloudiness: note on a novel case of frequency, Proc. Roy. Soc., London, 62, 287-290.

Snow, J.W., Bunting, J.T., d'Entremont, R.P., Grantham, D.D., Hardy, K.R., and Tomlinson, E.M., 1985: Space shuttle photograph assist in correcting meteorological satellite data. EOS Trans. Am. Geophy. Union, 66, 489-490, ADA 162059.

Snow, J.W., Grantham, D.D., Tomlinson, E.M., and Willand, J.H., 1986: The characterization of cumuliform cloud fields using space shuttle photography. Proc. Second Conf. on Satellite Meteorology, Remote Sensing and Application, Amer. Met. Soc., Williamsburg, VA, 228-232.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

Pattern analysis of growing season precipitation in south-central Canada

Peter J. Lamb and Michael B. Richman

Climate and Meteorology Section
Illinois State Water Survey
Champaign, Illinois 61820

1 Introduction

Summer precipitation is of considerable importance for crop production in the rich agricultural regions of south-central Canada that extend to the east of the Rockies, to the north of the Great Lakes, and through the St. Lawrence Valley. This precipitation is less the product of synoptic-scale weather systems than that of other seasons (Whittaker and Horn, 1981); it is partly convective and therefore more variable in space and time. An understanding of its intraseasonal and interannual fluctuations in terms of the relative importance of several possible contributing climate system processes -- for example, the local presence and external supply of water vapor, and the troposphere's predisposition to provide the precipitation-inducing vertical motion -- is clearly desirable, particularly since that knowledge is essential for the development of physically-based (as opposed to purely statistical) climate prediction schemes.

The difficulty of this task is enhanced by the precipitation concerned being partly convective. It requires that, before the desired climate system insight is sought, the characteristic patterns of summer precipitation be identified for the region. We here initiate consideration of the latter subject, by presenting the results of eigenvectorial climatic pattern analyses of three- and seven-day precipitation totals for the May-August periods of 1949-80. Longer time-scales will be treated in a subsequent communication.

The present paper complements a recent counterpart study for the central United States (Richman and Lamb, 1985, henceforth termed RL). The latter effort included assessment of the sensitivity of eigenvectorial results to several parameters that had hitherto been the subject of considerable qualitative concern, but little quantitative inquiry -- station density, domain shape, data form, dispersion matrix, eigenvectorial model and its linear transformation/rotation, and pattern faithfulness to interrelations among input data. RL reviewed extensively both the need for that methodological investigation and the previous neglect of short time-scales (and therefore time-scale dependence) in precipitation pattern analysis research. We here capitalize on the earlier methodological work, and only revisit the above issues when it is specifically warranted for the present study domain.

This inquiry culminates with a regionalization of south-central Canada into subareas within which three- and seven-day summer precipitation tends to be spatially coherent. It dovetails closely with a counterpart central United States regionalization presented in RL. We conclude by suggesting several varied potential applications of the present regionalization.

2 Data

Because our concern is with time scales of less than the calendar month, for which there existed no readily available historical data sets with the spatial resolution necessary for the treatment of convective precipitation, we have created a new data ensemble for use in this study and related investigations. It contains the individual daily precipitation amounts for 155 locations in central Canada south of 56°N and 15 United States stations along the Canadian border for the May-August periods of 1949-1980, which total 669,120 observations. The United States data are from a counterpart ensemble described in RL. The present station network is displayed in Fig. 1; it is bounded on the west by the Rocky Mountains, on the south and east by the United States, and on the north by the northern limit of agriculture and/or the availability of densely-spaced reporting stations.

The first step in the development of this data set was the acquisition of all daily airport and cooperative substation precipitation totals archived by the Canadian Climate Centre for the study region and period. After identifying the stations with the longest records, and carefully scrutinizing and quality controlling their data, it was concluded that a network resolution of 80-110 km was the finest practicable over most of the domain from the standpoints of spatial and temporal unifor-

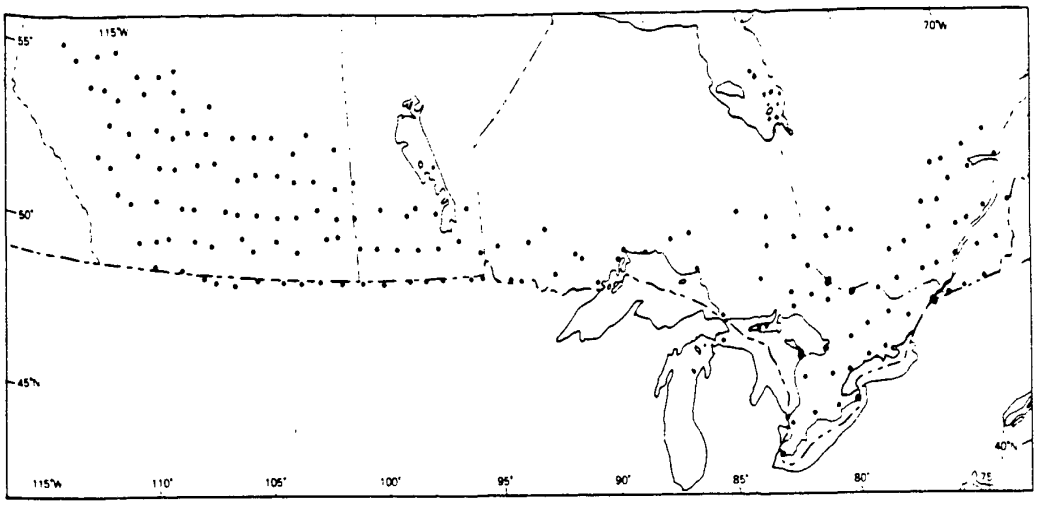


Fig. 1. Station network (170 locations) for south-central Canada daily precipitation data set for May-August periods of 1949-1980.

mity and data set completeness. A modified 1° latitude-longitude "target" grid (the meridional separation north of $49^{\circ}N$ was held constant at the $49^{\circ}N$ value) was therefore established, and the station nearest each intersection with the most complete and seemingly most reliable record was identified. These stations were located, on average, 24 km from the grid intersection concerned, and were chosen to form the present data set (Fig. 1). For days those stations did not possess observations (which averaged 23 percent of the total), the value from one of three adjacent "alternate" stations was substituted without adjustment. These alternate stations were usually situated within 30 km of their network station and were selected because their records, although often incomplete, covered the gaps in those of the network station. Most of the network stations made observations at 0800 LST, with the major exception being a few airport stations that reported at 0600 LST. The relatively even station distribution should minimize the potential errors that can affect eigenvectorial patterns extracted from irregularly spaced data (e.g., Karl et al., 1982).

The present paper utilizes square-root transformations of three-day (i.e., 1-3 May, ..., 29-31 August) and seven-day (i.e., 1-7 May, ..., 21-27 August) station totals computed from the above data set for each year during 1949-80. No three- or seven-day period received zero precipitation at all 170 stations. The square-root transformation was employed to reduce the positive skewness of the distributions of the short-period totals (Table 1). See RL for a review of the skewness problem. The square-root transformation was chosen on the basis of the experimentation with alternative transformations conducted by or surveyed in RL. It was found to effect (a) the greatest reduction in positive skewness by "pulling in" the right-hand tail of the distribution without increasing the spread on the left and (b) the largest enhancement of the variance explained by individual eigenvectorial patterns without altering their morphology.

3 Computational considerations

The pattern analysis of precipitation data can be performed using a range of statistical techniques. Eigenvectorial methods are employed here. The general advantages of this approach have been reviewed by Stidd (1967) and Gregory (1975). Its present adoption primarily resulted from our overall inquiry into south-central Canadian summer precipitation requiring not only the establishment of the major spatial patterns (or modes) of precipitation variation, but also formally related information on the relative importance of those patterns as functions of time. RL argued that the available alternative statistical techniques -- Fourier analysis, Tschebyceff polynomials, cluster analysis, Lund (1963) map typing -- do not permit the simultaneous related accomplishment of both of these objectives, and also posses other limitations.

Recent years have seen meteorological and oceanographic research utilize an increasing range of eigenvectorial techniques, a development that is exploited in this investigation. Because that development has been well documented in the literature, we have chosen not to provide complete discussions of the methods used. Instead, readers are directed to RL, and also the the many studies referenced there

Table 1. Skewness statistics for frequency distributions of raw (denoted by R) and square-root transformed (denoted by $\sqrt{R}$) three- and seven-day precipitation totals. The transformations were applied to totals expressed in inches. The means (M), standard deviations (SD), maxima (MX) and minima (MN) given were computed from the skewness values obtained for each of the 170 stations in Fig. 1 from the entire 1949-1980 sets of totals described in Section 2. See RL (Table 1) for skewness equation.

Data Form	3-Day Totals M	SD	MX	MN	7-Day Totals M	SD	MX	MN
R	+3.017	0.639	+4.990	+1.654	+2.003	0.491	3.764	+0.931
$\sqrt{R}$	+0.962	0.276	+1.605	+0.300	+0.413	0.228	0.915	- .132

that provide the necessary information, including appropriate theory. Several standard texts contain further relevant material (e.g., Thurstone, 1947; Rummel, 1970; Mulaik, 1972; Harman, 1976), as does the review paper by Richman (1986).

Our analyses were performed in an S-mode sense, one that seeks associations among stations over a set of time periods for a fixed parameter. In addition to computing the well-known direct (i.e., unrotated) principal components (PCs) of interstation correlation and covariance matrices, we also obtained from this input orthogonally (VARIMAX criterion) and obliquely (Harris-Kaiser Case II B'B criterion) rotated PCs and Common Factors (CFs). The most important difference between the often-used PC and lesser-known CF methods is that CFs are extracted from only the variance that is shared by two or more locations, and not the total variance at each site as is the case with PCs. The two methods therefore do not typically yield the same spatial loading patterns. The exclusion from CF analysis of variance that is unique to each station (due to small-scale processes and sampling/measurement errors) was considered attractive in the present context because of the partly convective origin of the precipitation.

We linearly transformed/rotated the PCs and CFs obtained because of the possibility that the unrotated solutions may not adequately capture the shape of the dominant individual modes of variability of the input data. As reviewed in RL, that problem stems from PCs and (to a lesser extent) CFs basically being data compression devices designed to maximize the variance fraction explained over the entire domain. The problem is most likely to occur when the data, as in our case, contain distinct clusters. The linear transformation/rotation of PCs and CFs seeks to avoid the above situation by positioning the PCs/CFs so that the variables have small projections on them. It accordingly produces spatial loading patterns that typically contain many near-zero values and a smaller number of larger values, and hence generally feature somewhat stronger anomalies covering smaller portions of the domain than occur without rotation. The orthogonal rotation used (VARIMAX) seeks to maximize the variance of the station loadings of individual PCs/CFs along the lines implied above. Although VARIMAX's orthogonality ensures that the time-dependent scores of the spatial loading patterns are uncorrelated, the spatial patterns themselves can be intercorrelated. The oblique rotation employed (Harris-Kaiser Case II B'B), however, relaxes the orthogonality constraint entirely in pursuit of the same general objective as VARIMAX.

4 Patterns given by unrotated solutions

Table 2 and Fig. 2 document the first eight unrotated PCs extracted from each data set for the 170 station network shown in Fig. 1. These unrotated PCs, unlike all subsequent ones, were suggested to contain non-random signals by application of the "scree" test to the eigenvalue series (Cattell, 1966; Walsh and Richman, 1981). This test considers that the last pronounced discontinuity in the eigenvalue series prior to its leveling off constitutes the division between the unrotated PCs that possess non-random signals (those with larger eigenvalues) and the ones that do not. The available alternative methods for determining this truncation point, such as Preisendorfer and Barnetts' (1977) Monte Carlo test (e.g., Rogers and van Loon, 1982) and the Kaiser (1960)/Guttman (1954) "lower bound criterion" or "K-G test" (e.g., Horel, 1981), suggested the retention of far too many unrotated PCs for the present cases (13-25 depending on the time-scale and test). Such "overfactoring"

Table 2. Principal Component results derived from interstation correlation matrices for square-root transformed three- and seven-day precipitation totals for all stations shown in Fig. 1 for the May-August periods 1949-1980. The numbering of the components/regions in Figs. 2-5 is consistent with the root numbers given here. The separation of the bracketed eigenvalues is less than the sampling error given by Kendall (1980, p. 27)/North et al. (1982) criterion. The cumulative explained variance for roots 1-8 is not given for the Harris-Kaiser rotated solution because the same variance can be accounted for by more than one root.

	3-Day 1312 Observations				7-Day 544 Observations			
Root No.	Unrotated Eigenvalue	Unrotated Percent Total Variance Explained	VARIMAX Percent Total Variance Explained	Harris-Kaiser Percent Total Variance Explained	Unrotated Eigenvalue	Unrotated Percent Total Variance Explained	VARIMAX Percent Total Variance Explained	Harris-Kaiser Percent Total Variance Explained
1	28.83	16.96	7.85	9.51	32.05	18.85	8.85	11.56
2	22.65	13.33	7.71	10.07	23.43	13.78	7.91	10.41
3	12.80	7.53	7.36	8.52	11.85	6.97	7.61	9.23
4	7.92	4.66	7.18	9.28	7.80	4.59	7.25	8.37
5	6.64	3.90	6.61	7.36	6.53	3.84	7.24	8.55
6	5.82	3.42	6.47	7.67	5.84	3.44	6.56	7.87
7	4.80	2.82	5.89	8.23	4.46	2.62	5.59	8.23
8	3.71	2.19	5.74	8.98	3.82	2.25	5.33	9.44
Sum 1-8		54.81	54.81	-----		56.34	56.34	-----

(Richman, 1981; Horel, 1981) by these other methods may be a potential problem when (a) the total number of PCs obtained is large (170 here) and (b) the concern (as here) is with pattern identification rather than just data compression. This question of the number of roots to retain is considered further in the Appendix.

Table 2 also indicates that the eigenvalue separations within both of its sets of unrotated PCs exceed the corresponding sampling errors given by the Kendall (1980)/North et al. (1982) criterion for all cases except seven-day roots 5-6. Furthermore, the eigenvalue separation for the latter root-pair failed that test very narrowly, while the separations for the above truncation point (i.e., between roots 8-9, not shown) both passed the test, as was recommended by North et al. (1982). This eigenvalue separation requirement guards against the possibility of the sampling errors for the station loadings of an unrotated PC being comparable to the magnitudes of the loadings of an adjacent PC. Such a situation can produce large intersample variability in the spatial loading patterns of individual (i.e., root number X) unrotated PCs (North et al., 1982). RL further confirmed the absence of sampling-error contamination of their unrotated PC patterns for the central United States by very closely replicating those patterns from subsets of the data. The extreme success of that procedure suggested it was not necessary to go to such lengths in the present case.

Table 2 indicates the first eight of the 170 total unrotated PCs account for approximately 55 percent of the entire domain variance, irrespective of the data time-scale. A review of other relevant studies in RL suggests this percentage is surprisingly high, given the short time-scales of the data, the large number of stations used (and hence PCs obtained), and the partly convective nature of the precipitation. It is also markedly above the counterpart values obtained for the central United States by RL (44-47 percent), which likely reflects a smaller convective contribution to the Canadian precipitation. Incidentally, when the above analyses were performed using untransformed data (not shown), the variance fractions explained by the first eight unrotated PCs were approximately 46 percent.

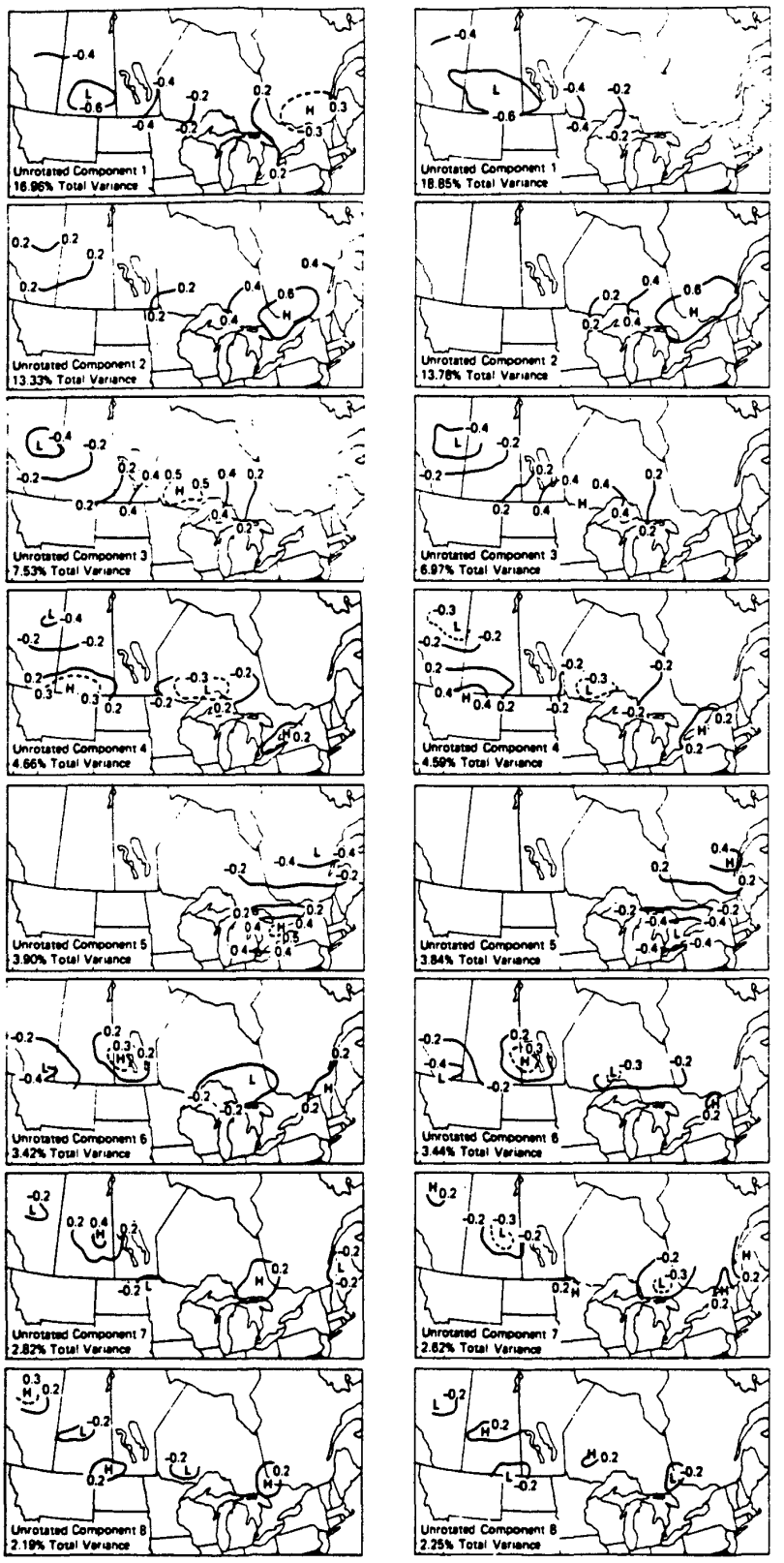


Fig. 2. Unrotated Principal Components 1-8 of the interstation correlation matrix derived from square-root transformed (a) three-day and (b) seven-day precipitation totals for all stations in Fig. 1 for the May-August periods of 1949-1980.

Fig. 2 presents the spatial loading patterns of the unrotated PCs introduced in Table 1. For the three-day time-scale, unrotated PC 1 (16.9% percent of total domain variance) features anomalies of opposite sign in the western (strongest) and eastern halves of the study region. The maximum station variance explained by this PC never exceeds 41 percent. Its large spatial scale and west-east sign contrast are surprising, given that the pattern is extracted from relatively short-period totals of rainfall that is partly convective, and the fact that the interstation correlation matrix contained no similar sign contrast (not shown). The next member of this three-day PC set (No. 2, 13.33 percent of total domain variance) also possesses a large spatial scale, being completely dominated by an anomaly of one sign that covers the entire domain and has a maximum along the Ontario-Quebec border. While three-day unrotated PC 3 (7.53 percent) features another west-east sign flip-flop, this time between centers in Alberta and southwestern Ontario, its anomalies are weaker (maximum station variance explained is approximately 30 percent) and of smaller spatial scale than those for roots 1 and 2. Further and progressive reductions in spatial scale and explained variance fraction (both station and total domain) occur between unrotated PCs 3 and 8 of this three-day set. The latter exhibit little structure and would seem especially difficult to interpret physically.

The seven-day unrotated PC spatial loading patterns and associated explained variance fractions (both station and total domain) are very similar to their three-day counterparts (Fig. 2), especially for roots 3-8. Note that the anomaly sign reversal between time-scales for some roots is arbitrary. The major difference between the three- and seven-day results in Fig. 2 is that the first seven-day PC does not contain an opposite sign anomaly in the eastern half of the domain. This time-scale difference was not characteristic of the central United States results reported in RL. A further contrast between the unrotated PC patterns in the two studies is that the Canadian ones are much less "Buell-like" (see RL, pp. 1133-1134 and their Fig. 2) than those for the United States. This probably results from the irregular shape of the present study domain.

Fig. 3a provides further insight into the nature of the unrotated PCs documented in Table 2 and Fig. 2. It plots the station loadings on one PC versus the station loadings on another PC, a procedure introduced in RL. The two examples shown in Fig. 3a adequately illustrate the range of patterns obtained for unrotated roots 1-8 for both time-scales considered. The most important feature of those patterns is the broad scatter of the data points, in one case (lower) across all four quadrants and in the other (upper) with a diagonal arrangement confined to the top two quadrants. For the unrotated PCs, the data points were never aligned along the axes or tightly clustered around the origin. This indicates that individual stations loaded highly on more than one unrotated PC, and that the analysis failed to identify individual areas of coherent precipitation (or clusters in the data). These results parallel those obtained in RL for the central United States, and further suggest the limited utility of the conventional direct (i.e., unrotated) PC solutions for capture of the individual modes of variation of short-period summer precipitation in central North America.

5 Patterns given by rotated solutions

In an attempt to overcome the above deficiency, unrotated PCs 1-8 for both time-scales were next rotated orthogonally to the VARIMAX criterion. The results are summarized in Table 2 and Figs. 3-5. This limiting of the rotation procedure to the number of unrotated PCs considered to be significant followed a relatively standard practice (see RL). The present choice of eight was also supported by experimental rotations of lesser and greater numbers of roots (see Appendix).

Table 2 readily shows that the cumulative variance fraction explained by each set of eight VARIMAX-rotated PCs is the same (to within at least two decimal places) as for the first eight unrotated PCs from which they were derived. This is because the rotation is orthogonal (see RL). However, Table 2 also clearly reveals that, while the variance accounted for by individual VARIMAX PCs decreases with increasing root number, this is less pronounced than for the unrotated solution. In comparison with their unrotated counterparts, VARIMAX PCs 1 and 2 explain much less variance, number 3 accounts for approximately the same, and numbers 4-8 explain more variance. This relatively even distribution of the explained variance among the VARIMAX PCs is consistent with that technique's inherent goal of identifying strong individual anomalies that accordingly tend to have relatively small spatial extents (see Section 3). This situation also means that it is inappropriate to use an equivalent of the eigenvalue separation (see Section 4) to test for the sampling errors of VARIMAX-rotated solutions (Horel, 1984). We instead confirmed the absence of such contamination by very

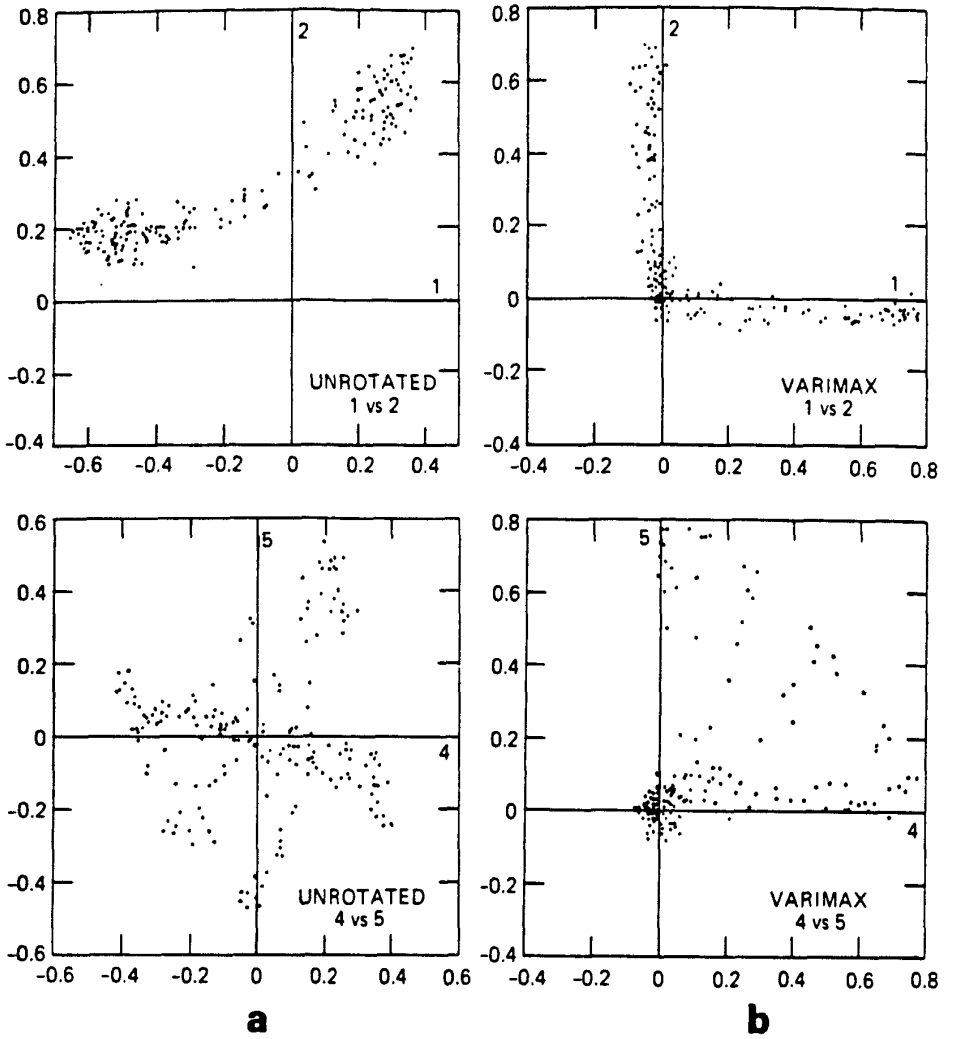


Fig. 3. Plots of station loadings for pairs of three-day Principal Components presented in Figs. 2 and 4. Unrotated examples appear in (a) and VARIMAX-rotated ones in (b). The root numbering is consistent with Table 2 and Figs. 2, 4, and 5. Some of the data points shown represent more than one pair of loadings (not identified).

closely replicating the results from subsets of the data, as will be illustrated in the next section.

Fig. 4 presents the spatial loading patterns of the three-day VARIMAX PCs introduced in Table 1. Although their seven-day counterparts were ordered somewhat differently, they collectively contained an essentially identical set of patterns (congruence coefficients of 0.990-0.995; see Appendix in RL) that accordingly are not shown here. The most prominent feature of Fig. 4 is the regional nature and strength of all eight of the VARIMAX PC patterns. The latter do not degenerate and lose their structure with increasing root number; number 8 is just as distinctive as number 1. Each VARIMAX PC identifies an anomaly that characterizes a different and relatively restricted part of the domain. The regions concerned are reasonably uniform in size and span areas equivalent to substantial portions of provinces (e.g., southern Quebec, central Alberta, southern Saskatchewan-Manitoba). Within most of each region, one VARIMAX PC explains at least 15 percent of the station variance; this fraction rises to central values of 50 percent for almost every region. Outside of its limited area of emphasis, each VARIMAX PC accounts for very little (< 4 percent) for the station variance. Consistent with this, the inter-PC loading correlation magnitudes were all less than 0.30, with a mean of only 0.21. These VARIMAX PCs are thus almost uncorrelated spatially (see Section 3).

Fig. 5 shows that the areas isolated by the VARIMAX PCs collectively span the entire study domain. In fact, what that display contains is a rather complete subdivision of the domain into regions within which three- and seven-day summer precipitation tends to be spatially coherent.

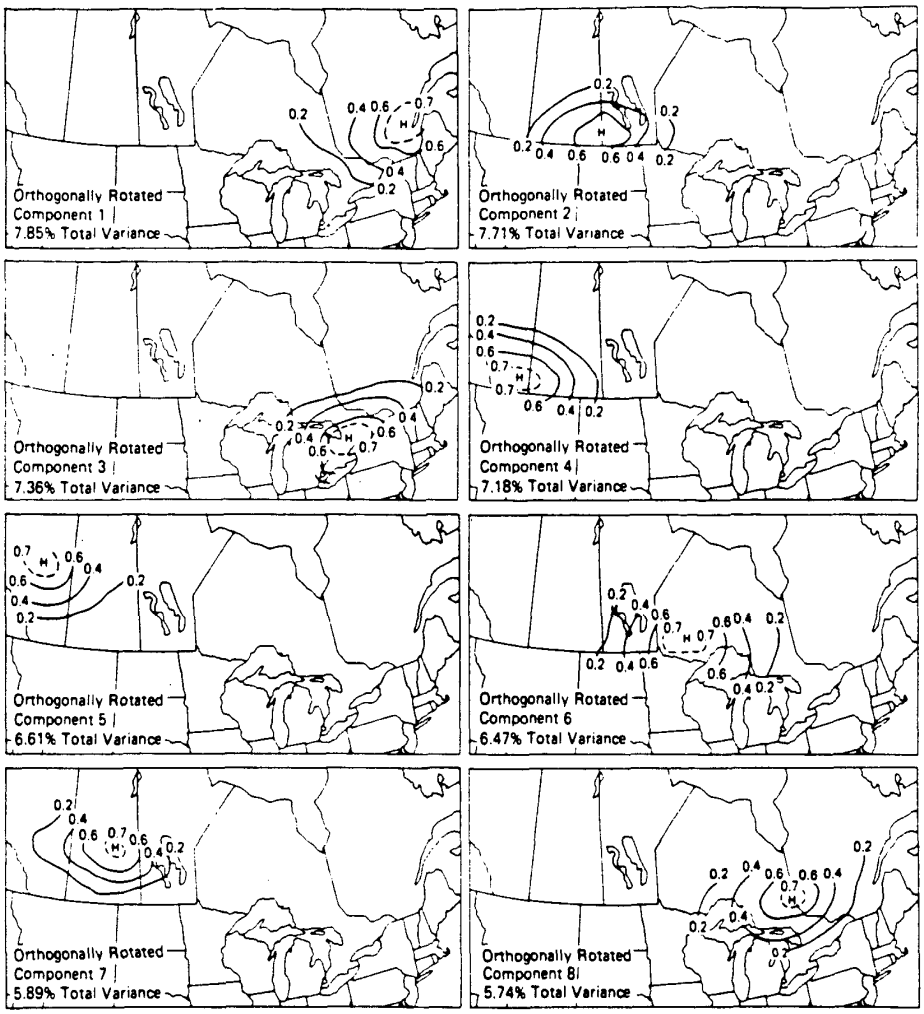


Fig. 4. VARIMAX orthogonally rotated Principal Components 1-8 of the interstation correlation matrix derived from square-root transformed three-day precipitation totals for all stations in Fig. 1 for the May-August periods of 1949-1980.

This regionalization includes very little overlap; only 34 of the 170 stations (20 percent) lie within the boundaries of two regions, and no stations belong to more than two regions. The validity (both statistical and physical) and utility of this regionalization are considered further below and in Sections 6 and 7.

The VARIMAX-rotated PC patterns in Fig. 4 are thus in pronounced contrast to the unrotated PC patterns (Fig. 2) on which they are based. In particular, the uniform size (medium) and strength (considerable) of the VARIMAX features differs greatly from the larger scale but weaker patterns of unrotated PCs 1-3 and the increasingly fragmented, smaller scale, and even weaker features of the remaining unrotated PCs. This contrast is further emphasized by Fig. 3. Panel (b) of that display plots, for two informative cases, the station loadings on one VARIMAX PC versus the station loadings on another VARIMAX PC. One case shown is for PCs that are centered a moderate (as opposed to the maximum possible) distance apart (1 versus 2; Fig. 4), while the other is for a pair of contiguously located PCs (4 versus 5; Fig. 4). In the former instance (Fig. 3b, top), the data points either cluster tightly around the origin or are aligned along the positive axes. Very similar or more pronounced such configurations were obtained for the other pairs of VARIMAX PCs that do not abut (not shown). Even when the latter is the case, however, this same basic structure is still apparent (e.g., Fig. 3b, bottom). It indicates that single stations almost always loaded highly on only one VARIMAX PC, and that the rotated analysis has isolated individual areas of coherent precipitation. As noted in Section 4 and exemplified by the very different patterns in Fig. 3a, the unrotated analysis failed to isolate such clusters in the data.

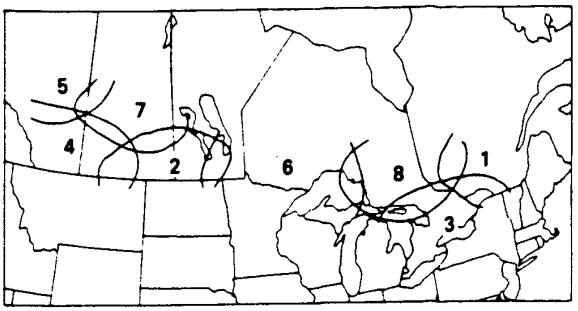


Fig. 5. Regionalization of south-central Canada for three-day summer precipitation on the basis of the VARIMAX-rotated Principal Component (PC) patterns in Fig. 4. The regional boundaries are the 0.4 loading isopleths for individual PCs, and the numbering of the regions is consistent with Table 2 and Figs. 3b and 4.

In an attempt to verify the foregoing VARIMAX PC results, two additional sets of computations were performed. First, obliquely-rotated (Harris-Kaiser Case II B'B criterion) PCs were extracted from the interstation correlation matrices for both time-scales (see Table 2). Their spatial loading patterns (not shown) essentially duplicated the VARIMAX ones in Fig. 4; congruence coefficient calculations yielded magnitudes of 0.977-0.990. Consistent with this, the average Harris-Kaiser inter-PC score correlation magnitudes were only 0.111 (3-day) and 0.103 (7-day), which correspond to mean angular separations between PCs of 83.6° and 84.1°, respectively. This clearly points to the dominant modes of variation of the study phenomenon being nearly orthogonal in time, and eliminates for the present study the possibility that "A drawback of an alternative (i.e., rotated) solution is its lack of uniqueness" (Horel, 1981). Further confirmation of the VARIMAX PC results was provided by a correlation-based Harris-Kaiser CF analysis for the three-day time-scale (not shown), for which the common variance was 74 percent of the total variance. The congruence coefficients between the loading patterns in Fig. 4 and their Harris-Kaiser CF counterparts were all in excess of 0.999.

All of the above rotated solution results are fully consistent with those obtained by RL for the central United States.

6 Further verification of rotated patterns

Because of the considerable potential utility of the regionalization in Fig. 5 (see next Section), the VARIMAX PC patterns (Fig. 4) on which it is based were subjected to several further forms of verification. The procedure sought to establish the statistical reliability (stable and reproducible?) and validity (consistent with interstation correlation input matrix?) of the patterns, and largely followed that introduced in RL for a similar purpose. It was limited to the three-day time-scale, due to the strong agreement between the three- and seven-day patterns reported in Section 5. Results appear in Figs. 6-8.

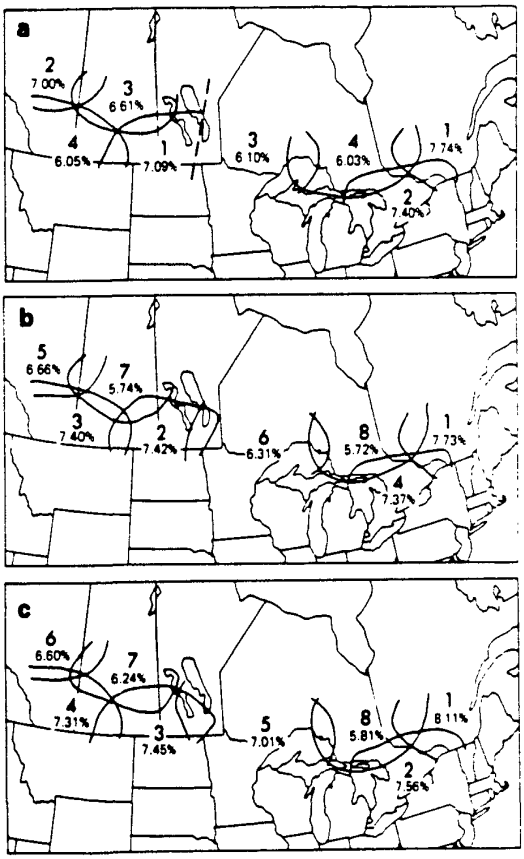


Fig. 6. Regionalization of south-central Canada for square-root transformed three-day summer precipitation on the basis of VARIMAX-rotated Principal Components (PCs) computed separately for (a) the western and eastern halves of the domain (broken line gives boundary) from the entire 1949-1980 data set, (b) the complete domain for 1949-1964, and (c) the complete domain for 1965-1980. Regional boundaries are the 0.4 loading isopleths for individual PCs, regions are numbered by root number, and the percent total variance explained by each root appears in parentheses. In each case, the number of roots retained and rotated was suggested by the procedure outlined in Sections 4 and 5.

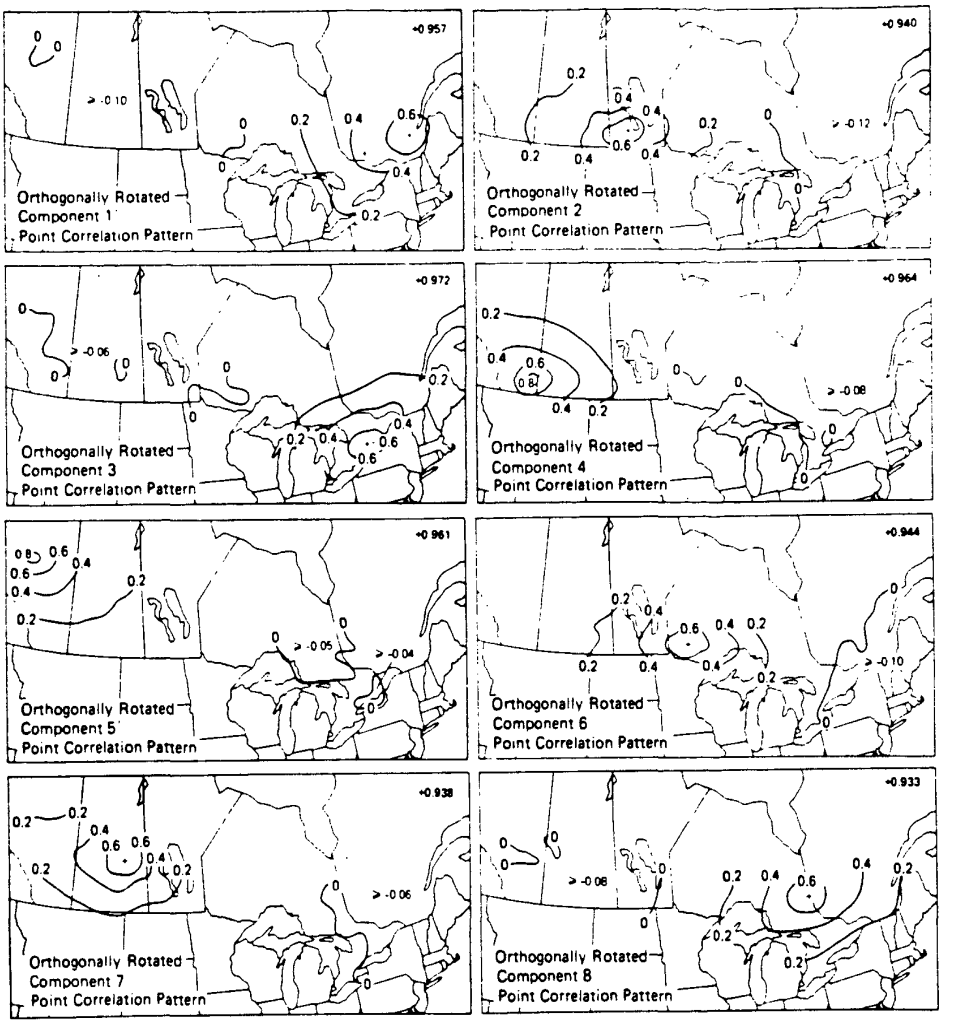


Fig. 7. Point correlation (teleconnection) patterns corresponding to the VARIMAX-rotated Principal Components (PCs) in Fig. 4. Cross locates station with the highest PC loading in Fig. 4. Correlation coefficients between that station and the remaining 169 stations are isoplethed; those with magnitude less than approximately 0.06 (0.08) are not significant at the 95% (99%) level according to the z-test. Number in upper right corner of each panel is congruence coefficient between the correlation field and corresponding PC loading pattern in Fig. 4.

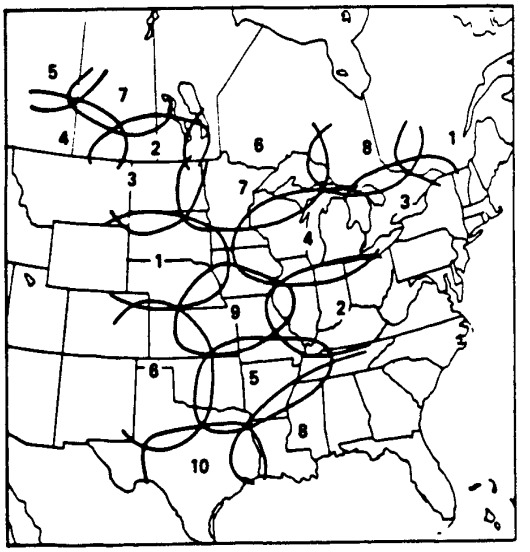


Fig. 8. Combined regionalization of south-central Canada with central United States for three-day summer precipitation. The Canadian pattern is taken from present Fig. 5 and the United States pattern from RL's Fig. 5a.

The first verification involved performing separate VARIMAX PC analyses on subsets of the data, and was strongly supportive of the patterns in Fig. 4. The latter emerged unequivocally when all (i.e., 1949-80) data for the western and eastern halves of the domain were considered separately (Fig. 6a), and when the data for the entire domain were halved in a temporal sense into 1949-64 and 1965-80 subsets (Figs. 6b, c). The congruence coefficients between the patterns in Fig. 4 and those summarized in Fig. 6a (Figs. 6b, c) ranged from 0.964-0.983 (0.927-0.997) with a mean of 0.976 (0.978).

The next verification quantitatively compared the VARIMAX PC spatial loading patterns in Fig. 4 with the interstation correlation matrix from which they were derived. For each VARIMAX PC, the station with the highest loading was identified. The three-day square-root transformed rainfall totals for each of those stations were then separately correlated with their counterparts for the remaining 169 network stations (Fig. 1). Next, the resulting fields of correlation coefficients were isoplethed to produce point correlation (or teleconnection) patterns, and those patterns were then related to their corresponding VARIMAX loading fields via the congruence coefficient used earlier. Results appear in Fig. 7. The point correlation patterns shown there are highly similar to Fig. 4's VARIMAX PC loading counterparts with respect to location, size, morphology, and magnitude. This is reflected in the congruence coefficients for all VARIMAX-point correlation pairs being in the 0.933-0.972 range (mean of 0.951), and strongly supports the relatively uniform size and strength of the VARIMAX patterns identified in Section 5. Clearly, this verification procedure has confirmed the validity of our VARIMAX PC results.

7 Summary and applications

This paper has presented the results of climatic pattern analysis of three- and seven-day summer (May-August) precipitation totals for south-central Canada. It is intended to complement a recent counterpart study for the central United States (Richman and Lamb, 1985, here termed RL). A range of eigenvectorial methods was applied to 1949-80 data for a fairly uniform network of 170 stations that extends to the east of the Rockies, to the north of the Great Lakes, and through the St. Lawrence Valley. The major objectives were to identify the characteristic spatial patterns of the above precipitation (including any dependence on eigenvectorial model) and to suggest some potential applications of the results.

The entire domain variance fraction cumulatively explained by the first eight unrotated Principal Components (PCs), and also the eight orthogonally-rotated (VARIMAX criterion) PCs derived from them, was approximately 55 percent for square-root transformations of both three- and seven-day totals (Table 2). The unrotated PC that the two sets of regional boundaries are almost perfectly aligned at the international border, a particularly encouraging result. Furthermore, the cross-border regions that are so delineated (Fig. 8) appear very meaningful in agricultural system or ecosystem senses -- southern Saskatchewan-Manitoba is linked with North Dakota and northeastern Montana (cereal production), southwestern Ontario with northern Minnesota-Wisconsin (largely forested), and southeastern Ontario with Michigan and southern Wisconsin (mixed agriculture).

As for Section 5, all of the above findings are fully consistent with those obtained by RL for the central United States. The final verification employed pursued this theme of Canadian-United States consistency to a logical end by comparing Fig. 5's regionalization with its RL counterpart (Fig. 5a there). Fig. 8 shows spatial loading patterns (Figs. 2 and 3a) possessed moderately strong large-scale (e.g., one-third to half domain) anomaly features through root 3, but thereafter decayed progressively on both counts. The higher roots exhibited little structure and seemed especially difficult to interpret physically. While these results parallel those obtained by RL for the central United States, the Canadian patterns are much less "Buell-like" (see RL, pp. 1133-1134 and their Fig. 2), probably because of the irregular shape of the present domain.

The VARIMAX PC spatial loading patterns (Figs. 3b, 4, 5) were very different to their unrotated counterparts. Each emphasized a relatively strong anomaly in a different and somewhat restricted part of the domain. The areas concerned were reasonably uniform in size and spanned areas equivalent to substantial portions of provinces. Within most of each area, one VARIMAX PC explained at least 15 percent of the station variance; this fraction rose to central values of 50 percent for almost every region. Much lower station variance fractions tend to be accounted for by the unrotated PCs. In a further contrast to the latter, the VARIMAX PC patterns did not degenerate and lose their structure with increasing root number; number 8 was just as distinctive as number 1. All of these VARIMAX PC results are fully consistent with those obtained by RL for the central United States.

The VARIMAX PC loading patterns collectively yielded a rather complete regionalization of the domain into 8 subareas within which three- and seven-day summer precipitation tends to be spatially coherent (Fig. 5). That regionalization was strongly supported by (i) obliquely rotated (Harris-Kaiser Case II B'B criterion) PC and Common Factor (CF) analyses, (ii) further VARIMAX PC analyses on subsets of the data, (iii) quantitative comparison with the interstation correlation matrix from which it is derived, and (iv) the fact that it dovetails closely with its central United States counterpart in RL. This regionalization and its associated VARIMAX PC score time series are now providing a basis for the selection of case study seasons for climate system investigations of the type advocated in the introduction; see Lamb and Peppler (1985) for similar United States work. It is also likely that the regionalization will be of utility in the areas of crop yield modeling, 48-96 hour precipitation prediction, some types of climatic change studies, precipitation chemistry and mesometeorology research, and perhaps watershed- and basin-scale surface hydrology. Considerable detail in this regard appears in RL, to which reference is again made.

Acknowledgments

This study was supported by NSF Grant ATM 83-12653. We appreciated the opportunity to present most of the above material at the Twentieth Annual Congress of the Canadian Meteorological and Oceanographic Society in Regina, Saskatchewan, during 3-6 June 1986. It is a pleasure to acknowledge the courteous and helpful data acquisition assistance given by Mike Webb (AES, Downsview) and the technical expertise of Rebecca Runge and Linda Riggin, and the research assistance of Phyllis Stone.

APPENDIX

On the Retention and Rotation of Principal Components

Sections 4 and 5 noted the considerable difference in the number (8-25) of Principal Components (PCs) that standard tests suggested should be retained and rotated in the present study. This situation, coupled with the fact that the rotated PC spatial loading patterns are functionally dependent on the number of roots rotated, prompted the present quantitative assessment of the effects of rotating varying numbers of roots. After considerable experimentation, this assessment used a "mean absolute difference" coefficient given by

$$d_{AB} = \sum_{i=1}^{M} \left[\sum_{j=1}^{N} (|a_j - b_j|)/N \right]/M$$

where A is a PC loading vector, B is the interstation correlation vector for the station with the highest absolute loading for that PC (see Section 6), M is the number of PCs rotated (variable, 2-23 for the three-day time-scale used here), and N is the number of stations (constant, 170). The d_{AB} results are summarized in Fig. A1, along with the range of values and standard deviations associated with them.

Fig. A2 is an equivalent of Fig. 5 for rotations of between two and 10 PCs.

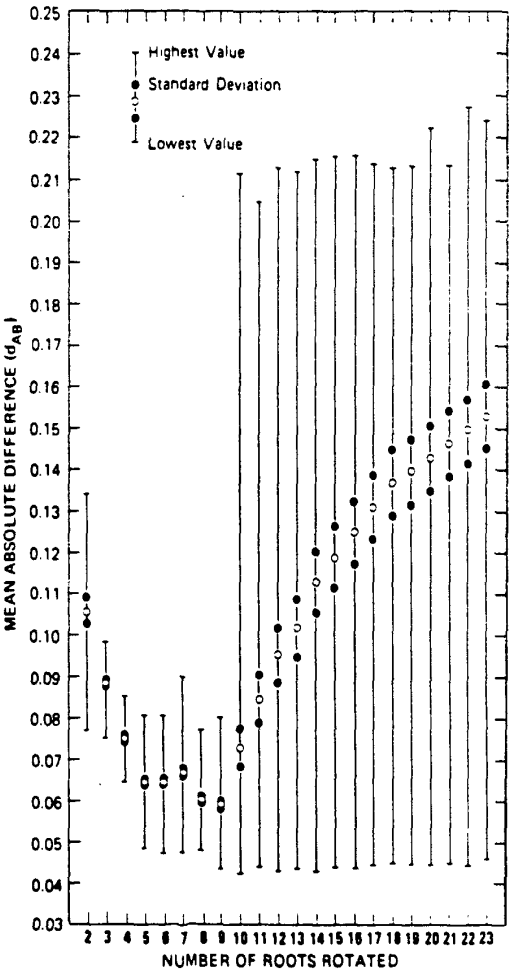


Fig. A1. Mean absolute difference coefficient (open circles, see text) as a function of number of Principal Components VARIMAX-rotated. Also shown are the range and standard deviation of the individual absolute differences associated with each coefficient.

The regionalization produced by the rotation of only two roots is incomplete, with most of western Ontario being excluded (see Section 5 for discussion of "completeness" in this context). This solution thus did not merely halve the domain. The regionalizations emanating from the rotations of 3-9 PCs are generally fairly complete, with the size of individual regions decreasing with increasing root number. However, the rotations of 10-23 PCs increasing yielded fragmented and weak spatial loading patterns for the higher roots (e.g., Fig. A2's 10 PC summary map; Fig. A3) that had no basis in the interstation correlation matrix (not shown) and which, in turn, would only break down the regionalizations given by the lower roots if utilized.

Fig. A1 shows that d_{AB} in general decreased from the rotation of two roots to the rotation of eight or nine, which reflects increasing PC coherency and enhanced PC scale-match with features of the interstation correlation matrix (see Fig. 7 for examples of the latter), and then increased logarithmically from the rotation of nine to the rotation of 23 roots. A

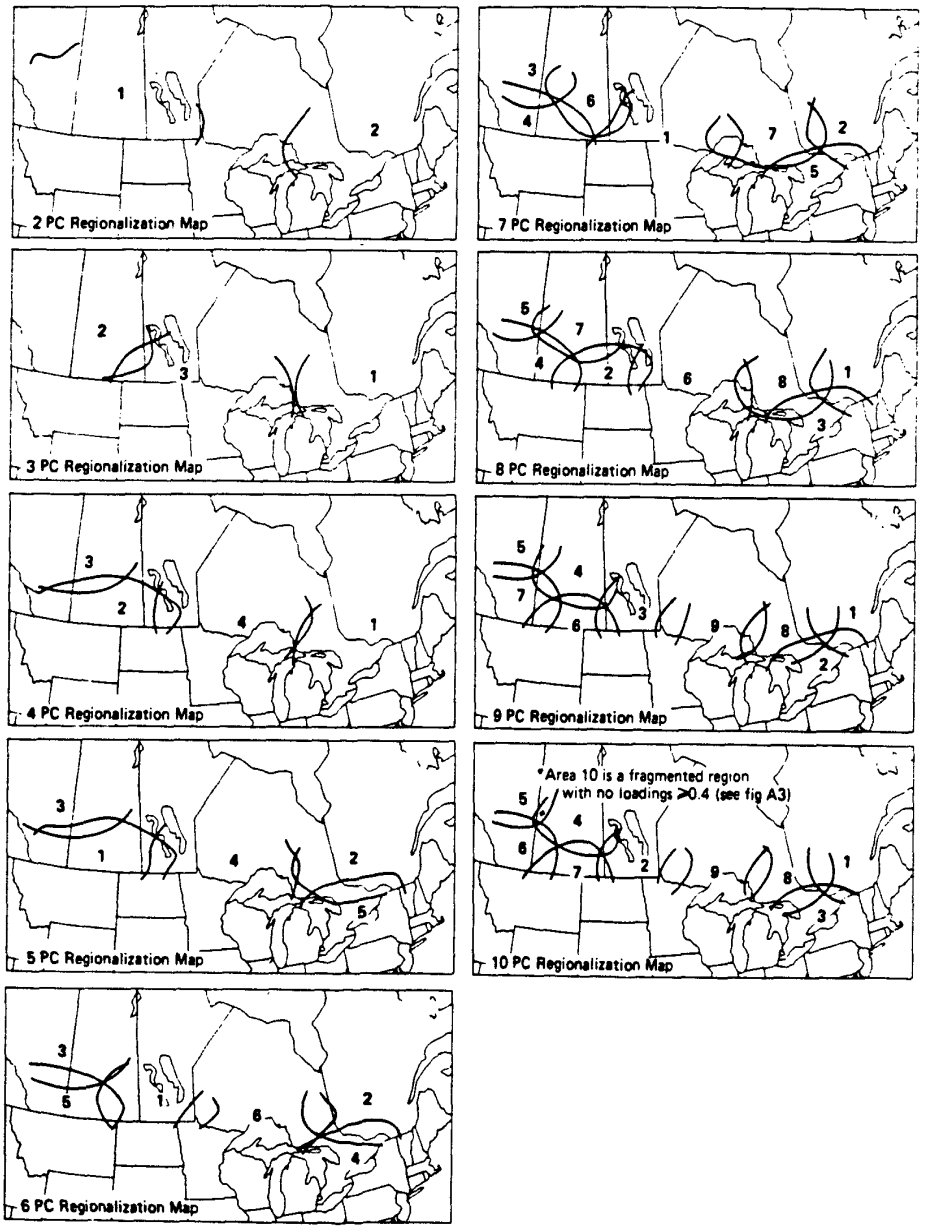


Fig. A2. Same as Fig. 5 for the VARIMAX-rotation of 2-10 Principal Components. The asterisk in the "10 PC Regionalization Map" locates the station with the highest loading on root number 10. The latter's loading pattern appears in Fig. A3.

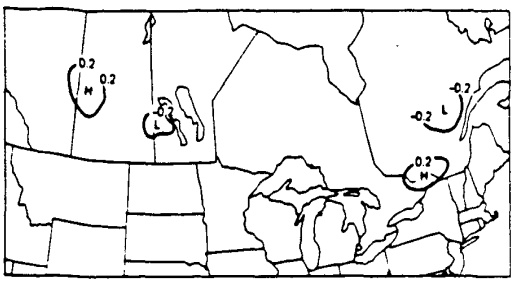


Fig. A3. Same as Fig. 4 for VARIMAX-rotated Principal Component number 10 when ten roots were rotated.

pronounced amplification in the range of absolute differences contributing to d_{AB} (see above equation) and their associated standard deviations occurred between the rotations of nine and ten PCs; both parameters were relatively constant for rotations of 10-23 PCs. The d_{AB} values for the rotations of eight and nine PCs are separated by less than a standard deviation and are therefore essentially indistinguishable.

The above results clearly suggest that the optimum number of PCs to retain and rotate in the present case is either eight or nine. Our choice of eight in Sections 4 and 5 was prompted by (i) its support from the application of the "scree" test to the eigenvalue series reported in Section 4, (ii) its smaller range of absolute differences contributing to d_{AB} in Fig. A1, and (iii) the fact its regionalization (Figs. 5, A2) matched better with the United States counterpart in RL (see Section 6).

This Appendix has thus presented a combined statistical-physical approach to the problem of deciding the number of PCs to retain and rotate. We believe it is superior to depending solely and in a purely statistical sense on the traditional tests referenced in Section 4. It accordingly may be of use in other studies.

References

Cattell, R. B. 1966. The scree test for the number of factors. Multivariate Behavioral Res. 1: 245-276.

Gregory, S. 1975. On the delineation of regional patterns of recent climatic fluctuations. Weather. 30: 276-287.

Guttman, L. 1954. Some necessary conditions for common-factor analysis. Psychometrika. 19: 149-161.

Harman, H. H. 1976: Modern Factor Analysis (Third Revised Edition). The University of Chicago Press, Chicago, 487 pp.

Horel, J. D. 1981. A rotated principal component analysis of the interannual variability of the Northern hemisphere 500 mb height field. Mon. Wea. Rev. 109: 2080-2092.

Horel, J. D. 1984. Complex principal component analysis; theory and examples. J. Clim. Appl. Meteor. 23: 1660-1673.

Kaiser, H. F. 1960. The application of electronic computers to factor analysis. Educational and Psychological Measurement. 20: 141-151.

Karl, T. R., A. J. Koscielny and H. F. Diaz. 1982. Potential errors in the application of principal component (eigenvector) analysis to geophysical data. J. Appl. Meteor. 21, 1183-1186.

Kendall, M. 1980. Multivariate Analysis. Charles Griffin, High Wycombe, 210 pp.

Lamb, P. J. and R. A. Peppler. 1985. Tropospheric static stability and central North American summer rainfall during 1979. Proceedings of Ninth Annual Climate Diagnostics Workshop, United States Department of Commerce, Washington, D.C., 274-283.

Lund, I. A. 1963. Map pattern classification by statistical methods. J. Appl. Meteor. 2: 56-65.

Mulaik, S. A. 1972. The Foundations of Factor Analysis. McGraw Hill, New York, 453 pp.

North, G. R., T. L. Bell, R. F. Cahalan and F. J. Moeng. 1982. Sampling errors in the estimation of empirical orthogonal functions. Mon. Wea. Rev. 110: 699-706.

Preisendorfer, R. W., and T. P. Barnett. 1977. Significance tests for empirical orthogonal functions. Preprints of Fifth Conference on Probability and Statistics in Atmospheric Sciences, American Meteorological Society, Boston, 169-172.

Richman, M. B. 1981. Obliquely rotated principal components: an improved meteorological map typing technique. J. Appl. Meteor. 20: 1145-1159.

Richman, M. B. 1986. Rotation of principal components. J. Clim. 6: 293-335.

Richman, M. B. and P. J. Lamb. 1985. Climatic pattern analysis of three- and seven-day summer rainfall in the central United States: some methodological considerations and a regionalization. J. Climate and Applied Meteorology. 24: 1325-1343.

Rogers, J. C. and H. van Loon. 1982. Spatial variability of sea level pressure and 500 mb height anomalies over the Southern Hemisphere. Mon. Wea. Rev. 110: 1375-1392.

Rummel, R. J. 1970. Applied Factor Analysis. Northwestern University Press, Evanston, IL, 617 pp.

Stidd, C. K. 1967. The use of eigenvectors for climatic estimates. J. Appl. Meteor. 6: 255-264.

Thurstone, L. L. 1947. Multiple Factor Analysis. The University of Chicago Press, Chicago, 535 pp.

Walsh, J. E. and M. B. Richman. 1981. Seasonality in the associations between temperatures over the United States and the North Pacific Ocean. Mon. Wea. Rev. 109: 767-783.

Whittaker, L. M. and L. H. Horn. 1982. Geographical and seasonal distribution of North American cydogenesis, 1958-1977. Mon. Wea. Rev. 109: 2312-2322.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

SUBSTANTIATION OF THE CLIMATE MONITORING DATA SET COMPOSITION

G.V.Gruza
Hydrometeorological Centre of the USSR

The last decade has seen the problem of the organization and development of a system of climatic and environmental monitoring to take shape [10].

The main goal of a climate monitoring is to summarize data on the current state of the climatic system and to estimate anomalous behaviour of this state in terms of probability. Within the framework of a climate monitoring system, it would be useful to reveal the physical causes of the observed anomalies in the global meteorological processes and to estimate the scales of possible variations in future.

As a matter of definition we look into certain principal requirements for a climate monitoring system. First of all, such a system must function on a set time basis, subject to the existed regulations (in terms of the product composition and terms of issue, users network, feedback etc.).

Secondly, monitoring should provide an adequate representation of the states of the climatic system over all portions of its time-space variation spectrum. Therefore time-space smoothing in such systems seems inevitable, and on different scales, from a few weeks to decades, and from local to global scales. (Accordingly, we categorize monitoring into global, regional and local, with the following temporal characteristics: month - season - year - decades).

Still another requirement assumes that the quantitative data aquired on the monitoring objects be informative and complete. We imply by this not only that the spacial and temporal characteristics of an object be represented fully but also that the quantitative parameters be informative enough to include, in addition to the spatial (temporal) mean, the spatial (temporal) variability, the degree and character of the anomaly, etc. Below we give a set of statisticals which, taken together seems to provide a rather complete and comprehensive description of the regional features of meteorological fields. We believe the set is reasonably acceptable, at least as a basis for field parametrization in a surface monitoring system.

Finally, the following principle is inherent in monitoring systems: a current state represented by current values of a given system of parameters must be necessarily analysed against the background of long-term variations of these parameters. That implies that adopted parametrization of an object should be supported by its historic (long-term) time series as well as its statistical generalizations (means, variances, distribution functions, etc.). Such an approach would permit one to pass on to the probability characteristics of the current state and to estimate how anomalous.

All the estimates given below were obtained on the basis of empirical data on the monthly mean air temperature in the nodes of a regular grid with a step of 5° in latitude and 10° in longitude for 1891-1982. (This data file was analysed in detail and described in works [1,5,6,7,8, 11]. Further data (from 1983 onwards) have been aquired in the Hydrometeorological Centre of the USSR by a similar technique, in the quasi-on line mode and arrive in our system with a month's lag.

1. STATISTICAL CHARACTERISTICS OF CURRENT AIR TEMPERATURE FIELDS IN THE NORTHERN HEMISPHERE AS FIGURES OF MERIT OF SURFACE MONITORING

For the purpose of climate monitoring it has been suggested to use a set of parameters permitting characterization of different scale anomalies in comparable units [7].

The set of parameters for sea-level air-temperature monitoring is summarized in Table 1.

STATISTICAL CHARACTERISTICS OF AIR TEMPERATURE FOR CLIMATE MONITORING AND THEIR IDENTIFICATION IN THE MACHINE BULLETIN

Table

Statistical characteristics	Area of averaging		
	Point	Latitudinal circle	Region
Mean monthly value	T	ZLT	ZLP T
Mean monthly anomaly	7G T	ZL 7GT	ZLP 7GT
Mean (over space) quadratic deviation of anomalies	-	ZL 7GT	ZLP 7GT
Quadratic mean (over space) of normalized anomaly	-	ZL WGT	ZLP WGT
Probability (in a long-term series) of non-exceeding the mean (over the region) anomaly	PG 7GT	PG ZL 7GT	PG ZLP 7GT

Here we explain their designations.

In a data file T_{gmfl}, g - codes the year, m - the month, f - the latitudinal circle, and l - the meridian.

For simplicity, we introduce the averaging operators first for a two dimentional matrix $\{X_{ij}\}$, where i - is the running index for the operator, and j - is the summary index which combines all the other coordinates:

1. Averaging $E^i(X_{ij}) = \sum_i P_i X_{ij} = E^i_j X$

2. Deviation from the mean (anomaly) $V^i(X_{ij}) = X_{ij} - E^i_j X = V^i_{ij} X$

3. Quadratic averaging $Q^i(X_{ij}) = \sqrt{\sum_i P_i X^2_{ij}} = Q^i_j X$

4. Standard deviation $S^i(X_{ij}) = Q^i(V^i_{ij} X) = S^i_j X$

5. Normalized deviation $W^i(X_{ij}) = V^i_{ij} X / S^i_j X = W^i_{ij} X$

6. Empirical variance $D^i(X_{ij}) = (S^i_j X)^2 = D^i_j X$

7. Empirical probability $P^i(X_{ij}) = \int_{-\infty}^{x} f^i_x di = P^i_j X$

Now return to the temperature data file. The indices over which the operator is taken (i.e., a statistical generalization is performed), are superscripts. It may be any subgroup of indices (coordinates) of the elements of the initial file, so the temporal averaging is performed over the indices g and m and the spatial averaging over f and l. For example,

$$E^g(T_{gmfl}) = E^g_{mfl} T = \frac{1}{NG} \sum_{g=1}^{NG} T_{gmfl} = EGT$$

is the long-term mean and

$$E^{lm}(T_{gmfl}) = E^{lm}_{gf}T = \frac{1}{NL \cdot NM} \sum_{l=1}^{NL} \sum_{m=1}^{NM} T_{gmfl} = ELMT$$

is the annual-latitudinal mean.

Thus, the variation range of the superscripts in an operator designation defines the region of averaging, while the subscripts indicate the coordinates retained in the result, i.e., they define the variation range of the argument. By way of example, E^{g}_{mfl} stands for the long-term mean taken in grid node f,l and in month m. The abbreviations in the right-hand side of the formilae are used to identify the respective statistical characteristics in the computer tables of monitoring bulletin.

The expressions for the other operators over the same averaging range are as follows.

Quadratic averaging: $$Q^{g}(T_{gmfl}) = Q^{g}_{mfl}T = \sqrt{\frac{1}{NG} \sum_{g=1}^{NG} (T_{gmfl})^2} = \sqrt{E^{g}\left\{(T_{gmfl})^2\right\}} = QGT$$

$$Q^{lm}(T_{gmfl}) = Q^{lm}_{gf}T = \sqrt{E^{lm}_{gf}(T^2)} = QLMT$$

The current (for a given year-month-node) value of deviation from the respective long-term mean is

$$V^{g}(T_{gmfl}) = V^{g}_{gmfl}T = T_{gmfl} - E^{g}_{mfl}T = VGT$$

and from the annual-latitudinal mean is

$$V^{lm}(T_{gmfl}) = V^{lm}_{gmfl}T = T_{gmfl} - E^{lm}_{gf}T = VLMT$$

Empirical variances

$$D^{g}(T_{gmfl}) = D^{g}_{mfl}T = E^{g}\left\{(V^{g}_{gmfl}T)^2\right\} = Q^{g}(V^{g}_{gmfl}T) = DGT$$

and

$$D^{lm}(T_{gmfl}) = D^{lm}_{gf}T = E^{lm}\left\{(V^{lm}_{gmfl}T)^2\right\} = Q^{lm}(V_{gmfl}T) = DLMT$$

estimate respectively the interannual variability in a fixed month-node of the grid and intermonthly intralatitudinal variance of point temperature values for a certain year-latitude.

As usual, root-mean-square (standard) deviations alter only the scale of estimates for D:

$$S^{g}(T_{gmfl}) = S^{g}_{mfl}T = \sqrt{D^{g}_{mfl}T} = Q^{g}_{mfl}(V^{g}T) = SGT \ ;$$

$$S^{lm}(T_{gmfl}) = S^{lm}_{gf}T = \sqrt{D^{lm}_{gf}T} = Q^{lm}_{gf}(V^{lm}T) = SLMT$$

Normalized anomalies or referred deviations:

$$W^{g}(T_{gmfl}) = W^{g}_{gmfl}T = \frac{V^{g}(T_{gmfl})}{S^{g}(T_{gmfl})} = \frac{T_{gmfl} - E^{g}_{mfl}T}{S^{g}_{mfl}T} = WGT \ ;$$

$$W^{lm}(T_{gmfl}) = W^{lm}_{gmfl}T = \frac{V^{lm}(T_{gmfl})}{S^{lm}(T_{gmfl})} = \frac{T_{gmfl} - E^{lm}_{gf}T}{S^{lm}_{gf}T} = WLMT$$

Finally, by an empirical probability of a current temperature value we mean the result of integral probability transformation, which is interpreted as a probability of non-exceeding the current value on a set of all element values from the averaging range

$$P^{g}(T_{gmfl}) = P^{g}_{mfl}T = \int_{-\infty}^{T} f^{g}_{T}(x)dx = PGT \ ;$$

$$P^{lm}(T_{gmfl}) = P^{lm}_{gf}T = \int_{-\infty}^{T} f_T^{lm}(x)dx = PLMT$$

Here, $f_T(x)$ is the temperature distribution density in the averaging range (i.e., over the set of the superscript values).

The quantity $P(T)$ implies a non-linear transform of T treated as a random variable with a uniform distribution on the interval $[0,1]$. It enables various probability distributions to be reduced to a single scale. It may be computed either from a previously estimated empirical distribution (or from a regular series), or proceeding from an a priori accepted distribution law according to the empirical estimates of parameters $\bar{E}$, S. For temperature, it would be natural to assume a normal distribution $f^g_T(x)$ so that

$$P^g(T_{gmfl}) = 0{,}5\ [1 + S_g(W^g_{gmfl}) \cdot ERF(|W^g_{gmfl}T|)],$$

where

$$ERF(x) = \frac{1}{\sqrt{2\pi}} \int_0^x 1^{-\frac{t^2}{2}} dt$$

These definitions and equations may form a bases for a language of descriptive meteorological statistics[9].

Returning to the parameters in Table 1, we recall that here "region" stands for a geographical region, covering a set of points (f,l) over which a respective statistical operator is taken. Fig.1 presents all regions examined in this monitoring system. They cover the northern hemisphere in such a pattern as to meet the requirements of monitoring in all the three spatial scales - global, regional, and local.

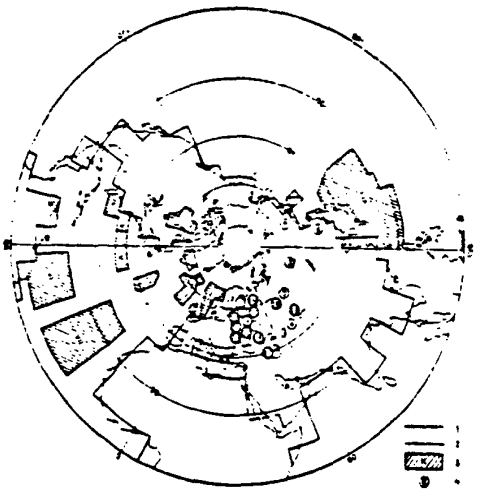


Fig.1. Map of regions of spatial averaging in the bulletin of surface climate monitoring:
1 - conventional boundaries of Europe-Asia-Africa;
2 - conventional boundaries of USSR, Canada, USA;
3 - zones of the ocean-atmosphere intense heat transfer (K = I-VIII);
4 - physiographic regions of the USSR (K = 1.19).

2. LONG-TERM TEMPERATURE STATISTICS FOR THE NORTHERN HEMISPHERE.

The main feature of diagnostic research into climate monitoring is that the current state of the climatic system must be analysed in comparison with the respective long-term time series and their statistics. There-

fore, for all the parameters shown in Table 1, first of all the time series (1891-1982) have been computed and from them the long-term statistics E^g, S^g, P^g etc. have been estimated. (Actually, we have derived a more expanded set of such estimates for all basic meteoelements and summarized them as the reference material [2,3,4,10]; these data will be regularly updated in future as new information is acquired).

Figure 2 shows, as space-time isopleths, the long-term mean for latitudinal and intralatitudinal standards and the zonal intensity of anomalies (root-mean-square normalized anomaly). They readily indicate some specific features of the zonal temperature distribution familiar from other works, namely: the displacement of calendar seasons to the north and south from ~50°N, especially pronounced for the summer maximum in subtropics; the zonal variation maximum in moderate latitudes (~65°-70°N) throughout the year; the anomaly maximum of the processes in moderate latitudes (40°-60°N) insignificantly increased in the transitional seasons (on the whole, the annual variation for the anomality is absent).

We note only the revealed appreciable growth of the zonal inhomogeneity and anomalous behaviour of the processes in these regions (with the exception of the summer in the polar regions) throughout the 92-year period

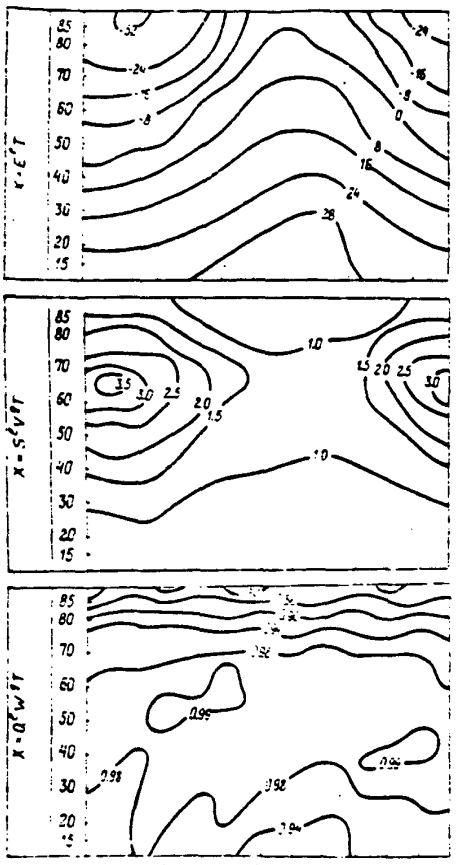


Fig.2. Latitudinal-annual variation of long-term zonal mean characteristics of mean monthly air temperature fields, the Northern Hemisphere, 1891-1982:
E^g - mean latitudinal temperature;
$S^g \nu$ - intralatitudinal standard of anomalies;
$Q^g w$ - latitudinal quadratic mean of normalized anomalies.

under examination; its intensity, however, is not the same in various seasons.

Examine Fig. 3 which plots the interlatitudinal cross-correlations of the characteristics under study (for the month of January as an example). For the purposes of climate monitoring, the diagonal elements of correlation matrices, containing "one-point"correlations of different characteristics, are of special interest.

A striking feature here is an intimate correlation between zonal inhomogeneity and zonal

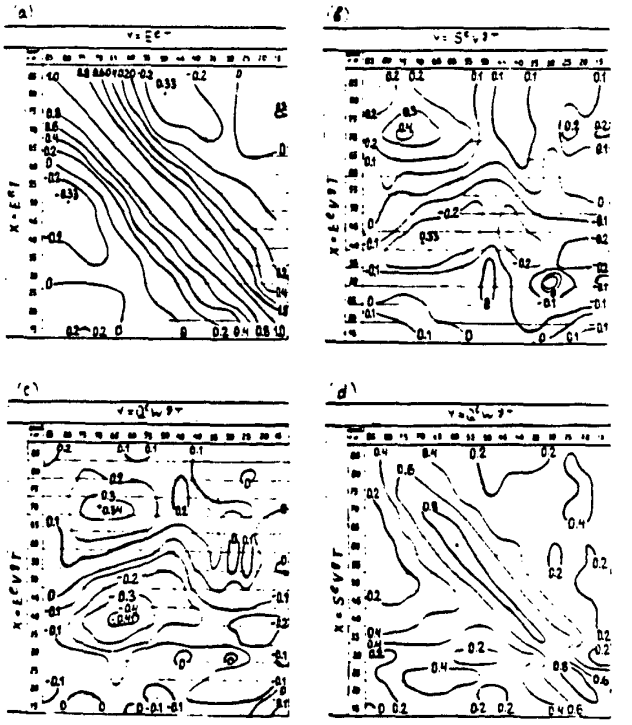


Fig.3. Interlatitudinal correlation $R(Y_\varphi, X_\varphi)$ of temperature characteristics for the Northern Hemisphere latitudinal zones, 1891-1982, January:
E^g - mean latitudinal temperature,
$S^g \nu$ - intralatitudinal standard of anomalies,
$Q^g w$ - latitudinal quadratic mean of normalized anomalies.

mean anomaly (Fig.3d), which somewhat abates in polar latitudes and is absent for each of these characteristics when coupled with the zonal mean anomaly (Fig.3d,c).In the 35°-40°N zone these characteristics exhibit significant negative correlation with latitudinal mean temperatures of practically the whole of the hemisphere, with the maximum being at approximately 65°N.

Another interesting (and apparently new) fact is a significant negative correlation revealed between zonal temperatures in polar and moderate latitudes. Taken as a whole, the data presented in Fig.3 confirm the validity of parallel use of all these characteristics in a monitoring system, because they complement each other, rather than replace and together provide multifacetal description of large-scale temperature fields.

3. TEMPERATURE REGIME IN 1984

Consider the principal findings of surface air temperature monitoring (at sea level) in the northern hemisphere taking the 1984 data as an example.

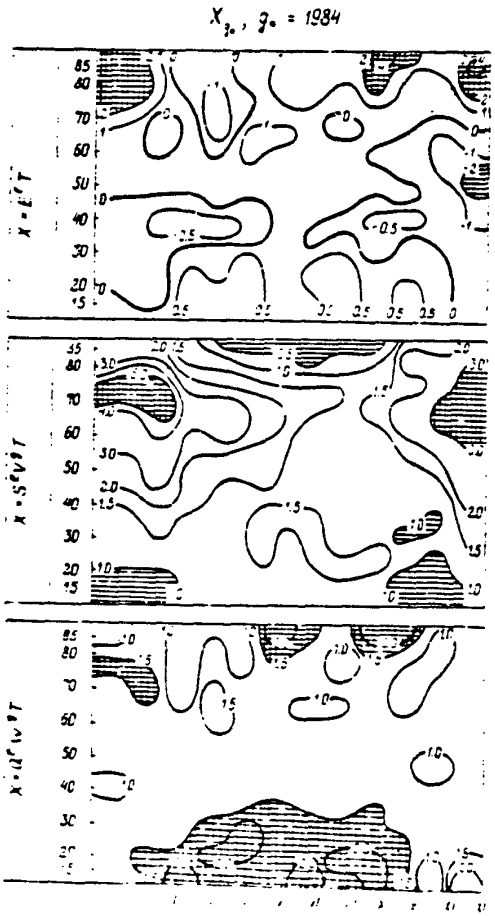


Fig 4. Latitudinal-annual variation in temperature characteristics of the Northern Hemisphere in 1984:
$\bar{T}^{g}$ – latitudinal mean temperature,
$S^{g}T$ – intralatitudinal standard of anomalies,
$Z^{g}T$ – latitudinal quadratic mean of normalized anomalies.

The zonal structure peculiarities of the 1984 temperature fields are pronounced in Fig.4, showing time-latitude isopleths of three spatial (zonal) statistics, $E^{1}V^{g}T$, $S^{1}V^{g}T$, and $Q^{1}W^{g}T$. Compared with the respective long-term isopleths in Fig.2, a discrete nature of increased anomaly regions may be noticed (both in time and space). The principal feature of the field is apparently a northward displacement and increase in the winter maximum of the intralatitudinal field inhomogeneity and an enhanced spatial inhomogeneity of the field in low latitude during most of the year. The last implication though might be attributed to low accuracy of operational data over subtropical latitudes of the Pacific.

REFERENCES

1. Angell J.R., Gruza G.V., 1984: Climate variability as estimated from atmospheric observations. "The Global Climate." edited by J.Houghton; Cambridge University Press, 25-36.
2. Apasova E.G., Gruza G.V.,1982: Data on climate structure and variability. Precipitations. The Northern Hemisphere. Obninsk: VNIIGMI-MCD, in : "Data in climate structure and variability.", 212pp.
3. Gruza G.V., Kleschenko L.K., 1982: Data on climate structure and variability. Geopotential of the 500 gPa gPa surface and some characteristics of circulation. The Northern Hemisphere. In: "Data on climate structure and variability", 218pp.
4. Gruza G.V., Kleschenko L.K., Rankova E.Ya., 1981: "On the prognosis of monthly mean fields of air temperature over the Northern Hemisphe re using automized scheme of group

analogues. Meteorology and hydrology, N2, 28-39.
5. Gruza G.V., Kleschenko L.K., Timofeeva T.P., 1982: On the variability of temperature and circulational regimes of the atmosphere of the Northern Hemisphere. Meteorology and hydrology, N3, 8-20.
6. Gruza G.V., Rankova E. Ya., 1979: Data on climate structure and variability. The sea level air temperature. The Northern Hemisphere. In: "Data on climate structure and variability", 203 pp.
7. Gruza G.V., Rankova E.Ya., 1980: Structure and variability of the observed climate. Air temperature over the Northern Hemisphere. Hydrometeoizdat, 72 pp.
8. Gruza G.V., Rankova E.Ya., Semenyuk E.A., 1985: Climatic variability of characteristics of large-scale temperature anomalies. Meteorology and hydrology, N5, 67-80.
9. Gruza G.V., Reitenbakh R.G., 1982: Statistics and analysis of hydrometeorological data. Leningrad, Hydrometeoizdat, 216 pp.
10. Izrael Yu.A., 1979: Ecology and natural environment control. Leningrad, Hydrometeoizdat, 376 pp.
11. Robock A., 1982: The Russian Surface Temperature Data Set. (1891-1980). Journal of Applied Meteorology, 21, December, 1781-1785.

Third International Conference on Statistical Climatology
June 23-27, 1986, Veinna, Austria

NON-PARAMETRIC METHOD FOR CONSTRUCTION OF TEST REFERENCE YEAR

Lučka Kajfež-Bogataj
Edvard Kardelj`s University of Ljubljana
Agronomy Department

ABSTRACT

In the present paper methodology based mainly on nonparametric methods is presented for construction of test reference year consisting of typical months. As an example test reference year for Ljubljana is added based on observation period 1961-1980.

INTRODUCTION

At theoretical studies of environmental impact on natural and man governed processes which take place in open area classical climatological data do not satisfy up to date demands. Dynamical modelling of energy consumption of buildings (Cehak, 1982; Kletter, 1982; Lund, 1974; Olgyay, 1962), dry matter production of plant canopy (Sivakumar, 1984; Hočevar, Kajfež-Bogataj, 1984; Band, 1981), and others, respectively needs representative climatological data on hourly bases. They are usualy named test reference year.

Known models of test reference year can be divided into two groups viz. synthetical one, and historical one. Test reference year constructed by the synthetical method uses time series spectral analysis and similar other analysis of data. On the contrary,the one constructed by hystorical way takes properly selected data, which emerge from real measurements. It has some advantages in comparison with the first: Elaboration of data is fast and simple, daily courses of various meteorological elements are simultaneous and they are not masked, statistical analysis of some meteorological phenomena (days with fog, rain) in synthetic way has little sense, and so on.

The proper selection of observed data which must be representative values is a hard job. At it till now mostly parametric methods were used (Kletter, 1982). In our work nonparametric methods for their selection are used and described together with an example: Test reference year for Ljubljana constructed by the typical whole months from the observation period 1961-1980.

METHOD

Many meteorological variables do not always satisfy the conditions for parametric tests: neither being drawn from normally distributed populations, nor these population have same variances. So we searched for other possibilities.

The nonparametric statistical techniques are often called "distribution-free", one of their primary merits being that they do not assume that the analysed scores were drawn from a population distributed in a certain way, e.g. from normally distributed population. They may be used with scores which are not exact in any numerical sense, but simply ranks. Another advantage of these techniques is their computational simplicity.

The nonparametric tests in spite of having smaller power efficiency than parametric ones are very suitable, not requiring specific conditions about the population parameters. Therefore these have been used.

The selection of particular typical months of the year according to their meteorological characteristics was made in the following way: Meteorological characteristics of particular month of the year were compared with the climatological values for the period. As a typical month the one was chosen which was according to its meteorological characteristics the closest to the climatological mean value of particular month. The procedure of this selection took different phases which will be described in particular.

Classification of months

At first all the months of the year were ranked according to 23 meteorological characteristics (table 1). 16 were taken into account at all months of the year and some of them at winter or at summer months only.

Table 1: Meteorological elements used at ranking of months

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	x	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x	x	x	x	x	x	x	x
3	x	x	x	x	x	x	x	x	x	x	x	x
4	x	x	x	x	x	x	x	x	x	x	x	x
5	x	x	x	x	x	x	x	x	x	x	x	x
6	x	x	x	x	x	x	x	x	x	x	x	x
7	x	x	x	x	x	x	x	x	x	x	x	x
8	x	x	x	x	x	x	x	x	x	x	x	x
9	x	x	x	x	x	x	x	x	x	x	x	x
10	x	x	x	x	x	x	x	x	x	x	x	x
11	x	x	x	x	x	x	x	x	x	x	x	x
12					x	x	x	x	x			
13	x	x	x	x	x	x	x	x	x	x	x	x
14	x	x	x	x	x	x	x	x	x	x	x	x
15	x	x	x	x	x	x	x	x	x	x	x	x
16	x	x	x	x	x	x	x	x	x	x	x	x
17	x	x	x	x	x	x	x	x	x	x	x	x
18				x	x	x	x	x	x	x		
19					x	x	x	x	x			
20	x	x	x								x	x
21	x	x	x	x	x				x	x	x	x
22	x	x	x							x	x	x
23	x	x	x	x	x					x	x	x

(The numbers have the following meaning:
Air temperature: 1. mean, 2. maximal, 3.minimal, 4. at 7 a.m., 5. at 2 p.m., 6. at 9 p.m., 7. sunshineduration, 8. global radiation, 9. wapor pressure, 10. precipitation, number of days with: 11. fog, 12. thunderstorm, 13. cloudiness $< 0{,}2$, 14. cloudiness $> 0{,}8$, 15. precipitation $\geq 0{,}1$ mm, 16.precipitation $\geq 1{,}0$ mm, 17. precipitation $\geq 10{,}0$ mm, 18. maximal air temperature $\geq 25^{\circ}C$, 19. maximal air temperature $\geq 30^{\circ}C$, 20. maximal air temperature $< 0^{\circ}C$, 21. minimal air temperature $\leq 0^{\circ}C$, 22. minimal air temperature $< -10^{\circ}C$, 23. snow cower.

Statistical scores (in our case deviations of meteorological elements from mean values) were ranked according to their value from the largest to the smallest. Each element of each month was ranked from 1 to 20. If some months had equal deviations we take into account ties. The worst month has the largest deviation and so the highest rank. The months were classified to the described method into four groups, viz. very bad (vb), bad (b), good (g) and very good (vg) (Table 2). All but very good months were excluded at this phase.

Table 2: Classification of months into groups good (g), bad (b), very bad (vb), very good (vg)

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980	g	vb	vg	vb	vb	b	b	g	b	g	b	vb
1979	g	b	vb	b	vb	vb	b	g	vg	vg	vg	vb
1978	b	g	b	b	vb	b	vb	b	b	b	vb	g
1977	b	vb	vb	b	vg	g	g	g	vb	b	g	g
1976	b	b	vb	b	b	vb	g	vb	b	g	b	g
1975	vb	vb	g	vg	b	vb	vg	g	vb	vg	g	vg
1974	b	vb	g	vg	vg	vb	g	vb	g	vb	b	vb
1973	b	vg	b	vb	vb	g	vg	vg	b	b	b	g
1972	b	vb	b	vb	g	vg	vb	vg	vb	b	g	vg
1971	g	b	vb	b	g	b	vb	vb	vb	vb	vg	b
1970	vb	vg	b	vb	b	b	g	b	b	g	vb	b
1969	vg	vb	g	b	vb	g	b	b	g	b	vb	vb
1968	b	b	b	vb	g	g	vg	g	g	b	g	vb
1967	vg	b	b	g	b	g	vb	b	b	b	vg	b
1966	g	vb	vg	b	g	b	vb	b	b	vb	b	g
1965	g	vb	vg	b	b	vg	b	b	vg	vb	b	g
1964	vb	vg	b	g	g	vb	g	g	g	b	b	b
1963	vb	vb	vb	g	g	g	g	b	b	vg	vb	vb
1962	g	g	vb	vg	b	vb	vb	vb	g	g	vb	vb
1961	g	b	vb	vb	b	b	vb	g	vb	vb	vb	b

Selection of typical months

From the very good months we have to choose the typical one for each month of the year. To do that we could use several methods; using not all meteorologic characteristics, giving more weight to particular ones and so on, respectively. We choose the following method: We take into account all the used meteorological characteristics used at ranking and add a new time interval decade at three elements, air temperature,sunshine duration and global radiation. The cause for that is the yearly course of these elements. So we skip giving the weight to these elements "a priori" but we still accentuate their meanning (for instance: at air temperature we got three additional independent ranks for deviations from mean decade temperature). This was important from statistical test point of view which followed in two directions: We enlarged the samples and got additional differences between the ranks of months.

At selection of appropriate statistical tests we take the nonparametric ones which are according to the nature of the problem and to the mentioned advantages the most suitable (Siegel, 1956). All the very good months were compared among themselves in pairs. When statisticaly significant differences showed up we eliminate the worse one and compare the better one with the next.

The mentioned comparison of months in pairs was made by Wilcoxon matched pairs signed-ranks test (Wilcoxon, 1949) which is based on direction and on magnitude of the differences within pairs.Scores are usualy absolute numerical values but we used for them the ranks of deviations.This does not violate the demand the statistical score to be at ordinal scale which is basic at this test, thou this attempt has not been found in the literature. At practical work it has proved to be very suitable.

RESULTS

According to the described method we selected the typical months of the year. The level of significance was less than 0.05 except for three months where it was less than 0.15 what is still rather acceptable for such kind of studies. The selected typical months for Ljubljana together with the spare ones (for the case if the typical month data for same meteorological elements are incomplete) are given in table 3.

Table 3: Test reference year for Ljubljana

Month	Typical	(Spare)
January	1967	(1969)
February	1970	(1964)
March	1966	(1965)
April	1975	(1974)
May	1974	(1977)
June	1972	(1965)
July	1975	(1973)
August	1975	(1972)
September	1979	(1965)
October	1963	(1970)
November	1977	(1971)
December	1975	(1977)

From it can be seen that four standard months belong to the year 1975. From that we can speculate that the whole year 1975 can be in some way a typical one which was not the case. Other months of the year 1975 are not close to the typical ones but to the extreme ones!

CONCLUSION

The methodology for construction of test reference year based on nonparametric tests has proved to be rather simple and therefore useful. Its use on data for Ljubljana gives the selection of typical months of the year with the level, of significance less than 0.05 at majority of months.

At the end we would like to add few thoughts about the proper use of the test reference year from the meteorological point of view. Data of typical months should not be used for shorter time period as the month which was the elementar time interval used in the work. Weather is namely governed by local and advective processes. Sometimes prevail the first and sometimes the last ones. Weather for shorter time period than a month in our latitudes can therefore not be representative and so the use of such data can not be justified either.

LITERATURE

Arhiv Hidrometeorološkega zavoda SRS.

Band et al, 1981: Appplication of a photosynthesis model to an agricultural region of varied climates: California, Agric. Meteorol. 24, 3. 201-218.

Cehak, K., 1975: Ein "Test-Referenz-Jahr" für Energieberchnungen in der Raumphysik. Wetter und Leben, 27, 3/4, 254-258.

Hočevar, A. in Kajfež-Bogataj, L., 1984: Aplikacija modela neto-fotosinteze na klimatsko različnih območjih Slovenije, Zbornik BF 43, 9-23.

Kletter, B., 1982: Eine österrreichische Version eines Test-Referenz-Jahres für die Anwendung in der Bauphysik. Archives for meteorology, Geophysics, and Bioklimatology. Ser. B, Vol. 30 No. 3, 283-301.

Lund, H., 1974: Meteorological Data for Design of Building and Installation, a Reference Year. Danish report No. 89, Danish Building Research Institute.

Olgyay, V., 1962: Design with Climate. Princeton University Press. Princeton, New Jersey, 190 str.

Siegel, S., 1956: Nonparametric statistics, International Student Edition, Mc Graw-Hill Book Co., New York, 312 str.

Sivakumar, M., 1984: Crop productivity in relation to interception of photosynthetically active radiation. Forest and agricultural meteorology, Vol. 31, No. 2, 131-143.

Thom, H.C.S., 1971: Some methods of climatological analysis, WMO, No.199, TP. 103, Geneva, 53 str.

Quenzel, K.H., 1969: Meteorological Data, Forster Verlag,Zurich.

Wilcoxon, F. (1949): Some rapid approximate statistical procedures Stanford, Conn., American Cyanamid co.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

OBJECTIVE PRECIPITATION FORECAST USING A SINGLE-STATION RADIOSONDE PROFILE

Pandžić,K., Hidrometeorological Institute,
Zagreb (Yugoslavia)

1. INTRODUCTION

Although in prognostic reports precipitation is a primary parameter,it is not often possible to forecast quantitatively rainfall amounts from synoptic maps,particularly in small area.Therefore,the need arises for prognostic models to forecast precipitation over some area.Models based on termodynamic equations are rather complicated and they require a great number of informations for initial conditions.Contrary,statistical models are often simpler and thus more attainable to forecasters.

A very simple statistical procedure was used by Druyan and Sant (1978). They were using single-station radiosonde profile for forecasting 12 h precipitation amount.For this purpose,they were using a graphical method,with four predictors.The first one is related to atmospheric saturation in the layer 1000-600 hPa,the second in the layer 1000-850 hPa,the third one to the height of maximum saturation layer and the fourth one to temperature on 700 hPa surface. Latter Druyan (1982) used the numerical method instead of the graphical,based on linear regression equations with the same predictors.Model was applied to the aerological data at Bet Degan (Israel) and the testing of model gave encouraging results.

In this paper regression equations are made for forecasting 12 h precipitation using aerological data at Zagreb (Yugoslavia).For this purpose,two-year time-series of aerological data is used.Verification of the model is done by the independed data set.Further step is made by replacing the second and the fourth Druyan's predictors by the other two (the 3 h surface pressure tendency, ΔP, and the temperature difference between 500 and 850 hPa surfaces, ΔT).

2. DESCRIPTION OF THE METHOD

We start by the method of Druyan (1982),but introducing some modification in terms of predictors as well as the calculation of regression coefficients.

a) Druyan's predictors

As indicated above,Druyan's first two predictors are related to humidity both in an absolute and a relative sense,in the layer from 1000 to 600 hPa. The parameter which takes into consideration the humidity in this form is

$$I = W/L \qquad (1)$$

where W is the mixing ratio (g/kg) and L is the difference (kPa) between the considered level and its corresponding lifted condesation level (Figure 1).

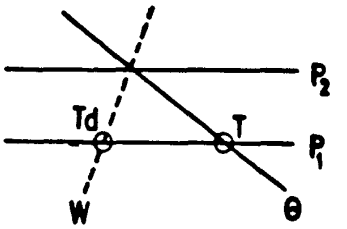


Figure 1. Part of the adiabatic diagram by which it is possible to determine the depth of the layer L= P1-P2 and a mixing ratio W from the equation (1).

The first predictor is taken as the arithimetic mean of a quantity I on nine successive levels separated in 50 hPa layers starting from 1000 hPa surface (or from the ground surface if 1000 hPa does not exists) to 600 hPa surface. The second predictor is the arithmetic mean of this quantity on the first four layers (1000-850 hPa). Since the distribution of aerological data performs only a small numbers of standard isobaric surfaces (1000,850,700,500,...,10 hPa), it is necessary to make an interpolation for eaach 50 hPa if we use data from TEMP dispatches. For these depthes, it can be presumed that temperature and dew point are linear function of ln P, where P is the pressure on some level. Thus

$$T = T1+(\ln P-\ln P1)*(T2-T1)/\ln (P2/P1) \qquad (2)$$

where, indecies 1 and 2 refer to two neighbouring standard levels.

The mixing ratio W can be calculated from the dew point and the pressure at some level. Thus

$$W = (0.622*e/(P-e))*10\uparrow 3 \qquad (3)$$

where the vapor pressure e (hPa) is

$$e = 10\uparrow(9.4-2353/Td) \qquad (4)$$

and Td is the dew point.

Parametar L can be replaced by the vertical distance Z between some level and the corresponding lifted condesation level. This distance can be approximated by presuming that both temperature and dew point lapse rates (γ and γ_d) are constant with height for parcel rising to its lifted condesation level. Thus, we have

$$T = T1- \gamma *Z$$

and

$$Td = Td1- \gamma_d *Z$$

where T1 is the temperature and Td1 is the dew point on the considered level. Since $\gamma-\gamma_d \approx 0.8$ °C/100 m and at condesation level T is equal to Td, then

$$Z = (T1-Td1)/0.8=1.25*(T1-Td1) \qquad (5)$$

where Z is expressed in units of 100 m. Since L is the same order of magnitude as Z we can replace L by Z in equation (1). Then,

$$I = W/Z \qquad (6)$$

where we take Z=1 for $0<Z<1$. From the quantity I Druyan derives the first two predictors:

$$\bar{I}9 = (1/9) * \sum_{i}^{9} (W/Z)i \qquad (7)$$

and

$$\bar{I}4 = (1/4) * \sum_{i}^{4} (W/Z)i \qquad (8)$$

The third predictor Ptr is determined as the pressure midway between the levels showing the lagest upward decrease in I. As the fourth predictor is the air temperature on 700 hPa surface (T700).

b) New predictors

Since Zagreb is placed in the middle latitude area it can be presumed that the precipitation phenomenon is related to the pressure depression and to the atmospheric instability. Therefore, we include two new predictors for forecasting 12 h precipitation: 3 h surface pressure tendency (ΔP) and the temperature difference between 850 and 500 hPa surface (ΔT).

c) Transformation of predictors to principal components

In order to simplify the procedure for obtaining the prognostic equation

of 12 h precipitation it is convenient to transform the predictors to some uncorrelated variables for which the next condition is valid

$$r_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} \qquad (9)$$

where r_{ij} is the correlation coefficient between new variables.Here the transformation into principal components is applied which satisfies the condition (9).In the case of uncorrelated predictors,the regression coefficients in prognostic equation are equal to the correlation coefficient between the predictand (precipitation amount) and predictor variables (principal components).The procedure for variable transformations to the principal components and their application in the regression analysis are described by Kendall (1976) and Fulgosi (1979).

3. DATA

Radiosonde data for aerological station Zagreb is used.Two-year series of continuous radiosonde observations (1975-76) for 0000 GMT and 1200 GMT are considered.Besides the radiosonde data,the coressponding series of 12 h precipitation is used.Data for January and July 1977 are used for testing the method.

4. RESULTS

Here we consider the forms of prognostic equations and the results of method testing.Preciding the presentation these results the correlation coefficients among 12 h precipitations and predictors are given in Table 1.

Table 1. The correlation coefficients between 12 h precipitation (R) and Druyan's (first four predictors) as well as the additionall two for Zagreb.

	Predictors	$\bar{I}9$	$\bar{I}4$	Ptr	T700	ΔP	ΔT
R	Bet Degan	0.53	0.36	-0.17	-0.48		
	Zagreb	0.35	0.31	-0.11	0.04	-0.03	0.04

From this table is obvious that the two new predictors (for Zagreb) have smaller correlations with 12 h precipitations than the first three (Druyan's predictors).Howewer,the signs of these coefficients are as expected taking into consideration the locations for which these calculations were made (southern part of the Pannonian lowland).For comparison,the results for Bet Degan are presented as given by Druyan (1982).The comparison of Druyan's results with those for Zagreb indicate the essential difference among the correlation coefficients for the first and the fourth Druyan's predictors,while for the others the agreement is rather well.Since the first and the second predictors (for Zagreb) have similar relation to precipitation,their contribution is not independet,therefore the second predictor can be replaced by the new one (e. g. ΔP).The fourth Druyan's predictor does not have a logical explanation for Zagreb,and thus it is replaced by the predictor ΔT which is related to the atmospheric vertical stability.Prognostic equations and their testing is described in the following text.

a) Prognostic equations

Using data which are described in Section 3 and theory of principal components,the prognostic equation for 12 h precipitation on the basis Druyan's predictors is expressed by

$$RPF1 = \bar{R} + 0.994\ \bar{I}9 - 0.058\ \bar{I}4 - 0.001\ Ptr + 0.043\ T700 \qquad (10)$$

where $\bar{R}$ is the monthly mean 12 h precipitation (annual course for the precipitation as well as for other predictors are taken into consideration), while the other predictors presented in (10) are the departure from their monthly means.

If we replace the second and the fourth predictors, succesively, by ΔP and ΔT, the following equation is obtained

$$QPF2 = \bar{R} + 0.618\ \bar{I}9 - 0.171\ \Delta P - 0.006\ Ptr + 0.157\ \Delta T \qquad (11)$$

The results of testing equations (10) and (11) are presented in the next section.

b) Testing of the method

The testing of equations (10) and (11) is made using the corresponding radiosonde and precipitation data for January (winter) and July (summer) 1977. Alltogether 124 cases are considered. Since the prognostic function deviate from zero (most in positive direction), it is taken that 12 h periods with QPF<1 have no precipitation and periods with QPF>1 have. The results of the skill forecast for indicated precipitation are presented in Table 2.

Table 2. The results of skill forecast of 12 h precipitations on basis equations: a) (10) and b) (11).

a)

Observation	Forecast: Rain	Forecast: No rain
Rain	46	30
No rain	13	35

b)

Observation	Forecast: Rain	Forecast: No rain
Rain	42	21
No rain	17	44

If the sums of the underlined diagonal elements in that table are divided by the total number of cases we get the percentage of skill forecast precipitation events, which for equation (10) is 65 %, and for the equation (11) is 69 %. Further analysis indicates that the values QPF>2 are followed by the precipitation occurrence in 86 % cases, and the values QPF>3 in 100 % cases. However, such a movment of boundary makes worse the skill forecasts of 12 h periods with no precipitation.

If we consider 12 h precipitation and prognostic functions only for those cases when the precipitation is observed, we can not notice stronger stochastic relation. Thus, with the above-indicated predictors we can get the information about the precipitation occurrence in Zagreb but not about its amount.

5. CONCLUSION

Druyan's predictors for forecasting 12 h precipitation amount behave differently in the regions of Zagreb and Bet Degan. It is especially expressed for the fourth predictor T700. However, the problem of the precipitation forecast for Zagreb could be solved by introducting the new predictors. Thus, in this paper the second and the fourth predictors are replaced by the other two. After that substitution, skill forecast of precipitation occurrence is increased from 65 to 69 %. However, these results are not satisfacting because the forecast of 12 h precipitation amount is not realised but only forecast of the precipitation occurrence. If we wish to have better results, further research on this problem is required.

6. REFERENCES

Druyan, L.M., and Sant, Y., 1978: Objective 12 h precipitation forecasts using a single radiosonde. Bull. Amer. Meteorol. Soc., 59, 1438-1441.

Druyan, L.M., 1982: Automated precipitation forecasts using a single-station radiosonde profile. Bull. Amer. Meteorol. Soc., 63, 499-502.

Fulgosi, A., 1979: Faktorska analiza. Skolska knjiga. Zagreb.

Kendall, M.G., 1976: The advanced theory of statistics, Volume 3, Charles Griffin. London.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

CLUSTER ANALYSIS OF WIND DIRECTION DISTRIBUTIONS

JELKO URBANČIČ
University of Ljubljana
Ljubljana, Yugoslavia

1. INTRODUCTION

In this paper the results of cluster analysis of wind direction characteristics for 85 meteorological stations in SR Slovenia - Yugoslavia are presented. The eight segment wind roses obtained from 30 years long climatological data set is the base for this analysis. The data are based on anemograph data or measurements using wind vane. Using the latter, the wind speed cannot be usable for any reliable statistical analysis, so we consider the wind direction characterisics only. These are represented by the wind roses.

The aggregation of the 85 meteorological stations into the natural groups helps us to fulfill the following two purposes:
- to detect the patterns of the climatic characteristics in the studied region
- to change the wind observation station - network in the certain area by abandoning some of the very similar meteorological stations within the same groups, or change the measurement equipment.

This analysis presents the first exploratory step, which should be followed by other sophisicated matematical and statistical analysis to reach the final goal.

2. MEASURE OF DISSIMILARITY

In our work we pay attention to the frequencies in the wind rose, neglecting the wind with less than 0.5 m/s or Beaufort force less than 1. In general, various wind roses have different number of events. This fact must not affect the dissimilarity measure, otherwise the data must be normalized to avoid the bias.

The dissimilarity measure should fulfill the prescribed clustering criterion and the metric axioms, where the triangular condition is desirable, but not neccessary. In our study the clustering criterion expresses the relationship between the observed elements. Therefore the correlation coefficient is the base of the dissimilarity function. The following dissimilarity function between elements X and Y is used:

$$d (X,Y) = 1 - r (X,Y)$$

The transformation of the correlation coefficient stated above implies the minimal dissimilarity between positive linear dependance of the two

elements, the dissimilarity 1 when the elements are independant, and dissimilarity 2 when dependance is linear and negative. It should be pointed out that the correlation coefficient computed on only eight pairs of data is unreliable from the statistical point of view, what diminishes the quality of the dissimilarity function. But the metric criteria are satisfied for all nonzero wind roses, and the dissimilarity function can be applied in the analysis.

3. CLUSTERING METHODS

We use the following frequently applied hierarchical clustering methods to obtain the natural groups: Maximum, Average, Ward and General. (See e. g. Lance, Williams, 1967, Sneath, Sokal, 1973, Everitt, 1974). This metods do not require a specification of the number of clusters in advance, and they are computationally quick. Initially every element is taken as a subgroup of size one. At each stage in the hierarchy two subgroups are pooled, until finally every element is in the same group. At each step two subgroups which have minimal dissimillarity function are pooled together.

The hierarchical methods assume the hierarchy in the population and search it, although in general this condition is not fulfilled. In such case the use of one of the nonhierarchical clustering methods is recommended.

4. PRESENTATION OF THE RESULTS

The results of clustering using the four hierarchical methods were compared. For the presentation the results obtained by the Ward method were chosen. In Figure 1 the dendrogram (the plot of the aggregation tree) obtained by this method is presented. The length of the branches in that tree is proportional to the dissimilarity between the elements and dissimilarity between subgroups and groups. Succesive association of the branches represents the aggregation of the elements into new subgroups and groups. The longest branches representing the highest dissimilarities can be cut off at a certain level and the groups are obtained. We decide to make cut-off that gives 7 groups. They are denoted by using letters A-G.

Different clustering methods naturally give us different results of aggregation. Comparing them we obtain the information about the stability of the elements. These are stable if they are members of the same groups when using different methods, while unstable elements change the membership within groups. In that way, it can even happen that a subgroup become unstable. Real data often contain some unstable elements. The separation is successful when their number is much smaller than the number of elements in a particular group. We applied all four, in paragraph 3 mentioned hierarchical methods and comparing their results 6 unstable elements were found. In further analysis these are removed from the original groups. In the dendrogram in the Figure 1 they are marked by asterix. The wind roses of the meteorological stations groups A-G are presented in Figure 2. The last line denoted by U contains the plots of wind roses of the unstable elements. The characteristics of the groups can be easily found out. The group A has frequently blowing southwest winds, group B all winds from the west side. Group C has dominant east or east-west directions and group D Northwest. Group E is characteristical because of frequent southeast component and F because of northeast. Group E contains the roses with dominant north-south winds.

The elements within particular group are not geographically connected. The

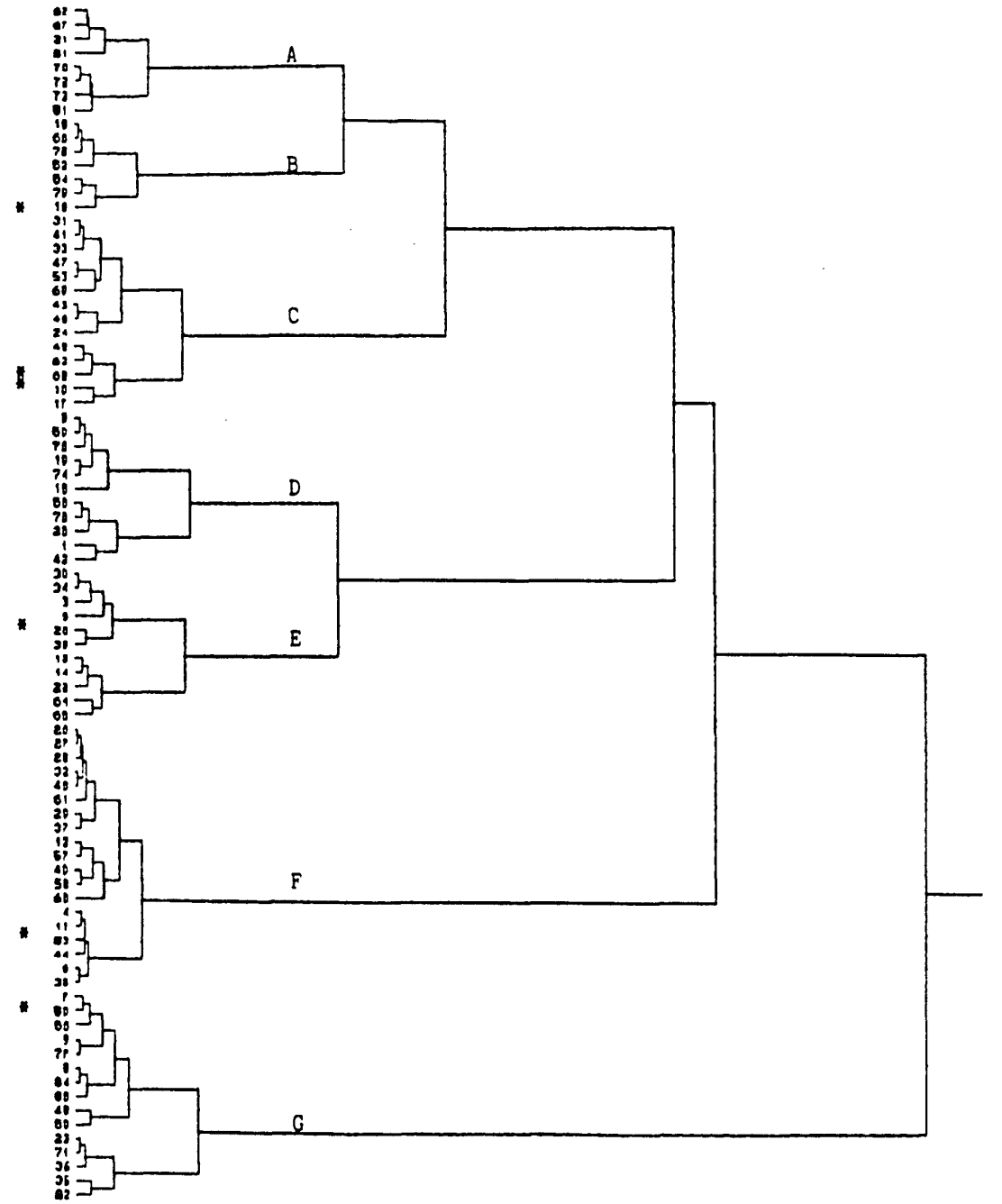


Figure 1: Dendrogram of wind roses from 85 meteorological stations obtained by Ward method.

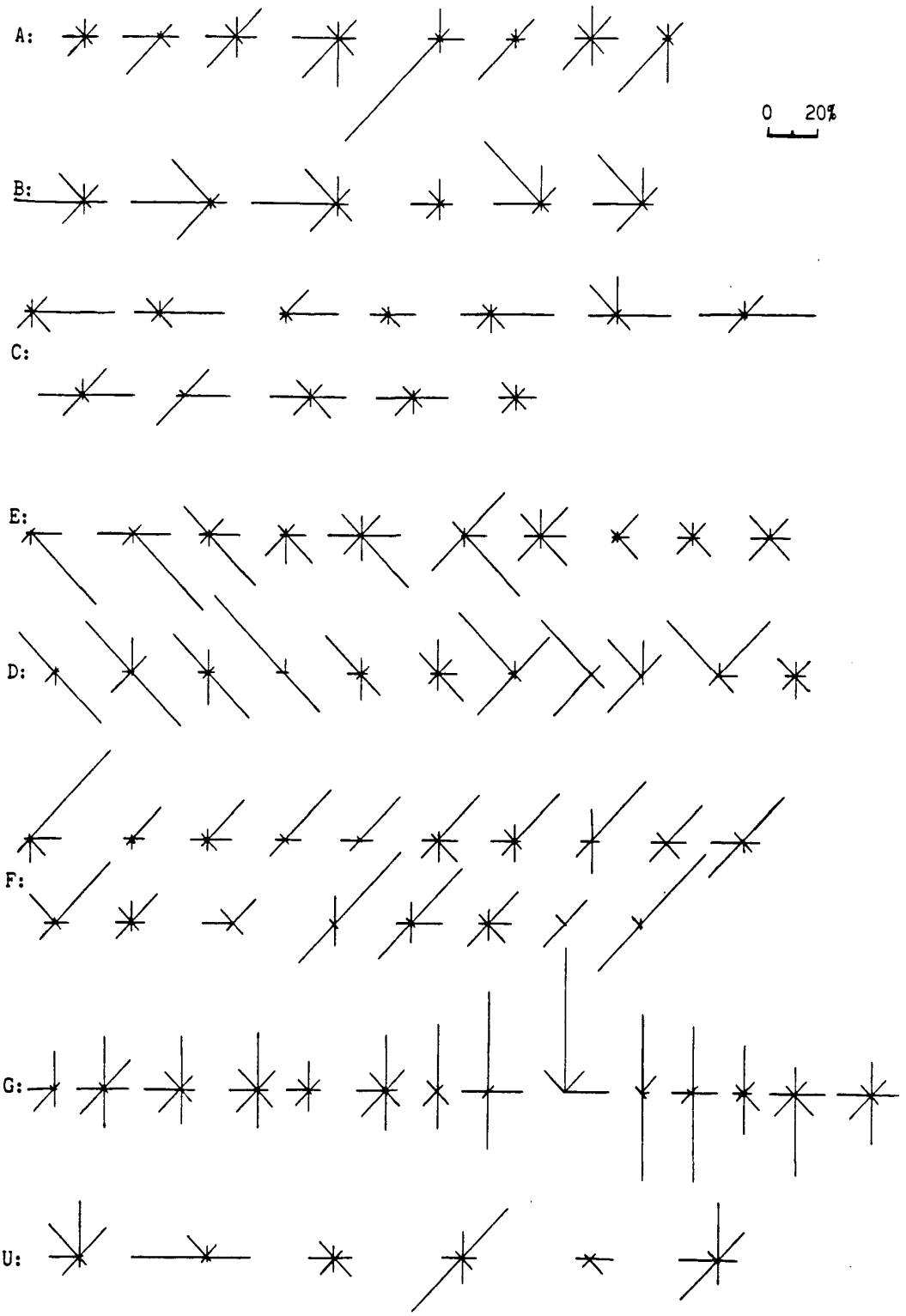


Figure 2: Wind roses for 85 stations in Slovenia - Yugoslavia presented in groups A-G and unstable elements U in the last row.
(Σ n + C = 100%)

wind roses of particular group are more common in certain area e. g. group F is characteristical for Karst region where Bora (from northeast) frequenty blows.

In the dendrogram in Figure 1 the very similar wind roses can be found. When they are similar to each other we can try to confirm the similarity by statisical methods and then make a decision about possible abandoning of the wind measurements in one of the stations. Good examples are the stations 26 and 27 with very similar wind roses. They are located only about 10 kilometers appart.

5. CONCLUSION

The results show that the cluster analysis is a successful method in searching the structure of the population of wind roses in a certain region. Its advantages are simplicity and low price. The two purposes stated in the introduction are successfully fulfilled, gropups of meteorological stations with similar wind roses are detected. The imperfections (correlation coefficient and the assumption of hierarchy) seem in this case not to affect significantly the results. For further analysis nonhierarchical clustering methods are proposed as well, as some other dissimilarity measures for different clustering criteria. At the end we would like to point out that this methodology presents the first step of analysis with the aim to explore the structure of the population of the wind roses.

6. ACKNOWLEDGMENTS

This work was done by means of program package CLUSE written by V. Batagelj from University of Ljubljana. K. Košmelj from the same university helps us giving many useful advices. We thank them for the all the help.

7. REFERENCES

- LANCE, G.N., WILLIAMS W.T., 1967: A General Theory of Classifactory Sorting Strategies. The Computer Journal, Vol. 9, No. 4, 373-380

- EVERITT, B., 1974: Cluster Analysis. Heinemann Education Books LTD., London

- SNEATH, P.H.A., SOKAL, R.R., 1973: Numerical Taxonomy. W.H. Freeman and Company, San Francisco

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

SEASONAL VARIATION OF RAINFALL IN BELGIUM

Sneyers R.* and Goossens Ch.**

*Inst. Royal Met. de Belgique, 3 av. Circulaire, 1180 Bruxelles, Belgique

**Institut d'Astronomie et de Géophysique G. Lemaître,
Université Catholique de Louvain, 2 Chemin du Cyclotron,
1348 Louvain-la-Neuve, Belgique

INTRODUCTION

Owing to the oceanic influence, the mean annual precipitation is moderate throughout Belgium (Fig. 1). It ranges from 1400 mm in the Ardennes to 700 mm along the coast and in Limburg. The continental influence is felt only in the south of Belgium.

As for most of the north-western European countries, in Belgium there are no well-defined wet or dry seasons and rainfall is rather evenly distributed all over the year. Nevertheless, an examination of the differences between the mean of the monthly normals and the annual mean shows that some months tend to be wetter than others, the period January to June being generally drier than the period July to December, the maxima occurring in summer and the minima in spring.

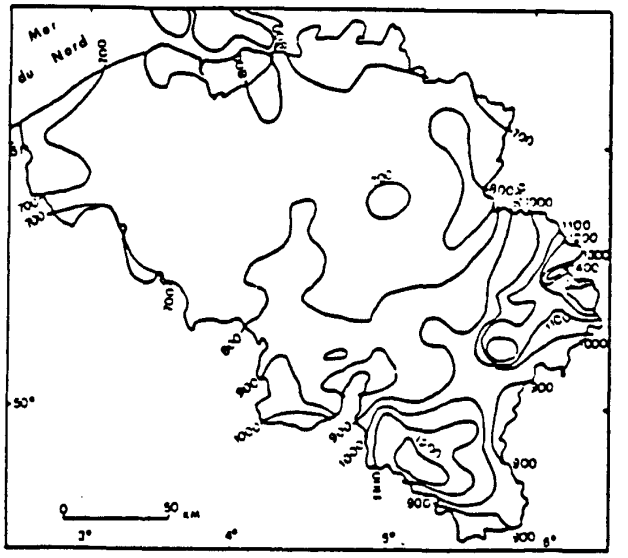


Figure 1. Distribution of the mean annual precipitation amounts as reduced to the 1833-1975 period (extracted from Dupriez and Sneyers, 1979).

In order to study the variations of the annual cycle of precipitation in Belgium, a principal components analysis has been applied to the monthly normals of the rainfall observing network (observations), as reduced to the period 1833-1975 (Dupriez and Sneyers, 1978). From among the 360 observing stations in the Belgian rainfall measurement network, 120 stations (variables) well distributed throughout the country have been selected.

METHODS AND SELECTION OF THE SIGNIFICANT COMPONENTS

The first step in the analysis is the reduction of the dimensionality of the specification of the climate at every station to the minimum possible without significant loss of information. Principal components enable this to be done. The principal components analysis model may be put in the form :

$$x_k = w_{k1}\, y_1 + w_{k2}\, y_2 + \ldots + w_{km}\, y_m$$

where each of the m observed variables is described linearly in the form of m new uncorrelated components y_j. This method has been described extensively in the literature (Sneyers and Goossens, 1986).

In order to determine the principal components having a variation significantly different from a random evolution, a selection rule, called rule N, explained by Overland and Preisendorfer (1982) has been chosen. These authors suggested a Monte-Carlo method technique in order to select the eigenvalues for which the meteorological signal is above the level of noise. The level of noise is simulated by repeated sampling of principal components computed from a spatially and temporally uncorrelated random process. Let λ_j be the eigenvalues extracted from the observed data set and δ_j the set of eigenvalues produced by the rth sampling. The following normalized eigenvalue statistics can be defined :

$$T_j = \lambda_j \Big/ \sum_{j=1}^{n} \lambda_j \qquad j=1, \ldots, r$$

$$U_j = \delta_j \Big/ \sum_{j=1}^{n} \delta_j \qquad r=1, \ldots, 100$$

The decision of rule N is, if $T_j > U_j$, y_j is declared to be significant.

The application of the rule N (Table 1) (Overland and Preisendorfer, 1982) shows that only the first component (Fig. 2) is significantly different from noise.

Actually, the graphical representation of the variation of the second and of the third principal component (Figs 3 and 4) shows that, apparently, also this variation is not random. The application of rule N thus seems to be too severe.

Therefore, in order to select those principal components which are significantly different from a random variation (noise), the 240 remaining stations of the Belgian rainfall measurement network have been subdivided into two series of 120 stations, every time well distributed throughout the country, and each series has been subjected to a principal components analysis. The eigenvalues of the new series of stations obtained in this way are given in table 1. From this we see that the values are coherent up to the fourth eigenvalue.

Table 1. Normalized eigenvalues of the correlation matrix m = 120, n = 12. Application of the rule N for the selection of significant components ($T_j/U_j^{95} > 1.00$).

		j					
		1	2	3	4	5	6
Series 1	T_j (x100)	80	12	6	2	0.14	0.06
Series 2	T_j (x100)	79	13	6	2	0.15	0.08
Series 3	T_j (x100)	80	12	6	2	0.29	0.26
	T_j/U_j^{95}	4.53	0.76	0.41	0.13	0.01	0.005

The computation of the values of the principal components made it possible next to determine the correlation coefficient between same order components taken two by two, given that a high correlation warrants the significant character of the considered component.

The obtained correlation coefficients (table 2) show that the coherence of the principal components is ensured up to the fourth.

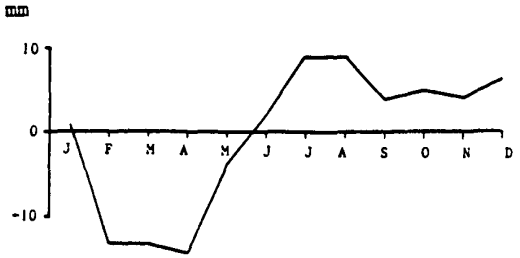


Figure 2. Seasonal variation of the first principal component.

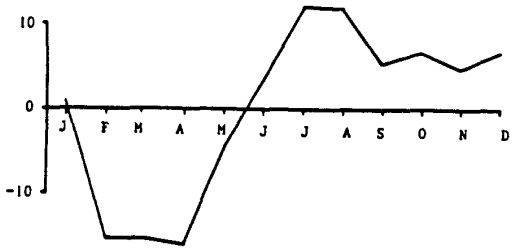


Figure 3. Seasonal variation of the second principal component.



Figure 4. Seasonal variation of the third principal component.

Table 2. Correlation coefficients between the principal components of the same order of the series of stations 1, 2 and 3 taken two by two.

Component N°	Series 1-2	Series 1-3	Series 2-3
1	1.00	1.00	1.00
2	0.99	1.00	0.98
3	0.99	1.00	0.98
4	1.00	0.99	0.99
5	-0.99	-0.01	0.09
6	0.79	0.14	0.20
7	0.27	-0.12	-0.18

PHYSICAL EXPLANATIONS

The first component, which explains 80% of the total variance, is positively correlated with all the stations (Fig. 5). Its seasonal variation (Fig. 2) has the same behaviour as the differences between the mean of the monthly normals and the annual mean. The maximum values occur in summer and the minimum values in spring. A secondary maximum is observed in December. The first component corresponds thus to a uniform field over the whole region, and represents the annual cycle of precipitation. It may be linked to the general atmospheric circulation.

The second component explains 12% of the total variance. It is negatively correlated with the coastal zone and the south of Belgium, and positively with the central part (Fig. 6). Its seasonal variation (Fig. 3) shows maximum values in summer. This component seems thus to be linked to the thundery showers during this season.

The third component explains 6% of the total variance (Fig. 7). It divides Belgium into two zones, the west region which is positively correlated with it and the east region, negatively. Its seasonal variation (Fig. 4) shows a maximum in autumn. It characterizes the influence of the proximity of the sea.

The fourth component explains 2% of the total variance. However, since the geographical distribution of the correlation of the stations with the fourth component does not show any regional organization, we can assign its physical significance on a local scale only.

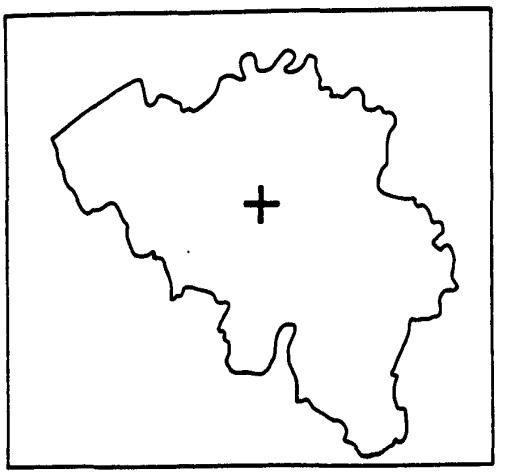


Figure 5. First principal component associated with the monthly normals of precipitation in Belgium explaining 80% of the total variance. Correlation with the stations of the network.

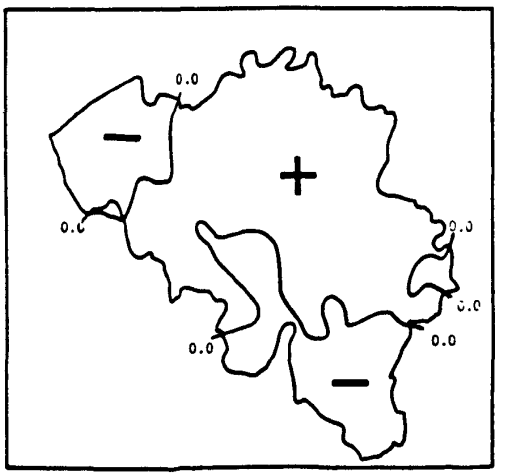


Figure 6. Second principal component associated with the monthly normals of precipitation in Belgium explaining 12% of the total variance. Correlation with the stations of the network.

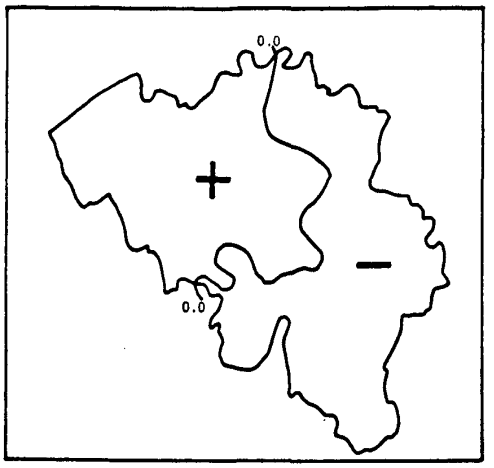


Figure 7. Third principal component associated with the monthly normals of precipitation in Belgium explaining 6% of the total variance. Correlation with the stations of the network.

CONCLUSIONS

Principal components analysis has been applied to the monthly normals of rainfall observing network, as reduced to the period 1833-1975. These variables are correlated with one another and this method gives the possibility to describe them in term of a smaller number of uncorrelated variables, four in this case. This reorganization of the data make them more easily to be understood.

The first component represents the annual cycle of precipitation and may be linked to the general atmospheric circulation. The second shows maximum values in summer and seems thus to be linked to the thundery showers during summer. The third shows a maximum in autumn and characterizes the proximity of the sea. The fourth shows only local variations.

REFERENCES

Dupriez G.L. et Sneyers R., 1978. Les normales du réseau pluviométrique belge. Institut Royal Météorologique. Publ. A, 101.

Lebart L., Morineau A., Tabard M., 1977. Techniques de la description statistique, Dunod, Paris.

Overland J.E. and Preisendorfer R.W., 1982. A significance test for principal components applied to a cyclone climatology. Mon. Wea. Rev. 110, 1, pp. 1-4.

Sneyers R. and Goossens Ch., 1986. Principal components analysis. Application to climatology and meteorology. To be published at the WMO, Geneva.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

IMPLEMENTATION OF AN INFORMATION SYSTEM OF DATA SETS OF THE CLIMATOLOGICAL FIELD

Barbara Benesch
Deutscher Wetterdienst

INTRODUCTION

All over the world there are many different institutions disposing of a great number of data which are important for climatological research, analysis and monitoring of the climate. To be able to make a better use of those data it is necessary to collect relevant information and make them available to all those who are interested in them. Thus in many cases not only a doubling of work could be avoided which is time-consuming and entails financial expenses, but also the appropriate data could be used for an investigation.

In order to propagate the existence of all those data a World Climate Information Referral Service - called INFOCLIMA - is being implemented within the framework of the WMO World Climate Data Programme. In an INFOCLIMA catalogue the most important characteristics for the description of data sets (WMO 1985) and stations have been listed.

Via the Technical Commissions the WMO Members are invited (draft resolution 4.1/2 EC-XXXVIII) to prepare national handbooks on climate data sources. Such national information systems are existing for example in Canada (Canadian Climate Center 1982) and the USA (National Climatic Data Center 1983). The data set description covers the following items:

- period
- geographical area
- elements and statistics
- storage medium
- explanations

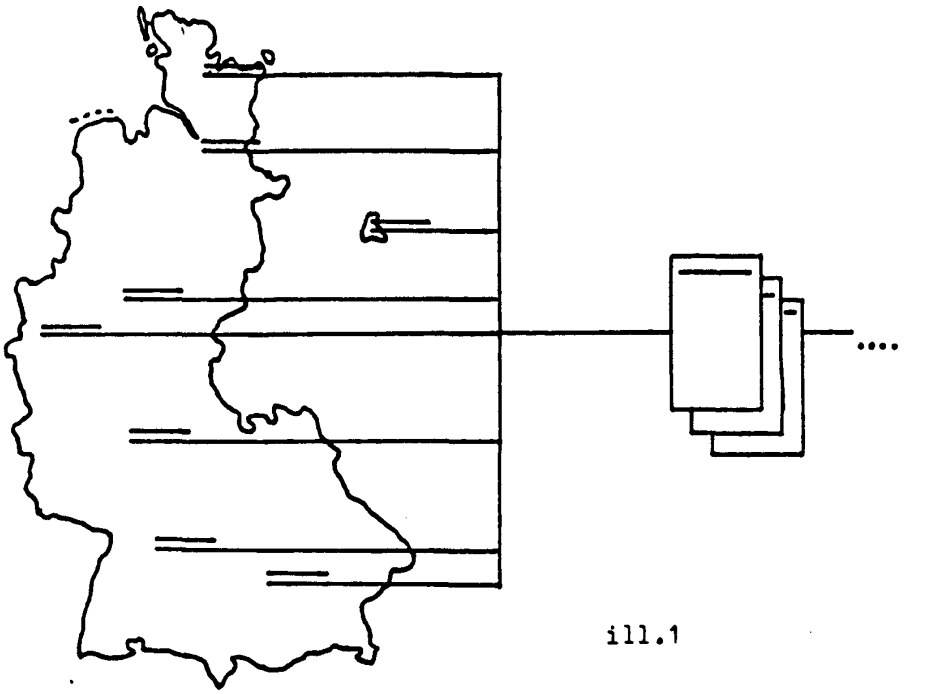
Numerous examples are provided in form of tables and graphs.

The following paragraphs contain a survey of the development and a description of the national information system of the Federal Republic of Germany (Deutscher Wetterdienst 1985).

INTRODUCTION OF INFOKLIF

Within the National Climate Research Programme of the Federal Republic of Germany a project called INFOKLIF Information System for National Climate Research has been initiated by the Deutscher Wetterdienst in November 1984 for a timeperiod

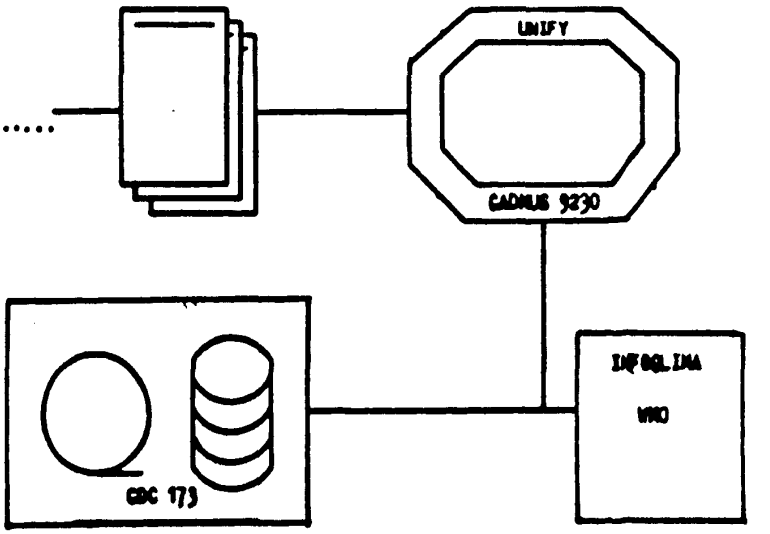
of two and a half years. INFOKLIF is sponsored by the Federal Ministry for Research and Techniques and aims at implementing an EDP-based information system on a national level in order to obtain detailed information about data sets of the climatological field. A questionnaire had been addressed to as many institutions as possible to obtain the necessary information (ill. 1) which have subsequently been processed.



ill.1

HARDWARE, SOFTWARE

The information system INFOKLIF containing the description of data sets is implemented with a microcomputer (ref. ill. 2).



ill.2

It is a CADMUS 9230 of Messrs PCS. The CADMUS is a multi-user- system with a working storage of 1 MB as well as a Winchester drive with 65 MB and a steamer tape of 16 MB. The operating system MUNIX is provided by the CADMUS 9230, apart from this the relational data base system UNIFY is installed. Inquiries for the data base have to be translated into the structured query language SQL.

The microcomputer is connected with the computer system CYBER CDC 173 of the Meteorological Computer Centre of the Deutscher Wetterdienst in order to use EDP-stored information on data sets of the Deutscher Wetterdienst.

There is a co-ordination and exchange of information between INFOKLIF and INFOCLIMA implemented by WMO. The project leader of INFOKLIF is at the same time Rapporteur on INFOCLIMA of the CCL Working Group on Climate Data Management so that use is made of experience already gained and relations on an international level are existing.

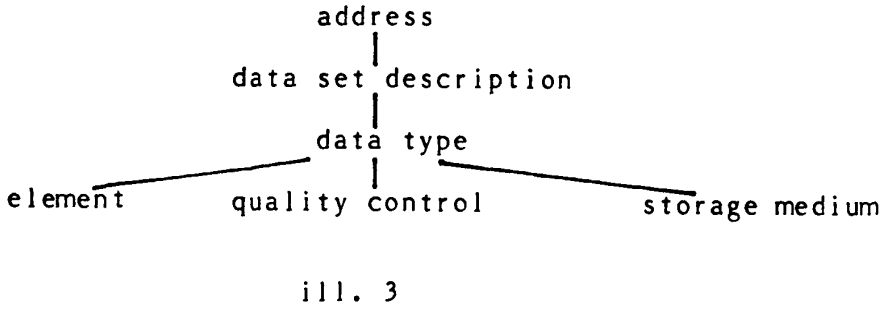
PRESENT STATUS OF IMPLEMENTATION

At the beginning of the project INFOKLIF an inquiry had been made among approximately 100 institutions in the Federal Republic of Germany in order to obtain information on the available data, and the requests and ideas of users of such an information system. On this basis a data base structure has been developed for the description of data sets. Data have been compiled in data sets under the aspect that the elements which are measured, observed or computed and their further processing should be as homogenous as possible. Illustration 3 shows the data base structure. The data base is classified by many-one relationships. Fields on a lower level may appear more than once.

E.g.: many data sets can be stored by one institution

o r

the data of one set can be stored on different storage media.



ill. 3

Some relevant explanations:

A global description is made of a data set stored in and administered by an institution.

The following details are provided:

- short description of the data set
- geographical area and/or geographical co-ordinates where the data are originating from
- period of which data are available; it is also marked when data are available of different stations which have started and/or ended the measurements at different times
- information about completeness of data; as these information refer to the entire data set, it is not possible to describe gaps at individual stations
- information on accessibility of data, whether for example costs are arising by their provision
- information on type of data; if stations are meant, accuracy of measurement and number of stations are indicated; in case of results of model computations the spatial distribution is given or in case of satellite data information provided on the satellite, as for example wave-length used
- the elements of the data type are listed, supplemented by information on further processing; like information on statistical evaluations or type of data acquisition as registration or computation

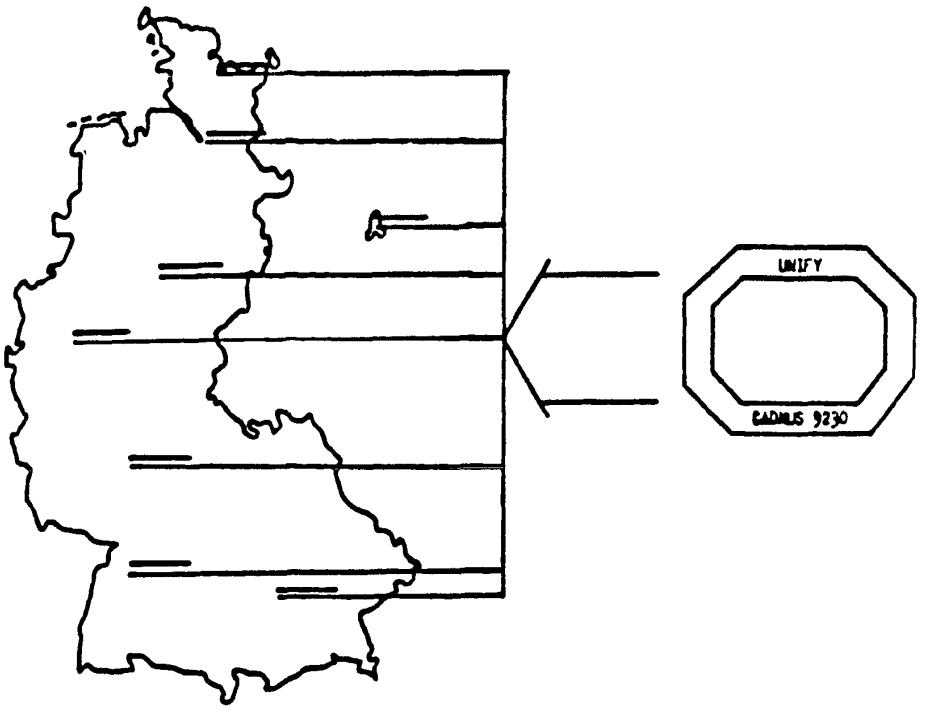
- information about quality control refer to the entire data set; it is not possible to give information for each element
- information on the storage medium; extent and number of magnetic type and annual accessions.

If necessary the period for the quality controls made and the storage medium can be indicated so that changes occurred during the years may be considered.

The information data base still is in a phase of development and testing. Approximately 100 data set descriptions have been stored. Within nine months the extent will have increased even if we have to assume that not all data sets will be described which are stored at institutions of the Federal Republic of Germany.

INQUIRIES ADDRESSED TO THE DATA BASE

As already mentioned information on data sets stored in the data base INFOKLIF shall be accessible to all users and parties interested, ref. ill. 4.



ill.4

In the following methods will be introduced how to acceed to the contents of the information data base. As already mentioned all requests have to be transferred into the language SQL. With SQL any question may be answered.

This method is suited for those only who are acquainted with this language and who know the data base structure with its relationships.

On the other side we assume that frequently certain standard questions will be asked. So, e.g. questions on period, geographical area and element. Therefore a selection of possible inquiries is offered by screen menu. Thus certainly a great part of the inquiries will already be answered. The described method is possible on-line only.

Parallel a printed catalogue is being prepared in which the data set descriptions will be compiled. This catalogue will be similar to already existing information catalogues, where the search for the requested information is made easier by detailed lists of contents and divisions according to certain criteria. It is planned to issue the first edition in one year approximately.

PROJECTED EXTENSIONS

The information stored in the data base do not suffice to meet all requests or answer all inquiries. So, e.g. no information can be provided on individual stations or measuring equip-

ment as requested by the users. Up to now INFOKLIF is focused on the description of data sets. In a second phase INFOKLIF shall be extended to station and space related information.

SUMMARY

In the Federal Republic of Germany information on climate-relevant data on a national level are being compiled. The aim is to draw attention to the existing basic data which are valuable for climate research.

In Austria there is a publication of working results of the meteorological-climatological research. (Vienna 1984) Prof. Dr. Cehak writes in the foreword of this bibliography

Quotation: "As soon as the existence of the bibliography was known, inquiries on keywords started, which could rapidly be answered. They proved that in scientic circles there is indeed a requirement for such a publication."

Further on: "When we now (1984) are in a position to hand over the bibliography to the public after many years of work we not only hope that it will be of use to investigators in and users of meteorology and climatology but that others may be inspired to prepare a similar publication for other important parts of meteorology."

INFOKLIF may be considered as one of these parts. Stimulated by this statement I would like to ask other countries to implement corresponding information systems, and at the same time I would like to support emphatically the request of WMO.

LITERATURE

- WMO, June 1985: INFOCLIMA-Catalogue of climate system data sets (Interim Edition).
- Canadian Climatic Center, Downsview, Ontario December 1982: Handbook on climate data sources of the Atmospheric Environment Service.
- National Climatic Data Center, Asheville, N.C., July 1983: Selective guide to climatic data sources.
- Deutscher Wetterdienst Offenbach 1985: Aufbau eines Informationssystems zur Klimaforschung. Jahresbericht 1985, Deutsche Wetterdienst-interne Veröffentlichung.
- Zentralanstalt für Meteorologie und Geodynamik in Wien 1984: Publikation Nr. 285.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

AN EVALUATION OF WIND DIRECTION PERSISTENCE AND PERIODICITY

Stanton E. Tuller
Department of Geography
University of Victoria
Victoria, B.C., Canada

It is a well-known fact that wind directions measured at a particular point are not independent and random. Some directions occur much more frequently than do others. The measured wind direction is a result of the interaction of the pressure gradient with the local relief and land surface type. Pressure gradients often show a tendency to return to certain preferred patterns. Diurnal and seasonal cycles are not uncommon. Local land use and topography are also relatively constant.

One of the important features of wind direction for both applied climatology and forecasting is its degree of persistence over time. Another is its periodicity or tendency to return to a similar direction at regular time intervals. The purpose of this study is to present some of the preliminary results of a larger project that demonstrate how simple quantitative measures can be used to illustrate the patterns of wind direction persistence and/or periodicity that can occur over time.

STATIONS AND METHODS

Data from three stations that represent different topographic situations common along the coast of British Columbia, Canada, were selected for presentation in this study (Figure 1, Table 1). Cape St. James represents a small island with no large land mass in the immediate vicinity. Victoria International Airport is on flat land, slightly inland from the coast and under the immediate influence of a large land mass. Spring Island represents a transition between these two. It is on a small island immediately offshore from a large land mass. In simple terms the influence of land on the station is least at Cape St. James and greatest at Victoria International Airport.

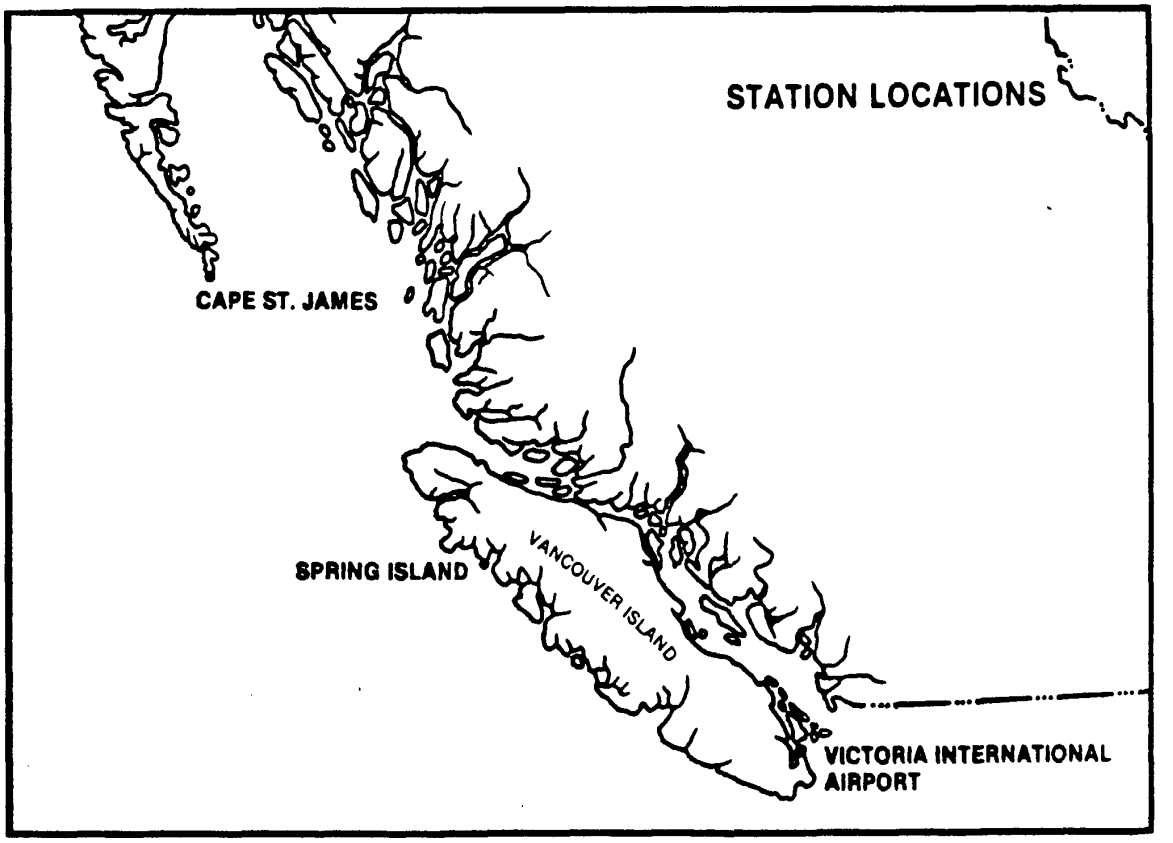


Figure 1: Stations

Between 19 and 23 years of hourly data were used (Table 1). Wind direction is recorded to 16 compass points (22.5°) plus calms and missing observations.

Table 1: Stations and Data Characteristics

Station	Years of Record	Total Records	Missing Observations		
			Number	Percent	Calms (percent)
Cape St. James	1957-1975	166,536	135	0.08	0.7
Spring Island	1953-1975	201,600	97	0.05	8.7
Victoria International Airport	1953-1975	201,600	35	0.02	5.5

Wind direction change or difference over time lags or separations of 1 to 192 hours (8 days), 29 to 31 days and 59 to 61 days were determined for each station. The differences can range from 0 to 180°. Both the mean and the relative cumulative frequency distribution of wind direction differences were computed. A scaled mean was then determined via:

$$\bar{x}\ \text{scaled} = (90^0 - \bar{x})/90^0$$

where is the mean wind direction difference over a given time lag. The scaled mean ranges from 1 for no wind direction change to -1 for a complete wind reversal (180° change). A change of 90° gives a scaled mean of 0. Thus, the closer the value to +1 the more consistent are the wind directions.

Calms and missing observations were included in the analysis. It was felt that calm can be considered a wind direction and calms are an important part of the record. Their elimination would remove useful information even though calms are not a sizable proportion of the observations (Table 1). The following rules were applied. If both observations were calm there is no change in wind direction and the difference is 0°. If one observation was missing the most frequently occurring (modal) direction was used in its place. This is the direction most likely to occur if each observation was independent of the others. The mid-point of the range of possible differences, 90°, was used when one observation was calm or both were missing. It was thought that the mid-point value would be the least prejudicial and introduce the least bias towards either association or independence of wind directions over time.

The scaled mean and the relative frequency of wind direction differences $\leq 22.5^0$ were graphed against time lag. A large scaled mean or high relative frequency of small direction differences indicates consistent wind direction for a given lag. The relative frequency of differences $\leq 22.5^0$ was chosen because this is the limit of resolution in the recorded wind directions and a change of this magnitude can usually be considered to be of minor importance for practical purposes.

The value of the scaled mean or relative frequency of wind direction differences $\leq 22.5^0$ gives a useful indication of the pattern of wind direction changes over time. Further information is gained by testing the independence of wind directions separated by different time lags. The general form of the null hypothesis is that the scaled mean (or distribution) of the population of wind direction differences, of which the values we observe comprise a sample, is equal to the scaled mean (or distribution) expected under the assumption that the wind direction observed at one time is independent of that observed at another time. If we must reject the null hypothesis a likely explanation for the lack of independence is either persistence or periodicity of wind directions over time.

The "independent" wind direction difference distribution and scaled mean were computed from the relative frequency distribution of observed wind directions. If f_k and f_l are the relative frequencies of wind directions k and l the expected relative frequency of occurrence of wind direction k at one time and l at some later time is $f_k.f_l$, if each observation is independent of the others. Summing the values for wind direction differences of 0° and 22.5° gives the relative frequency of differences $\leq 22.5^0$ expected under the null hypothesis. The scaled mean was determined from the relative frequencies of all possible differences 0° to 180°. The use of the actual long term recorded

frequency distribution is more conservative and gives less likelihood of rejecting the null hypothesis than does the assumption that each direction would occur with equal frequency.

The relative frequency of differences $\leq 22.5^{0}$ was used in a binomial distribution test. If the wind directions are persistent or periodic rather than independent we would expect a higher proportion of small wind direction changes than found in the independent distribution. A one-tailed test was used. The alternative hypothesis is that the proportion of small differences in observed wind directions is greater than that expected if the wind direction observed at one time was independent of that observed at some earlier time. The .005 critical values were computed and graphed using the normal approximation to the binomial distribution. The small significance level was employed to limit the chance of type 1 or alpha error. The number of trials, n-j, where n is the number of hourly observations and j is the time lag in hours exceeds 160,000. Therefore, the normal distribution should be a reasonable approximation of the binomial distribution.

Likewise, although the wind directions themselves are not normally distributed the direction differences should be by virtue of the central limit theorem with the large sample size available in this study. .005 critical values in a one-tailed test were used to test the equality of means. The alternative hypothesis is that the scaled mean of the observations is larger (indicating more consistency of wind direction) than that expected if each observation was independent of that recorded at some earlier time.

RESULTS

The graphs of the relative frequency of small wind direction changes and the scaled mean wind direction difference over time lags of from one hour to two months show three distinct patterns (Figures 2 and 3). The decline over the first 12 hours is very rapid at Victoria International Airport and the curves are marked by pronounced diurnal cycles. Cape St. James on the other hand displays slower rates of decline and virtually no diurnal cycle. Spring Island is intermediate displaying moderate diurnal cycles with the amplitude being about one-third that of Victoria International Airport.

The null hypothesis of no difference between the scaled mean or distribution of observed wind direction differences and that expected if each observation was independent of that observed at another time can be rejected if the scaled mean or relative frequency of wind direction changes $\leq 22.5^{0}$ lie above the critical value. The observed scaled mean at Victoria International Airport falls below the .005 critical value at a lag of 13 hours and again in the 32-40 hour period. The relative frequency $\leq 22.5^{0}$ first becomes less than the .005 critical value during the 33-40 hours lag period. We cannot reject the null hypothesis and thus have no reason to believe the observed wind directions are not independent on the basis of the information available to us at time lags centered on multiples of 24 hours but offset 12 hours. We can reject the null hypothesis at the .005 level for lags centered on multiples of 24 hours. Wind directions separated by time lags of one day tend to display some consistency or repetition of direction. A wind direction pattern of this type might be termed "periodic."

The small island station, Cape St. James, has observed values of the scaled mean wind direction difference and relative frequency of small differences that are consistently above the .005 critical values through lags of about 60 days. We can reject the null hypothesis and claim the alternative that there is a greater scaled mean and relative frequency of small wind direction changes than expected if each observation was independent of that recorded at some earlier time for time lags out to about 60 days. Beyond this lag we cannot reject the null hypothesis. There is no cycling above and below the critical value as there is at Victoria International Airport. The wind direction pattern at Cape St. James might be designated "persistent."

The binomial distribution and equality of scaled means tests differ somewhat in their results at Spring Island. The relative frequency of wind direction differences drops below the .005 critical value at 153 hours (6.5 days). The observed scaled mean difference remains above the critical value until 1425 hours (59.5 days). Spring Island shows properties of both persistence (the long delay before observed values fall below the critical value) and periodicity (the cycling of the curve above and below the critical value).

A likely explanation for the differences in pattern among stations is the relative importance of the diurnal sea-land breeze cycle. We would not expect Cape St. James to be influenced by the sea-land breeze cycle because of the lack of a large nearby land mass to generate local diurnal pressure variations. Instead its wind pattern is dominated by larger, more consistent, synoptic scale pressure distributions. Victoria International Airport is in a topographic situation where the sea-land breeze cycle is a part of the local wind regime. Spring Island is near, but not on, a major land mass and receives a subdued sea breeze effect.

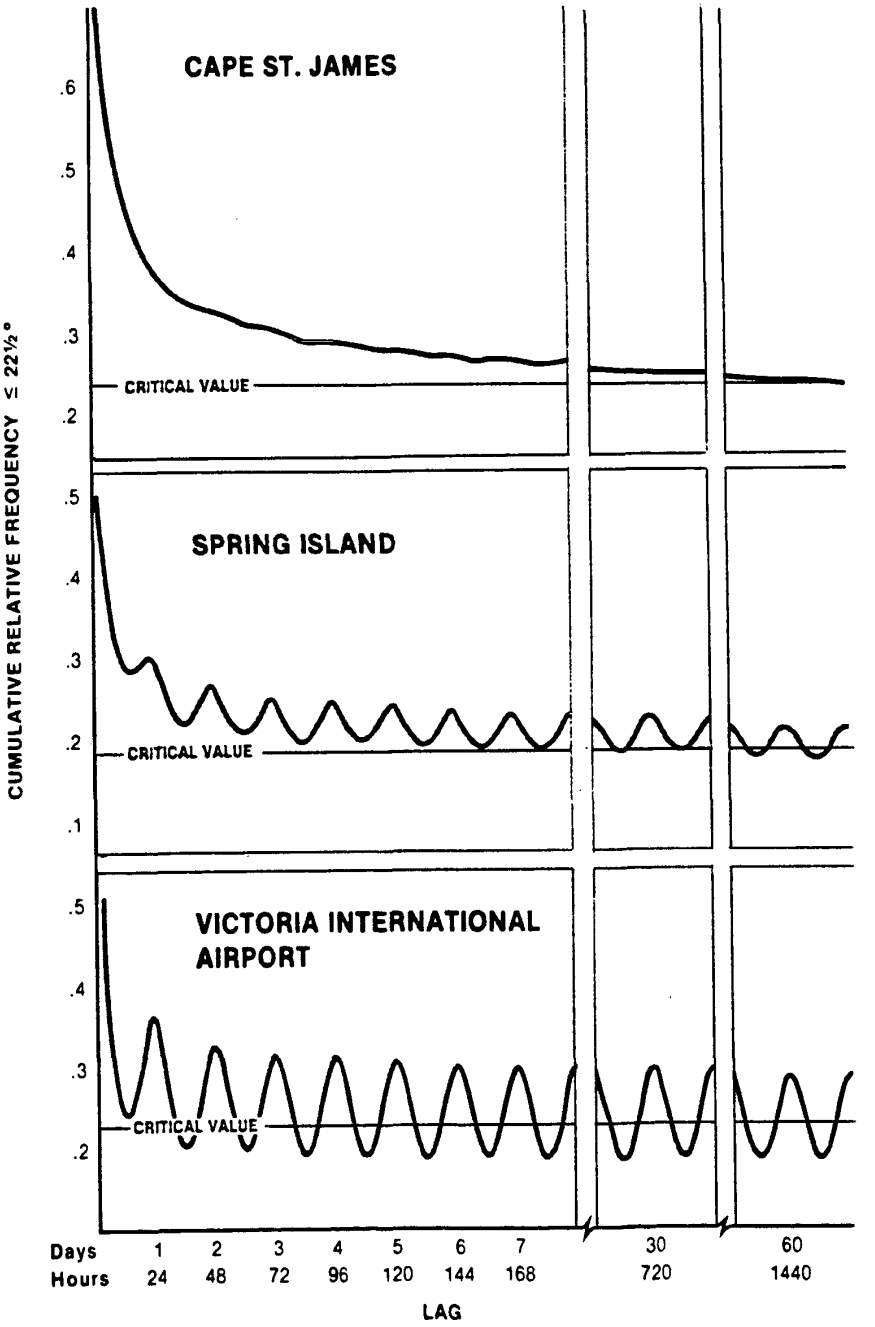


Figure 2: Relative cumulative frequency of wind direction differences $\leq 22.5°$ as a function of time lag.

CONCLUSIONS

One of the benefits of quantitative and statistical techniques in climatology is that they can clarify patterns that may exist but not be readily apparent in a mass of data consisting of numerous, individual observations. The two simple techniques discussed here help illustrate three different patterns of wind direction persistence and/or periodicity that exist at stations on the west coast of Canada. The parametric equality of means test and the non-parametric binomial distributions test give similar results. The conclusion that wind directions observed over short to moderate time separations frequently are not independent but show some consistency based on the persistence of topographic influences and the persistence or periodicity of the pressure gradient forces is not new. This study simply shows how a couple of common and simple, quantitative techniques can be used to help demonstrate and confirm this fact.

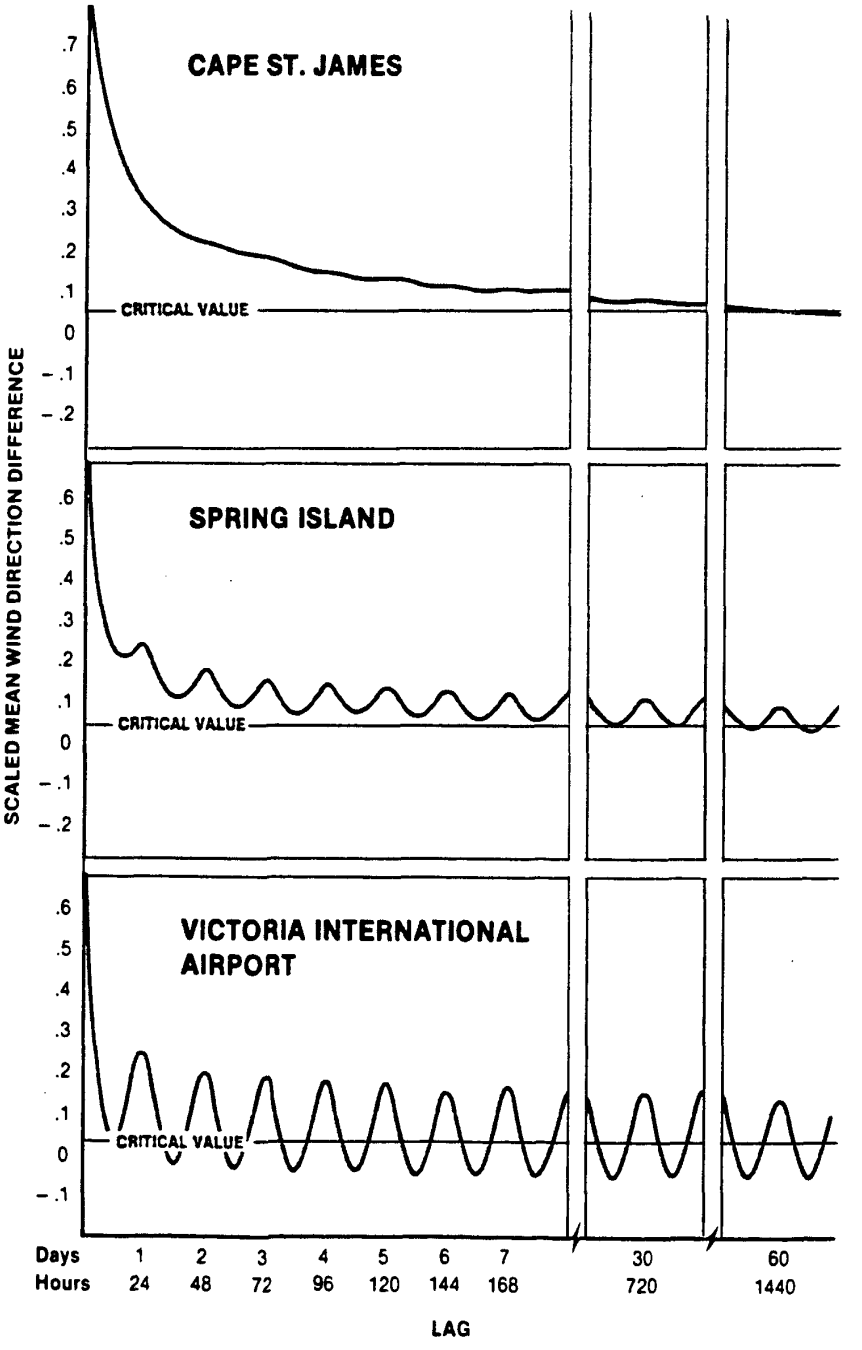


Figure 3: Scaled mean wind direction difference as a function of time lag.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

USING DIRECTIONAL DATA METHOD FOR ANALYSIS ON TYPHOON MOVEMENT AND ITS PROBABILITY DISTRIBUTION FEATURES

Jiuen Shi
Beijing Institute of Meteorology, Beijing

Xiaoming Chen Jingtian Xiang
Institute of Applied Mathematics, Academia Sinica, Beijing

1. INTRODUCTION

In the domain of natural sciences, such as meteorology, biology, geography, astronomy, etc, many of the problem of great interest are those regarding the magnitudes of angular values of which the values expressing the wind directions or the directions in which the animal disappearing migrants move away are the examples. The data of directions as mentioned above are called the directional data. The application of the method of mathematical statistics on the problem of directional data brings on the statistics of directional data. The analysis of the directional data statistics which is so characteristic of itself must be different from that of ordinary data statistics and therefore a methods of processing the directional data have to be employed. In recent years, the study of directional data statistics has attracted more and more attention in the world and this technique has met with success in meteorology, metrology, biology, geography, medical sciences, astronomy and some other fields of study that are essential to national economy. In our country the present study of directional data statistics is only at the outset and so the application of this technique is still a blank yet to be filled. In this paper, a description of using the method of directional data statistics is presented as to the statistical analysis of the moving directions of typhoons after their formation over China Sea or West Pacific, together with a representation of statistical distribution of the moving directions of the typhoons in several coastal areas. Furthermore, an estimate is made as to the greatest probability of the moving direction of the typhoon. It has been proved that a nature of being variant exists among the statistical

distributions occurring in different areas. A series of results of investigating the moving directions of typhoons has thus been obtained by analysing quantitatively the conditions.

2. Data Processing

Informative records about the moving directions of typhoons that occurred in four areas in the region of northwestern Pacific during the months from July to September in the years 1949-1980 were selected for the statistical analysis.

Table 1 The region numbers, latitude and longitude ranges associated with the moving directions of typhoons

Region no.	Latitudes	Longitudes
I	15°-20°N	115°-120°E
II	15°-20°N	125°-130°E
III	20°-25°N	125°-130°E
IV	20°-25°N	130°-135°E

It was stipulated that a typhoon moving due west was regarded as being moving in zero direction. For a typhoon track in a certain region, the angle θ which the line of the track makes with the X-axis in a coordinate system or the value of arc tangent Y/X is taken to be the direction of the typhoon (Fig.1). For example (measuring unit: radian. Latitude)

When X=3.1 Y=1.8, $\theta = t_g^{-1}\frac{Y}{X} = 30.1^{\circ}$

X=-20 Y=2.5, $\theta = 90^{\circ} + t_g^{-1}\frac{X}{Y} = 128.6^{\circ}$

X=2.0 Y=-0.5, $\theta = t_g^{-1}\frac{Y}{X} = 14.0^{\circ}$

The directional data of typhoons in four regions have been sorted out in accordance with above-mentioned principles. The numbers of sample size are 87, 125, 81 and 80 respectively.

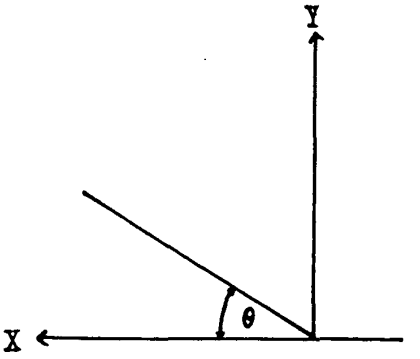


Fig.1 Diagram showing the determination of the moving direction of a typhoon

3. Fitting of a von Mises distribution and goodness of fit

In order to make a statistical analysis of the moving direction of a typhoon, we will first introduce an important type of distribution——the von Mises distribution. Its position in the distribution curve family of directional data is equivalent to the normal distribution in a linear distribution family [2] [4].

The density function of the type von Mises distribution is

$$g(\theta;\mu_0,k,1) = \frac{1}{2\pi I_0(k)} \exp\{k\cdot\cos 1(\theta-\mu_0)\} \qquad (1)$$

$$0<\mu_0<\frac{2}{1}\pi\ ,\ k>0,\ 1\geqslant 1$$

where

$$I_0(k) = \frac{1}{\pi}\int_0^{\pi} e^{k\cos\theta}\, d\theta$$

which is the modified Bessel function of the first kind and order zero. μ_0 is the position parameter, K is the scale parameter, and 1 is the coefficient of amplification.

The von Mises distribution is unimodal (in the case of multimodal, it can be treated as an unimodal seperately), and is symmetrical about $\theta = \mu_0$. The mode is at $\theta = \mu_0$ and the antimode is at $\theta = \mu_0+\pi$. The ratio of the density at the mode to the density at the antimode is given by e^{2k} so that the larger the value of k, the greater is the clustering around the mode.

The parameters μ_0, k, 1 can be estimated by the sample $\theta_1, \theta_2, \cdots, \theta_n$ from von Mises distribution as $M(\mu_0,k)$ [3] [4].

(1) $\hat{\mu}_0 = X_0$

where X_0 is called the mean direction of the sample,
It is also the solution of the equation

$$\begin{cases} C = R\cos X_0 \\ S = R\sin X_0 \end{cases} \tag{3}$$

where $C = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i \qquad S = \frac{1}{n}\sum_{i=1}^{n}\sin\theta_i$ (4)

$R = (C^2 + S^2)^{\frac{1}{2}}$

R is the length of resultant vector.

(2) $\hat{K} = A^{-1}(R)$ (5)

where $A(\cdot) = J_1(\cdot)/I_0(\cdot)$, $I(K) = \frac{1}{\pi}\int_0^{\pi}\cos\theta\, e^{k\cos\theta}d\theta$

The relation between $\hat{K}$ and R is rather complicated. If R is known, however the value of K can be obtained by consulting the statistical table. Both $\hat{K}$ and $\hat{\mu}$ are maximum likelihood estimates.

(3) $l\left(\frac{\max\theta_i - \min\theta_i}{360}\right)^{-1}$ (6)

when $l=1$, the expression (1) is exactly the von Mises distribution. In case $l \neq 1$, the transformation $\theta' = l\theta$ is made and maximum likelihood estimation can be applied to μ and K.

Let θ be von Mises distribution and the density function is $g(\theta; \mu_0, K, 1)$. In case $\mu_0 \neq 180^0$, ease of consulting the table concerned can be effected by letting $\nu = \mu_0 - 180^0$ and a translational transformation is taken,

$$\theta^* = l\theta + \nu \tag{7}$$

The example in this paper the calculated values of $\hat{\mu}_0$ in the three regions I, II, III were found to be 19.5, 22.4, 37.1 respectively. The parameters θ_i^* ; $\hat{\mu}_{0i}^*$; $\hat{k}_i$ and l_i etc were then found out from expression (7) and tabulated in table 2.

Table 2. Some principle parameters of typhonic directional data

Region number (group serial number)	I	II	III
min. θ_i , max, θ_i	-16.7, 133.4	-15.6, 115.6	-13.5, 129.3
coefficient of amplification L	2.35	2.70	2.40
Coefficient of translation	134.3	114.0	91.0
$\hat{R}$	0.705	0.625	0.505
$\hat{\mu}_o^*$	180	180	180
$\hat{K}$	2.04	1.62	1.16

It will be seen from table 2 that in the three regions memtioned above, the mean direction $\hat{\mu}_o$, the lengths R of the resultant vectors and the scale parameters $\hat{K}$ are different in magnitude respoectively from each other. This accounts of the fact that the typhoons in the regions of rather low latitudes or to the west, the modes are to the west, and to which the typhoons concentrate is considerable. If the latitudes and longitudes of the regions are to the east or to the north, then there will be increased possibilities that the mean directions are to the north, the probability that the typhoons concentrate in the neighbourhood of the mean directions declines.

After having evaluating the von Mises distribution parameters corresponding to each section of data. We used χ^2-method to test the goodness of fit

$$\chi^2 = \sum_{i=1}^{m} (m_i - np_i) / np_i \qquad (8)$$

where m is number of groups, m_i is the actual frenquency, np_i is the theoretical frequency.

The results showing in table 3 indicate that all the directional data in the groups I, II, III had the degrees of confidence that exceeded 90% and were accepted as instances of von Mises distribution. But group IV was not the case.

Table 3. The goodness of fit test on the direstional data in four regions

region number	I	II	III	IV
group number(m)	12	12	8	8
$\chi^2_{m-s-1}(\alpha)$	$\chi^2_9(0.9)=4.168$	$\chi^2_9(0.9)=4.168$	$\chi^2_5(0.9)=1.61$	$\chi^2_5(0.9)=1.61$
χ^2-goodness of fit statistic	3.645	3.581	1.149	7.512
acceptable or not	yes	yes	yes	no

4. The confidence Intervals of Typhoon modal directions and the Probability distribution of θ

The mean direction $\hat{\mu}_0$ of the confidence interval (a, b) can be calculated as follows:

By using the given conditions and estimated values, we can fine out $\bar{K}(=nKR)$, and from the calculated $\bar{K}$ and the table concerned we obtain δ (according to a certain degree of confidences[3]. We have

$$a = \mu_0 + \delta - \pi, \qquad b = \mu_0 - \delta + \pi \tag{9}$$

This is also an arc which comprises the confidence interval of μ_0 or is expressed by

$$\text{arc}(\mu_0 + \delta - \pi \ , \mu_0 - \delta + \pi)$$

from which we obtain the confidence interval (a^*, b^*) of μ_0^* or

$$a^* = \frac{a-\nu}{l}, \qquad b^* = \frac{a-\nu}{l} \tag{10}$$

The results of corresponding evaluations in various regions are tabulated in table 4

Table 4 The confidence interval $\alpha=0.05$ of the mean direction of typhoons

region number	I	II	III
μ_0^*	(15.1, 23.7)	(20.4, 28.5)	(30.3, 43.9)
μ_0	(170, 190)	(169.1 191.0)	(167.6, 196.4)

Furthermore, the probabilities at the intervals in each direction of the movements of typhoons are tabulated in table 5. It will be shown in the table that the probabilities of moving directions of the typhoons at the intervals in each direction are varied. This may be attributable to the various geographic locations and the typhoon samples which differ in composition in each region.

Table 5. The probabilities at the intervals in each direction of the movements of the typhoons

region number	I	II	III
$P(a<\theta\leq b)$ $=P(a1+1)<\theta^{*}\leq b1+1)$	$P(-15<\theta\leq 57)=0.170$ $P(5<\theta\leq 35)=0.574$ $P(35<\theta\leq 65)=0.186$ P(other)+0.070	$P(-15<\theta\leq 10)=0.185$ $P(10<\theta\leq 40)=0.565$ $P(40<\theta\leq 70)=0.189$ P(other)=0.061	$P(-10<\theta\leq 20)=0.216$ $P(20<\theta\leq 50)=0.430$ $P(50<\theta\leq 80)=0.231$ P(other)=0.123

Lastly, by using the rank method of nonparametric test [3] it has been proved that marked differences exist in the distributions of the moving directions of the typhoons in the three regions I, II, and III.

5. Problems and Discussions

(1) The moving directions of the typhoons in the region IV is not von Mises distribution. This may be attributable to the fact that the location of the said region is more to the north or to the east than are the other three regions. The track of the typhoon when moving through the region is of a rather complex form. The influence of large-scale circulation on the movement of a typhoon is also very complicated.

(2) The statistical analysis of directional data can be applied to not only some evidently directioned data, such as those of typhoons, radar weather observations, wind directions, etc, but also to some kinds of data that are not in general expressible in any angular form, e.g., the air temperature in a 24-hr-weather observation, the summear rainstorms, etc. The reason for this is that these later kinds of data can be changed into certain forms of directional data

through some transformation process, and thereby can be analysed.

(3.) Like the linear regression model, the directional data can also fitting of the regression model. The asymptotic solution of the parameters associated with the model can be obtained by iterative method. Further investigation into the predition of typhoons will be discussed in another paper dealing with the subject.

6. References

—— Chen. L. S. and Ding, Y. H., 1979: An Introduction to North West Pacific typhoon. Academic Press, Beijing.

—— Shanghai typhoon Institute, 1984: Basic data of North-West Pacific typhoon (1949-1980), Meteorological Press, Beijing.

—— Mardia, K. V., 1972: Statistics of directional data, Academia Press, New York.

—— Watson, G. S., 1983: Statistics on Spheres, John Wiley and Sons.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

A STATISTICAL MODEL FOR PREDICTING THE DURATION OF MEIYU RAIN BASED ON MINIMIZING MEAN SQUARE ERROR OF PREDICTION

Shanxian Yu Xiaoyun Chen

Department of Computer, Zhejiang Meteorological Administration, Genshamen, Hengzhou, P.R. China

1. INTRODUCTION

Using a statistical model established by classical multiple regression to predict a varying state of system, some deficiencies are usually existed, especially in predicting abnormal years, the prediction error are often considerable. This is caused by the input variables chosen, especially by using a fixed model to predict a varying state of system. the error would increase with predicting projection, the predictors used in the model only reflect a mean state of historical sample, the model is not capable to represent the real state of system.

In order to remedy the above deficiencies, D.M. Allen (1971) has suggested the criterion of mean square error of prediction MSEPx its main point is placed at the mean of square of prediction. Given a predicted point x, the expectation of squares of prediction deviation is only related in independent variables selected on this point. Using this criterion, a subset of independent variables can be selected to minimize the expectation of squares of prediction deviation on point x. For different predicted point, the predicting model is different, the predictor in this model can optimally reflect the state of predicting system.

2. CRITERION OF SQUARE ERROR OF PREDICTION MSEPx AND CALCULATING METHOD

We suppose that there are a sample with N individuals composed of M independent variables X and dependent variables Y. When we select a set of independent variables, it is equivalent to X multiplied by matrix Q.

For example, when we select independent variables x1, x2, x3, it is equivalent to X multiplied right by

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 0 & 0 \end{pmatrix}$$

where the rows of Q equal to columns of X, then we select a design matrix of model as XQ. When the number of independent variables is M, the number of independent variables which have been selected is P, therefor Q is MP order matrix, its rank is P. The independent variables which have not selected in equation are equivalent to X multiplied left by

$$R = \begin{pmatrix} 0 & 0 & 0 & \cdots\cdots & 0 \\ 0 & 0 & 0 & \cdots\cdots & 0 \\ 0 & 0 & 0 & \cdots\cdots & 0 \\ 1 & 0 & 0 & \cdots\cdots & 0 \\ 0 & 1 & 0 & \cdots\cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & & 1 \end{pmatrix}$$

the matrix of independent variables which have not selected is XR, the order of R is M(M-P), its rank is M-P and satisfies $R'Q=0$, then

$$Q(Q'X'XQ)^{-1}Q' = (X'X)^{-1} - (X'X)^{-1}R(R'(X'X)^{-1}R)^{-1}R'(X'X)^{-1}. \quad (1)$$

Let

$$Z' = x'(X'X)^{-1}R(R'(X'X)^{-1}R)^{-1}R' \quad (2)$$

Following formula (1) the value of prediction on point x is

$$\hat{Y}_Q = (x - Z)'\hat{\beta}$$

where $\hat{\beta}$ is the least square estimation of β in complete model. In order to calculate the mean of square of prediction deviation $D=\hat{Y}_Q-X'\beta$ on point x, then

$$V_{ar}(\hat{Y}_Q)=(x-Z)'(X'X)^{-1}(x-Z)\sigma^2=(x'S^{-1}x+ZS^{-1}Z-2x'S^{-1}Z)\sigma^2 \qquad (3)$$

where $S = X'X$. since

$$Z'S^{-1}Z=x'S^{-1}R(R'S^{-1}R)^{-1}R'S^{-1}R(R'S^{-1}R)^{-1}R'S^{-1}x$$

$$=x'S^{-1}R(R'S^{-1}R)^{-1}R'S^{-1}x=Z'S^{-1}x,$$

it follows

$$V_{ar}(\hat{Y}_Q)=(x'S^{-1}x-Z'S^{-1}Z)\sigma^2 .$$

Further

$$E(\hat{Y}_Q-x'\beta)=E(\hat{Y}_Q)-x'\beta=(x-Z)'\beta-x'\beta=-Z'\beta$$

Thus, we get the mean of square of prediction deviation on point x :

$$MSEPx(Q)=E(D^2)=E((\hat{Y}-x'\beta)^2)=V_{ar}(\hat{Y}_Q)+(E(\hat{Y}_Q-x'\beta))^2$$

$$=\sigma^2(x'S^{-1}x-Z'S^{-1}Z)+(Z'\beta)^2$$

substitute $\hat{\sigma}^2$ and $\hat{\beta}$, the least square estimation in complete model, for σ^2 and β in above formula, discard the $\sigma^2 x'S^{-1}x$ which is independent on Q, we can get estimation of MSEpx :

$$\widetilde{MSEPx}(Q)=(Z'\hat{\beta})^2-Z'S^{-1}Z\,\hat{\sigma}^2$$

The key point for calculating MSEPx is that calculate Z, we left multiply formula (1) by x' and right multiply (1) by X'X, further according to (2) we obtain

$$Z'=x'-x'Q(Q'X'XQ)^{-1}Q'(X'X) \qquad (4)$$

Z can be calculated by scanning operation which is described as follows: Suppose M-order square matrix $A = (a_{ij})$, take a_{ii} as the axis, denote as S_iA, definite a new square matrix $C=(c_{ij})$, where:

$$C_{jk} = a_{jk} - a_{ik}a_{ji}/a_{ii} \qquad j\neq i,\ k=i$$

$$C_{ji} = -a_{ji}/a_{ii} \qquad j\neq i$$

$$C_{ij} = a_{ij}/a_{ii} \qquad j\neq i$$

$$C_{ii} = 1/a_{ii}$$

there suppose $a_{ii}\neq 0$

Now calculate Z, the variables which have been selected are denoted as i1, i2, ..., ip, the others which have not selected are denoted as j1, j2, ..., jr.

Denote

$$C = S_{i1}S_{i2}\cdots S_{ip}(X'X) \quad ,$$

then we have

$$(1),\quad Z_l = x_l - \sum_{k=1}^{p} x_{ik}C_{ikl} \qquad l=j1, j2, \ldots, jr \qquad (5)$$

$$(2),\quad Z_{i1} = Z_{i2} = \ldots = Z_{ip} = 0$$

It can be proved that the Z obtained by calculating with (4) or (5) are the same. The possible MSEPxs corresponding the minimum among the MSEPxs for all predictors are calculated, predicting model $\hat{Y}_Q=(x-Z)'\hat{\beta}$ is selected.

3. EXAMPLE

3.1, Selecting Predictor

According to some papers of study on Meiyu rain which occurs always during the plum mature period over the Yantse River the duration of Meiyn rain is mainly related to circulation type on 500 hPa level.

Following these study, we take a correlation computation of monthly mean mesh point on 500 hPa height field from November to March, the predictors whose correlation coefficient is higher than 0.4 are selected, namely

X1, monthly mean 500 hPa potential height of area covering 10°-15°N, 50°-65°E in December.

X2, monthly mean 500 hPa potential height of area covering 45°-65°N, 160°E-175°W in Janury.

X3, monthly mean 500 hPa potential height of area covering 15°-35°N, 20°W-25°E in February.

X6, monthly mean 500 hPa potential height of area covering 10°-25°N, 55°-70°E in March.

The same correlation computation of sea surface temperature in Northwest Pacific Ocean from November to March is made in same way as above, two predictors with the correlation coefficient 0.48 and 0.63 are selected, namely

X4, 35°N, 150°-155°E monthly mean sea surface temperature in November.

X5, 45°N, 150°=155°E monthly mean sea surface temperature in December.

3.2 Results of calculation

As an example, we use the data of Meiyu rain duration in North of Zhejiang Province of China from 1953 to 1985 and above six independent variables to set up two statistical models by stepwise regression and MSEPx. Having deleted one year individual the predictors are made by two models. The F-values of F test for both receiving and rejecting variables are 8.5, the predicting values are denoted as $\hat{y}'(k)$, and $\hat{y}(k)$ respectively and the errors of prediction are denoted as $\hat{e}'(k)$ and $\hat{e}(k)$ respectively, the results of calculation are shown in Table 1.

From the results shown as Table 1, we find that the trend of duration of Meiyu rain predicted by two models is similar, but in abnormal years, the predicting accuracy of model established according to criterion of MSEPx is much better.

The other advantage of using criterion of MSEPx is less independent variables in model but the improved precision of prediction. For example, in 1957, the model established according to criterion of MSEPx only have one predictor, the duration of prediction is 19 days, the duration of Meiyu rain predicted by stepwise regression model composed of five

Table 1 Results by Both Stepwise Regression and MSRPx

k	year	y(k)	$\hat{y}(k)$	$y^*(k)$	$\hat{e}(k)$	$e^*(k)$	k	year	y(k)	$\hat{y}(k)$	$y^*(k)$	$\hat{e}(k)$	$e^*(k)$
1	1953	33	29	29	-4	-4	18	1970	32	31	16	-1	-16
2	1954	78	76	52	-2	-26	19	1971	31	36	35	5	4
3	1955	34	29	29	-5	-5	20	1972	13	15	15	2	2
4	1956	45	45	30	0	-15	21	1973	14	14	13	0	-1
5	1957	20	19	14	-1	-6	22	1974	38	29	29	-9	-9
6	1958	7	10	16	3	9	23	1975	30	29	29	-1	-1
7	1959	10	15	22	5	12	24	1976	30	35	44	5	14
8	1960	18	21	21	3	3	25	1977	19	11	11	-8	-8
9	1961	8	18	23	10	15	26	1978	11	12	11	1	0
10	1962	21	18	18	-3	-3	27	1979	20	25	25	5	5
11	1963	20	23	23	3	3	28	1980	41	34	34	-7	-7
12	1964	5	6	6	1	1	29	1981	9	2	-5	-7	-14
13	1965	14	6	5	-8	-9	30	1982	25	30	30	5	5
14	1966	31	27	28	-4	-3	31	1983	30	20	20	-10	-10
15	1967	20	25	25	5	5	32	1984	19	28	28	9	9
16	1968	19	27	26	8	7	33	1985	12	17	24	5	12
17	1969	19	15	15	-4	-4							

predictors is 14 days, the errors compared with observations are one day and six days respectively. Again in 1959, model of criterion of MSEPx contains one predictor by which 15 days are predicted, model of stepwise regression with five predictors predicts 22 days, the errors compared with observations are 5 and 10 days respectively. Particularly in 1954, the duration of Meiyu rain lasted 78 days, two models contain four predictors but the accuracy of prediction have a remarkable difference, errors caused by model of MSEPx are two days, errors causec by stepwise regression model are 26 days.

4. CONCLUSION

The results mentioned above show as follows:

a. utilizing the criterion of MSEPx to establish a model for predicting duration of Meiyu rain, The score of yes or no prediction and predicting accuracy especially in abnormal years are improved.

b. The model established according to criterion of MSEPx can better reflect system state on predicting point. In this model, deleting some independent variables which in deed influence dependent variables but are not easy to estimate the regression coefficients if they are received in model equation, When rejecting them, the precisions of estimating regression coefficients are increased instead. This indicates that it is advantageous that rejecte the independent variables from model which are related to dependent variables but difficult to control.

5. REFERENCE

Chen Xiru and Wang Songgui, 1984 : Modern Practical Regression, Guangsi Renmin Press, (in Chinese).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

A SEQUENTIAL TEST METHOD FOR CORRELATION COEFFICIENTS AND ITS APPLICATION

Ding Yuguo

Nanjing Institute of Meteorology, Nanjing, China

1. INTRODUCTION

The current method of the correlation coefficient statistical test is usually used for the single sampling under a certain sample size n. It is based on a single try. For example, if the test hypothesis H_0 is $\rho = \rho_0$ (ρ_0 may not be zero), we can use another statistics

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{1}$$

When α level of the significance is given, the normal variable U test may be used for (H_0) $\rho = \rho_0$.

The above test method is only a significance test for deciding whether H_0 is accepted or rejected. The observations of meteorological variables are usually time dependent. When a random sample of n items is given, The sample correlation coefficient (SCC) often has random fluctuations, depending on the lapse of observational time. Hence it cannot be used to infer the population correlation coefficient (PCC) from the arbitrary single sampling.

There are two categories of errors in applying the test method. Generally, if $u > u_\alpha$, then (H_0) is rejected, but if $u < u_\alpha$, it is not safe to accept H_0. In the statistical test, the critical values of the two categories of error probability must be determined, i.e.,

$$P(|u| > u_\alpha \mid H_0) = \alpha \tag{2}$$
$$P(|u| < u_\alpha \mid H_1) = \beta \tag{3}$$

so that a reliable decision may be made for (H_0) and the alternative hypothesis (H_1). In fact, it is impossible to get the critical value of the classical test from a given α. In order to solve the problems mentioned above, the sequential test for correlation coefficients is presented in this paper.

2. THE PRINCIPLE AND METHOD

In formula (1), Z follows the normal distribution. According to Fisher (1943), approximately, we have

$$E(Z) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} = \mu \tag{4}$$

$$Var(Z) = \frac{1}{n-3} = \sigma^2 \tag{5}$$

Suppose we have two sets of observations from variables X and Y, $x_1, x_2, \ldots, x_t, \ldots, x_n$ and $y_1, y_2, \ldots, y_t, \ldots, y_n$. Considering (H_0) $\rho = \rho_0$ against (H_1) $\rho = \rho_1$, we may write the following statistics

$$\mu_1 = \frac{1}{2} \ln \frac{1+\rho_1}{1-\rho_1} \tag{6}$$

$$\mu_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} \tag{7}$$

However, under (H_0) and (H_1), we have the normal distribution density $f(Z, \mu_1)$ and $f(Z, \mu_0)$ for the variable Z.

From the probability ratio test developed by Wald (1943), the independent observations $z_1, z_2, \ldots, z_i, \ldots, z_m$ follows the same normal distribution and have the joint probability density ratio

$$T_m = \ln \frac{\prod_{i=1}^{m} f(Z_i, \mu_1)}{\prod_{i=1}^{m} f(Z_i, \mu_0)} = \frac{1}{\sigma^2} \left[(\mu_1 - \mu_0) \sum_{i=1}^{m} Z_i + \frac{m}{2} (\mu_0 + \mu_1)(\mu_0 - \mu_1) \right] = \sum_{i=1}^{m} t_i \tag{8}$$

where T_m is the joint statistics from m times of sampling. The basic idea of the sequential test is to design a function, i.e., a test rule, which must be considered for taking Z_1 in each sampling. The function or test rule generally include the following three decisions for each of the successive sampling.

i) Accepting the hypothesis
ii) Rejecting the hypothesis
iii) Continuing observations by making one more sampling.

When α and β are given, the critical values can be written as follows

$$A_o = \ln A = \ln\frac{1-\beta}{\alpha} \tag{9}$$

$$B_o = \ln B = \ln\frac{\beta}{1-\alpha} \tag{10}$$

As Z_m is given, the following test rule may be used: if $T_m = \sum_{i=1}^{m} t_i \geq A_o$, then (H_o) is rejected against (H_1) and if $T_m = \sum_{i=1}^{m} t_i \leq B_o$, then (H_o) is accepted against (H_1), but if $B_o < T_m < A_o$, then sampling must be made m+1 times. It has been proved that the probability of making a decision is equal to 1 by finite times of sampling (see Fisz, M., 1958), that is, by the sequential test, the decision can be made after several samplings.

For convenience, from (1),(8), and by simplification, we have

$$T_m = (n-3)\left\{\frac{1}{4}(c_1-c_o)\left[\sum_{i=1}^{m} d_i - \frac{m}{2}(c_o+c_1)\right]\right\} \tag{11}$$

where $c_o = \ln\frac{1+\rho_o}{1-\rho_o}$, $c_1 = \ln\frac{1+\rho_1}{1-\rho_1}$, $d_i = \ln\frac{1+r_i}{1-r_i}$, $i = 1,2,\ldots m$.

Furthermore, the formula may be written as the recurrence formula

$$T_m = T_{m-1} + \Delta T_m \qquad m = 2,3,\ldots \tag{12}$$

where

$$\Delta T_m = (n-3)\left[\frac{1}{4}(c_1-c_o)\left(d_m - \frac{1}{2}(c_o+c_1)\right)\right] \tag{13}$$

Obviously, (12) shows the probability ratio of Z obtained from m times of sampling is the sum for the probability ratio obtained from m-1 times of sampling and the increment ΔT_m relating to the mth sampling. When $m > 1$, the recurrence computation may be made by using (12) and (13).

From the above mentioned, the sequential test of the SCC has indirectly used the sequential test of variable Z, where each Z_i corresponds to a pair of records of variables X and Y, which is a sample with size n. Thus, the sequential sampling of Z is equivalent to a sequential sampling of the PCC under the size n. Suppose that the sequentially moving correlation calculation (the sample size n) is made in the time domain. Then, by applying (1_2) or (13) for i=1,2,...,m, the indirect decision of rejecting or accepting the hypothesis can be made. It is obvious that the above test of the SCC seems to be a "dynamic" correlation test. The test method is more advantageous and objective than the classical correlation test by using a single sampling. As the values of α and β are given, this method is more precise and efficient than the classical test, and its sample size is more economical. Particularly, this test method is good for samll samples.

3. CASE STUDIES

Example 1

The records (1873-1904) in pair for January average temperature of Shanghai and the yearly relative number of the sunspots are given. The hypothesis (H_o) is $\rho = \rho_o = 0.0$ and the alternative hypothesis (H_i) $\rho = \rho_1 = 0.3$. Given $\alpha = \beta = 0.05$, the sequential test is made.

First, records over an arbitrary period of time for n = 10 (1873-1882) is taken and the correlation coefficient $r_1 = 0.2910$ is determined. Then, from (12), we have the statistics

$$T_1 = (n-3)\left\{\frac{1}{4}(c_1-c_o)\left[d_1 - \frac{1}{2}(c_o+c_1)\right]\right\} = 0.3140$$

On the other hand, from (9) and (10), the critical calues are as follows

$$A_0 = \ln\frac{1-\beta}{\alpha} = \ln\frac{1-0.05}{0.05} = 2.944$$

$$B_0 = \ln\frac{\beta}{1-\alpha} = \ln\frac{0.05}{1-0.05} = -2.944$$

And $B_0 < T_1 < A_0$. Thus, a further sampling must be made. Suppose a new sampling is to slide in the time domain, e.g., the next sampling (n=10) is made at 2-year intervals and the next sample correlation coefficient calculated is $r_2 = 0.4108$. Then, by using (12) and (13), we obtain

$$T_2 = 0.6108$$

$$T_2 = T_1 + \Delta T_2 = 0.314 + 0.6108 = 0.9248$$

However, $B_0 < T_2 < A_0$

Further samplings should go on in the same way until m = 5, and then we have

$$T_5 = 3.0239 > A_0$$

Now we can reject the hypothesis $(H_0): \rho = \rho_0 = 0.0$ against the alternative hypothesis $(H_1): \rho = \rho_1 = 0.3$, and Table 1 shows the above sampling process.

TABLE 1. The sequential test process in example 1.

m	r_m	d_m	T_m	T_{m-1}	T_m	Comparison of m with A_0 and B_0	Decision	Critical value
1	0.2910	0.5993	0.3140		0.3140	$B_0 < T_1 < A_0$	Undecided	A_0=2.944
2	0.4108	0.8731	0.6108	0.3140	0.9248	$B_0 < T_2 < A_0$	Ditto	B_0= -2.944
3	0.5461	1.2256	0.9927	0.9248	1.9175	$B_0 < T_3 < A_0$	Ditto	$\alpha = \beta = 0.05$
4	0.5126	1.1325	0.8918	1.9175	2.8093	$B_0 < T_4 < A_0$	Ditto	
5	0.2485	0.5076	0.2147	2.8093	3.0239	$T_5 > A_0$	H_0 rejected	

It is shown in Table 1 that the sequential test can be used to make two possible judgements in which the critical values α and β are given for (H_0) and (H_1). In other words, the test is discriminative for (H_0) and (H_1). In this example, the moving SCC for n=10 is taken over 18 years (1873-1890). This shows that the population correlation coefficient of n=10 is perhaps close to $\rho = 0.3$.

In contrast, the U test for the classical sampling of r is also made under $\alpha = 0.05$ and (H_0) $\rho = 0$. Using the average value of the above sequential sampling $\bar{r} = 0.4018$ as the representative value in the time period, i.e., supposing r = 0.4018 is the single sampling value in the time period (1873-1890), the transform variable Z = 0.4258 and the statistics

$$U = \frac{Z-\mu_0}{\frac{1}{\sqrt{n-3}}} = \frac{0.4258}{\frac{1}{\sqrt{7}}} = 1.1265$$

Obviously, $U < U_\alpha = 1.96$,
and we cannot reject (H_0). If $\rho = 0.3$ is used as (H_0), the calculation result by using U test is still $U < U_\alpha$. It follows that the discrimination from the single sampling test seems to be lower than that from the sequential test under n=10. In order to examine the sample size which is required by the sequential method or classical mithod, we may estimate n for $U_\alpha = 1.96$ and r = 0.4018. From $U_\alpha = 1.96 = (Z-\mu_0)\sqrt{n-3}$, we have

$$n = \left(\frac{1.96}{0.4258}\right)^2 + 3 = 24.$$

It is shown that at least 24 years is required in the sample for rejecting (H_0) $\rho = 0$ under $\alpha = 0.05$ and $r = 0.4018$. However, in the sequential method only 18 years is required.

Example 2.

The records in pair for July rainfall and yearly rainfall of Shanghai are given. Let $n = 20$, (H_0) $\rho = 0.3$ and (H_1) $\rho = 0.6$. The sequential test is made for the SCC r under $\alpha = \beta = 0.05$. The whole process of the sequential test is given in Table 2 (omitted).

It may be seen that in the final decision to reject H_0, $m = 5$, and in fact, the sampling time domain of the records is still 18 years (moving at 2-year intervals).

In the same way, the classical U test is made for the SCC by using the mean value of the above sampling $r = 0.5402$ as the representative value in the time domain. The result shows that in the single sampling the rejection of H_0 is made when $n \geqslant 47$ and $r = 0.5402$. It follows that under the same hypothesis the sequential test needs a smaller sample for the SCC than the classical test.

4. EXAMINATION OF THE STABILITY OF THE SCC

The fluctuation of the SCC is a subject that has drawn much attention and study so far. Although this subject has been under discussion by some researchers (Lin, X. C., 1978, Zhu, S.M., 1982), there are still some problems, such as the stability of the SCC, to be solved. The sequential test can prove to be useful to check the stability of the SCC.

From (11) and (12), the statistics T_m is a function of c_1, c_0, and d_i. It is equivalent to the function of H_i, H_0 and r_i ($i = 1, 2, \ldots, m$). When the SCC increases linearly, T_m will change with the times of sampling for different values $\rho_1 - \rho_0$ (corresponding to $c_1 - c_0$). This is shown in Fig. 1 (omitted).

Furthermore, if the SCC remains constant, then T_m is constant, see (13), and T_m shows a linear trend. As r persists for quite long, the final decision may be made. On contrary, if the SCC shows fluctuations, the rate of increase or decrease of T_m will be small, and as a result, the inference of H_0 and H_1 will be deferred.

In view of the above fact, we may calculate the number of times of the critical sampling under the persistent correlation and given α, β.

In table 3 the number of times of the critical sampling are given under $\alpha = \beta = 0.05$ and two differences of c_1 and c_0, i.e., $\rho_1 - \rho_0$.

TABLE 3.a. Minimum times of sampling for $c_1 - c_0 = 0.5$.

d / n	H_0 rejection/H_0 acception		
	$c_0+0.5/c_1-0.5$	$c_0+0.4/c_1-0.4$	$c_0+0.3/c_1-0.3$
10	14	22	67
15	8	13	39
20	6	9	28
25	4	7	21
30	4	6	17

TABLE 3.b. Minimum times of sampling for $c_1 - c_0 = 0.9$.

d / n	H_0 rejection/H_0 acception		
	$c_0+0.9/c_1-0.9$	$c_0+0.72/c_1-0.72$	$c_0+0.54/c_1-0.54$
10	4	7	21
15	3	4	12
20	2	3	9
25	1	2	7
30	1	2	5

In table 3, if the difference of c_1 and c_0 is great, then the samll persistence of the SCC

is required. For example, if $c_1 - c_0 = 0.9$ (i.e., $\rho_1 - \rho_0 \approx 0.37 - 0.42$), the SCC of n = 30 generally needs 1 or 2 times of sampling for rejecting or accepting H_0, but if $c_1 - c_0 = 0.5$ (i.e., $\rho_1 - \rho_0 \approx 0.23 - 0.25$), then the SCC of n = 30 needs at least 4 times of sampling. From the above mentioned, the stability of the SCC may indirectly be cheked by the number of times of sampling which is required in the persistent correlation.

A more detailed table similar to the above may be worked out by chosing proper factors or analyzing the correlation of variables.

The change of partial correlation coefficients has some influence on the stability of the multiple regression. By using the sequential test, the stability of each factor can be discriminated.

If Fisher's Z variable is used in the partial correlation, we need only to rewrite the variance of variable Z as

$$\sigma_z^2 = \frac{1}{n-(k-2)-3} = \frac{1}{n-k-1}$$

Generally, the partial correlation coefficients $r_{yi_1 \cdot i_2 i_3 \ldots i_k}$ are in the linear regression with k independent variables. Thus, from (11)-(13), we have

$$T_m = (n-k-1)\left\{\frac{1}{4}(c_i-c_o)\left[\sum_{i=1}^{m} d_i - \frac{m}{2}(c_o+c_1)\right]\right\}$$

$$\Delta T_m = (n-k-1)\left\{\frac{1}{4}(c_1-c_o)\left[d_m - \frac{1}{2}(c_o+c_1)\right]\right\}$$

$$T_m = T_{m-1} + \Delta T_m$$

where all the symbols are the same as in the preceding formulas. The sequential test of the multiple regression equation based on the selection of the partial correlation will be discussed in another paper written by the author.

5. CONCLUSIONS

i) The sequential test of the SCC is more advantageous than the classical correlation test, for α and β have been considered for the two categories of errors and only a small sample is required. The meteorological record is in fact a time series, so that the effect of the test is equivalent to increasing the length of records. That is, the test method is particularly fit for a short sample series.

ii) In the regression analysis, the choice of factors depends mainly on the SCC and the partial SCC. As the small samples may produce false high correlation, the sequential test can be used to discriminate the stability of the SCC. As a result, the discriminability for the factors can be improved by means of the sequential test.

Acknowledgment. The author wishes to extend special thanks to Prof. C.S. Yao for his careful reading of the manuscript and many helpful suggestions for the improvement of this paper.

REFERENCES

1. Hald, A., 1955: Statistical Theory with Engineering Applications, New York, John Wiley (2nd).

2. Lin, X.C., 1978: The Unstability of the Correlation Coefficient in Statistical Forecasting of the Weather (in Chinese), Atmosphere Science, 2, 1, 55-63.

3. Zhu, S.M., 1982: The Stability Analysis Method of Correlation Coefficient and Its Application, (in Chinese), J. Meteorology, Vol. 40, No.4, 497-501.

4. Yao, C.S., 1963: Climatic Statistics, 22, Science Publ., Beijing, China.

5. Chen, X.R., 1981: A Introduction of Mathematic Statistics, Science Publ., Beijing, China (in Chinese).

6. Ding, Y.G., 1980: A Preliminary Research for the Influence of Meteorological Element Fluctuation of Frequencies upon the Correlation (in Chinese), Journal of Nanjing Institute of Meteorology, No. 2.

7. Fisz, M., 1958: Wahrscheinlichkeitsrechnung und Mathematische Statistik, VEB Deutscher Verlag der Wissenschaften, Berlin.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

A PROBABLE AVENUE OF THE EFFECT OF THE POLAR ICE ON THE NORTH PACIFIC SUBTROPICAL HIGH

FANG ZHIFANG

(Chengdu Institute of Meteorology, Chengdu,
Sichuan, People's Republic of China)

Fang (1984) suggested that there obviously exists a negative correlation between the polar ice in the Bering sea, the Beaufort sea and near the Alaska gulf during spring and winter and the subtropical high in the North Pacific during summer. Yoshino and Xie (1983) indicated that the current makes very important contribution to the interaction between sea and atmosphere. However, so far some problems have not been clear to us such as how the polar ice in spring and winter has an influence on the subtropical high in summer, what the probable mechanism of their influence is and what role the ocean current plays in them.

In this paper, interaction of ice-sea-atmosphere is statistically studied, We present a probable avenue of effect of polar ice on the North Pacific subtropical high via the effect of ocean currents on the sea surface temperature (named SST).

1. DATA AND METHOD

As a basis of this study, we make use of the monthly mean data of polar ice covered area within 160°E - 110°W and north of 50°N, including the Arctic ocean, the Bering sea, the Beaufort sea and the Alaska gulf, from November to July of following year for 25 years (from 1953 through 1977).

The areal index data of the North Pacific subtropical high at 500hPa are provided by the National Meteorological Administration of China.

Monthly mean sea surface temperature at 245 grid points within 120°E - 80°W and 50°N - 0°N from 1953 through 1977 is applied. We select SST in January, March, May and July to represent SST conditions from winter to summer because of the huge thermal inertia of ocean water. Considering direction of ocean current and distribution characteristics of water mass, the North Pacific being closely related to the polar ice is divided into three regions (see Fig.1), in which region A is the current region of Northeastern Pacific and its current flows from northwest to southeast; region B refers to the California cold current from north to south, then with a turn to southwest; region C is located at the downstream of the California cold current and belongs to north side of north equatorial current, its current flowing from east to west.

The respective lag correlations are calculated between the polar ice and sea surface temperature in the North Pacific from November to July, and between sea surface temperature in the North Pacific from November to July and areal index of subtropical high of the North Pacific lagging zero to eleven months.

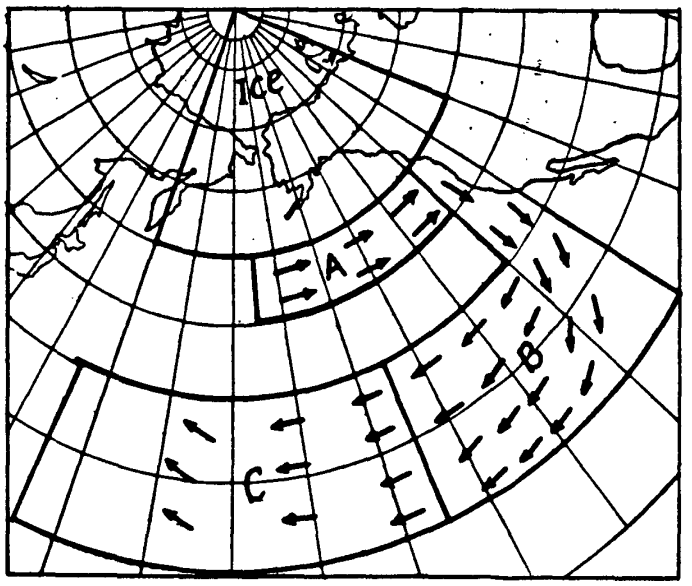


Fig.1. Distribution of regions A, B and C in the North Pacific

2. RESULTS

Effect of polar ice in winter and spring on SST of the North Pacific

The lag correlation between the polar ice from November to July of the following year and SST from January to July is analysed.

The lag correlation between the polar ice in November and December and SST from January to July is not obvious. It is shown that the effect of the polar ice in November and December on SST from January to July is not significant. However, the lag correlation between polar ice from January to May and SST from January to July shows that all the monthly correlation fields have a common character, that is, the polar ice of each month during January - May has a negative influence on SST in regions A and B first, then SST in region C only in July. This result indicates that the negative correlation region tends to be propagated towards down-stream of the ocean current with time. The analysis also indicates that from January to May the polar ice is closely related to the lagging SST field; especially the effect of the polar ice in April and May on SST in May and July is more evident.

In order to show the effect of the polar ice on SST in spring and winter, the correlation between the polar ice in May and SST in May and July is shown in Fig.2. In Fig.2(a), there is a correlation center in regions A and B, and region A and region B are combined via isocorrelation -0.3. It is interesting to see that the direction of isocorrelation coincides with direction of the Northeast Pacific current and the California current obviously. The correlation extent in regions A and B is wider than that of the effect of the polar ice from January to March on SST in May, and correlation region in region A is close to the latitude circle of 50°N. The values of two correlation centers reach to -0.55 and -0.75 respectively, and -0.75 is the maximum among all the correlation coefficients between the polar ice and SST.

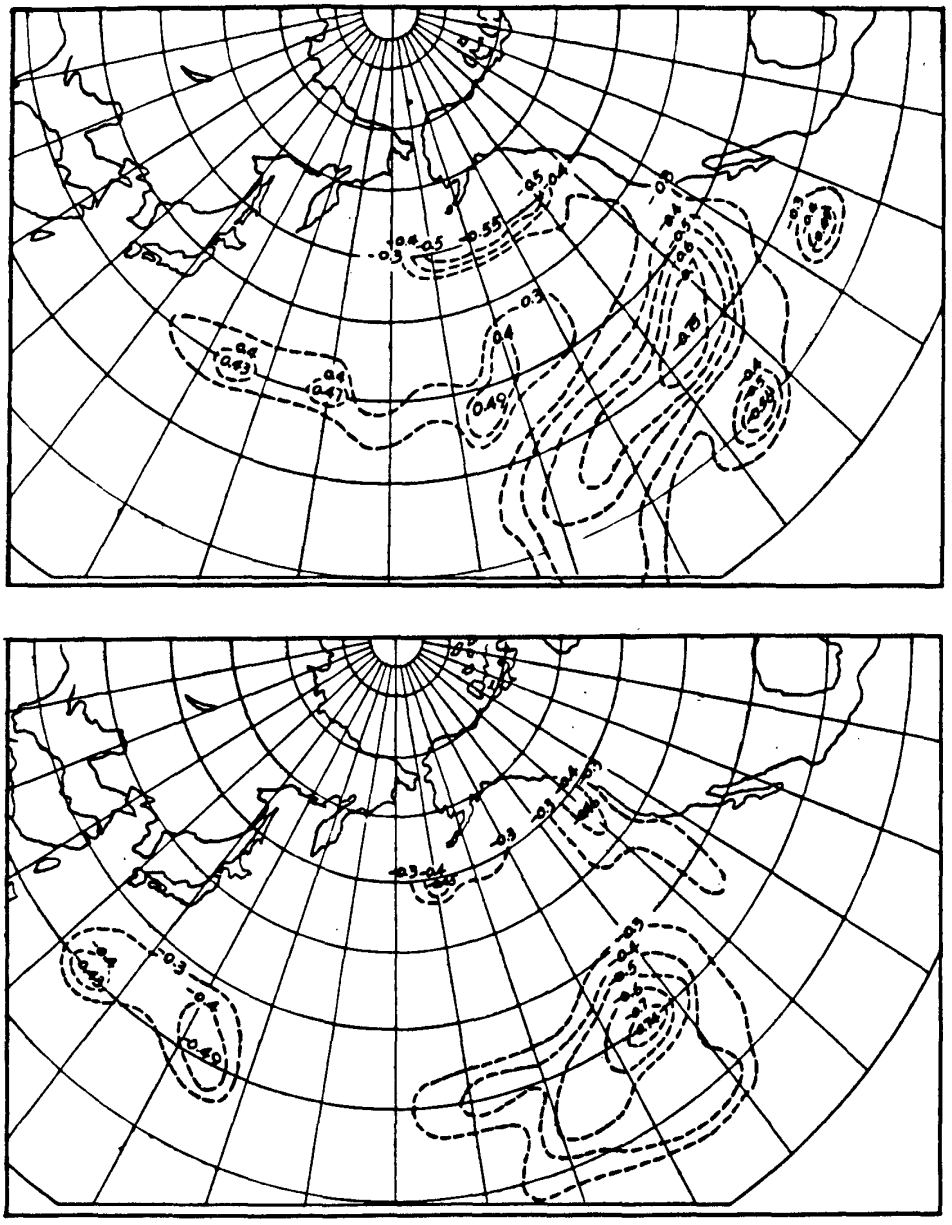


Fig.2. The lag correlation between polar ice in May and SST in May (Fig.2(a)) and July (Fig.2(b))

Fig.2(b) demonstrates the lag correlation field between polar ice in May and SST in July. It shows that region A seems to be moved to east and nearly reaches to west coast of U.S.A. But region B is moved to west and reaches to region C, and its center value is -0.74. It is noteworthy that both the correlation regions seem to be moved toward downstream with the ocean current.

We think that the polar ice in this region has an apparent influence on SST in the Northeast Pacific because of absorbing heat during the ice melting period, that is, if the extent of the polar ice in April and May is larger, then SST in region A in May is obviously lower, and the pack-ice and the lower SST region tend to move toward downstream with ocean current. As a result, the polar ice in April and May has also an apparent influence on SST in the California cold current region. In July, the correlation region reaches to the east side of the north equatorial current, but the strength of correlation is weakened. The effect of the polar ice in June and July on SST in July is much smaller than that from January to May. Their correlation coefficients are small and their distribution is irregular because polar ice moves back to the north of the Bering Channel and it is difficult to have an influence on SST in the North Pacific via the Bering Channel.

The effect of SST in California cold current region from March to July on subtropical high in the North Pacific

The lag correlation between SST in March, May and July and areal index of subtropical high lagging zero through eleven months in the North Pacific is analysed. The result of analysis shows that the correlation between SST in some regions and subtropical high is apparent, especially regions A, B and C. In these lag correlation fields, the common character exists, that is, there are closely positive correlations between SST of each month in regions B and C and subtropical high in June, July and August. But in other months, their correlations are not significant. It is emphasized that the effect of SST in the California cold current region in May on subtropical high is the most evident. In July, the correlation values between SST in regions B and C and subtropical high are smaller than those in region B in March and May. Interestingly, the extent of the correlation region between SST and subtropical high basically coincides with that of the polar ice and SST.

In order to show the effect of the polar ice on subtropical high, the lag correlation between SST in May and subtropical high in July is demonstrated in Fig.3. Analysis shows that there is apparently positive correlation between SST in region B in May and subtropical high in July, and its maximum value reaches to 0.62. However, the extent of positive correlation is basically in agreement with that of the California cold current. Comparing Fig.3 with Fig.2(a), we find that the correlation region between SST in region B in May and subtropical high is almost coincided with that of SST and the polar ice in May. For the other months, a similar phenomenon exists. According to analysis described above, we think that the polar ice from January to May has an influence on SST in the California cold current region from March to May. If the area of the polar ice is larger, then SST is lower. Owing to the thermal character of ocean, there is lag of effect of SST on subtropical high. Therefore, effect of SST on areal index of subtropical high in the North Pacific generally lags several months and causes a weaker subtropical high from June to August, and vice versa.

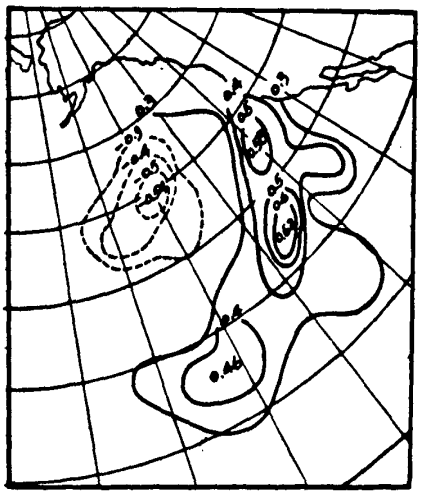
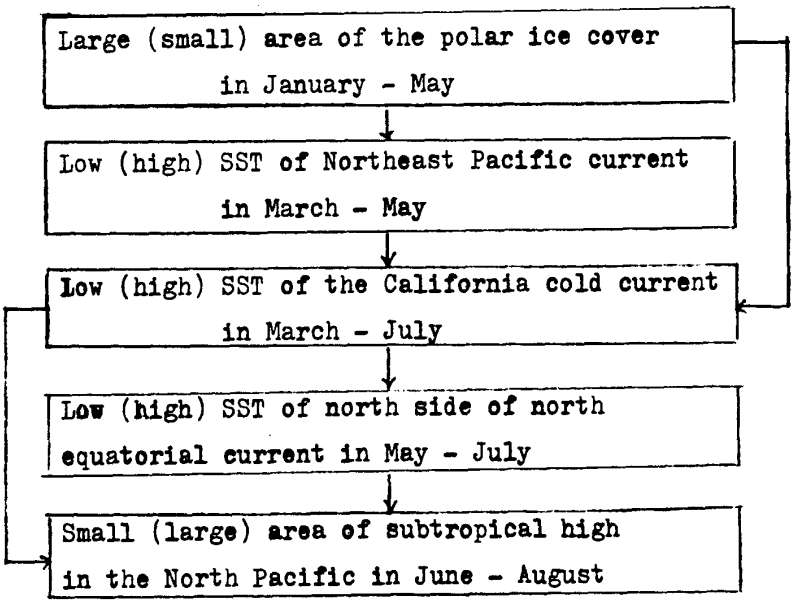


Fig.3. The lag correlation between SST in May and Subtropical high in July

3. SUMMARY

According to the analysis mentioned above, we can come to the following conclusion: the polar ice from January to May, especially in April and May, has an influence on SST in the current region of the Northeast Pacific first, then in the California cold current region and north side of the north equatorial current. All of them emerge in negative correlations. However, effect of SST in these regions from March to May on strength and area of subtropical high from June to August is evident and their correlations are positive. Therefore, we consider that one way to the effect of the polar ice in spring and winter on subtropical high in summer is the effect of the polar ice on SST in these regions via the California cold current etc. first, and then SST has an influence on subtropical high. All the procedure forms an influence mechanism of ice-sea-atmosphere. It may be illustrated by the following flow chart:



In this mechanism, there is lag of effect of the polar ice on sea and sea on atmosphere because of the thermal character of ocean. Owing to seasonal variation of the polar ice, seasonal variation of an influence mechanism also exists. In spring and winter (January - May), especially April and

May, when ice is melting, effects of the polar ice on SST in spring and summer (March - July) and subtropical high in summer (June - August) are clear, but effects of polar ice in winter (November and December) and summer (June and July) on SST and subtropical high are not significant. The ocean currents play an important role in all the influence procedure of ice - sea - atmosphere. The currents transfer the cooling ocean water caused by the melting polar ice, and as a result, SST is decreased. Furthermore, the lower SST has an influence on the atmosphere via the interaction between sea and atmosphere. Therefore, in the correlation fields mentioned above, the lag correlation regions which coincide with the ocean currents are formed.

REFERENCE

- Fang Zhifang, 1984: Interaction between the subtropical high and the polar ice in the Northern Hemisphere, The proceedings of Beijing International Symposium on Climate(to be published).
- M. M. Yoshino and Simei Xie, 1983: A preliminary study on climate anomalies in East Asia and sea surface temperature in the North Pacific, Science Reports of the Institute of Geoscience, University of Tsukuba, Section A, 4(1983).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE KUTZBACH EOF'S OF SUMMER MONSOON CIRCULATION AND RAINFALL FIELDS OVER INDIA

P.P. KAMTE, S.V. SINGH AND R.H.KRIPALANI
INDIAN INSTITUTE OF TROPICAL METEOROLOGY, PUNE-411008, INDIA.

INTRODUCTION

The EOF's have been applied in meteorology/Climatology well over since past 30 years for identifying the important modes of variation. They are the linear combinations of original variables, such that each ordered function explains the maximum variance of the remaining variance. In meteorology/climatology the variables can be grid point or station data fields of one or many weather elements. A combined EOF analysis of meteorological fields can enable us to study the important linear relationships among these fields, including the geographical distribution of strength of these relationships. Alternatively, one can extend the usual EOF analysis procedure to examine the sequential evolution of one or many fields. Weare and Nasstrom (1982) analysed the sequential evolution of SST fields by this way and named the method as Extended EOF analysis. Recent examples can be found in Love (1985), Teixeira and Reiter (1985), Brinkmann (1985) among others. It may be noted that to some extent this method works like poor man's cross spectral analysis such that the dominant periods if any in the sequential evolution of fields can be identified. We call this procedure as Kutzback EOF (KEOF) analysis after Kutzback (1967) who first adopted traditional EOF analysis for studying the inter-relationships between sea level pressure, surface temperature and precipitation over North America. In this paper we describe some results of studies on sequential intra-seasonal evolution of monsoon circulation and rainfall fields over India.

DATA

5-day average grid point data of 70hpa height fields and mean sea level pressure anomaly (MSLPA) fields for respectively 21 and 30 years and weekly rainfall anomaly field for 20 years are considered (Fig. 1).

RESULTS

The first two KEOF, found to be complimentary and in quardature to each other, show northward propagation of circulation and rainfall anomalies across Indian region with a recurrence period of about 40-days (For more details see Singh and Kripalani, 1986a). The evolutionary variance explained by first two or three KEOF's are given in Table below. The first two KEOF's related to the 40-day oscillation explain about 50, 45 and 22 % of the variance of 70hPa height, MSLPA and rainfall fields.

Table : Variance explained by KEOF's

Field	KEOF 1	2	3
70 hPa height	31.3	18.5	-
MSLPA	24.3	21.1	12.3
Rainfall	11.4	10.3	-

A complete cycle of evolution of rainfall fields as deduced from the loadings of first two KEOF's is shown in Fig.2. The movement of these anomalies is depicted by solid arrows. It is noted that the prominent rainfall anomalies (Loadings of EOF) first appear over the southeast

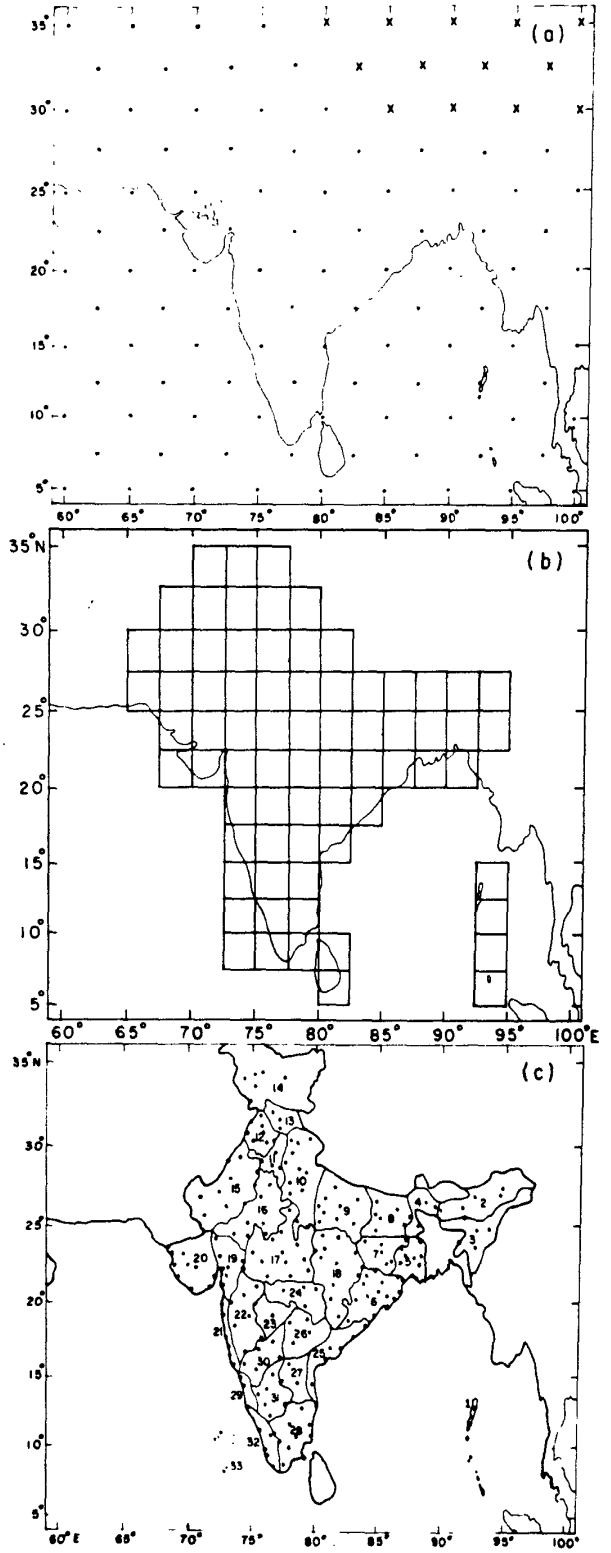


Fig. 1 : (a) The 98 grid points of the 70 hPa height fields. Dots represent the 25 points considered for the KEOF analysis

Fig. 1 : (b) The 72 blocks for the MSLPA fields shaded blocks are considered for the KEOF ana-lysis.

Fig. 1 : (c) The 33 meteorological subdivisions considered for representing the rainfall fields.

peninsula and move north to north-northwestward upto about 20°N lat. at a rate of about 0.5° Lat./day. Thereafter they move in a northeasterly direction. The anomalies are more pronounced near about 20°N latitude and weaken substantially as they reach extreme northeast part of the country. Similar results are obtained when rainfall data averaged over uniform blocks are considered (Singh and Kripalani, 1986b). In addition to the northward propagation of circulation and rainfall anomalies, the third KEOF of MSLPA field also showed the westward movement of pressure anomalies along about 15°N lat. with a recurrence period of 15 days (See Singh and Kripalani, 1986a).

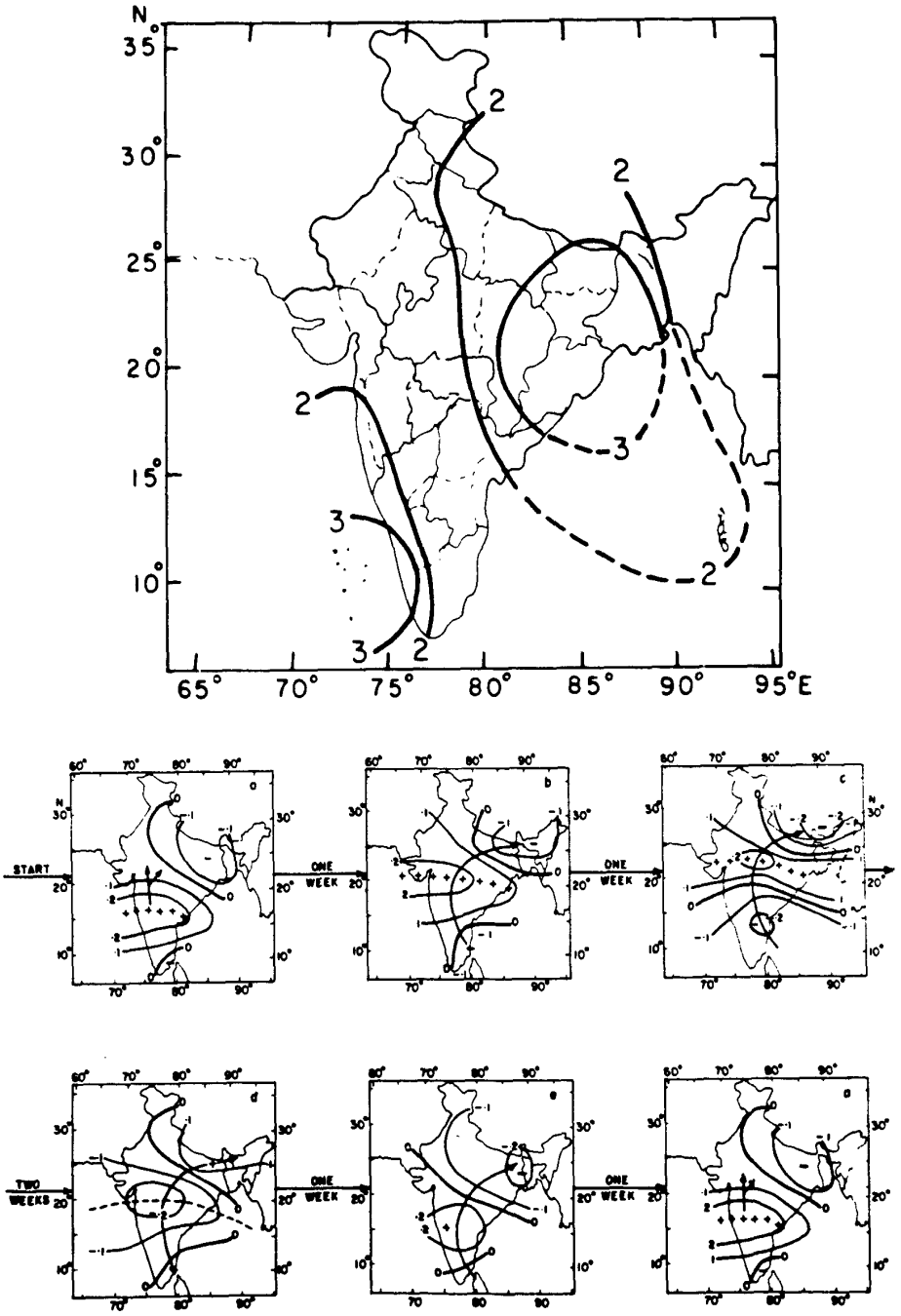


Fig. 2 : The most dominant feature of the sequential evolution of the weekly rainfall anomaly fields as composited from the first two KEOFs. The arrows show the general direction of movement of the centres of loadings in the next week.

APPLICATION IN FORECASTING

The principal components corresponding to 5 different phases of the 40-day cycle shown in Fig. 2 are correlated with the rainfall anomalies in the successive week for all the 32 subdivisions (See Fig. 1). The correlation coefficients exceeded .4 in absolute value for about 10 subdivisions in the west-central parts of the country. Further multiple regression equations are developed to predict the weekly rainfall anomaly by using these 5 principal components and the antecedent week's rainfall anomalies as predictor. These regression equations showed Heidke skill scores ranging .15 to .20 in forecasting weekly rainfall in three categories, viz. Excess, normal and deficient, for the west-central parts of the country, when verified on independent sample of about 50 cases (Kripalani et al. 1985).

FUTURE WORK

Recently we have obtained daily rainfall data of 365 stations for a period of 80 years. The KEOF's will be utilised for studying the evolutionary patterns of rainfall fields averaged for regular size blocks after suitable filtering. It may also be employed for studying the evolution of wind fields over Indo-Pacific region and the pressure and wind fields over the Southern Indian Ocean and adjoining region.

REFERENCES

Brinkmann, W.A.R., 1985 : Association between summer temperature and precipitation patterns over the great lakes region and water supplies to the great lakes. J. Climatology, 5, 161-174.

Kripalani, R.H., Bansod, S.D. and Singh, S.V., 1985 : The spatio-temporal evolutionary features of weekly monsoon rainfall over India, Mausam (Submitted).

Kutzbach, J., 1967 : Empirical eigen vectors of sea level pressure, surface temperature and precipitation complexes over North America. J. Appl. Meteor., 6, 791-802.

Love, G., 1985 : A study of the linear relationships between the monthly mean fields of sea surface temperature, mean sea level pressure and cloudiness over the north west Pacific. J. Met.Soc. Japan,201-209.

Singh S.V., and Kripalani, R.H., 1986a : Application of Extended Empirical orthogonal function analysis to inter-relationships and sequential evolution of monsoon fields. Mon. Wea. Rev., (accepted).

Singh, S.V. and Kripalani, R.H., 1986 b : The south to north progression of rainfall anomalies across India during the summer monsoon. PAGEOPH, 123, (to appear).

Teixeira, L. and Reiter, E.R., 1985 : Sea surface temperature, rainfall and southern oscillation : Relationship over Tropical America. Archiv. Meteor. Geophys. Biokl. A34, 85 - 109.

Weare, B.C., and Nasstrom, J.S., 1982 : Examples of extended empirical orthogonal function analysis. Mon. Wea. Rev., 110, 481-485.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

RAINFALLS AND WEATHER TYPES IN SOUTH WESTERN FRANCE

J.L. CHAMPEAUX, P. PERIS,
Centre National de Recherches Météorologiques
Toulouse, France

and

P. BENICHOU
Météorologie Nationale, Paris, France

1. INTRODUCTION

Climatological normals (particularlymean and frequency of precipitation) are the consequence of several situations corresponding to weather types. It is interesting to characterise the precipitation distributions for each weather type. To define a meteorological situation, we have choosen the geopotential 700 HPA which is one of the more explanatory levels of pressure for precipitation.

2. AUTOMATIC CLASSIFICATION OF THE GEOPOTENTIAL 700 HPA

Data are available for geopotential 700 HPA fields at OOH and 12 H a.m. from 1971 to 1980 (7184 data fields). The 8 by 8 point grid (fig. 1) is centered in South Western France ; the grid includes the nearby Atlantic zone and a large part of Western Europe which permits us to have a good image of the meteorological situation.

To classify these 7184 data fields, we have used the dynamic clusters method in nonhierchical clustering (DIDAY 1970) with the adaptative City-block distance :

$$d(x,y) = \sum_{j=1}^{p} \lambda_j |x_j - y_j|$$

with $x = \{x_1, \dots, x_j, \dots x_p\}$ → $\in$ Data fields

$y = \{y_1, \dots, y_j, \dots, y_p\}$ → $\in$ Data fields

$\lambda = \{\lambda_1, \dots, \lambda_j, \dots, \lambda_p\} \in \mathbb{R}^p$

This allows partitioning of the data fields. At each iteration, the data fields are agglomerated to the nearest center of gravity of the class. The principal limits of the method are the choice of the number of classes and the selection of the initial cluster centers.

The number of classes has been limited to 15. The classification was obtained after a large number of random initial cluster centers were tested. In general, the classes have seasonal characteristics (fig. 2) : for example, class 8 is typically a winter cyclonic situation and class 14 is a summer anticyclonic situation.

The analysis of geostrophic wind at Bordeaux (center of the working area) shows that each class is characterized by a well determined wind direction. For example, for the class 4, 98% of the situations are included in a wind sector 275°-365° with 95% of the situations having a wind speed greater than 10 m/s (fig. 2).

We shall label as strong patterns a group of data fields which always appear in the same class regardless of how the classes are obtained. These strong patterns correspond to characteristic meteorological conditions.

3. PRECIPITATION ANALYSIS FOR EACH WEATHER TYPE

For the precipitation analysis, daily precipitation values recorded in Meteorologie Nationale files during the period 1971-1980 were used. 774 precipitation gage sites were selected from the network of South Western France (fig. 3).

For each weather type, we analysed conditional precipitation parameters : frequency of precipitation, daily mean precipitation, frequency of high intensity precipitation, and quantiles of precipitation in order to better describe the distributions.

The class 4 (fig. 4), autumn and winter type, associated with a Norway-Sicily thalweg, is a rainy pattern. Through the analysis of frequency and daily mean rainfalls (fig. 7, 8) we can make several interesting comments :

The limit between oceanic and Mediterranean climates is clearly marked. We note the blocking effect of Pyrenees on North-West circulation. In addition, a precipitation wedge appears distinctly 20 kms inland. This wedge is found in all West, North-West circulations ; the sea-forest transition is a probable explanation of this phenomenon.

Using the strong pattern concept, we can extract from class 4 extremely characteristic meteorological patterns (strong patterns 24 and 25). The strong pattern 24 (fig. 5) corresponds to a clearer zonal circulation which explains more marked rainfalls on the west shoulders of Massif Central and less important rainfalls in the Pyrenees (fig. 9). In this pattern, the frequency of precipitation is greater than 90%. The strong pattern 25 (fig. 6) is a more meridian circulation ; it accentuates the blocking phenomenon of Pyrenees and shows decreased rainfalls on the Massif Central (fig. 10).

4. CONCLUSION

This climatological study took place within the HAPEX-MOBILHY land surface processes research program. The consideration of only annual means is not sufficient to correctly study surface hydrology. The type of study reported here may be seen as an interesting intermediate step between case studies and

classic climatology. It may also prove an original means of comparison between observations and mesoscale model outputs for some weather types.

5. REFERENCES

André J.C., Goutorbe J.P., Perrier A (1986) : HAPEX-MOBILHY : A hydrologic Atmospheric Experiment for the study of water budget and evaporation flux at the climate scale. AMS 67 n° 2.

Diday E (1970) : La méthode des nuées dynamiques et la reconnaissance des formes. Cahiers de l'IRIA.

Diday E (1971) : Optimisation en classification automatique non hiérarchique. IRIA.

CLASSE	NATURE *		SEASON	WIND DIRECTION		WIND SPEED (m/s)		
				DD°	SAME SECTOR **	<10	10-20	> 20
1	A	81%	JUNE - OCTOBER	---	---	86%	14%	---
2	I	74%	JANUARY - MAY	020	75%	54%	41%	5%
3	D	75%	JANUARY - MAY	230	71%	57%	39%	4%
4	D	83%	OCTOBER - MARCH	320	98%	5%	58%	37%
5	A	73%	MAY - SEPTEMBER	280	96%	23%	60%	17%
6	A	79%	JUNE - SEPTEMBRE	230	88%	49%	48%	3%
7	I	76%	NOVEMBER - MAY	270	96%	17%	60%	23%
8	D	90%	DECEMBER - MARCH	270	77%	25%	50%	25%
9	D	83%	DECEMBER - APRIL	240	80%	41%	50%	9%
10	A	60%	SEPTEMBER - JANUARY	340	83%	44%	53%	3%
11	D	88%	NOVEMBER - APRIL	010	61%	49%	46%	5%
12	A	74%	DECEMBER - MAY	---	---	71%	28%	1%
13	D	73%	DECEMBER - MARCH	270	94%	10%	51%	39%
14	A	70%	JULY - SEPTEMBER	300	67%	79%	21%	---
15	I	62%	MAY-JUNE-SEPT-OCT.	230	87%	47%	47%	6%

Fig. 2: Description of 15 weather types

* : A : Anticyclonic ; D : depression ; I : intermediary

** : Same Sector is reference direction ± 45°

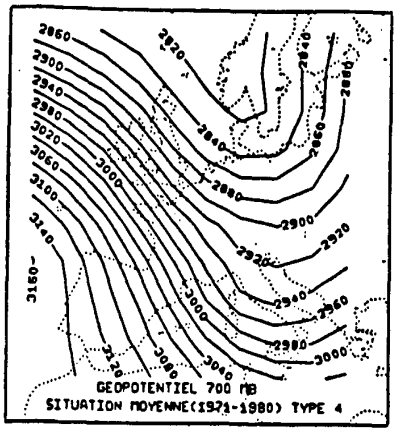


Fig. 4 : Mean data field GEO 700 HPA
Class 4

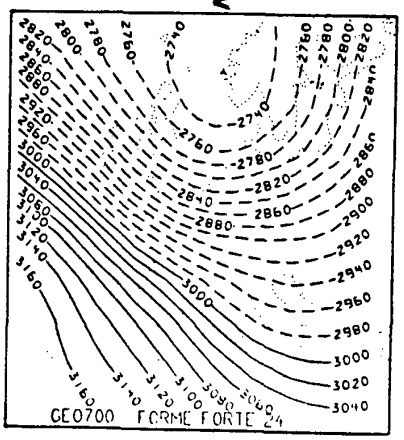


Fig. 5 : Mean data field GEO 700 HPA
Strong pattern 24

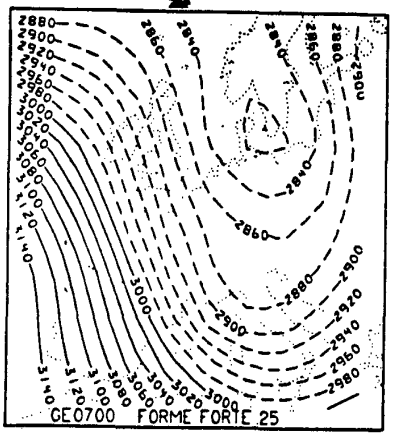


Fig. 6 : Mean data field GEO 700 HPA
Strong pattern 25

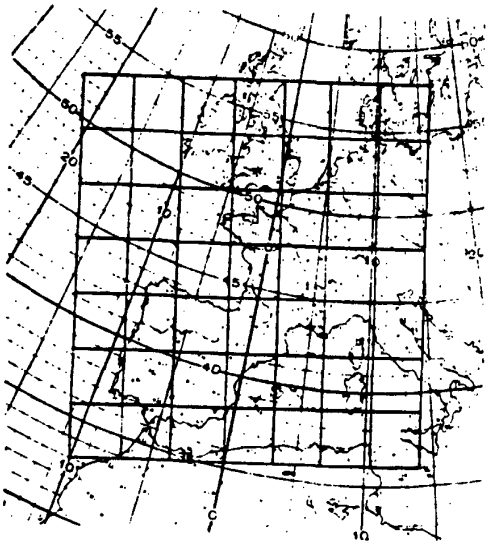


Fig. 1 : 8 by 8 grid

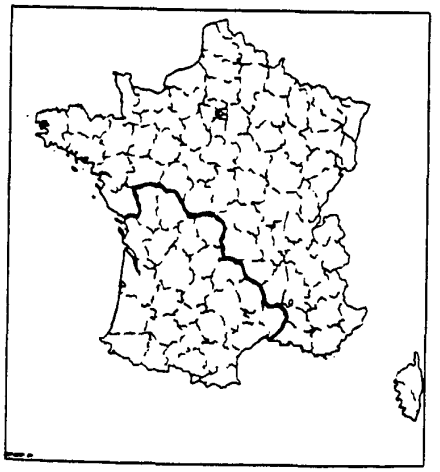


Fig. 3 : Network of South Western France

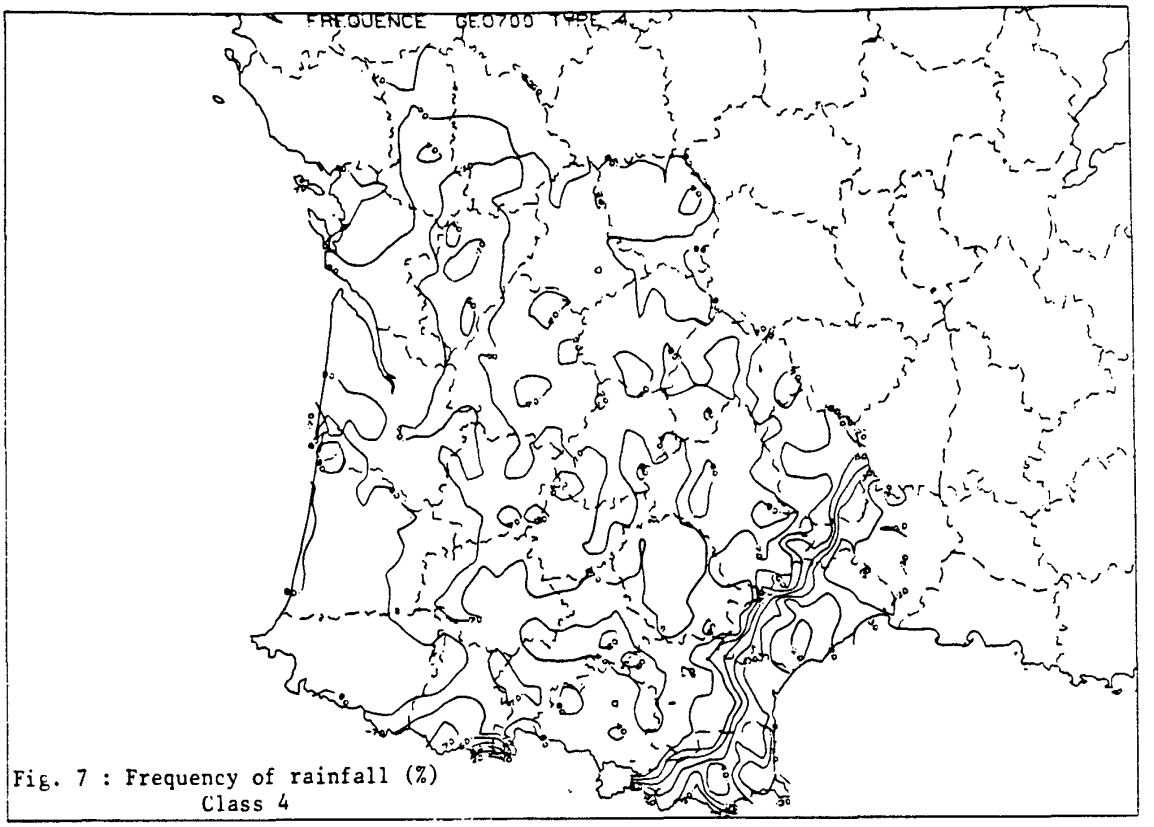


Fig. 7 : Frequency of rainfall (%)
Class 4

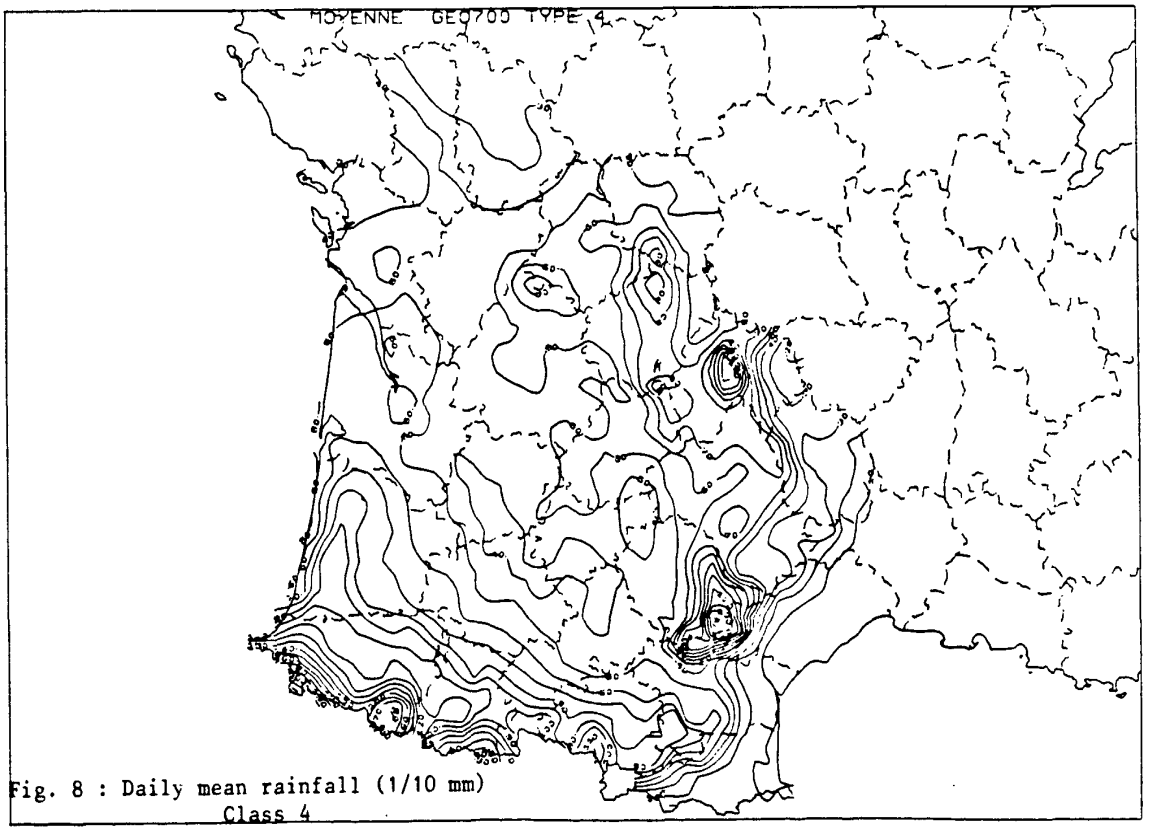


Fig. 8 : Daily mean rainfall (1/10 mm)
Class 4

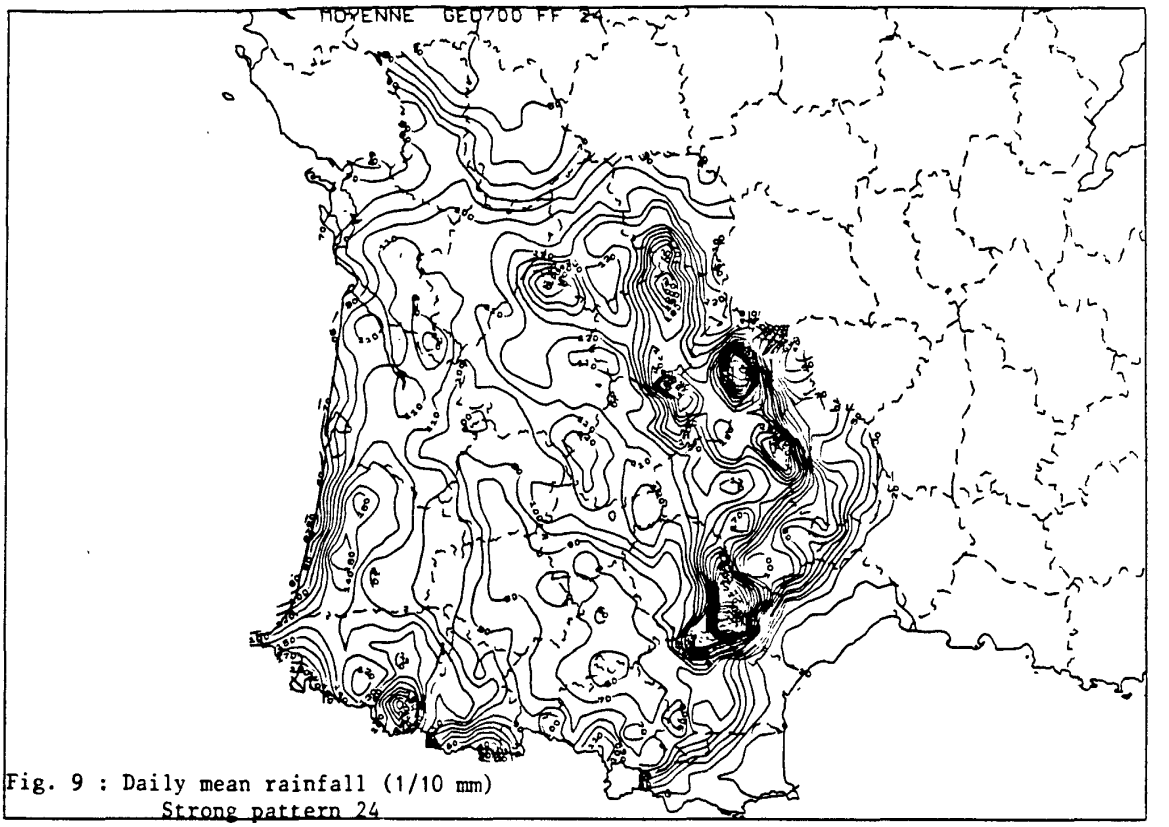


Fig. 9 : Daily mean rainfall (1/10 mm)
Strong pattern 24

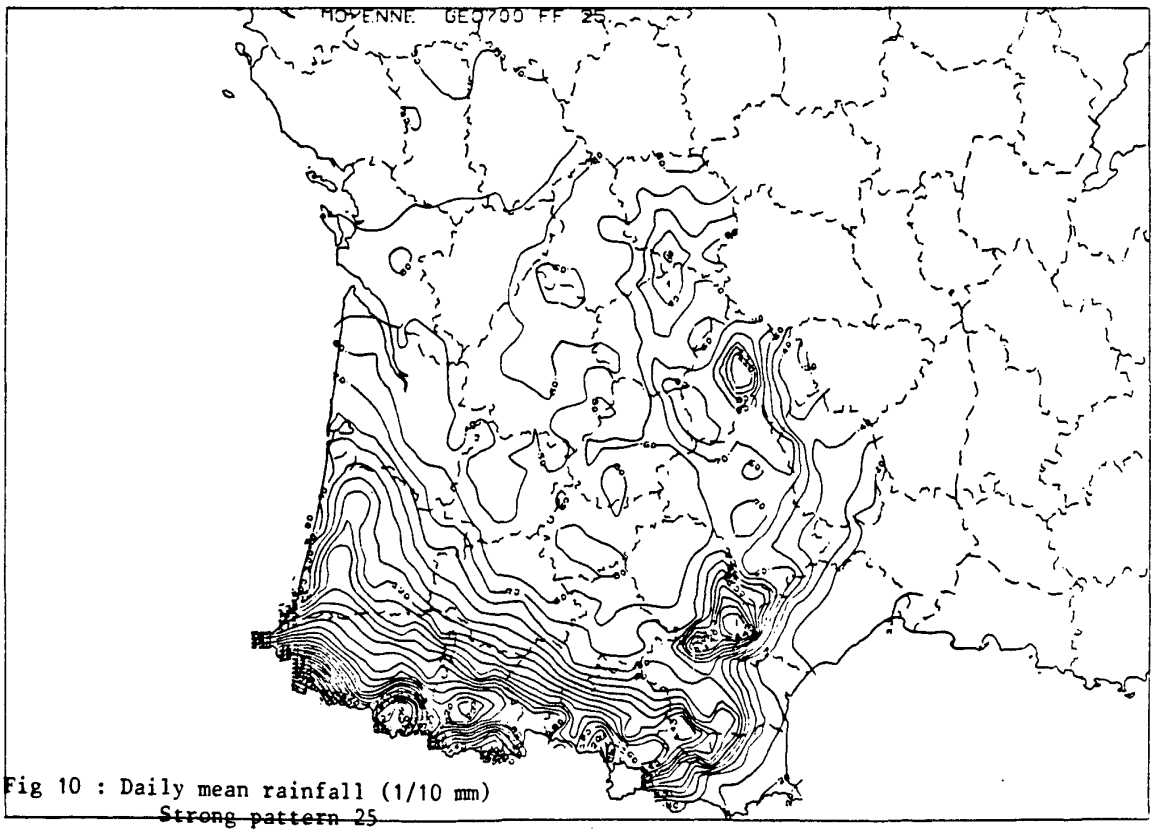


Fig 10 : Daily mean rainfall (1/10 mm)
Strong pattern 25

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE POTENTIAL PREDICTABILITY OF 70hPa HEIGHT AND RAINFALL FIELDS OVER THE INDIAN AND ADJOINING REGION DURING THE SUMMER MONSOON SEASON.

R.H.Kripalani, P.P.Kamte and S.V.Singh,
Indian Institute of Tropical Meteorology, Pune : 411 005, India.

INTRODUCTION

The potential predictability of monthly or seasonal averages can be studied by comparing the interannual variability of the observed monthly or seasonal means to the natural variability of the time averages of day-to-day values. Pioneering work in this field has been done by R.A.Madden (Madden, 1976). A similar approach was used by Shukla and Gutzler (1983) to study the predictability of seasonal northern hemispheric 50hPa contour heights and by Nicholls (1983) to study the temperature and pressure over the Australian region. One of the principal results of these studies is that the predictability is dependent on season and geographical location, being higher for the tropics than for the middle or higher latitudes.

The southern limit for most of these studies has been 20°N. Availability of sufficient length of daily 70hPa contour height and mean sea level pressure anomaly grid point data over the Indian region enabled us to analyse the potential predictability of seasonal mean circulation fields over the Indian region. We followed the approach of Shukla and Gutzler (1983) or method of Trenberth (1984 a) which is based on simple analysis of variance for studying the potential predictability of bi-monthly (July-August) means of 70hPa contour heights, and mean sea level pressure anomaly. The potential predictability of July-August means of rainfall amounts over the homogeneous meteorological regions has been studied by following an approach similar to that of Katz (1983). The choice of averaging over the two months, July and August, is determined by the factors that (i) this period is sufficiently long to be beyond the limit of day-to-day predictability, (ii) the meteorological fields and the weather remain almost stationary during this period of established monsoon and (iii) the external conditions like sea surface temperature remain more or less constant. For data used refer to the data section and Fig. 1 of Kamte et al. (in this issue).

RESULTS

For the given degfees of freedom an F-value (a test for signal to noise ratio) is significant at $\lesssim$ 1% level. The value 2 of F also signifies that the signal-to-noise ratio (i.e. F-1) is equal to 1 and that half of the observed interannual variability is potentially predictable. The higher the value of F, the more predictable the field is. However the null hypothesis of no signal should not be accepted fof the cases of F $\lesssim$ 2 since the confidence limits on F are quite large (Trenberth, 1985; Hayashi, 1982).

The results for the circulation fields are presented in Fig. 1 .

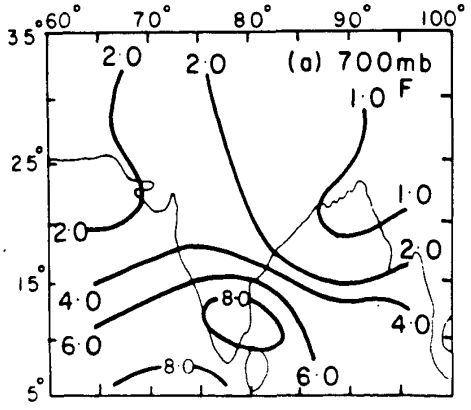


Fig. 1 (a) The F-values for 70hPa height fields.

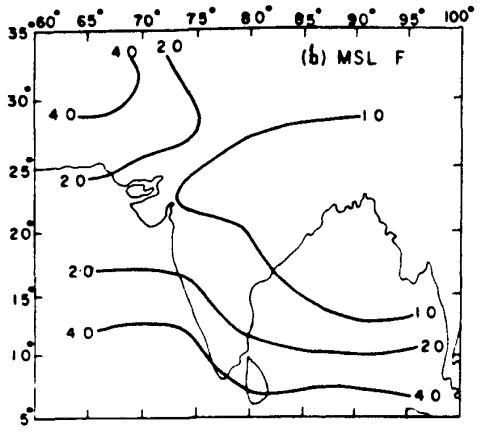


Fig. 1 (b) : The F-values for mean sea level pressure anomaly fields.

The most important feature of the distribution of F-values for 70hPa heights (Fig. 1 a) is the general decrease with the increase in the latitude. The values are very high, exceeding 8 at places, south of 15°N latitude. There is also a very sharp gradient between 15° and 20°N latitude. The area north of 20°N can be divided in 3 longitudinal sectors. Over the middle sector, of about 10° longitude width, the value of F exceeds 2 and the F-values are less than 2 over the other two sectors. Northeast of the head of the Bay, the values of F are even less than 1 meaning that the field over this region is unpredictable. This pattern suggests that the variation in fields near or south of 15°N have a very important bearing on the predictability of the monsoon circulation. It may be noted that the main storm/synoptic activity during July-August lies between 20 to 25°N latitude, where the effective time between the independent observations (τ) is higher. This is contrary to the results obtained for the extratropical regions where smallest τ are found in the regions of main storm tracks in both the hemispheres (Trenberth, 1985). This may be due to the longer time scale fluctuations, viz. quasi-biweekly and 40-50 day scale, noted in the monsoon activity. Although the life time of individual storms is few days, the monsoon remains active or weak for longer periods.

More or less similar results are obtained for mean sea level pressure anomaly fields (Fig. 1b). Over major parts of the country, the values of F are less than 2 with a minimum near the head of the Bay of Bengal. F-values are more than 2 south of 15°N and over Northwest India. A comparison of the spatial patterns of the F-ratio for 70hPa contour height and mean sea level pressure anomaly reveals that the F-values differ near 15°N. This is found to be mainly due to the difference in autocorrelations at lags greater than 1. For the mean sea level pressure anomaly the autocorrelation falls off more slowly over and near the head of the Bay of Bengal and relatively faster over the remaining parts of the country. For further details (See Singh and Kripalani (1986)).

The values of F for rainfall exceed 3 over a contiguous region consisting of East M.P., Orissa, Bihar and Bengal lying to the east of 80°E longitude and exceed 2 over the west coast and the surrounding regions. Thus over large parts of India more than 50 % of the interannual variability is potentially predictable. There is indeed ample evidence that a few regional or remote parameters are able to explain more than 50 % of the variance of the monsoon rainfall and the results obtained are consistent with this experience. It may be noted that the temporal dependence in the daily rainfall is more for the sub-divisions lying in the western parts of the country (Singh et al., 1981). Also the statistical models developed for predicting 5-day and weekly rainfall have shown higher forecast skills for this region.

However certain conflicts in the results and possible lacunae in the methodology may be noted. Where as the circulation fields show little potential predictability around the normal position of monsoon trough (15°N - 25°N Lat.), the rainfall shows potential predictability over the Eastern sector. This may imply that the circulation and rainfall fluctuations are independent of each other over the monsoon trough region. Yet our studies have shown that the circulation and the concurrent rainfall fields over the above region are significantly related, strongly so over the western parts. The results reported here will suggest that these specification relationships cannot be used for prediction of rainfall by first predicting the circulation fields by some (possibly numerical) method as the circulation fields exhibit no potential predictability over the monsoon trough region. Besides several regional and global large scale rainfall over the West-Central parts of the country (Singh et al. 1986) whether these discrepancies are due to inevitable errors in data (particularly 70hPa height data),, choice of the circulation parameters, irregular size and shape of the rainfall subdivisions or methodology is a matter of further investigation. Recently Trenberth (1984a,b) has noted some bias in procedure employed by us for estimating the autocorrelations and noise. Yamamoto et al (1985) has recommended the use of the minimum rather than average intraseasonal variance for estimation of the noise. In view of these difficulties the potential predictability of the tropical circulation and rainfall will be further studied by using different source and type of data and by employing various methodology. These results will be reported elsewhere.

REFERENCES

Hayashi, Y., 1982 : Confidence intervals of a Climatic Signal. J.Atmos. Sci., 39, 1895-1905.

Katz, R.W., 1983 : Statistical procedures for making inferences about precipitation changes simulated by an Atmospheric General Circulation Model. J. Atmos. Sci., 40, 2193-2201.

Madden, R.A., 1976 : Estimates of the natural variability of time averaged sea-level pressure. Mon. Wea. Rev., 104, 942-952.

Nicholls, N., 1983 : The potential for long range prediction of seasonal mean temperature in Australia. Aust. Met. Mag., 31,203-207.

Shukla, J., and D.S. Gutzler, 1983 : Interannual variability and predictability of 500 mb geopotential heights over Northern Hemisphere. Mon. Wea. Rev., 111, 1273-1279.

Singh, S.V., Kripalani, R.H., Shaha, P., Ismail, P.M.M., and Dahale, S.D., 1981 : Persistence in daily and 5-day summer monsoon rainfall over India. Arch. Met. Geoph. Biokl., Ser. A., 30, 261-277.

Singh, S.V., and Kripalani, R.H., 1986 : Potential predictability of lower tropospheric monsoon circulation and rainfall over India. Mon. Wea. Rev., 114.

Trenberth, K.E., 1984 a : Some effects of finite sample size and persistence on meteorological statistics. Part I : Auto Correlations. Mon. Wea. Rev., 112, 2354-2368.

----------, 1984 b : Some effects of finite sample size and persistence on meteorological statistics. Potential predictability. Mon. Wea. Rev., 112, 2369-2379.

----------, 1985 : Potential predictability of geopotential heights over the Southern Hemisphere. Mon. Wea. Rev., 113, 54-64.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

POLAR LOWS IN THE NORWEGIAN AND THE BARENTS SEA. A CLIMATOLOGICAL STUDY

I. Kanestrøm, T. Ese and K. Pedersen
Institute of Geophysics, University of Oslo, Norway

INTRODUCTION

The purpose of this paper is to present some results concerning air flow pattern during weather situations with polar lows. A polar low is a meso-scale low or an irregularity in an otherwise uniform cold air flow blowing southward from the polar region. The horisontal extension of the low is 100 - 500 km. Strong winds, gale or stronger, is usually associated with the lows. Due to the small horisontally scale, the lows can usually not be forecasted by a common weather forecast model. Therefore, fishermen and other may be overtaken by polar lows. Consequetly the lows may represent a great safety risk for the human activety in the Norwegian Sea. The objective of this project is to look for some possibility to forecast polar lows.

CRITERIA CHARACTERIZING THE FLOW PATTERN DURING FORMATION OF POLAR LOWS

Our data base, geopotential heights and surface pressure are supplied by NCAR, and the data of observed polar lows are supplied by The Norwegian Meteorological Institute. In the periode 1971 - 82 (1980 missing) altogether 74 polar lows are identified.

The geographical distribution of polar lows at the time of discovery, makes it reasonable to divide the events into two groups, a western group consisting of lows first observed west of longitude 5^0E, and a eastern group consisting of those observed east of this longitude. The boundary line between these groups coincides with the position of open water west of Spitsbergen as seen from the average position of the ice edge in February (Fig.1).

Our analysis is restricted to polar lows in the Norwegian Sea and the Barents Sea. For all cases we found a strong northerly air flow over the oceans. In some cases, however, a low was located about the Greenwich meridian in the North Sea with a strong southerly flow in

the eastern part of the domain considered. This indicates that it is necessary to characterize the flow patterns of interest with more than one parameter. Seven parameters are chosen as follows:

1: $P(20^0W,70^0N) - P(5^0E,70^0N)$ 5: $P(15^0E,75^0N) - P(50^0E,75^0N)$
2: $P(5^0E,70^0N) - P(30^0E,70^0N)$ 6: $P(20^0W,75^0N) - P(50^0E,75^0N)$
3: $P(20^0W,70^0N) - P(30^0E,70^0N)$ 7: $P(10^0E,80^0N) - P(30^0E,70^0N)$
4: $P(20^0W,75^0N) - P(15^0E,75^0N)$

Here P is the sea level pressure. Consequently the parameters can be interpreted as a measure of the geostrophic wind speed.

The values of these parameters (criteria values) should be chosen in such a way that at least one of the pressure differences observed during formation of polar lows should be equal to or larger than the corresponding criterion value. The method used to select the criteria values may be illustrated by the following reasoning. The pressure difference no.1 has a near Gaussian distribution about its mean value. For days with values of this parameter above the mean value, polar lows may occur. Looking at the tail with extremely high values of this parameter we find that nearly all days when such a large pressure gradient occurs, polar lows are observed. The criteria values are given in Table 1, and the criteria is fulfilled by 72 out of the 74 observed polar low events. As seen from the table, the parameter values vary relatively little with the season.

TABLE 1

Criteria for the three seasons winter(Dec-March), spring (April-June), fall (Sep.- Nov.)

Season/Par.No.	1	2	3	4	5	6	7
Winter	22.5	22.5	31.5	22.5	22.5	31.5	25.5
Spring	21.5	21.5	30.0	22.0	22.0	33.0	30.0
Fall	21.5	21.5	30.0	21.5	21.5	33.0	33.0

STATISTICAL ANALYSE

The climatological values of the pressure differences for the period 1946 - 79 are given in Table 2. These values show that the northern air flow for days with polar lows is considerably stronger than the climatological values.

TABLE 2

The climatological values of the pressure differences

Season/Par. no.	1	2	3	4	5	6	7
Winter	5.0	-3.0	2.1	8.1	-3.0	5.1	3.5
Spring	1.9	-0.4	1.5	2.6	-0.4	2.2	1.9
Fall	3.6	-2.4	1.2	5.9	-1.6	4.3	2.1

The flow pattern for the period 1973-79 is analysed in some detail. In Table 3 the yearly mean numbers of days with pressure differences (00Z or 12Z) that fulfil the criteria are given. The first line covers the cases where only one of the seven pressure differences meet the requirement, and the second line gives the distribution when the limit given for two or more parameters are surpassed. As seen from the table, the parameters no.1 are most frequently fulfilled. This shows, as also indicated by table 2, that in most cases the strong northerly airflow is located in the western part of the sector, close to the east coast of Greenland. About 75% of all events are picked out by these two paremeters.

TABLE 3

The yearly mean number of days with pressure differences taken from NCAR data that fulfil the criteria given in table 2.

Par. No.	1	2	3	4	5	6	7
Number of days	14.1	1.1	0.8	10.9	2.4	2.4	1.6
Number of days	31.4	4.7	11.7	28.8	4.4	10.4	8.7

The air flow condition given by the criteria may be met at several subsequent times of observation. Such a time interval will in the following be named an episode. More than one polar low may occur in this time interval. For the period 1973-79, the mean distribution of episodes over the year is found to be;

J	F	M	A	M	J	J	A	S	O	N	D
4.4	3.3	4.1	2.9	1.7	1.0	0.7	1.3	2.9	3.1	5.6	5.0

As seen, the distribution has a secondary minimum in February. This is in agreement with the results reported by Wilhelmsen (1985). The

maximum is found to be in November. The mean number of episodes are found to be 36 pr. year, with an average duration of about 40 hours. As seen, the number of possible episodes is much larger than the number of polar lows observed in the period 1971-82. For the years 1983-84 and 1984-85, the analysis is to a higher degree based on satellite images. This may be the reason why more polar lows are identified these years. For 1983-84 and 1984-85 the numbers of observed polar lows are 32 and 28, respectively. These numbers agree very well with the number of episodes forecasted for the periode 1973-79.

VARIATION WITH TIME

Our data base contains mean sea level pressure from 1899 to 1983 (1980 missing). Counting the days with a pressure distribution which meet our criteria, an increase of days with time is found. Using linear regression the number of days N as a function of n (n counts the subsequent 5 years period) is found to be

$$N=63.6 + 7.58\ n,$$

with a correlation coefficient $r = 0.90$. The regression is significant within the 95 per cent confidence limit. A question may be raised. Is this increase with time caused by the change in the analysis technique of the weather map on which our data are based? However, the observed pressure for December month at one station at the east coast of Greenland (70^0N, 22.5^0W) and at one station in Norway (71^0N, 25^0E), indicates that the increased frequency with time is real and not caused by a change in the weather map analysis technique.

COMPOSITE AND ANOMALY PATTERNS OF THE SYNOPTIC MAPS

In order to make composite pictures we find it convenient just to divide the polar low events into the two geographical groups as described earlier. For the key days, mean maps, composite maps are computed for each group. In the following, we will discuss only the western sector maps for the winter season.

The composite SLP map (mean map) is given in Fig.2. The "normal" weak trough in the Norwegian Sea is changed to a deep low pressure located at 79^0N and 5^0E with a centre value of 982 mb. The pressure ridge over Greenland is stronger than normal, 1020.8 mb. Also the Azores high is changed with a center values of 1028 mb at 45^0N,25^0W.

Fig.3 shows the anomaly patterns. In the western sector the main feature seems to be the large negative anomaly associated with the Norwegian Sea Low and the positive anomaly over the Greenland Ridge.

The northerly geostropic wind component between these two pressure systems is strong. We also notice a strong positive anomaly at 50^0N, 30^0W. The Aleutian Low is deeper and the Siberian Monsoonal High is stronger than normal.

The anomaly patterns for the 500 hPa are similar to those found at the surface. However, the values of the anomalies are remarkably large, the negative anomaly close to the Greenwich meridian has a value of 230 m. The composite thickness pattern for the 700-1000 hPa layer shows that a cold air through is located in the northerly flow over the Norwegian Sea. From the thickness anomalies we find the northerly air flow to be 8^0C colder than normal in the 500-1000 hPa layer. We also find an other marked anomaly. The air in the southerly flow west of the Atlantic Heigh pressure centre is about 6^0C warmer than normal. Inspecting the surface flow patterns and the 700-1000 hPa thickness patterns, one finds a strong baroclinicity south of the ice edge. North of the ice edge the air in the northerly flow is rather barotropic.

THE ZONALLY AGERAGE FLOW

The zonally averaged geostrophic wind is computed for the surface and for the 500 hPa level. It is found that north of about 70^0N the zonal wind is more easterly at the surface and less westerly at 500 hPa when polar lows are observed. than normal. In a belt south of about 60^0N at the surface, and 65^0N at the 500 hPa level, the zonal winds are more westerly, than normal. It is of interest to look at this increased shear at the 500 hPa level between the latitudes 82.5^0N and 62.5^0N. The zonal wind at these latitudes and the wind shear are given in Table 4. Also the corresponding values for each winter month are given.

One notices the remarkable difference in the zonal wind and in the shear between February and the other months. February seems to have an air flow that differs from that which favours formation of polar flows.

TABLE 4

The zonally averaged geostrophic wind (m s^{-1}) and shear (at 500 hPa)

Latitude	West	East	Normal	Dec.	Jan.	Feb.	March
82.5^0	0.2	-0.8	2.4	2.4	1.8	3.0	2.0
62.5^0	9.0	6.8	6.1	6.5	5.8	5.4	6.7
62.5^0-82.5^0	8.8	7.6	3.7	4.1	4.0	2.4	4.7

References

Wilhelmsen,K. 1985. Climatological study of gale-producing polar lows near Norway. Tellus, 37A, pp.451-459.

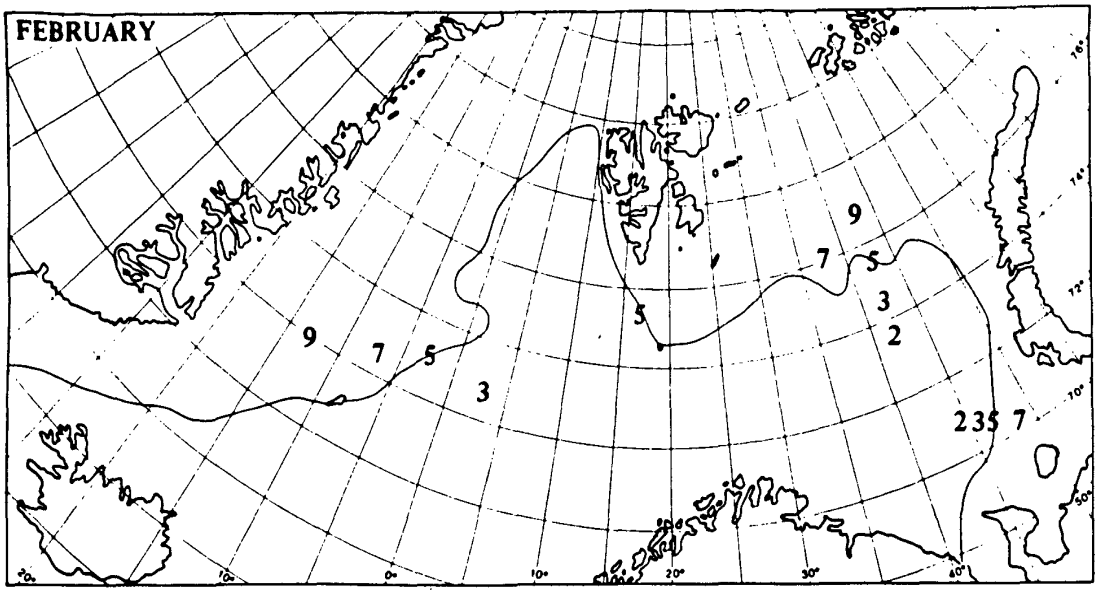


Figure 1: The averaged position of ice edge in February.

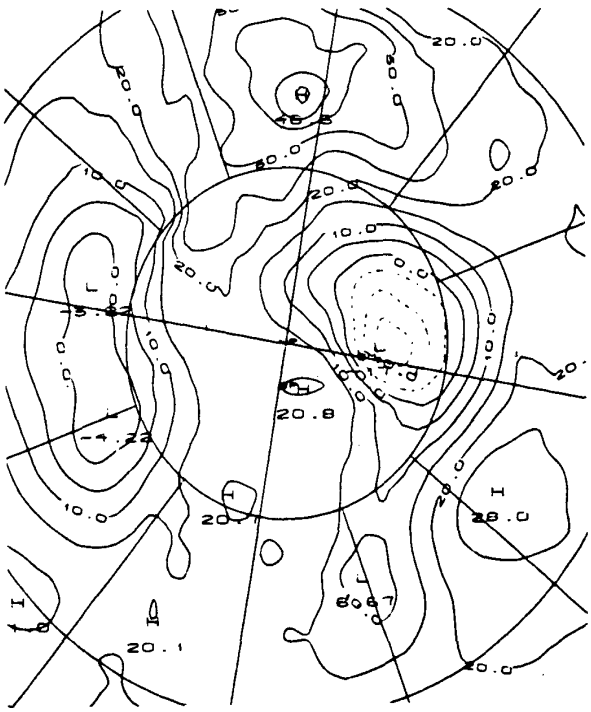


Fig.2: The composite sea level map (contour interval 5 mb).

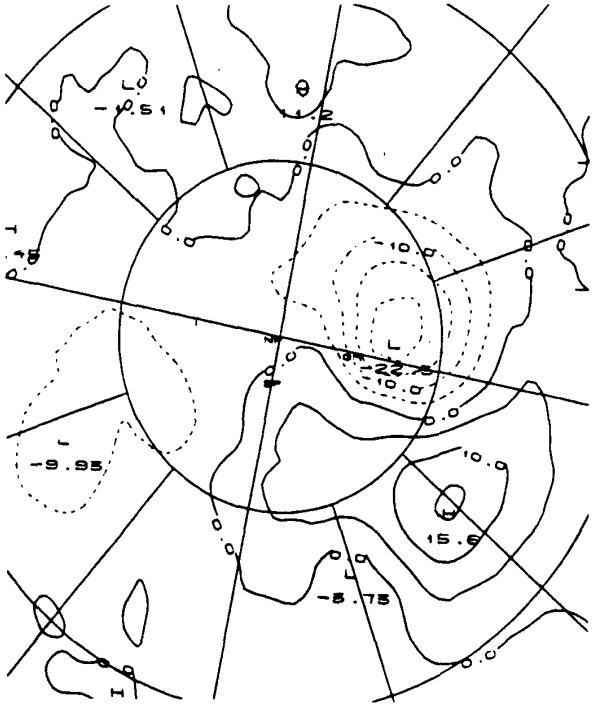


Fig.3: The sea level anomaly field (contour interval 5 mb).

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

MAXIMUM WINDS AND RAINFALL IN TAIWAN DURING A TYPHOON

H. P. CHU H. Y. TSENG *CHIN-PIAO PU*

Air Navigation & Weather Services *Air Traffic Services Division*

Civil Aeronautics Administration

Taipei, Taiwan, Republic of China

1. INTRODUCTION

Taiwan is situated in the Northwest Pacific Ocean off the South-eastern coast of the Chinese mainland and lies in the course of typhoon, therefore it is attacked by typhoon very frequently. According to the statistics of the past 80 years (1905–1984), there are averagely about 23 tropical storms occurring in the Northwest Pacific and the South China Sea each year, and among them about 3.4 would hit Taiwan (1945–1984). In summer and autumn especially, the opportunity of being invaded by typhoon is most frequent; while in winter (December to March) typhoon disappears completely in Taiwan area because of the influence of the East Asian continental high pressure.

Taiwan is a narrow and elongated island with the Central Mountain Range of some 3000-meter height running through the whole island from north to south and dividing the island into two parts: the eastern part and the western part. The eastern part is a longitudinal valley with a few plains, while the western part is a more extensive coastal plain. Because of the influence of the special topagraphy, and also the different path of typhoon center passing through Taiwan or the waters around, the precipitation and wind velocity in various places during typhoon invasion are therefore widely different, and disasters caused in various places are also apparently different. The meteorologists in Taiwan, such as Chi (1978), Chu (1983, 1982) Shieh (1986), Tseng (1983), Wang (1982), etc., had successively made analyses, investigations, and studies on the matters concerning the intensity, precipitation and moving path of the typhoons occurred in the Northwest Pacific. Generally, when Taiwan is under typhoon's attack the magnitudes of wind and rain-fall in various places not only have close relation to the circulation structure of the typhoon itself, but also have relation to its moving path and whether or not the typhoon is accompanied with any other weather system when it is invading Taiwan. Wang (1982) ever made a detailed discussion in this concern in his study on typhoon forecasting.

According to the investigation over Taiwan area heavy rain disasters in the recent 20 years, if the hourly precipitation of a place is higher than 15mm or the daily precipitation is higher than 130mm, usually a disaster could be expected. In summer and autumn when typhoon invades, the maximum precipitation of a place in a short time is often much higher than those figures, therefore the downpour brought forth by a typhoon has attracted very serious attention.

2. THE OCCURRENCE FREQUENCIES OF TROPICAL STORMS INCLUDING TYPHOON IN THE NORTHWEST PACIFIC AND THE SOUTH CHINA SEA AND OF THE TYPHOONS THAT INVADE TAIWAN

According to the analysis and statistics we made for the past 80 years (1905–1984), the occurrence frequency of typhoon in the Northwest Pacific and the South China Sea is shown in Table 1. Frequency of tropical storms (including typhoons) by month.

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	TOTAL	REMARK
TS/TY	27	14	26	46	76	111	309	345	348	271	195	89	1,860	1905–1984
TS/TY	16	8	20	28	41	64	160	194	184	153	104	47	1,018	1945–1984
TY	11	2	10	21	30	38	93	122	127	115	71	28	679	1945–1984
INVADED TAIWAN				2	3	12	29	32	37	13	7		135	1945–1984

Table 1. Frequencies of tropical storms (including typhoons) by month
TS: Tropical storm
TY: Typhoon

a. The occurrence frequency of tropical storms in the Northwest Pacific and the South China Sea

In the vast oceanic area of the Northwest Pacific and the South China Sea within 5°N to 40°N, 180° to 105°E, totally there were 1860 tropical storms including typhoon occurred in the past 80 years (1905~1984), on the average about 23 occurred in a year. In terms of month, the highest frequency appears in months from July to October, and nearly 70% of the tropical storms taking place in these months; while in months from January to April, the that it is only 6% of the total amount.

Because the weather reports made in the earlier years from 1905 to 1944 are incomplete, the data statistics has to be done by using data of the recent 40 years (1945~1984). According to the latter statistics, totally there were 1018 tropical storms, the yearly mean frequency is 25.5 and that it is 4.5 more than the 21 of the old statistics made from 1905~1944.

If in terms of typhoon (a tropical storm with a maximum wind velocity of 32.7 m/sec or higher near its center), totally there were 679 occurred in the 40-year period from 1945~1984, the yearly mean frequency is about 17 per year.

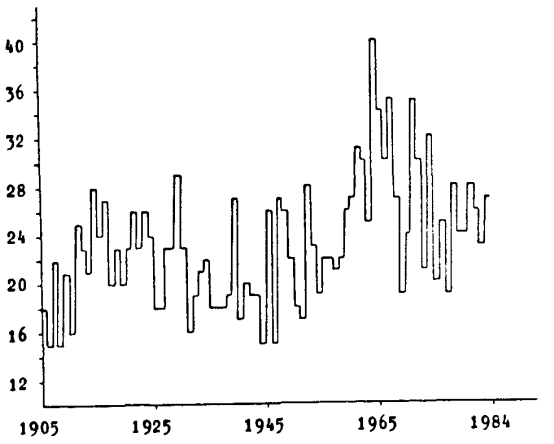


Fig. 1 Frequency of tropical storms (including typhoon) by year

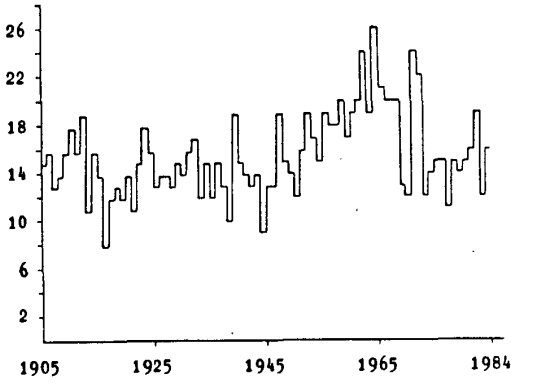


Fig. 2 Frequency of typhoon by year

b. The yearly occurrence frequency and cyclic variation of the tropical storms

According to the 80-year statistics as shown in Fig. 1. and Fig. 2., the highest frequency appearesin 1964 and the number is 37; the lowest frequency appears in 1906, 1908, 1945 and 1947 and the number is 15 each year.

In terms of typhoons, the highest frequency also appears in 1964 and the number is 25; the lowest frequency appears in 1916 and the number is 8; the highest frequency is three times the lowest one and obviously that is very irregular.

In the past, it was considered that the yearly occurrence frequency of tropical storm or typhoon should have a periodicity. Although a sine wave curve can be drawn based on the data of Fig. 1. and Fig. 2, the periodicity is not very regular, and considerable discrepancy also exists between the maximum and the minimum.

c. The tropical storms that invade Taiwan

The tropical storms (including typhoon) that invade Taiwan are referred to as with their centers pass through the island or the inshore waters around (20°~26°N, 118°~124°E); or those make strong

wind with a maximum velocity of 17.2 m/sec in any place of this island or bring forth heavy rain fall.

Because the observation data of the earlier years are incomplete, we take only the data from 1945 to 1984 for statistics. The statistics for the past 40 years (Table 1.) shows that a total of 135 ever invade Taiwan in this period, or around 3.4 per year on the average; the highest frequency appears in 1959, and the number is 7 Generally, the frequency is 2 to 5 per year, but in 1964 there was no typhoon over Taiwan.

From Table 1, we know that typhoons invade Taiwan most frequently in July, August and September each year. The 40-year statistics shows that a total of 98 ever hit this island in these three months, and this number is about 72% of all the typhoons that hit this island. If in terms of month, the highest frequency appears in September, the total number is 37; and the next high frequency appears in August, the number is 32; from December to the next March the Northeast monsoon is prevailing, no typhoon is seen in the vicinity of this island. As for April and May, there were only 5 tropical storms get to Taiwan in these two months. In other words, during the periods of winter and spring no typhoon or few typhoons are seen in Taiwan, while in mid-summer and early-autumn this island is most likely to be invaded by typhoon.

3. THE MAXIMUM WIND VELOCITY IN VARIOUS PLACES OF TAIWAN UNDER THE INFLUENCES OF TYPHOON OR TROPICAL STORM

In order to understand the status of maximum wind velocity in various places of Taiwan under the influences of typhoon or tropical storm, we divide, according to the terrain, this island into four parts, namely the north part, the central part, the south part and the east part (See Fig. 5), for preparing the statistics of April to November each year. As shown in Table 2, in the north part the strongest wind velocity is 50.7 G 66.0 m/sec (1962, 8, NORA); the second strongest wind is 41.7 m/sec (1948, 7); the weakest wind is only 16.0 m/sec (1961, 5, BETTY). In the central part, the strongest wind velocity is 41.2 G 54.1 m/sec (1956, 9, DINAH). In the south part, the strongest wind velocity is 43.0 G 55.8 m/sec, while the weakest wind also appears in May (14.5 m/s). As to the east part, the strongest wind velocity is 46.7 m/sec (1945, 9), and second strongest is 45.0 m/sec (1944, 8).

From the statistical data mentioned above, we can see when Taiwan is under the invasion or influence of a typhoon, due to the blocking effect of the Central Mountain Range, the wind velocity in the central part of this island is weaker than those in the other parts; while in the east part, because of its opening direct to the typhoons coming from the Northwest Pacific, generally the maximum wind velocity appears in this part in each month is higher than that of the other parts, a strong wind of 40.0 m/sec can even appear in May.

	MAY	JUN	JUL	AUG	SEP	OCT	NOV
N. Taiwan	SE/16.0	NW/18.3	WNW/41.7	ENE/50.7	NNW/36.0G52.5	NNE/24.2G36.0	ENE/19.0
C. Taiwan	N/21.6G32.9	NE/18.0	N/21.6G32.9	N/25.7G36.0	NNW/41.2G54.1	NNE/13.3	N/16.3
S. Taiwan	N/14.5	S/38.0	SE/41.2G51.5	SSW/43.0G55.8	S/39.2	N/36.2	E/34.5G38.6
E. Taiwan	NNW/19.2	NE/40.0	NNE/43.8G59.2	N/45.3G59.2	NNE/46.7	N/31.3	ENE/36.0

Table 2. Maximum Wind Velocity (m/s)

4. THE MAXIMUM DAILY PRECIPITATION IN TAIWAN UNDER TYPHOON'S INVASION OR INFLUENCE

The maximum precipitation in various places of Taiwan under typhoon's invasion or influence is shown in Fig. 3.

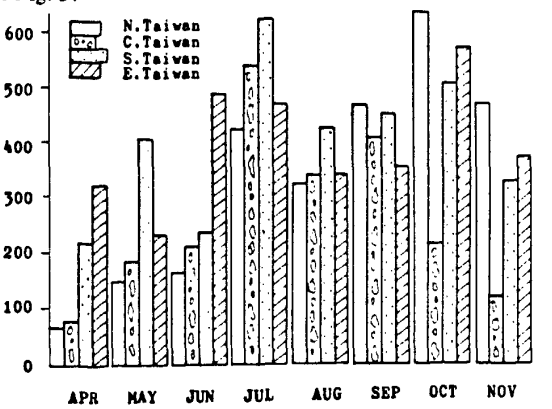


Fig. 3 Daily Maximum Precipitation (mm)

a. The maximum daily precipitation of Northern Taiwan under typhoon invasion

From the precipitation statistics, we can see that in the Northern Taiwan the 624.0mm maximum daily precipitation is the highest (1973, 10, NORA), and the second highest appears in September and October respectively, and the number is about 460mm.

b. The maximum daily precipitation in Central Taiwan under typhoon invasion

From Fig. 3, we can see that in Central Taiwan the maximum daily precipitation is 397.4 mm, and the second highest is 336.4 mm. But the above mentioned precipitation is only the highest maximum daily precipitation on the plain of the Central Taiwan. If in terms of mountain area, then the maximum daily precipitation should be most spectacular, as in Mountain Ali precipitation as high as 874.3 mm have been experienced (in 1963), this number is more than twice as the maximum daily precipitation 394.7 mm on the plain in the same part.

c. The maximum daily precipitation in Southern Taiwan under typhoon's invasion or influence

Under typhoon invasion or influence, the Southern Taiwan of the maximum daily precipitation 621.5 mm is the highest, precipitations over 400 mm are also seen in May, August, September and October respectively.

d. The maximum daily precipitation in Eastern Taiwan during typhoon invasion

In Eastern Taiwan, the maximum daily precipitation 484.8 mm appeared in June is the highest, precipitations over 400 mm are also seen in July and October.

e. Comparison between the maximum daily precipitation and the monthly mean precipitation under typhoon invasion

Generally, in Northern Taiwan winter is the rainy season, but the precipitation is not high; when spring is fading away and summer is coming on (this period is called "Mei-yu season"), heavy rain often occurs. In summer and fall, except some occasional local thunder-showers appear in the afternoon, most of the precipitation comes from the typhoon rain. Except April, the maximum daily precipitation brought by typhoon is higher than the monthly mean precipitation, especially in October the difference is very remarkable.

As for the Central, Southern and Eastern Taiwan, because every month's maximum daily precipitation is higher than the monthly mean precipitation during the typhoon season, so we can easily see that how tremendous the effect of typhoon rain fall is in a short time in various places of these three parts.

f. The maximum hourly precipitation in Taiwan during typhoon invasion

According to statistics and analysis, should the hourly precipitation of a place in Taiwan exceeds 15 mm, disaster could be expected. Fig. 4. clearly shows the maximum hourly precipitation appears in various parts of Taiwan in the past 40 years (1945~1984). From this figure we can understand that in each part of Taiwan the extreme maximum hourly precipitation of each month is all very high particularly in September in the northern part, in August in the central part, in July and August in the Southern part, and in April, June, September and November in the eastern part, the maximum hourly precipitations are all over 21mm; the maximum hourly precipitation in the southern part is 163.3 mm; on July 29, 1947 in the eastern part the most spectacular precipitation is 149.3 mm.

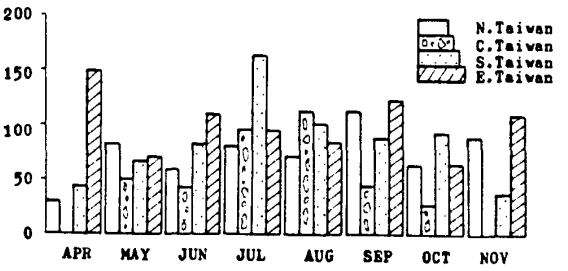


Fig. 4. Maximum hourly precipitation (mm)

Also, judging from the statistics for the maximum hourly precipitations of various parts, it seems that the maximum hourly precipitation tends to appear earlier and earlier by month from the north part, central part, south part to east part.

5. THE RELATIONS BETWEEN THE MOVING PATH OF THE TYPHOON WHICH INVADES TAIWAN AND THE MAXIMUM WIND AND RAINFALL OF VARIOUS PLACES

The wind and rain brought to a place by a typhoon not only have direct relations to the size of the typhoon's circulation structure, but also have close relation to the movement and speed of the typhoon center through this island. In addition, whether or not there is another weather system accompanying before or after the landfall do play an important role too in the matters of wind and rain. Shieh (1986), Tsai (1982), Wang (1983) and Yu (1982), had successively made statistics, analysis and study in the characteristics of the typhoon that affects Taiwan as well as the wind and rain caused during typhoon passage over various places of this island.

In general, the typhoons which invade Taiwan mainly come from the Pacific, their centers cross over the Central Mountain Range, or sweep over the inshore waters to the north, or pass through the Bashi Channel to the south of Taiwan. There are some typhoons heading northwestward in the Pacific sweep over the waters east of Taiwan (within 124°E). And there are also a few typhoons coming from the South China Sea, or typhoons moving westward through the Bashi Channel and entering into the South China Sea, then turning to the northeast and finally passing over Taiwan. Figure 5 shows the moving paths of the typhoons that affected Taiwan in the past 80 years (1905~1984).

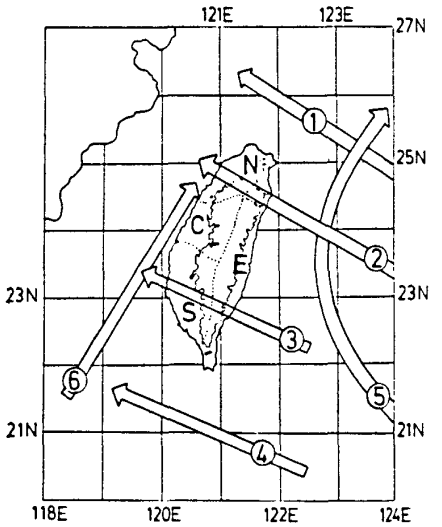


Fig. 5 A schematic diagram for the six typhoon track categories.

a. The statistics of the moving paths of the typhoons that strike Taiwan

Figure 5. shows that 12% of the typhoons on Track 1; 23% on Track 2; 22% on Track 3; 18% on Track 4; 14% on Track 5; and 4% on Track 6; the rest 7% of the typhoons are not shown in Fig. 5 and are with their unique tracks.

b. The relations between the maximum wind and rain in various places when a typhoon invaes Taiwan

Because the Central Mountain Range has a blocking effect on the typhoon's movement, therefore wind and rain of extremely different intensity appear in various places when this island is under typhoon attack. Usually, a typhoon coming in the east from the Pacific crosses over the Central Mountain Range or heads northward in the Taiwan Strait would cause considerably strong wind and heavy rainfall in the south and central parts, for examples, on August 8, 1959 when typhoon "Ellen" passed over the central part, the daily precipitation of Taichung was as high as 669.2 mm; and on August 11, 1974 when typhoon "Lucy" passed through the same part, the 24-hour precipitation of the Mountain Ali area was even up to 874.3 mm; while the daily mean precipitation of Central Taiwan was only 375.6 mm. When the center of a typhoon sweeps over the northern end of this island or the inshore waters north of the island (that is the commonly called "North-westerly Typhoon"), usually it would cause very strong wind and unusually heavy rain in Northern Taiwan, for example, on October 13, 1978 when typhoon "Ora" swept over the inshore waters of Keelung, it brought a maximum 24-hour precipitation of 472.3 mm to the Ilan area, while the yearly mean precipitation of Northern Taiwan is only 334.6 mm.

As to the maximum wind velocity (10-minute mean) in various places during typhoon passage, the strong wind ENE/50.7 G 66.0m/sec appeared in Ilan on August 5.1962 when typhoon "Opal" passing through the north part of this island is considered to be the highest. Whereas, in the off-shroe island Lanyu, a spectacular maximum wind velocity of up to 77.4 m/sec were experienced on May 26, 1961 under the influence of typhoon "Baddy".

6. TYPHOON DISASTERS IN TAIWAN

When a typhoon strikes a place, other than the strong wind velocity and the unfavourable wind direction would cause serious damage to the crops and buildings, the heavy rain brought forth in a short time or excessive precipitation accumulated for a long time would also cause flood disaster that has great adverse influences in many respects. In Taiwan, because of special terrain and short rivers, the local strong wind and heavy rain often make the disasters worse. Ac-

cording to Shieh (1986) and Yu (1982) analyses and statistics made on the disasters caused by typhoon in Taiwan, the degree of disaster is not in direct proportion to the size of typhoon, but is closely related to the typhoon's moving path, speed and other accompanying weather system. Table 3. shows the statistics of the loss caused by typhoon from 1961 to 1982.

	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	TOTAL
No. of invaded typhoon	1	2	4	18	23	18	9	3	78
No. of people killed/ missing		34	70	382	236	1,096	280	6	2,104
No. of people injured		224	224	705	2,231	3,170	355	57	6,966
Agricultural damage total loss (1000 NT$)	1,732	1,416,935	772,684	9,997,513	5,927,516	18,944,249	5,171,608	955,297	43,187,714
Houses completely destroyed		3,003	5,462	11,081	11,243	46,412	5,980	1,818	84,999
Houses halfly destroyed		2,937	6,174	36,217	19,324	64,680	5,846	2,203	137,381

Table 3. Nautral Disaster Caused by Severe Weather Associated with Tropical Storm (1961–1982) (After Shieh 1986)

a. Agricultural loss (in NT$):

In the past 22 years, typhoons ever caused Taiwan to incur a great loss of about NT$4,320 million, approximately NT$200 million each year. In terms of individual part's loss, the south part takes the lead, NT$ 1,820 million; next is the central part, NT$ 1,530 million; then the north part, NT$630 million; and the east part brings up the rear, NT$330 million. In terms of monthly loss, the highest is in September, NT$ 1,890 million; the second highest is in July, NT$ 1,000 million; and the lowest is in April, only about NT$ 1.7 million that is because typhoon seldom strikes Taiwan in this month, and/or the intensity of typhoon in this month is comparatively weak.

b. Casualties:

From 1961 to 1982 in Taiwan area, a total of 9,070 casualties occurred in typhoons, about 412 casualties per year on the average. The heaviest casualties occurred in September, totally 4,266; the second heaviest in August, totally 2,467; and the lightest in November, just only 63. From the statistics stated above, it is understood that the heaviest casualties occur in the month that typhoons hit this island most frequently.

c. Damage to hourses and buildings:

In the 22-year period, a total of 84,999 houses and buildings were completely destroyed, while 137,381 were half-destroyed.

In recent years, because of the in-time typhoon warning and the improvements in construction, there is a tendency that the casulaties and the house/building damage are lowering. However, the agricultural loss keeps increasing because of the increase of economical farming, while loss in other respects is also increasing with the increase of national income.

7. CONCLUSION

From the above stated, we learn that from 1905 to 1984 a total of 1860 typhoons occurred in the Northwest Pacific and the South China Sea, about 23 per year on the average. Since the early-year typhoon observation is incomplete, so we refer to the statistics for the years from 1945 to 1984. In this period a total of 1018 tropical storms occurred in the Northwest Pacific and the South China Sea and among them 679 ever developed into typhoons (maximum wind velocity of 32.7 m/sec near the center). In the past 40 years, a larger part (98) of the 135 typhoons which invaded Taiwan appeared in July to September that means on the average 3.4 typhoons hit this island each year, and among them about 2.5 that appeared in July to September. From December to the next March, no landfall typhoon to this island is experienced.

When there is a typhoon invading Taiwan, the blocking effect of the Central Mountain Range not only makes wind velocity reduce sharply, but also causes the precipitation distribution in various places very uneven. In general, there are strongest wind and heaviest rain

appearing in Eastern Taiwan before typhoon center hits that area; when Northern Taiwan is under the attack of a northwesterly typhoon the wind and rain appearing there will be most violent; when there is a typhoon appearing in the inshore waters of Eastern Taiwan or Northern Taiwan, or there is a typhoon in the Taiwan Strait heading northward, flood disaster will usually occur.

REFERENCES

- Chi, K.H., 1978: The 80 Years (1897–1976) Typhoon of Taiwan. Res. Report, CWB ROC (in Chinese).
- Chu, K.K. and Liu, K.Y., 1983: A Study of Spatial Distributions Of Heavy Rainfall Caused by Typhoons in the Western Taiwan, Paper, Meteor, Res., *6*, 101–116.
- _____, 1982: On the Distribution of Heavy Rainfall in Taiwan. Quart. J. Meteor., *92*, 30–55 (in Chinese).
- Hsu, C.H. et al., 1972: a Study of the Typhoon which Attack Taiwan. Meteor. Bull., *18*, 21–38 (in Chinese with English Abstract).
- Shich, S.L., 1986: A Survey of Meteorological Disaster over Taiwan Area. Atmos. Sci., *13*, 89–108 (in Chinese with English Abstract).
- Tsay, C.Y., and Chou, K.C., 1982: Typhoon Tracks and Distribution of Wind Speed and Precipitation in Taiwan. Proc. Conf. on Mesoscale Weather Systems, 209–221, CWB. May 6–8, Taipei Taiwan (in Chinese with English Abstract).
- Tseng, H.Y. and Tsay, C.D., 1982: A Case Study of Wind and Rain in Taiwan Area by Typhoon Andy. Proc. 3rd Conf. Atmos. Sci. 31–39, Natl. Sci. Counc. Dec. 1–3, Taipei Taiwan (in Chinese with English Abstract).
- Wang, S.T. and Wang, C.S., 1982: A Preliminary Study on the Mesoscale Heavy Rainfall Associated with Typhoon Circulation over Taiwan. Atmos. Sci., *10*, 99–114 (in Chinese with English Abstract).
- _____, et al., 1982: The Characteristics of Typhoon Precipitation in Taiwan. Proc. Conf. Mesoscale Weather Systerms 175–208, CWB. May 6–8, Taipei Taiwan (in Chinese with English Abstract).
- Wu, T.Y. and Wang, S.T., 1983: Circulation and Track Changes of Typhoons Encountering the Central Mountain Range in Taiwan: Field Observations and Laboratory Experiments. Proc. CCNAA AIT Joint Seminar on Tropical and Monsoon Meteorology., 117–128, Apr. 17–22, Taipei Taiwan.
- Yu, C.C., 1982: A Study on the Relationship Between Heavy Rain over Southwestern Taiwan and Typhoons Passing over Seas North of Taiwan. Atmos. Sci., *9*, 95–112 (in Chinese with English Abstract).
- _____, 1985: On the Strong Winds over Taiwan Area. Proc. The ROC-Japan Joint Seminar on Multiple Hazards Mitigation. 935–954, Mar. 20–25, Taipei Taiwan.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

ERROR MEASURES FOR EVALUATING THE REPRESENTATIVENESS OF DIVISIONAL CLIMATIC DATA

Anthony J. Brazel and Robert C. Balling, Jr.
Laboratory of Climatology and Department of Geography
Arizona State University
Tempe, Arizona 85287

INTRODUCTION

Many investigators (e.g., Diaz and Quayle, 1980; Karl and Koscielny, 1982; Diaz, 1983) are using divisional temperature and/or precipitation data in their empirical studies of climatic variations in the United States. These data from the 344 climate divisions presumably avoid the site specific inhomogeneities that can be associated with single station records. These researchers also assume that the areally-averaged divisional data are generally representative of the temporal variance in the raw station data. Given the wide utilization of the divisional data in the United States, surprisingly little work has been conducted to evaluate the general quality of these climatic data.

Accordingly, the primary purpose of this research was to introduce and operationalize a series of difference or error measures for effectively evaluating the agreement between individual station records and the derived divisional climatic data. The procedures described in this paper were developed from the methods recently introduced by Willmott (1981, 1982, 1984) for evaluating the performance levels of numerical climate models. Test statistics for each climate division (1) quantify the magnitude of temperature and/or precipitation differences or errors, (2) partition the errors into systematic and unsystematic components, (3) provide a standardized measure of the representativeness of the divisional data, and (4) identify individual stations contributing unusually large levels of error to their division.

METHOD

Many difference or error measures have appeared in the climate literature over the last decade. In the last few years, the root mean square error, RMSE, has become a more widely accepted statistic for describing differences between two arrays (Willmott, 1984). The RMSE is defined as:

$$RMSE = [N^{-1} \sum_{i=1}^{N} (P_i - O_i)^2]^{1/2} \quad (1)$$

where P_i is a predicted value, O_i is an observed value, and N is the number of cases. RMSE has the distinct advantage of allowing for further mathematical decomposition and statistical

interpretation. From the basic form of the RMSE, Willmott derived its systematic and unsystematic components, $RMSE_s$ and $RMSE_u$, respectively, and a dimensionless, bounded index of agreement, d.

The basic logic followed by Willmott can be extended to allow for an analysis of the representativeness of divisional climate data. By substituting station data, S, for the observed array, O, and the divisional data, D, for the predicted array, P, the RMSE can be extended for a climate division to take the form:

$$DRMSE = M^{-1}\sum_{j=1}^{M}[N^{-1}\sum_{i=1}^{N}(D_i - S_{ij})^2]^{1/2} \quad (2)$$

where DRMSE is the divisional root mean square error, M is the number of stations in a division, and N is the number of years of record.

The unsystematic divisional root mean square error, $DRMSE_u$, can be developed as:

$$DRMSE_u = M^{-1}\sum_{j=1}^{M}[N^{-1}\sum_{i=1}^{N}(D_i - \hat{D}_{ij})^2]^{1/2} \quad (3)$$

with $\hat{D}_{ij}$ equal to $a_j + b_j S_{ij}$ where a_j and b_j are parameters estimated from ordinary least-squares simple linear regression between the j^{th} station and the divisional data. The systematic divisional root mean square error, $DRMSE_s$, may be found through the relationship:

$$DRMSE^2 = DRMSE_u^2 + DRMSE_s^2 \quad (4)$$

In general, systematic errors in divisional data would be associated with a high correlation with the individual stations; the differences would arise from distance between the means (one time series is consistently higher or lower than the other) or differences in the variance levels causing the slope, b_j, to vary from 1.0. Unsystematic errors would stem from a low correlation between the divisional and station records.

An index of agreement, d, can be derived from Willmott's work as:

$$d = 1 - \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}(D_i - S_{ij})^2}{\sum_{j=1}^{M}\sum_{i=1}^{N}(|D_i - \bar{S}_j| + |S_{ij} - \bar{S}_j|)^2} \quad (5)$$

where $\bar{S}_j$ is the mean of the station data and is assumed to be error free. The term in the denominator represents a potential

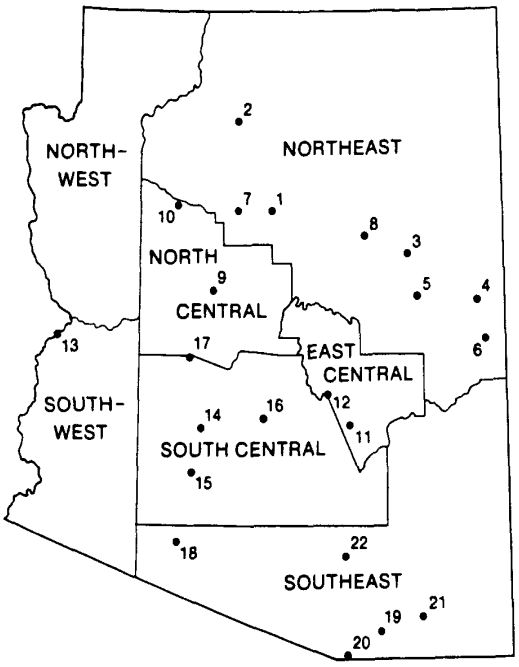


Figure 1. Bradley Station Network and Climate Divisions in Arizona

Table 1

Bradley (1982) Stations by Arizona Climate Division

Northwest Division
(no stations)
Northeast Division
1. Fort Valley
2. Grand Canyon
3. Holbrook
4. St. Johns
5. Snowflake
6. Springerville
7. Williams
8. Winslow
North Central Division
9. Prescott
10. Seligman
East Central Division
11. Miami
12. Roosevelt
Southwest Division
13. Parker
South Central Division
14. Buckeye
15. Gila Bend
16. Mesa Exp. Farm
17. Wickenburg
Southeast Division
18. Ajo
19. Canelo
20. Nogales
21. Tombstone
22. Univ. of Arizona

error variance that can exist between the divisional and station data. The resultant statistic d is dimensionless and bounded between 0 and 1.0; the ideal climate division should have a value of d near the 1.0 level of perfect agreement.

APPLICATION

The statistics described in this paper were initially developed to test the quality of the seven climate divisions in the state of Arizona. The complex topography of the state combined with the great diversity of climatic types produces a difficult part of the United States for constructing homogeneous climate divisions.

In this study, the annual divisional temperature (^{o}F) records from 1931 to 1982 were compared to the annual temperature data for stations selected by Bradley (1982) for their high quality long-term records. The 22 stations in Arizona (Table 1, Figure 1) met Bradley's criteria of (a) no more than five percent missing data, (b) fewer than an average of one station move per 15 years, (c) changes in location of less than 1.6 km and (d) little change in elevation. The relatively high quality records from these stations provided an exceptionally good base for testing the climatic divisions of Arizona.

The DRMSE, $DRMSE_s$, $DRMSE_s$, and d statistics (Table 2) revealed substantial variations across the five divisions containing more than one of the Bradley stations. The four stations in the South Central division produced the lowest DRMSE and the highest index of agreement. This South Central division encompasses the large, relatively flat portion of the Salt, Verde, and Gila River basins. Oppositely, the Southeast climate division produced the lowest index of agreement (0.243) and the highest unsystematic root mean square error. The Southeast division represents a large area of diverse topography; relief in the division varies from over 3200 m to less than 450 m. In general, the index of agreement appeared to vary directly with the homogeneity of the landscape and indirectly with the size of the division.

Table 2

Selected Statistics by Division

Division	Stations	DRMSE	$DRMSE_s$	$DRMSE_u$	d
Northeast	8	3.335	3.250	0.745	0.320
North Central	2	5.249	5.223	0.527	0.317
East Central	2	4.562	4.459	0.965	0.296
South Central	4	2.386	2.303	0.626	0.491
Southeast	5	4.761	4.659	0.980	0.243

The individual station RMSE, $RMSE_u$, $RMSE_s$, d, and Pearson-Product Moment Correlation, r, may be used to effectively identify the sources of error within each division. In the Southeast division for example, Nogales and Tombstone produced the lowest RMSE and the highest d and r values (Table 3). These two stations appeared to be particularly well represented by the annual temperature pattern for the Southeast division. Ajo's far northwesterly position in this large division appeared to be responsible for its large RMSE and low d and r values. The station at the University of Arizona in Tucson also produced a high RMSE and a low value of d; the negative correlation (but not significant at the 0.05 confidence level) between the station and divisional annual temperatures may be due to the strong urban effect at the Tucson location (Cayan and Douglas, 1984). However, the annual temperature data in the Southeast division did not appear to adequately capture the variance displayed at several other stations selected for outstanding data quality.

Table 3

Statistics for Bradley Stations within the Southeast Division

Station	Mean	RMSE	$RMSE_s$	$RMSE_u$	d	r
Ajo	71.52	8.538	8.475	1.034	0.162	0.304
Canelo	57.19	6.012	5.924	1.028	0.225	0.321
Nogales	61.57	2.163	1.978	0.877	0.642	0.589
Tombstone	63.53	1.050	0.537	0.903	0.699	0.555
Univ. Az.	68.78	6.042	5.949	1.058	0.264	-0.221

CONCLUSIONS

A set of statistics were developed in this study for evaluating the representativeness of divisional climatic data in Arizona. Over 50 years of divisional annual temperatures were compared to observations from a network of stations that had been previously selected for outstanding data quality. The statistics generated in this study (1) quantified the mean errors within each division, (2) partitioned the errors into systematic and unsystematic components, (3) provided a general measure of the representativeness of the divisional temperature data, and (4) identified stations that contributed unusually large levels of error to the division. The techniques introduced in this study appear to hold considerable promise in the design and evaluation of climatological data collection networks.

REFERENCES

Bradley, R.S., 1982: Climatic fluctuations of the western United States during the period of instrumental records. Contribution No. 42, Department of Geology and Geography, University of Massachusetts.

Cayan, D.R., and A.V. Douglas, 1984: Urban influences on surface temperatures in the southwestern United States during recent decades. J. Clim. Appl. Meteor., 23, 1520-1530.

Diaz, H.F., 1983: Some aspects of major dry and wet periods in the contiguous United States, 1895-1981. J. Clim. Appl. Meteor., 22, 3-16.

Diaz, H.F. and R.G. Quayle, 1980: The climate of the United States since 1895: spatial and temporal changes. Mon. Wea. Rev., 108, 249-266.

Karl, T.R. and A.J. Koscielny, 1982: Drought in the United States: 1895-1981. J. Clim., 2, 313-329.

Willmott, C.J., 1981: On the validation of models. Phys. Geogr., 2, 184-194.

Willmott, C.J., 1982: Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc., 63, 1309-1313.

Willmott, C.J., 1984: On the evaluation of model performance in physical geography. Spatial statistics and models, G.L. Gaile and C.J. Willmott, eds. Reidel Publishing, Dordrecht, Holland, 443-460.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

MESOSCALE CLIMATOLOGICAL STUDIES OF THE BLACK HILLS - AN ANALYSIS OF THE SPATIAL REPRESENTATION OF TEMPERATURES

André A. Doneaud and James R. Miller, Jr.

Institute of Atmospheric Sciences
South Dakota School of Mines and Technology
501 East St. Joseph Street
Rapid City, South Dakota 57701-3995
Tel: (605) 394-2291

1. INTRODUCTION

Rising from out the monotony of the Great Plains, stretching away westward from the Missouri, the Black Hills is an outlying group of mountains located near the southwest border of the State of South Dakota; these mountains are higher than the Alleghenies, steeper than the Rockies, older than the Sierras, and in wealth of natural element, mineral resource, and rare metal, the "richest hundred miles square on Earth."

Lying principally within parallels 43 to 45 degrees north latitude, and meridians 103 to 104 degrees, about 200 km north-northwesterly and southerly by about 105 km in width, are the Black Hills. The major mountain ridge is oriented north-south. The highest peak rises to 2207 m from the 1300 m plateau in Wyoming (to the west) and the 900 m Great Plains (to the east).

Conventional climatological analyses do not take sufficient advantage of the variations with elevation of meteorological parameters (i.e., temperature) in the preparation of maps (Pielke and Meering, 1977; Climates of the States, 1974, Vols. 1 and 2). As such, equal temperature lines are crossing mountain ridges neglecting the natural stratification of animal and vegetable species with elevation. The works by Pielke and Meering (1977) and by Alemán and García (1974) are two examples known to the authors where terrain effects have been explicitly considered in the preparation of climatological maps. Information presented in this manner will have useful applications in such areas as ecology, forestry, and water management, as well as in weather forecasting.

As the major mountain ridge is oriented north-south, significant temperature and lapse rate differences will occur between the western and the eastern side of the mountains in relation to the general circulation.

In this paper, we utilize and extend the methodology, as discussed by Alemán and García (1974), by Conrad and Pollak (1962), and by Pielke and Meering (1977) to describe monthly and seasonal mean temperatures over the Black Hills using standard climatological data and terrain contours.

2. DATA

Daily maximum and minimum temperatures from 28 meteorological stations located around the Black Hills were used. Thirty years (1950-1980) of data were considered. Daily averages were computed using the maximum and the minimum values. The difference between the highest (1895.9 m) and the lowest (707.1 m) meteorological stations was 1188.8 m.

3. ANALYSIS

Monthly and seasonal averages were computed from the daily mean temperatures. Thirteen stations are located west and 15 stations are located east of the mountain divide (Table 1). In processing the data, difficulties related to missing data and changes in station locations were removed using known statistical techniques. First, linear regressions were applied to each month's and season's data in the form of:

$$T(Z) = B + MZ \qquad (1)$$

The results are displayed in Fig. 1, where R is the correlation coefficient; R^2, the coefficient of determination; and S, the standard error of estimates. The vertical lines are confidence limits for the 30-year average values for a given station. The vertical lines are drawn only in case the vertical extension of the line does not exceed the frame of the figure. It is evident from Fig. 1 that R is low in winter only. The highest R is found in spring (R = -.87); then, respectively, in summer (R = -.86); and in fall (R = -.80). It is hypothesized that the strong temperature versus elevation correlation is alterated in winter by several factors like the low level topospheric inversions, the snow, and the subsidence effects, as well as the effects of

TABLE 1

Tabulation of the Meteorological Stations as a Function of Elevation

LOCATION	ELEVATION (meters)	SIDE
WASTA	707.1	E
SCENIC	855.9	E
NEWELL	870.2	E
ORAL	902.2	E
BELLE FOURCHE	919.6	E
ANGOSTURA LAKE	957.1	E
RAPID CITY WSO	968.8	E
FORT MEADE	1005.8	E
EDGEMONT	1054.6	W
HOT SPRINGS	1077.5	E
ARDMORE	1082.0	
COLONY (WYO)	1088.1	
SPEARFISH	1109.5	E
DEVILS TOWER 2	1177.1	W
REDBIRD (WYO)	1185.7	W
MORRISEY (WYO)	1249.7	W
NEWCASTLE (WYO)	1315.2	W
DEVILS TOWER	1318.9	W
ALVA 5 SE (WYO)	1338.1	
DEADWOOD	1423.4	E
PACTOLA DAM	1438.7	E
SUNDANCE (WYO)	1447.8	W
UPTON 13 SW (WYO)	1456.9	W
MOUNT RUSHMORE	1577.3	E
LEAD 1 E	1625.2	E
CUSTER	1655.1	W
DEERFIELD DAM	1789.2	W
DEERFIELD DAM	1895.9	W

latent heat release and upslope. The upslope effect is more frequently encountered during winter when the jet stream forms short waves traveling far south of South Dakota generating easterly winds over the Black Hills. Two monthly linear regressions are presented in Fig. 2. Again, a better fit is found during springtime (R = -.90 in April); in July, R = -.84.

The scatter of the data (the 28 station data) either in Fig. 1 or in Fig. 2 suggests that a quadratic form may be more appropriate to relate temperature with elevation in the Black Hills. As such, quadratic regressions in the form of:

$$T(Z) = A + BZ + CZ^2 \qquad (2)$$

were applied to each month's and season's data.

Figure 3 displays the results for the spring, summer, and fall periods. The nonlinear regression shows a better fit of the scatter of the data than the linear fit. The coefficients of determination show only slight improvement. Up to a height of 1.2 km, the temperature decreases slowly with the elevation, while above 1.2 km, the temperature decrease with the elevation is sharper. This is due to the fact that over the High Plains (below 1.2 km), as a result of the semi-arid character of the climate, the maximum temperature (and, as such, the daily average temperature) is high during the spring, the summer, and the fall, keeping the lapse rate* low for a 500 m increase in elevation.

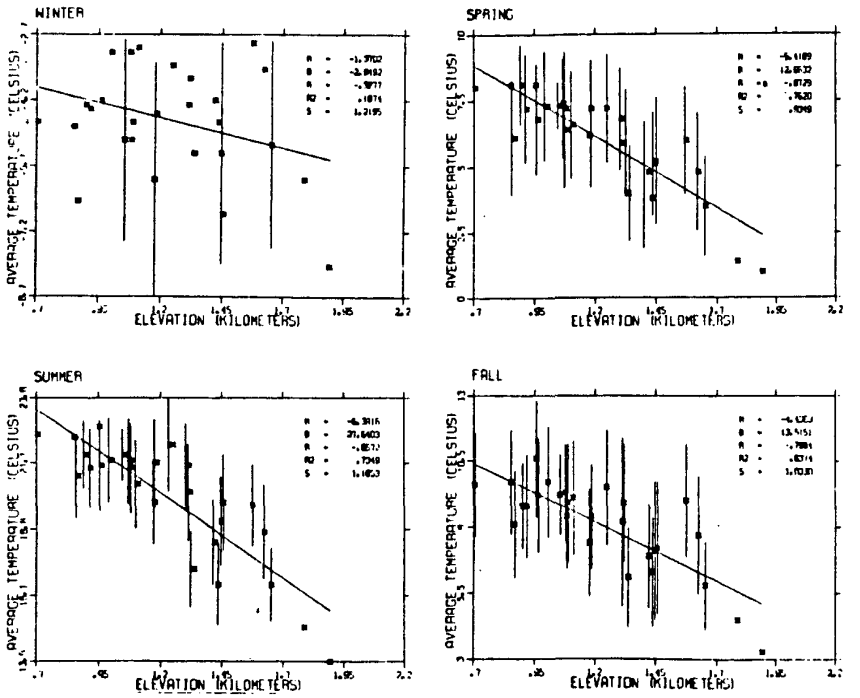


Fig. 1: Linear regression and scatter of the data for the seasons of winter, spring, summer, and fall. The legend on the top right side of each figure is explained in the text.

At least three outliers could be visually identified in all three graphs in Fig. 3. They are labeled 1, 2 and 3, and represent the stations ALVA 5 SE (WYOMING), PACTOLA DAM (SOUTH DAKOTA) and MOUNT RUSHMORE (SOUTH DAKOTA), respectively. The first two stations lie far below the temperature regression line, while the third station lies above the regression line. The outliers are closer to the regression curve in April when the influence of the topography and the vegetation is reduced.

Figure 4 displays similar quadratic temperature versus elevation regression curves computed on a monthly

*The lapse rate is defined here as the temperature variation per km elevation ($C = \Delta T/\Delta h$ °C km^{-1}).

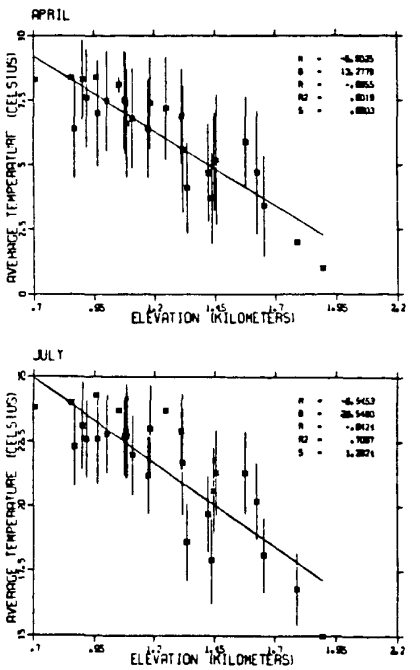


Fig. 2: Linear regression and scatter of the data for the months of April and July.

basis. The same aforementioned three stations are outliers. Table 2 summarizes some of the results, given on a monthly basis. The coefficients of determination are always bigger (for every month of the year) when the second degree fit is considered. From Table 2 as well as from Fig. 4, it is evident that the month of April has the highest coefficient of determination. The months of April, May, and June have R^2 bigger than .81 emphasizing a strong temperature versus elevation quadratic correlation during the spring and the first part of the summer. The quadratic correlation is still high during the last part of the summer (>.73), including the month of September. Approaching the winter months, the correlation becomes weaker and is the weakest during January (the coldest month). That is shown by the stable lapse rate in winter and is consistent with frequent temperature inversions.

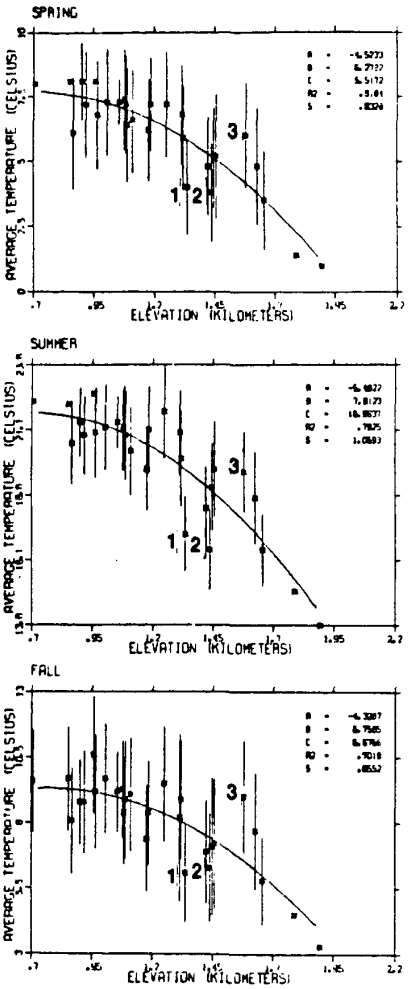


Fig. 3: Quadratic regressions and scatter of the data for the spring, the summer, and the fall seasons. The stations labeled (1), (2), (3) are the outliers discussed in the text. The legend on the top right side of the figures are explained in the text.

The monthly lapse rate increases from January (-.55) to July (-6.55) and then decreases to December (-1.90). The magnitude of the lapse rate found in this region is close to that generally quoted in climatological references, except for the winter months with much smaller magnitudes (Table 2). Baldwin (1973) gives a value of -6.01°C km^{-1}, while Blumenstock and Price (1967) in their paper on the climate of the island of Hawaii reported a decrease in mean monthly temperature upslope of +5.46°C km^{-1}, although they added that actual monthly means vary from -1.85°C km^{-1} to -10.93°C km^{-1} depending on local differences in cloudiness and exposure to the trade winds. No such big variations in lapse rates were found by Pielke and Meering (1977) in the Appalachians, with a maximum of -8.23°C km^{-1} in March 1960 and a minimum of -3.43°C km^{-1} in February 1962. Snow (1975) found lapse rates in the Columbia Andes ranging from -4.5°C km^{-1} at elevations <1 km to values of -7.0°C km^{-1} from 4-5 km.

The orientation of the mountain ridge from north to south, perpendicular to the westerly general circulation, suggested an analysis of the data by considering separately the western and eastern sides of the Hills (Table 1). As a consequence of the effects of the general circulation and indeed of the mountain topography and vegetation, the western side of the Hills shows a better goodness of fit than the eastern side. The .96 and .94 coefficients of determination found for spring and summer, respectively, represent an excellent goodness of fit for the western side of the Hills

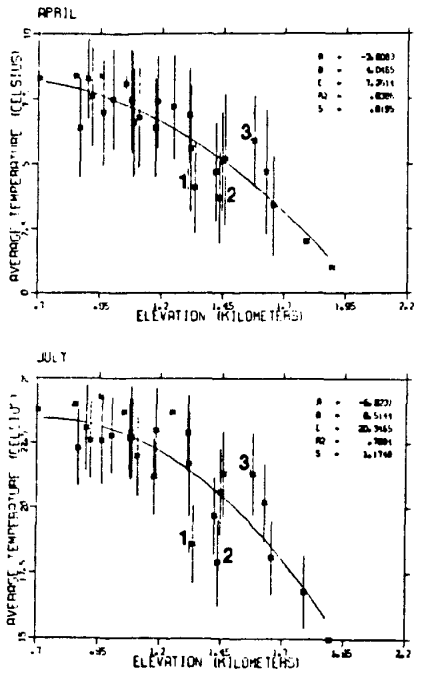


Fig. 4: Quadratic regression and scatter of the data for the months of April and July.

TABLE 2

Monthly Coefficients of Determination (R^2) for the Linear (A) and the Second Degree (B) Fits and the Monthly Lapse Rates (C) in °C km^{-1}

I			II			III			IV		
A	B	C	A	B	C	A	B	C	A	B	C
.02	.14	- .55	.13	.25	-1.66	.62	.77	-4.47	.80	.84	-5.80
V			VI			VII			VIII		
A	B	C	A	B	C	A	B	C	A	B	C
.79	.83	-5.99	.77	.81	-6.38	.71	.77	-6.55	.70	.77	-6.25
IX			X			XI			XII		
A	B	C	A	B	C	A	B	C	A	B	C
.68	.73	-5.69	.62	.68	-4.29	.50	.57	-3.33	.17	.30	-1.90

(Fig. 5). The scatter of the data is larger and the coefficient of determination is smaller for the eastern side of the Hills for both seasons ($R^2 \approx .80$). Three stations (Ardmore, Colony, and Alva) were eliminated from the analysis because of the difficulties to locate the station on the west or the east side of the Hills (Table 1). Figure 6 displays the regression curves for April and July for the western and the eastern sides of the Hills. The western side of the Hills is a little warmer for elevations up to 1.4 km, but the temperature difference is not significant. For higher elevations (>1.45 km), the eastern side of the Hills is much warmer than the western side. It is hypothesized that that is the result of a downslope warming and chinook (föhn) effects.

The temperature versus elevation data, the standard error of estimate, and the index of uncertainty in elevation Te (m) for the months of January, April, July, and October are presented in Table 3.

The index of uncertainty in elevation as defined here is the ratio between the standard error of the temperature estimates (°C) and the average climatological lapse rate (°C km^{-1}) for a given month in meters,

$$Te(m) = °C/°C\ km^{-1} \times 1{,}000$$

The goodness of fit of the quadratic regression, as described by the magnitude of the standard error of estimate, determines the resolution with which we can apply this climatological data to describe the mean temperature distribution over a limited geographic mountainous area. More confidence in the mean temperature for a given elevation is possible in spring and summer for both sides of the Hills and in the fall for the west side of the Hills, because the standard error is less and the average lapse rate is greater (Table 3). During the winter, such an estimate of the mean temperature cannot be made according to the values of the index of uncertainty and of the coefficient of determination (Table 2).

4. FINAL REMARKS

A method is presented for improving the analysis of monthly mean temperatures in mountainous terrain for a limited geographic area. The technique is based on plotting mean monthly temperatures as a function of elevation, fitting the data points with quadratic regressions, and plotting the estimated mean temperatures on a topographic map in place of elevation.

The highest correlations were found in the spring and the lowest in the winter. The quadratic regressions always have, for every month of the year, smaller standard error of estimates and larger coefficients of determination than the linear regressions.

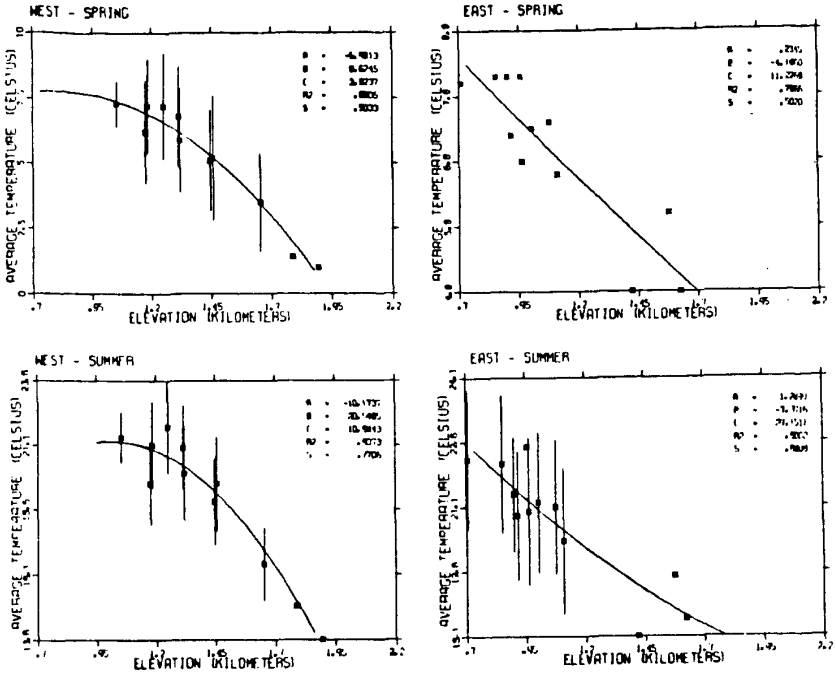


Fig. 5: Quadratic regressions and scatter of the data for the west and east sides of the Hills for the spring and the summer seasons.

By dividing the stations in western and eastern stations with respect to the mountain divide (ridge), it was found, as a consequence of the effects of the general circulation and of the mountain topography and vegetation, that the western side of the Hills shows a better goodness of fit of the temperature data than the eastern side. For elevations >1.45 km, the eastern side of the Hills is much warmer than the western side as a result of downslope and chinook warming effects in a much drier environment.

According to the goodness of fit of the quadratic regressions, as well as the standard error of estimates, more confidence in the mean temperature for a given elevation is possible in spring and summer for both sides of the Hills, and in the fall for the west side of the Hills. During the winter, such an estimate of the mean temperature cannot be made.

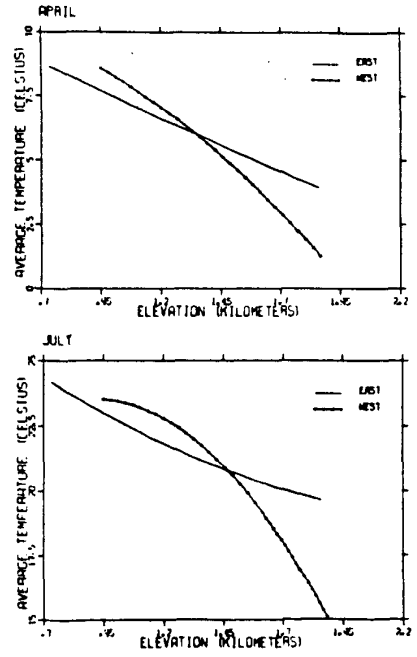


Fig. 6: Quadratic regressions for the spring and the summer seasons for the east and west sides of the Hills.

TABLE 3

The Temperature Versus Elevation Data, the Standard Error of Estimate, and the Index of Uncertainty in Elevation Te(m) for the Black Hills for the months of January, April, July and October

Elev.	January		April		July		October	
(km)	East	West	East	West	East	West	East	West
.700	-7.38	- 9.06	8.79	9.72	24.36	23.08	10.40	8.62
.950	-5.82	- 7.43	7.69	8.57	23.00	23.51	9.99	9.37
1.200	-5.01	- 6.62	6.62	7.06	21.83	22.81	9.51	9.21
1.450	-4.96	- 6.63	5.58	5.19	20.85	20.97	8.95	8.15
1.700	-5.66	- 7.45	4.56	2.95	20.05	18.00	8.33	6.19
1.950	-7.11	- 9.10	3.57	.36	19.44	13.89	7.62	3.32
2.200	-9.32	-11.56	2.61	-2.60	19.02	8.64	6.85	- .45
St. Er.	.88	1.38	.61	.48	.77	.86	.81	.71
Te(m)	1,600	2,509	105	83	118	131	211	83

The approach of displaying weather data as a function of elevation with high spatial resolution needs more study. At least the local effects of topographic mountain and valley features, accompanying discontinuities in vegetation, mountain lakes, and populated areas have to be taken into account.

Acknowledgments. Support for this research was provided by the National Science Foundation under Grant No. ATM-8219473. Mr. Richard Tax, a work study student at the SDSM&T, is thanked for helping in processing the data. Thanks are also given to Mrs. Sandra Palmer and Mrs. Joie Robinson for work on the manuscript.

REFERENCES

Alemán, P. A. M., and E. García, 1974: The climate of Mexico. World Survey of Climatology, Vol. 11, R. A. Bryson & F. K. Hare Eds., Elsevier. 420 pp.

Baldwin, J. L., 1973: Climates of the States, NOAA publication. 113 pp. [Available from the Superintendent of Documents, U.S. Govt. Printing Office, Washington, DC]

Blumenstock, D. I., and S. Price, 1967: The climate of Hawaii. Climates of the States, Vol. 2, 614-639. [Available from Water Information Center Inc., Water Research Building, Manhasset Isle, Port Washington, NY, ISBN 0-912394-09-9].

Climates of the States, Vol. 1, Eastern States (plus Puerto Rico and the U. S. Virgin Islands), 1974. Water Information Center, Inc., Water Research Building, Manhasset Isle, Port Washington, NY 10050, ISBN: 0-912394-09-9. 486 pp.

Climates of the States, Vol. 2, Western States, 1974: Water Information Center, Inc., Water Research Building, Manhasset Isle, Port Washington, NY 11050. ISBN: 0-912394-09-9.

Conrad, V., and L. W. Pollak, 1962: Methods in Climatology. Harvard University Press. 610 pp.

Pielke, R. A., and P. Meering, 1977: Use of mesoscale climatology in mountainous terrain to improve the spatial representation of mean monthly temperatures. Mon. Wea. Rev., 105, 1, 108-112.

Snow, J. W., 1975: The climates of northern South America. M.S. thesis, University of Wisconsin, Madison. 238 pp.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE SYNOPTIC CLIMATOLOGY OF MONTHLY PRECIPITATION AMOUNTS OVER THE UNITED STATES DURING WINTER IN RELATION TO THE SURROUNDING 700 MB HEIGHT FIELD

William H. Klein and Hal J. Bloom
Department of Meteorology
University of Maryland
College Park, MD 20742

1. INTRODUCTION

This paper extends research by Klein (1963,1965) on the synoptic climatology of precipitation over the contiguous United States (U.S.) in relation to the 700 mb circulation from 5-day mean data to the monthly time scale. Improvements are made by using a finer grid of height, a larger sample of data, an updated period of record, composite maps as well as correlation fields and a more reliable set of precipitation data. Statistical refinements include expressing amounts in terms of percentiles estimated from a gamma distribution, assessing significance of correlation coefficients by means of a Monte Carlo technique and establishing linearity in the relations between precipitation and circulation. Section 2 discusses the data and procedure used, section 3 the correlation fields, section 4 the composite maps and section 5 the linearity assumption.

2. DATA AND PROCEDURE

Because of the small scale and discontinuous nature of precipitation, it was averaged within 60 quasi-homogeneous climate divisions (CDs) developed by Englehart and Douglas (1985), each composed of approximately 15 stations with a mean area of 24,000 km^2 (Fig. 1). Since precipitation

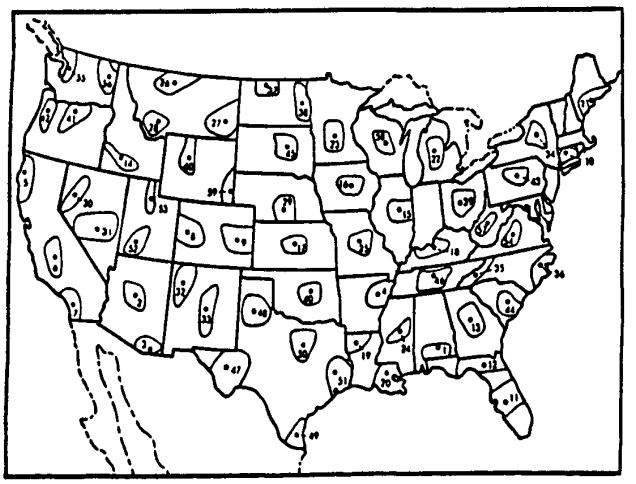


Fig. 1. Locations of the 60 climate divisions defined by Englehart and Douglas (1985) for which monthly precipitation amounts were studied. The divisions are numbered alphabetically by state. Their centers of gravity are given by the heavy dots.

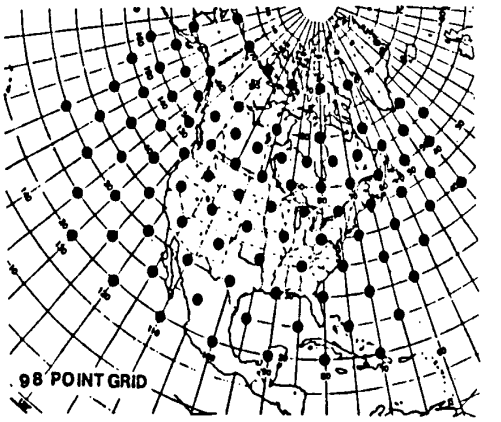


Fig. 2. Network of 98 grid points used to delineate the field of 700 mb height. The grid points are located at intersections of latitude and longitude ending in 0 (even) or 5 (odd).

amounts are positively skewed, the monthly totals were fitted to a gamma distribution (Thom, 1968) and classifed into 21 percentiles (PP) (Ropelewski and Jalickee, 1983). Use of PP gave results as good as those yielded by any other transformation of precipitation amount in specifying precipitation from the 700 mb height field (Klein and Bloom, 1986).

Monthly mean 700 mb height anomalies (MMHA) were computed at 98 standard intersections of latitude and longitude ending in zero (even) and five (odd) over North America and surrounding oceans (Fig. 2). The period of record for both heights and precipitation extended from January 1951 to December 1980, giving 30 years of monthly data. In order to reduce random fluctuations in a small sample, we pooled data for the months of Dec., Jan. and Feb., so that a total of 90 cases was available for the winter season. We then calculated simple linear correlation coefficients (r) between values of PP at each CD of Fig. 1 and MMHA at each grid point of Fig. 2.

Because we selected r maxima a posteriori, we applied a Monte Carlo technique, similar to that used by Lund (1970) and Walsh (1984), to find the true 95% significance level. We randomly shuffled the temporal order of our PP data, but kept the MMHA data unchanged in order to retain its persistence and intercorrelations. For 100 such trials, we extracted the extreme positive and negative r, ordered the values from lowest to highest, computed the cummulative frequency and selected that r exceeded only 5% of the time, as illustrated in Fig. 3 for the maximum positive r in North Dakota. In this example, and for 9 other CDs, an r exceeding .32 is required to be statistically significant at the 95% confidence level. This is considerably larger than the value of .20 given by Student's t-test for 90 independent cases.

3. CORRELATION FIELDS

In order to highlight regional differences in the synoptic climatology of PP, this section summarizes the r fields at all 60 CDs. Fig. 4 gives

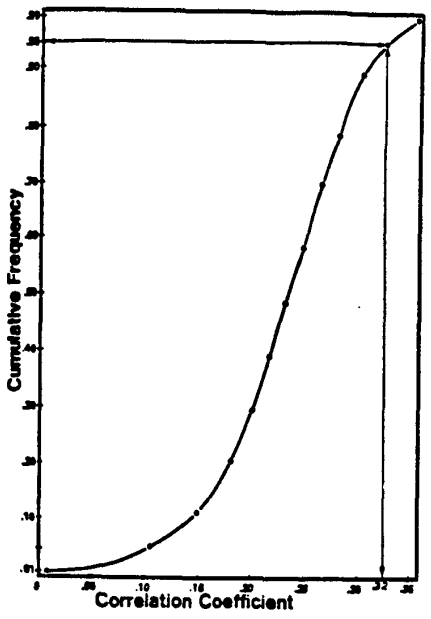


Fig. 3. The cumulative frequency distribution of correlation coefficients for North Dakota obtained with randomly shuffled precipitation data. The line at 0.95 shows that correlations must exceed 0.32 in this example to reject the null hypothesis with 95 percent confidence.

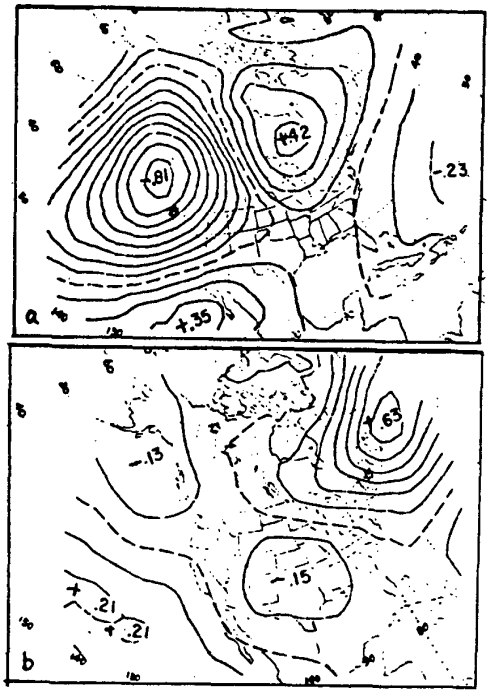


Fig. 4. Correlation fields between concurrent winter anomalies of 21 percentiles estimated from a gamma distribution at climate divisions for (a) northern California, (b) Maine (located by stars) and 700 mb height on grid of 98 points. The isopleths are drawn at intervals of 0.10 with the zero line dashed and central values labeled.

the correlations between PP and MMHA for two CDs at opposite ends of the U.S. The map for northern California (Fig. 4a) is dominated by a negative center of -.81, the highest $\underline{r}$ of either sign found for any CD. It indicates that below normal heights off the coast of Oregon are closely associated with wet weather in northern California, while above normal heights accompany dry weather. Fig. 4b gives the field of $\underline{r}$ for Maine, where the only significant center is a positive value of .63 about 800 km to the northeast. Above normal heights here are responsible for Atlantic maritime air and wet weather in Maine, while negative MMHA favors dry continental air.

Fig. 5 maps the maximum negative $\underline{r}$ for all CDs. Highest values (-.8) are found in western California, where wet weather is associated with orographic lifting of moist air from the Pacific. Weaker centers are found in the deep South, northern Plateau, Ohio Valley and Central Plains. Minima occur mainly east of the Continental Divide and in the Northeast, where some of the CDs (shaded) have their highest $\underline{r}$ below our critical level of ±.32. A similar analysis for the maximum positive $\underline{r}$ is shown in Fig. 6. Values range from .63 in Maine to .12 in North Dakota. They exceed .6 in the Northeast, Great Lakes and Tennessee, but are less than .32 (shaded) in the Gulf and South Atlantic Coasts, Great Plains and far Southwest. When the magnitudes in Fig. 5 are subtracted from those in Fig. 6, the pattern of Fig. 7 emerges. In most of the West, Great Plains and southeast coastal regions, negative differences indicate that the negative $\underline{r}$ center has more influence on precipitation than the positive center. Thirty percent of the CDs have positive differences (shaded). Here, PP depends more on the positive than the negative $\underline{r}$.

Fig. 8 shows schematically the joint location of positive and negative $\underline{r}$ centers associated with above normal PP in each of five geographical regions of the U.S. (Fig. 8f). It was obtained by assuming the CD to be at the origin of a polar coordinate diagram, plotting the distance and direction separately for positive and negative centers and analyzing in terms of homogeneous clusters. Wet weather in the eastern half of the U.S. is associated with positive MMHA 600-2800 km northeast of the reference CD and negative anomalies 1200-2900 km to the west. In both the Pacific Northwest and Southwest, heavy precipitation is accompanied by negative MMHA 800-1600 km northwest of the CD; the optimum locations for above normal heights are 2000-3000 km south of CDs in the Northwest and 2600-3500 northeast of CDs in the Southwest. In the Great Plains, wet weather tends to occur with negative MMHA 800-1600 km southwest of the CD and above normal heights 900-2400 km to the north-northeast. In a small region labeled 'Montana', high PP is associated with a negative center 200-1100 km to the southeast and a positive center 3000-4400 km to the north-northwest. In all areas except the East, the negative correlation center

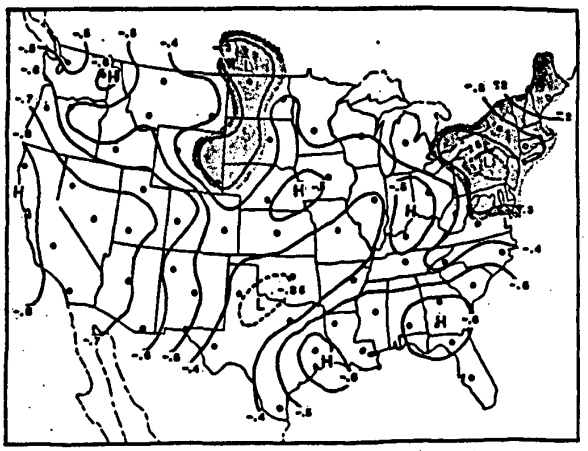


Fig. 5. Maximum negative correlation between concurrent winter anomalies of 21 precipitation percentiles and 700 mb height, as analyzed on each of 60 correlation fields. The shaded areas represent regions where the maximum negative correlation falls below the critical cut-off criterion of >.32.

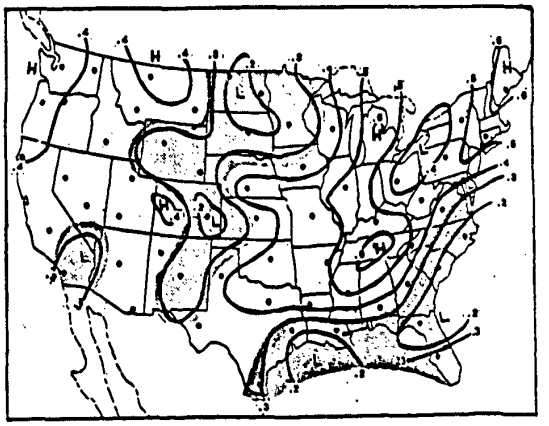


Fig. 6. Maximum positive correlation between concurrent winter anomalies of 21 percentiles and 700 mb height. The shaded regions represent areas where values fall below the critical cut-off criterion of >.32.

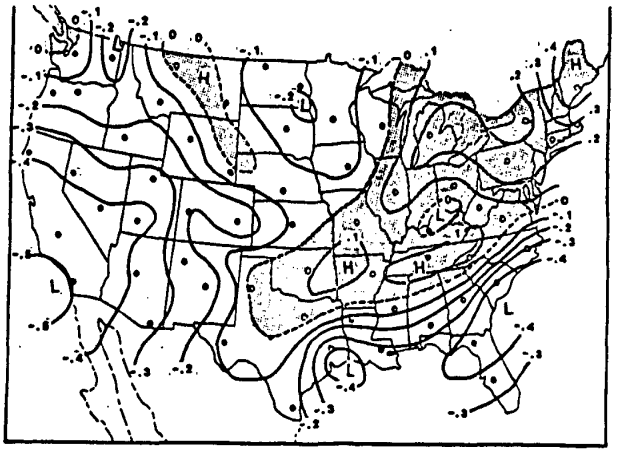


Fig. 7. Magnitude of the maximum positive correlation (Fig. 6) minus magnitude of the maximum negative correlation (Fig. 5). Shading delineates areas where the positive exceeds the negative correlation.

is closer to the reference CD than the positive center; in all regions except Montana, the negative r is west of the CD. In the eastern half, distances to the positive r centers are smaller than those to the negative centers, and wet weather is favored by negative MMHA to the west and positive anomalies to the east. Since r is assumed linear, preferred height configurations for dry weather in these five areas should be opposite to those shown in Fig. 8.

4. COMPOSITE MAPS

By neglecting spatial gradients in the standard deviation of 700 mb height (σ), Stidd (1954) showed that flow inferences usually drawn from

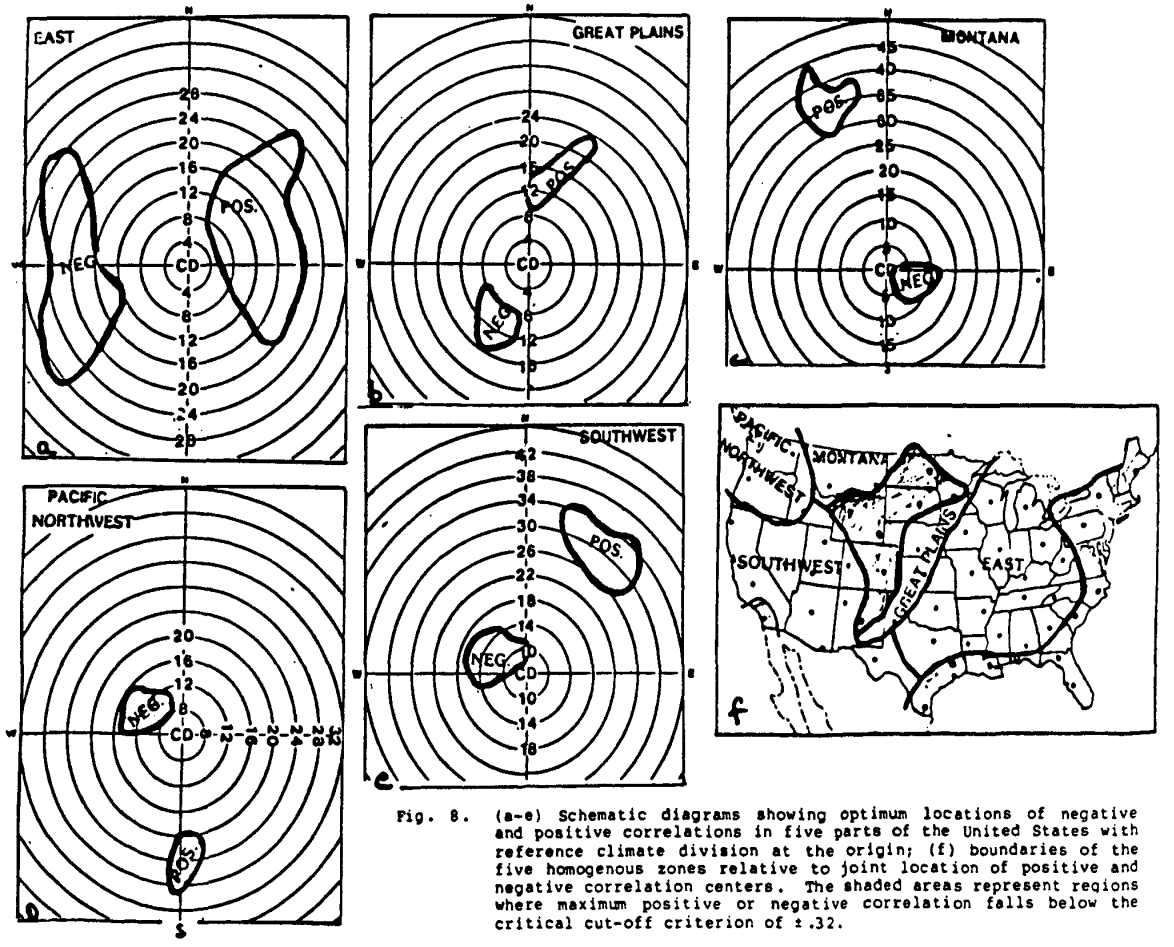


Fig. 8. (a-e) Schematic diagrams showing optimum locations of negative and positive correlations in five parts of the United States with reference climate division at the origin; (f) boundaries of the five homogenous zones relative to joint location of positive and negative correlation centers. The shaded areas represent regions where maximum positive or negative correlation falls below the critical cut-off criterion of ±.32.

conventional composite maps can be made directly from $\underline{r}$ fields. However, σ is not constant in space. Therefore we reinserted it into our correlation fields by multiplying each $\underline{r}$ by the value of σ at the same point, thereby producing a height anomaly map in meters (Kline and Klein, 1986). If the relations between precipitation and circulation are linear, our maps of $\underline{r\sigma}$ can be considered as composites for PP one standard deviation above normal at the reference CD. Furthermore, these composites are now based on all available data, rather than only the traditional extreme cases.

Sample composites for the two $\underline{r}$ fields shown in Fig. 4 are given in Fig. 9. Fig.9a reveals that the negative height anomaly center associated with wet weather in northern California is displaced about 300 km west of the negative $\underline{r}$ center in Fig. 4a towards higher σ. It also shows that high PP at this CD goes with below normal heights locally and anomalous geostrophic cyclonic southwesterly 700 mb flow. The composite for Maine (Fig. 9b) contains a strong center of positive height anomaly over the North Atlantic, northeast of its corresponding $\underline{r}$ center in Fig. 4b. It indicates that wet weather in Maine accompanies above normal heights locally and anticyclonic southeasterly anomalous flow.

The local components of anomalous 700 mb flow indicated by the $\underline{r\sigma}$ composites to be conducive to wet weather were plotted and analyzed for each of the 60 CDs, with results given in Fig. 10. The preferred components for high PP are southwesterly in the West and Florida, northerly in the northern slopes, southeasterly in the Great Plains and along the North Atlantic Coast, and southerly in portions of the Ohio, Mississippi, and Tennessee Valleys, the Great Lakes area and the Mid-Atlantic Coast. Dry weather in these regions is accompanied by the opposite direction of anomalous flow.

Figure 11 shows the moisture source for wet weather suggested by the anomalous flow on the $\underline{r\sigma}$ composites. This was obtained by following the isopleth passing through the reference CD upstream to its first

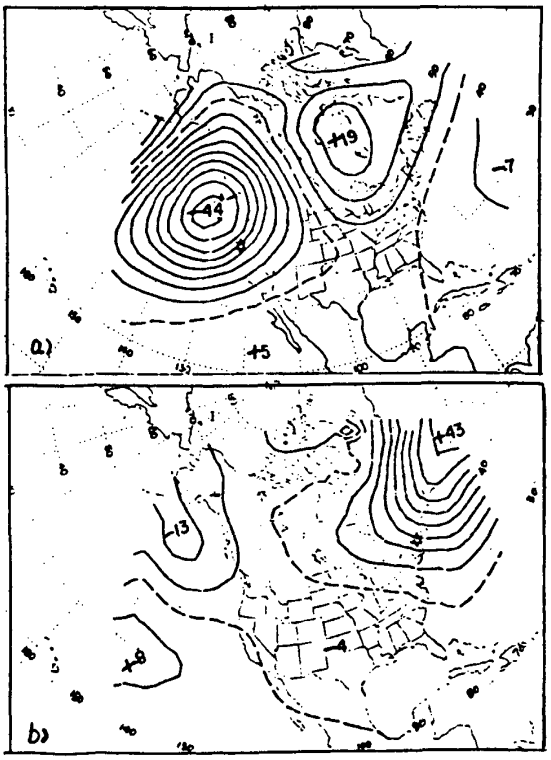


Fig. 9. The $\underline{r\sigma}$ composites showing monthly mean 700 mb height anomalies conducive to wet weather during winter at: a) northern California, b) Maine (located by stars). They were obtained by multiplying the correlation coefficients of Fig. 4 by the standard deviation of 700 mb height at the same point. Isopleths are drawn every five meters with zero line dashed and central values labeled.

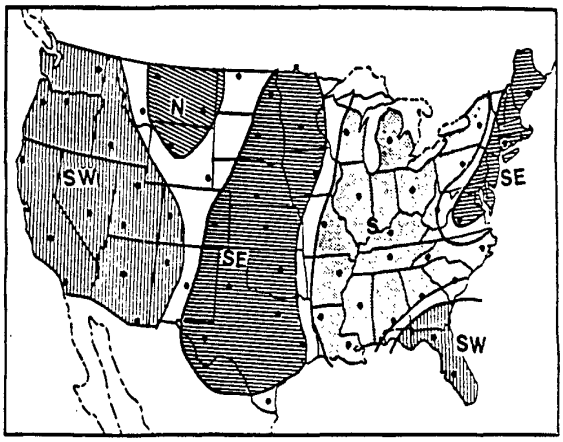


Fig. 10. Local component of anomalous 700 mb geostrophic flow indicated by composites to be conducive to wet weather during winter in different parts of the country. Flow is indeterminate in blank areas.

intersection with a large body of water. For nearly all CDs west of the Continental Divide, the moisture source is the Pacific Ocean. In the Northern Plains, the wet flow does not seem to originate in any one well-defined source of moisture. For most of the South and Midwest, anomalous 700 mb flow originates in the Gulf of Mexico for heavy precipitation. Finally, along the east coast from north Carolina to Maine, the principal source of moisture is the Atlantic Ocean at lower latitudes.

Figure 12 shows the local curvature of anomalous 700 mb flow conducive to wet weather, as determined from the $r\sigma$ composites. The preferred curvature is cyclonic over the western two-thirds of the country and also along the Gulf and South Atlantic coasts. However, anticyclonic curvature favors high PP in parts of a belt extending from southwest Texas through the Ohio and Tennessee Valleys to New England and the Upper Great Lakes. Since our method for construction of composites is linear, curvature for dry weather should be anticyclonic over most of the country, but cyclonic over parts of the Northeast, Ohio Valley and central Texas.

5. CONCLUDING REMARKS

So far we have assumed that our r fields and composites are linear in nature. In order to check the validity of this assumption, we ranked the PP at each CD and then averaged the MMHA separately for the upper and lower 30 percent of our 90-month sample. The resulting conventional composite maps are illustrated for southern California in Fig. 13. This figure gives three different levels of information for percentiles: a) conditions favoring above normal, b) conditions favoring below normal, and c) above-below normal (the composite difference chart). If the relations between PP and 700 mb heights were truly linear, then the maps in part a) and part b) should have the same shape and magnitude of centers but with opposite signs, while part c) should resemble our $r\sigma$ composite shown in Fig. 13d, but with twice the magnitude (Kline and Klein, 1986). The map for wet conditions (13a) is almost the complete opposite of the dry map (13b), while the composite difference map (13c) looks like twice our $r\sigma$ composite (Fig. 13d). Figure 13d also strongly resembles the conventional composite map for wet (Fig. 13a). Therefore, our linear assumption is valid for this CD. Since similar results were obtained elsewhere, we may conclude that the relations between monthly precipitation and the large scale pattern of 700 mb height are linear to a first approximation. As a result, our method of computing composites based on all available data gives results similar to conventional composite maps based on extreme cases only.

Additional findings may be summarized as follows:

1. In most of the U.S., heavy precipitation is closely related to a center of negative MMHA near or west of the reference point. However, northeast and south-central portions depend more on a positive center to the east.

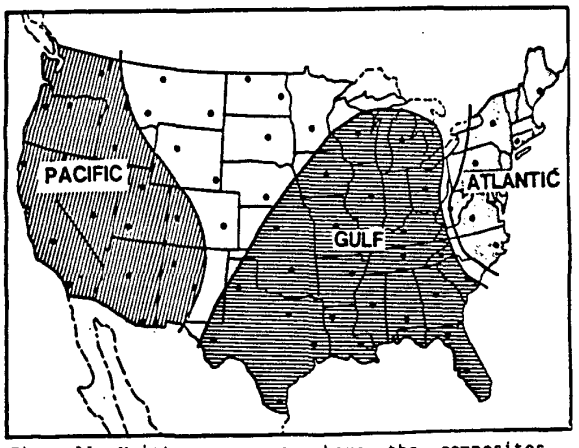


Fig. 11. Moisture source where the composites suggest that anomalous 700 mb flow originates for wet weather in winter. The moisture source is indeterminate in blank areas.

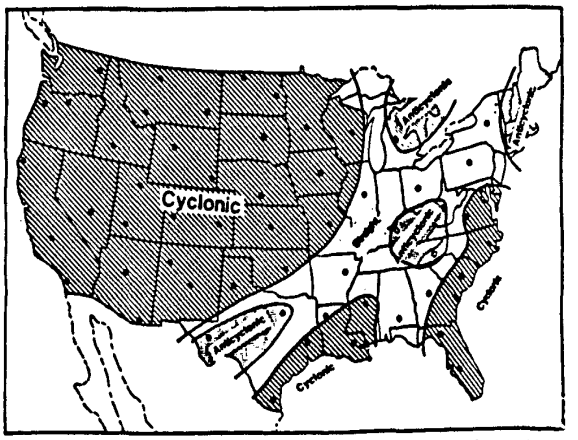


Fig. 12. The local curvature of anomalous 700 mb geostrophic flow indicated by the composites to be conducive to wet weather during winter. Curvature is straight in blank areas.

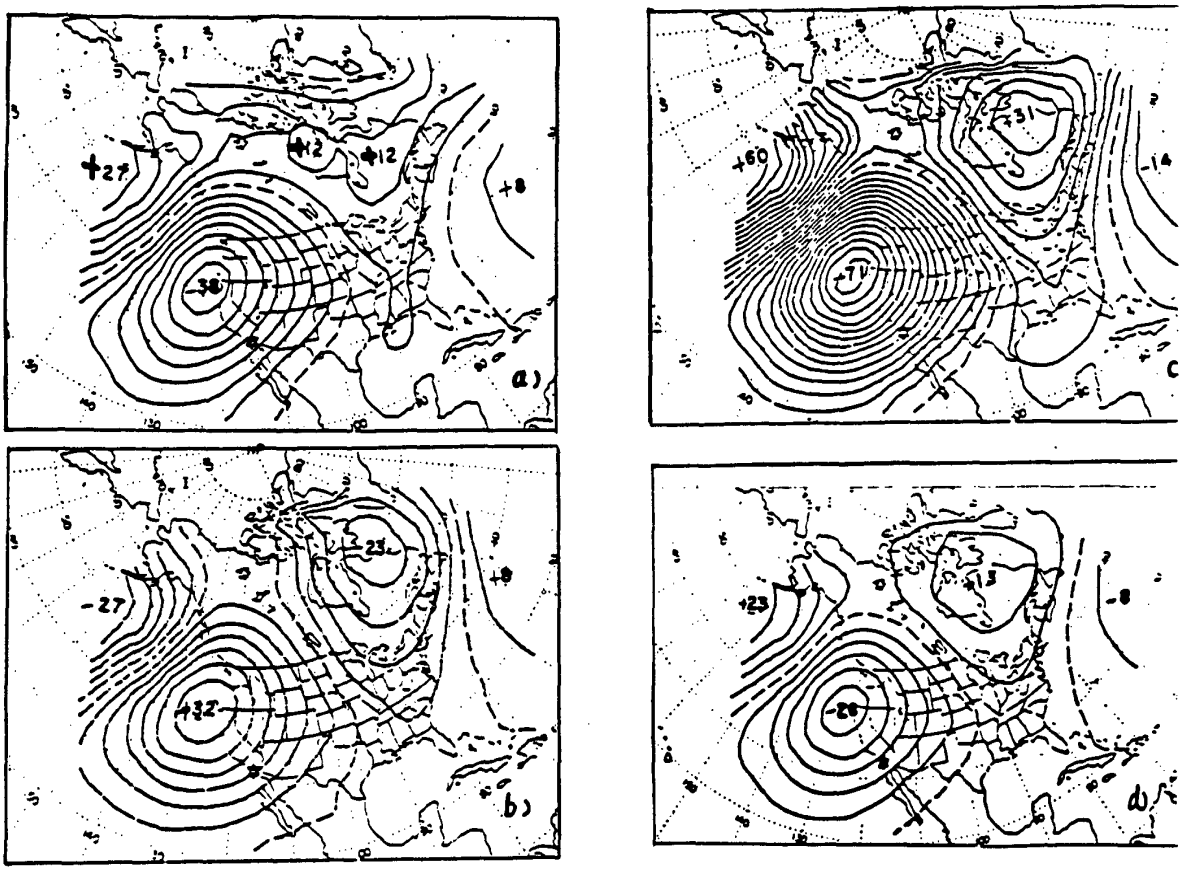


Fig. 13. Conventional composite maps for southern California showing mean 700 mb height anomalies during winter for a) months with high precipitation percentiles, b) months with low percentiles, and c) the difference: a)-b). The $r\sigma$ composite is shown in d). Isopleths are drawn every 5 meters, with zero line dashed and central values labeled in meters.

2. Anomalous 700 mb flow from moist maritime regions generally accompanies wet weather; light precipitation is associated with flow from dry continental sources.

3. The local curvature of anomalous 700 mb flow conducive to wet weather is cyclonic for most of the U.S. but anticyclonic over parts of the Northeast, Ohio Valley and Texas.

Acknowledgements

This material is based on work supported by the National Science Foundation under Grant ATM-8412817 as part of the Climate Dynamics Program. We are greatly indebted to Art Douglas and Phil Englehart at Creighton University for the use of their precipitation data. Valuable help was received from Bob Kaylor, Huug van den Dool and the Computer Science Center at the University of Maryland; as well as from Don Gilman, Chet Ropelewski and Ed Epstein at the Climate Analysis Center of the National Weather Service.

REFERENCES

Englehart, P. J. and A. V. Douglas, 1985: A statistical analysis of precipitation frequency in the conterminous United States, including comparisons with precipitation total. J. Clim. Appl. Meteor., 23, 350-361.

Klein, W. H., 1963: Specification of precipitation from the 700-millibar circulation. Mon. Wea. Rev., 91, 527-536.

_____, 1965: Application of synoptic climatology and short-range numerical prediction to five-day forecasting. Res. Paper No. 46, U. S. Weather Bureau, Washignton, DC, 109 pp, Department of Commerce, NOAA, 6009 Executive Blvd., Rockville, MD 20852.

_____, and H. J. Bloom, 1986: Specification of monthly precipitation over the United States from the surrounding 700 mb height field in winter. Mon. Wea. Rev. (submitted).

Kline, J. M. and W. H. Klein, 1986: The synoptic climatology of monthly mean surface temperature in the United States during summer in relation to the surrounding 700 mb height field. Mon. Wea. Rev., 114, 1231-1250.

Lund, I. A., 1970: A Monte Carlo method for testing the statistical significance of a regression equation. J. Appl. Meter., 9, 330-332.

Ropelewski, C. F. and J. B. Jalickee, 1983: Estimating the significance of seasonal precipitation amounts using approximations of inverse gamma function over an extended range. Eighth Conference on Probability and Statistics in Atmospheric Sciences, Hot Springs, AK, November 16-18, 1983, pp. 125-129.

Stidd, C. K., 1954: The use of correlation fields in relating precipitation to circulation. J. Meteor., 11, 202-213.

Thom, H. C. S., 1958: A note on the gamma distribution. Mon. Wea. Rev., 86, 117-121.

Walsh, J. E., 1984: Forecasts of monthly 700 mb height: Verification and specification experiments. Mon. Wea. Rev., 112, 2135-2147.

Third International Conference on Statistical Climatology
June 23-27, 1986, Veinna, Austria

THE SYNOPTIC CLIMATOLOGY OF MONTHLY MEAN SURFACE TEMPERATURE IN EUROPE IN RELATION TO THE 700 MB HEIGHT FIELD

William H. Klein and Runhua Yang
Department of Meteorology
University of Maryland
College Park, MD 20742

1. INTRODUCTION

Since World War II, numerous studies on the relation between surface temperature and the 700 mb height field have been reported for 5-day mean data (summarized in Klein, 1965). More recent work has emphasized the monthly time scale. For example, Klein and Kline (1984) and Kline and Klein (1986) assessed the synoptic climatology of monthly mean air temperature anomalies (MMATA) in the United States during winter and summer in relation to the surrounding 700 mb height field.

This paper extends the above research to a surface network in Europe and a height field covering most of the Northern Hemisphere. After a brief description of the data and methods used, we discuss the synoptic climatology of MMATA during winter and summer, respectively, based on analysis of two types of maps. The first consists of fields of simple linear correlation coefficient (r) between the MMATA at each station and the simultaneous monthly mean 700 mb height anomaly (MMHA) at each grid point. The second is a special type of composite map, constructed by multiplying r by the standard deviation of MMHA at the same point. Finally, we show the optimum distribution of height anomalies for abnormally warm or cold weather for winter and summer in each of several climatologically similar regions into which Europe is subdivided.

2. DATA AND METHODOLOGY

From WMO data tapes, surface monthly temperature data for the 30-year period: Jan. 1951-Dec. 1980 were extracted at 68 stations in Europe (arbitrarily defined as the area west of longitude 55° E and north of 30°N) (Fig. 1). The field of MMHA was represented by a grid of 134 intersections of latitude and longitude shown in Fig. 2. The data at these points were obtained from the U.S. Climate Analysis Center (CAC). In order to reduce random fluctuations in a small sample, we pooled the monthly data for winter, defined as Dec., Jan. and Feb., and for summer, defined as June, July and Aug. However, anomalies of monthly mean temperature and height were taken as departures from the 30-year mean of each month. The r fields for winter and for summer were then computed and analyzed separately for each station.

Because of the temporal and spatial intercorrelations in MMHA and MMATA, and because we are arbitrarily selecting the highest r from a field of 134 *a posteriori*, we applied a Monte Carlo significance test, similar to that used by Lund (1970) and Walsh (1984). We first randomly mixed the temporal order of the temperature data at 37 stations but kept the original MMHA set in order (thereby retaining its intercorrelation and persistence properties). We then computed simple correlation coefficients between the

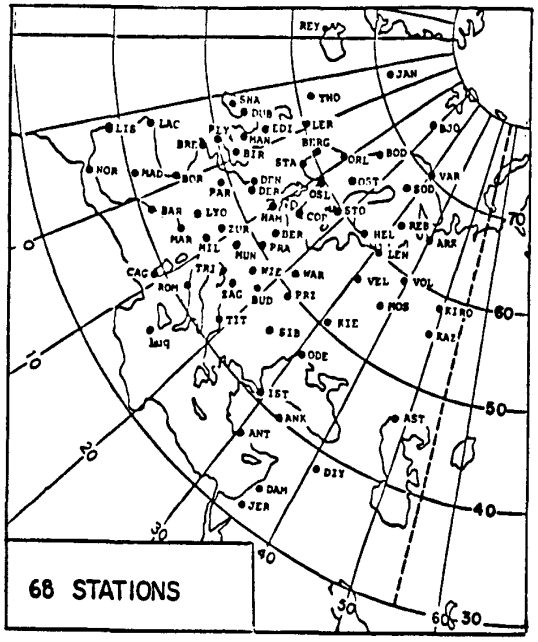


Fig. 1 Location and first 3 (or 4) letters of name of 68 surface stations for which monthly mean temperatures were available during 30-year period: 1951-80.

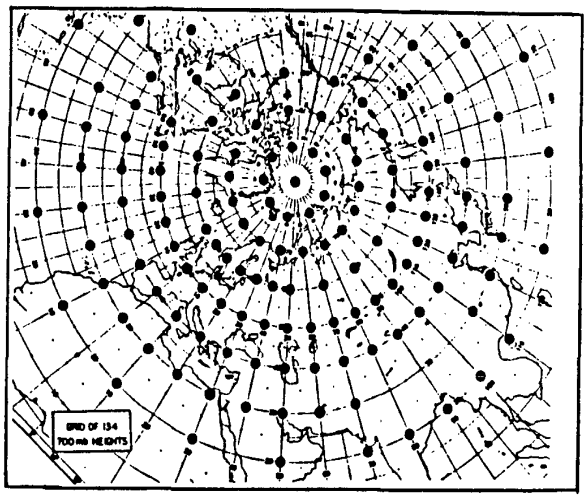


Fig. 2 Network of 134 grid points used to delineate the field of 700 mb heights.

mixed temperatures and MMHA at each grid point. Finally, we choose a maximum and a minimum r for each run. After mixing 100 times at each station, we found a pair of positive and negative correlation values at the 95th percentile. Their mean was ±0.36, and we defined this as the criterion for statistical significance of r. As expected, this value is much higher than the 0.21 given in classical tables for n=90 independent pairs of data at the 95% probability level.

Stidd (1954) showed that a field of correlation coefficients is strictly analogous to a geostrophic anomalous flow pattern if we neglect spatial variation in standard deviation of the heights (σ). In reality, σ is not constant in space. Therefore we reinserted it into the correlation fields by multiplying each value of r by σ at the same point. The resulting map is proportional to a chart of regression coefficients and thus to a composite difference chart. The latter is analogous to a composite map if the correlations are linear. Thus our chart of $r\sigma$ gives the optimum distribution of height anomalies for abnormally warm weather (actually one standard deviation above normal) at the reference station. It is therefore a composite based on all available data, instead of only a few extreme cases as customarily used.

If the relationship between the anomalies of height and temperature is linear, our composite should look like the conventional composite map for warmth. In order to test this assumption, we selected four stations with the highest or lowest absolute r in winter and summer. The MMATA for each station was ranked, and the upper and lower 30% defined as two classes: above and below normal. Then, the height anomalies corresponding to each of the two classes were averaged separately. The resulting conventional composite maps are illustrated in Fig. 3 for Oslo, Norway in summer. Note that the map for above normal temperatures (Fig. 3a) is almost completely opposite to the below normal map (Fig. 3b), while the composite difference map (Fig. 3c) looks just like our composite (Fig. 3d) but with about twice the magnitude. Note also the striking similarity between Fig. 3a and Fig. 3d. Therefore, the assumption of linearity is a good first approximation, and we shall henceforth use the term "composite" only to mean our special composite map.

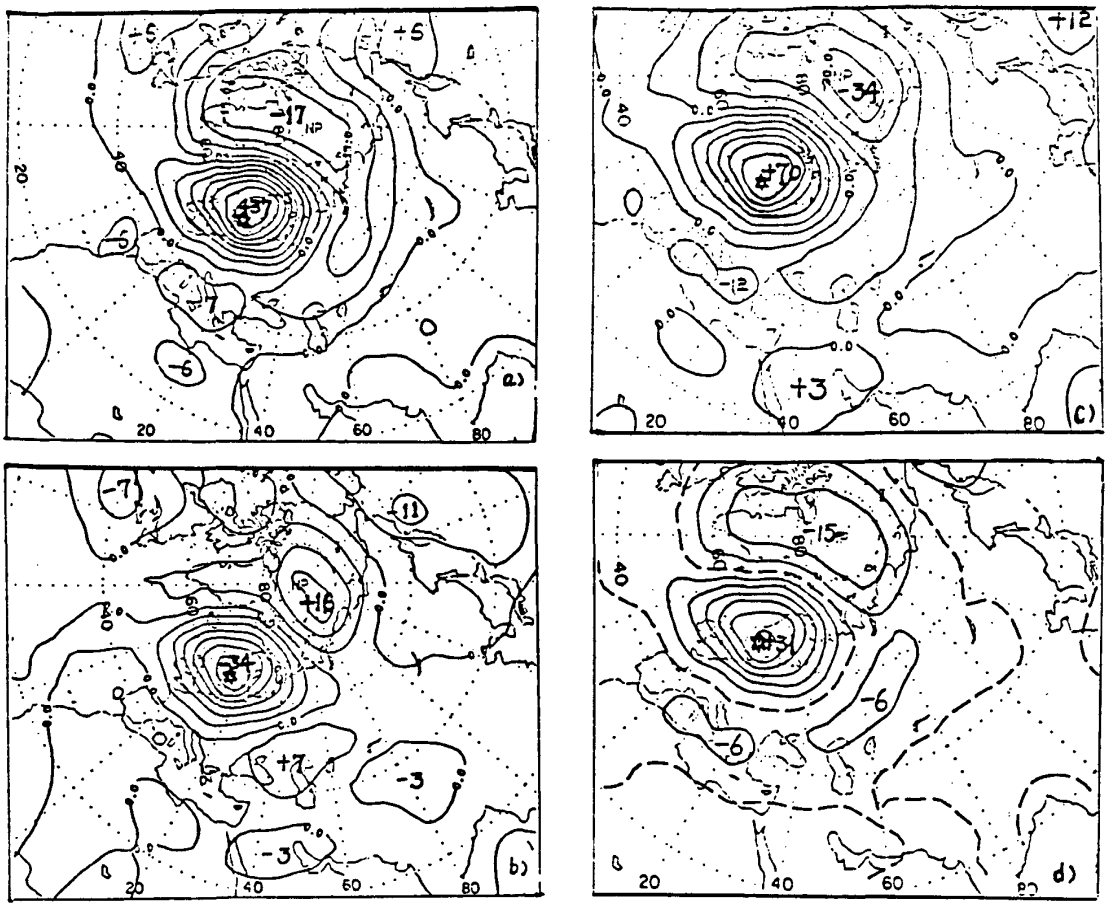


Fig. 3 Composite maps for Oslo, Norway (star) in summer for a) average MMHA in months with above normal MMATA, b) average MMHA in months with below normal MMATA, c) difference: a)-b), d) warm composite based on all cases. Isopleth interval is 5 meters, except 10 meters for c).

3. SAMPLE CORRELATION FIELDS AND COMPOSITE MAPS

Fig. 4 illustrates the fields of r at Vienna, Austria for winter and summer. The upper part shows that the dominant features during winter are a positive center (0.61) about 1200 km southwest of the reference station and a negative center (-0.54) about 2200 km northwest of the station. During summer (Fig. 4b), the dominant feature is a strong positive correlation center (0.72), just east of the reference station, while no other center on the map is statistically significant. This implies that equivalent-barotropic conditions dominate the lower troposphere.

Fig. 5a is our composite for Vienna in winter. It shows that above normal temperatures are favored by anomalous westerly flow of mild Atlantic air between a center of positive height anomaly in the Mediterrean and a negative anomaly center near Iceland. The optimum height pattern for cold weather is the same but with signs reversed; i. e., below normal temperatures accompany easterly anomalous flow of cold air from the continent and a blocking pattern in the European sector. The composite for Vienna in summer (Fig. 5b) shows that the strongest height anomaly center (+17) is about 400 km northeast of the station, so that southeasterly (northwesterly) anomalous flow produces warm (cool) weather in summer, compared to northwesterly (southeasterly) in winter.

4. SYNOPTIC CLIMATOLOGY OF MMATA IN WINTER

Correlation fields similar to those illustrated in Fig. 4 were constructed for all 68 stations of Fig. 1. Almost all stations in winter have two significant r centers. One of these is positive and approximately

1300 km from the reference station; the other is negative and about 2800 km away. The positive correlations are slightly greater in magnitude than the negative centers, as indicated by mean values of +.61 and -.57.

In order to find key regions where MMHA affects MMATA, the locations of maximum positive and negative centers were plotted on polar coordinate paper (with the reference city at the origin) and then subjectively grouped into homogeneous areas or clusters. Fig. 6a depicts two such clusters. The first is for two separate geographical regions: the Atlantic coast and the Mediterranean (Fig. 6b), where positive r centers are located from 600 km to 1800 km east or southeast of the reference stations. The second is for interior Europe with centers located from 800 km to 1700 km south of the stations. The centers for all but 7 of the 68 stations fell into these two homogeneous areas. Fig. 7 is similar to Fig. 6 but for clusters based on maximum negative r centers. The first, about 1700-2900 km northwest of the reference stations, contains the great majority of European cities. The second, 1800-3000 km west of the stations, contains a small number of northern islands and Arctic coastal stations denoted by 'far north'. Geographical locations of those homogeneous areas are shown in Fig. 7b.

By following the same procedure used in preparing Fig. 6, the European sector was divided into three climatologically similar clusters based on composites. These clusters (not shown) indicate that above normal temperatures usually are accompanied by above normal heights, warm advection and anticyclonic flow, while cyclonic curvature, cold advection and below normal heights prevail for below normal temperatures. Fig. 8a gives the location of upstream negative anomaly centers for Europe. Fifty-nine stations fell into the "Europe region," where the centers are 1600 km

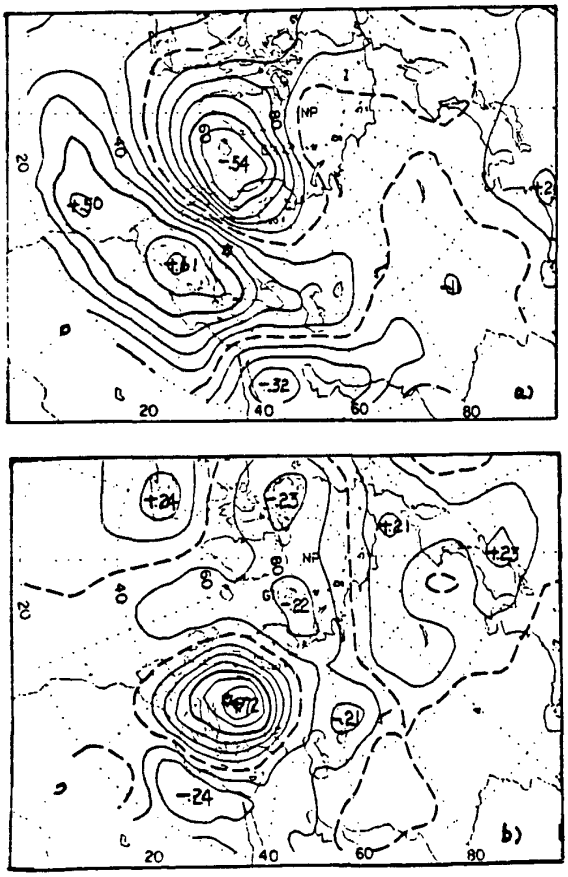


Fig. 4 Field of simple linear correlation coefficients between concurrent MMATA at Vienna, Austria (located by star) and MMHA at grid points of Fig. 2 in a) winter, b) summer. The isopleths are drawn at intervals of 0.10 with zero line dashed and central values interpolated.

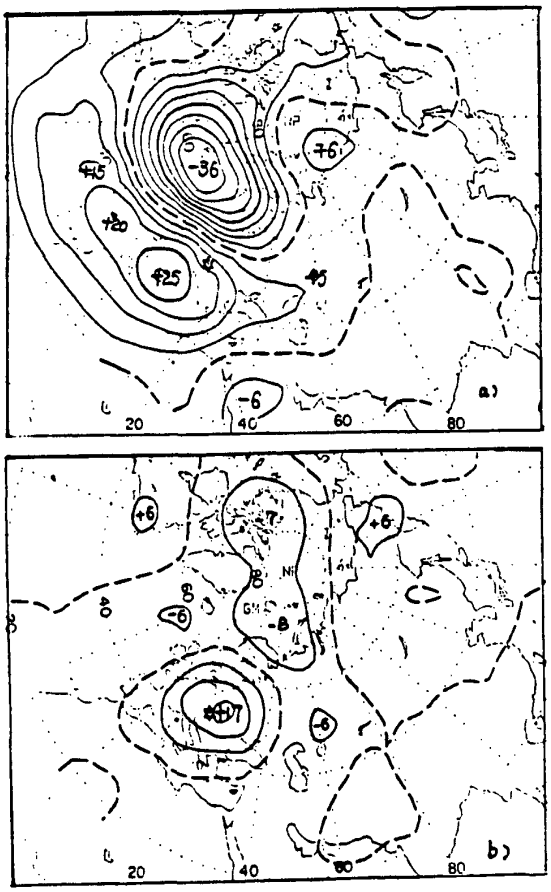


Fig. 5 Composites showing the patterns of MMHA conducive to above normal temperatures at Vienna, Austria in a) winter, b) summer. The isopleths are drawn every 5 meters, with zero line dashed and centers labeled.

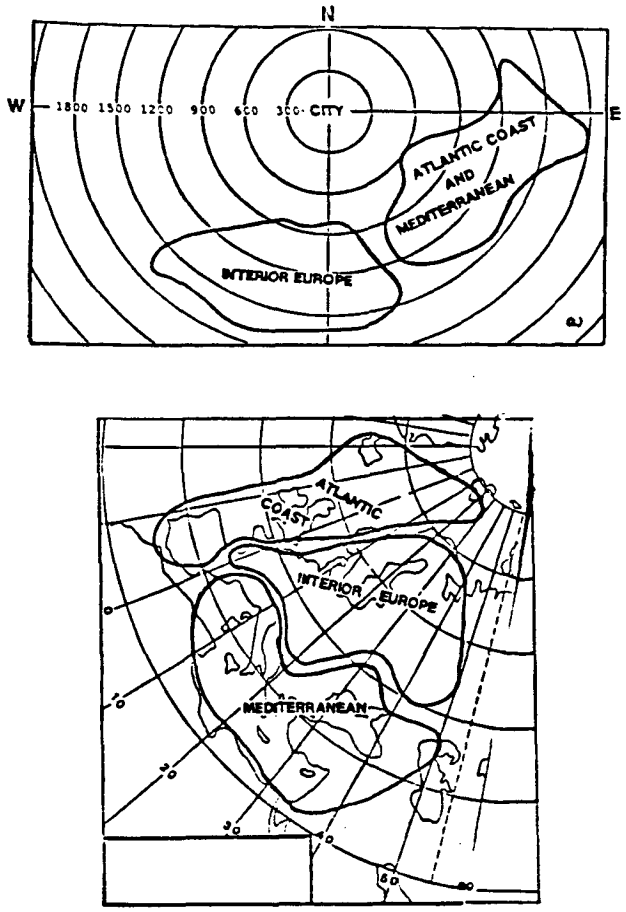


Fig. 6 a) Distance (km) and direction (north, south, east or west) from reference station (at origin) to center of maximum positive correlation between concurrent winter MMATA and MMHA, b) boundaries of corresponding homogeneous zones.

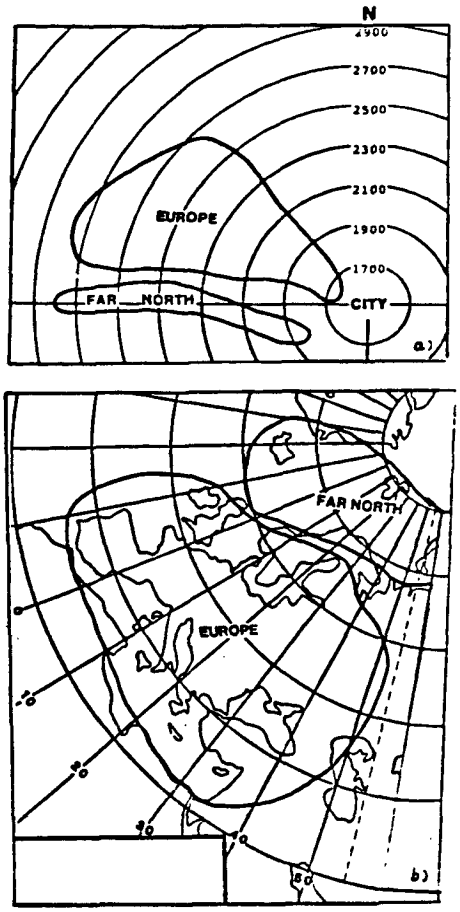


Fig. 7 a) Distance (km) and direction from reference station (origin) to center of maximum negative correlation between concurrent winter MMATA and MMHA, b) boundaries of corresponding homogeneous zones.

to 3500 km northwest of the station. A second cluster has negative anomaly centers from 1600 km to 2400 km west-southwest of a few stations in the northeast Atlantic. A third region, called "northmost Europe", has its negative anomaly centers about 2600 km to 3300 km west-northwest of the reference stations. Above (below) normal temperature in each region of Fig. 8b tends to accompany below (above) normal height centers in the corresponding cluster of Fig. 8a.

5. SYNOPTIC CLIMATOLOGY OF MMATA IN SUMMER

Compared to winter, the average value of the maximum positive r center in summer is higher, and the maximum positive center is much closer to the reference station. In addition, more than half the stations in summer lack significant negative centers, and less than half have their negative centers located west of the station, implying that the upstream circulation is no longer as important as in winter.

Fig. 9a gives the distribution of maximum positive r centers relative to the key stations in Europe for summer. Three clusters appear. The first, called "most of Europe," includes a majority of inland stations; here maximum positive centers are very close (100km-600km) to the reference station and located towards the east-northeast. The second cluster is for both the Atlantic coast and northwest Europe, with centers located 400-1100 km north-northeast of the reference stations. The third cluster includes only four stations over the Norwegian and Barents seas (NBS), with centers about 900-1350 km east-southeast of the stations. Their homogeneous areas are located in Fig. 9b.

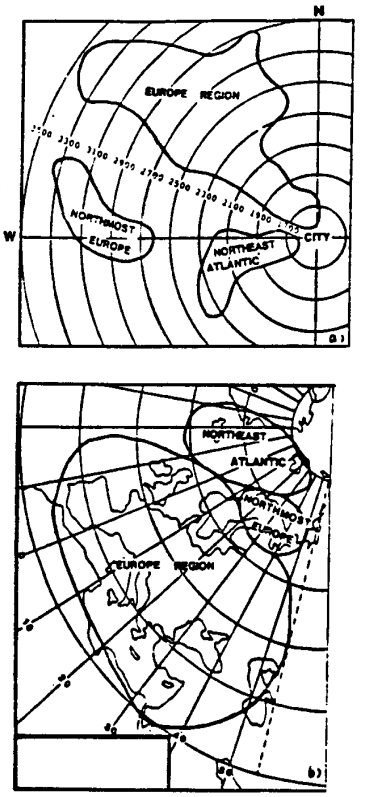


Fig. 8 Same as Fig.7 but for center of negative height anomaly.

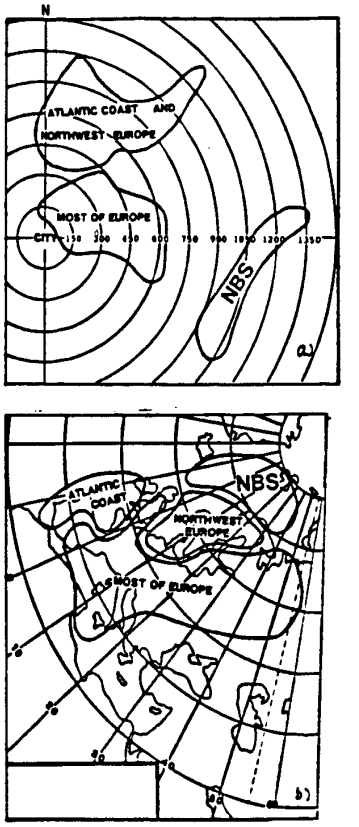


Fig. 9 Same as Fig. 6 but for summer.

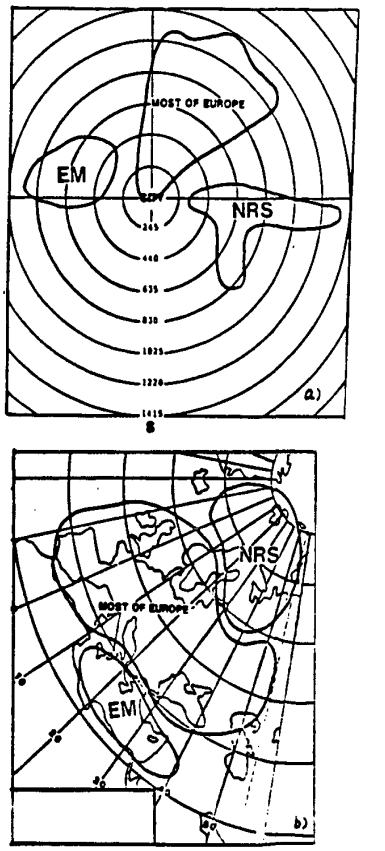


Fig. 10 Same as Fig. 8 but for positive anomaly centers in summer.

The polar coordinate diagram for maximum positive anomaly centers on our composites indicates three clusters in the European sector (Fig. 10a, b). The first, including stations in most of Europe, shows maximum positive centers northeast of the station at distances of about 50 km to 1200 km. Above normal temperatures in this region are likely to occur under anticyclonic, southeasterly anomalous flow around a positive anomaly center at 700 mb. Above normal temperatures in Northern Russia and Scandinavia (NRS) occur under anticyclonic, southerly 700 mb anomalous flow about 300 km to 1300 km west of a center of positive height anomaly. A third region includes only four stations in the eastern Mediterranean (EM), where above normal temperatures are associated with anticyclonic, northeasterly 700 mb anomalous flow around positive height anomaly centers located about 300 km to 900 km west-northwest of the stations.

The positions of the 700 mb height anomaly centers relative to the reference stations change from winter to summer. Positive centers are mainly located south, southeast or southwest of the station in winter but north, northeast or northwest in summer. The main reason for this seasonal change is the seasonal reversal of the thermal gradient between continent and ocean. During winter, southwesterly anomalous flow brings warm air from the Atlantic ocean to the colder continent, but in summer the anomalous flow brings warm air from the continent to the coast.

6. CONCLUSION

Our principal findings may be summarized as follows:

1) The properties of correlation fields for the European sector closely resemble those for North America. The MMATA is mainly affected by

two MMHA variables in winter, one with positive r, about 1300 km from the station; the other with negative r, about 2800 km northwest of the station. As a result, above normal temperatures usually accompany local anticyclonic, southerly flow around a nearby positive height anomaly center, and an upstream negative anomaly center that implies large scale advection of warm air. For below normal temperatures, the patterns are reversed.

2) During summer, the MMATA in Europe is mainly affected by only one MMHA center about 800 km from the station. Above normal temperatures usually accompany above normal heights with southerly or easterly anomalous flow, and conversely for cool weather.

3) The relations between surface temperature and the 700 mb circulation pattern are mainly linear. As a result, correlation fields can be readily converted to composite maps by multiplying r at each grid point by σ at the same point. These composites are more useful in subjective estimation since they resemble height anomaly maps routinely used to draw inferences about the probable temperature distribution. However, r fields are more useful in objective forecasting because they give a direct index of association between heights and temperature. A logical extension of the latter approach is to derive multiple regression equations by screening the entire field of 700 mb height (Klein and Yang, 1986). Equations of this type are currently being applied by CAC to objectively specify MMATA in the Northern Hemisphere from prognostic 700 mb maps.

Acknowledgements

This work was supported by the U.S. Climate Analysis Center, NOAA/Commerce Grant NA83AA-D-00048, AI, and the U.S. Department of Energy, Contract DE-AS0581EV-10539. We are grateful to CAC, NCAR, WMO and the University of Illinois for providing the basic data used in this paper. Valuable help was received from John Walsh, Huug van den Dool, George Huffman, Joann Kline, Hal Bloom and Robert Kaylor.

REFERENCES

Klein, W. H., 1965: Application of synoptic climatology and short-range numerical prediction to five-day forcasting. Weather Bureau Res. Pap. No. 46, 109 pp.(Library and Information Services Division, NOAA, U. S. Department of Commerce, Washington, DC 20852).

---- and J.M. Kline, 1984: The synoptic climatology of monthly mean surface temperature in the United States during winter relative to the surrounding 700 mb height field. Mon. Wea. Rev.,112, 433-448.

---- and R. Yang, 1986: Specification of monthly mean surface temperature anomalies in Europe and Asia from concurrent 700 mb monthly mean height aomalies over the Northern Hemisphere. J. Climatology, 6, (in press).

Kline, J. M. and W. H. Klein, 1986: The synoptic climatology of monthly mean surface temperature in the United States during summer in relation to the surrounding 700 mb height field. Mon. Wea. Rev., 114, June 1986.

Lund, I. A., 1970: A Monte Carlo method for testing the statistical significance of a regression equation. J. Appl. Meteor., 9, 330-332.

Stidd, C.K., 1954: The use of correlation fields in relating precipitation to circulation. J. Meteor.,11, 202-213.

Walsh, J. E., 1984: Forecasts of monthly 700 mb height: Verification and specification experiments. Mon. Wea. Rev., 112, 2135-2147.

Third International Conference on Statistical Climatology
June 23-27, 1986, Vienna, Austria

THE FREQUENCY OF HAILDAYS IN CENTRAL SWITZERLAND
A COMPARISON HAILPAD - INSURANCE DATA

H. H. Schiesser
Atmospheric Physics ETH
8093 Zürich, Switzerland

1. INTRODUCTION

Hail is a rare and often very limited occurence in time and space and shows a variability which exceeds most other weather phenomenon. An excellent review about the scales of hail is given by Changnon (1977). Because of that high variability it is difficult to get reasonable quantitative data for important hail variables as the frequency of haildays (spatial, temporal) and of hailfalls (covered area, duration) and their intensity (number and size of hailstones, kinetic energy and the related crop damage) which are necessary to establish a hail climatology for a certain geographical area. Possible data sources for such studies are the national weather service, the hail insurance and special networks from field experiments. The network of meteo-stations is usually very scarce and the insurance data provide a reliable cover of the investigated area only if this area is sufficientiy insured. An other problem is that only hailfalls are recorded which produced damages on insured crops. The special networks are limited in size but they provide a detailed picture of the hailstorms.

Grossversuch IV, a randomized hail suppression experiment, was conducted in Central Switzerland in the years 1977-81. It was designed (Federer et al., 1978/79) to test the hypothesis that seeding hail clouds would decrease the hail kinetic energy of the treated hailcells, measured by a 10 cm quantitative radar and a network of hailpads. The modification concept is meant to produce artificially more ice-nuclei in the hail embryo formation region which are then in competition for the supercooled water. Many small hailstones instead of some large ones should be the result.

To decide whether the cloud was hail-bearing or not, a quantitative seeding criterion was used. The criterion was effective meaning that all cells which produced hail on the ground have been detected (Waldvogel et al., 1979).

Fig. 1a presents an example of a cell moving through the experimental area. The radar contours are drawn in 10 minutes intervals and the rocket trajectories from different launching sites are shown. Fig. 1b shows the hailpad network. The pads which were hit by the storm are marked and the isolines of kinetic energy are drawn.

It is assumed that almost every hailfall (except a few very small and limited events which can fall through the mesh of the network) is recorded by at least one pad. In this paper the assumed "true" frequency of haildays of the pad network will be compared with the hailday data provided by the Swiss hail insurance company. The comparison will show whether the hailpad measurements cover really all haildays with crop damage and on the other hand how many haildays with no insurance records are counted by the pads. In addition the global hail kinetic energy of the paddays with and without insurance records will be compared. The results of this study can be taken as an estimate of how reliable insurance data, often used in practice, for hail climatological analysis are.

Because the statistical result of the confirmatory analysis of Grossversuch IV did not show any significant difference in the hail kinetic energy between seed and control cells (Federer et al., 1982), the two sets of haildays (seed and control) can be taken as one sample in this study.

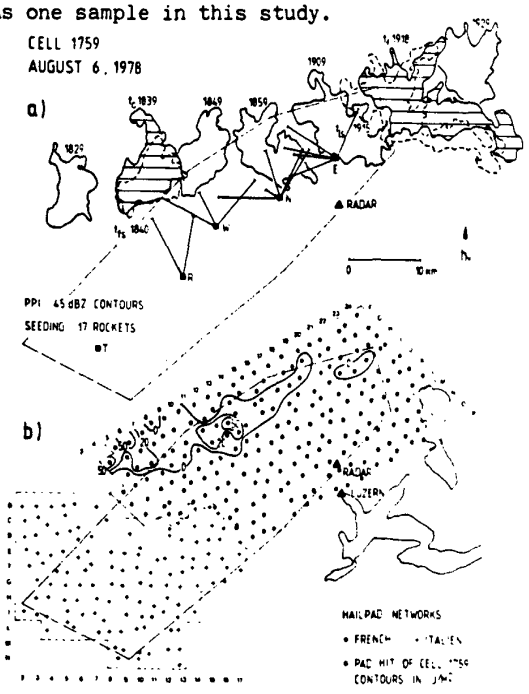


Fig. 1 Example of a storm as measured by radar (a) and the hailpads (b) during Grossversuch IV.

2. TEST AREA AND TIME PERIOD

The area investigated is limited to that part of the experimental area of Grossversuch IV which lies completely in the canton of Luzern and in which all communities are covered by the hailpad's network (see Fig. 2). 49 communities, covering an area of about 800 km^2, can be taken into account.

Fig. 3 shows the topography of the test area which can be subdivided in two geographically distinct areas, the flat - hilly country in the northeast (NE) with intensive crop farming and the mountainous southwest (SW) with forests and dairy farming.

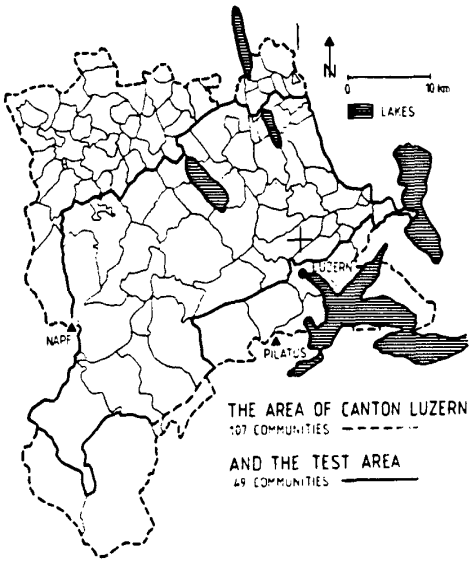


Fig. 2 The political boundaries of the 107 communities in the investigated area. The test area which is completely covered by the pad network is marked off (49 communities).

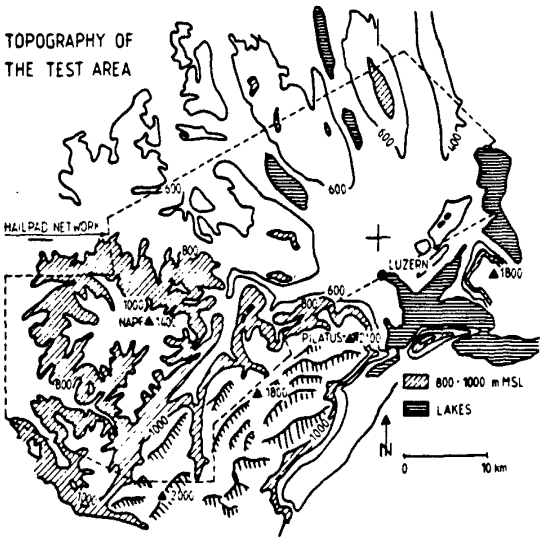


Fig. 3 The topography of the test area

The observed time period is between May 15 and September 15 during the years 1977-81. This period covers most of the growing season of the crop and of the hail season. Some hail events reported by the insurance have taken place before or after that period, but making up only 3.3% of the total.

3. THE DATA

3.1 Hailpad data

The data are derived from a regular network of 335 hailpads, covering an area of ~1300 km^2 (1 pad/3.9 km^2), see Fig. 1b. In the test area 207 pads (62%) are situated. The pads have a sensitive surface of 0.1 m^2. By quick and regular maintenance it is possible to obtain data on the production of hail assigned to individual cells (Mezeix and Doras, 1981) but in this paper only the haildays are considered. Information about the measurement of hail by pads and the design of networks are given by Mezeix et al. (1978) and Long (1978). A hailpad hailday is defined if one pad in the test area, region or community was hit by hailstones with $\geq$ 5 mm in diameter and it is labelled in the following as a PD (pad day).

3.2 Insurance data

The Swiss hail insurance company publishes the dates of days with hail damage claims on crop yearly for each community in Switzerland. If one single farmer makes a damage report for one single field, a hailday is declared for the community in which the field is situated, in the following labelled as ID (insurance day). The insurance itself does not check whether the date of the hailfall on the report is true or not; e.g. the farmer fills out the damage form next day to the event taking that date as the hailfall's date; or after reporting no damage is claimed anymore because of no or minor damage ($<$ 10%) not suitable for appraisal - all those days stay in the list. To eliminate the possibly false days the ID have been corrected using the seeding criterion of Grossversuch IV to discriminate the reported haildays in true and false.

If the seeding criterion was not fulfilled on a day with hail damage claims, in most of the cases the day could be attributed to a hailday before or after that date. Usually only one among several farmers in the same village reported a false date. If a single farmer claimed damage but no assessment has been made later, that day was eliminated. In very few cases, despite the seeding criterion fulfilled, a single claim was also attributed to a hailday before or after that date if there were any.

Table 1 shows all the symbols and their definitions used in this paper.

Symbol	Definition
ID_1	Insurance haildays of the canton Luzern (107 communities)
ID_2	Insurance haildays of the test area (49 communities)
ID_3	Corrected insurance haildays of the test area, region or community
PD	Pad haildays of the test area, region or community
ID_3^*	ID_3 not identical PD
TD	Total number of haildays (PD + ID_3^*)
E_G	Hailpad derived global kinetic energy of the haildays
SW	Southwestern part of the test area
NE	Northeastern part of the test area

Table 1 Symbols and their definitions

4. RESULTS OF THE COMPARISON

4.1 Haildays per community

Fig. 4 shows the long-term fluctuation of the yearly average frequency of ID_1 per community for the years 1929-85. The period of Grossversuch IV is marked. The experiment was fallen into an interval with obviously subaverage hailfall, except 1977. Two years before (1975) the worst hail season in the whole period happened (2.1 haildays above the long-term average of 1.4 days).

Table 2 presents the average frequency per community for single years and the whole period for the various hailday variables.

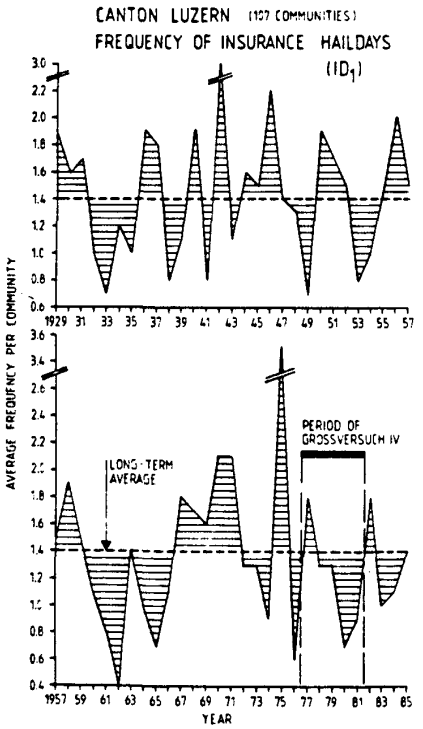


Fig. 4 The long-term fluctuation of the yearly average frequency of insurance haildays (ID_1) per community in the whole area of canton Luzern.

ID_x and PD are calculated using the formulas

$$\overline{ID}_x = \sum_1^N ID_x/N \qquad (1)$$

and

$$\overline{PD} = \sum_1^N PD/N \qquad (2)$$

N is the number of communities (49 or 107) and x the index 1,2 or 3 (see Table 1). $\overline{ID}_1$ (1.2) is less than the long-term average of 1.4. $\overline{ID}_2$ (1.43), the value of the test area, is higher than for the whole canton, indicating some regional differencies. $\overline{ID}_3$ (1.27), the corrected value of $\overline{ID}_2$, is about 10% lower than the reported days of the insurance meaning that about 10% of the haildays reported from single communities had not the correct date. The value of $\overline{PD}$, the pad hailday average, is 3.0 - 136% more than the insurance counts. It can also happen that some ID_3 are not identical with the PD (=ID_3^*). The pad network missed a smaller storm, in the average 0.38 days per community. If one adds $\overline{PD}$ and $\overline{ID}_3^*$, the best estimation of the real average ($\overline{TD}$) is obtained. The $\overline{PD}$ counts for 88.7% of the $\overline{TD}$ whereas $\overline{ID}_3$ only for 37.6%.

Days	Average frequency of haildays per community					
	1977	1978	1979	1980	1981	1977-81
$\overline{ID}_1$	1.80	1.30	1.30	0.70	0.90	1.20
$\overline{ID}_2$	1.96	1.80	1.59	0.88	0.94	1.43
$\overline{ID}_3$	1.80	1.63	1.49	0.65	0.76	1.27
$\overline{PD}$	4.27	3.73	3.08	1.94	1.98	3.00
$\overline{ID}_3^*$	0.39	0.78	0.33	0.14	0.24	0.38
$\overline{TD}$	4.66	4.51	3.47	2.08	2.22	3.38
% $\overline{PD}$ of $\overline{TD}$	91.6	82.7	90.3	93.3	89.2	88.7
% $\overline{ID}_3$ of $\overline{TD}$	38.6	36.1	43.7	31.3	34.2	37.6
% $\overline{ID}_3$ of $\overline{ID}_2$	91.8	90.6	93.7	73.9	80.9	88.8

Table 2 Number of haildays per community for the various hailday variables. The percentage of $\overline{PD}$ and $\overline{ID}_3$ to the best estimate ($\overline{TD}$) as well as the percentage of the corrected value ($\overline{ID}_3$) to the original one ($\overline{ID}_2$) are given.

4.2 Haildays of the whole test area

Fig. 5 shows a small-scale picture of the cumulative point frequency for the 5-year period derived from the hailpad network. The variation in space is remarkable and seems to be related to the topography (compare Fig. 3). The range is between 1 and 16 haildays per point (0.2-3.2 in the yearly average) with a high in the SW and a low in the NE. The average point frequency for the whole area attains 7.3 days (1.46 per year).

Table 3 presents the number of haildays for the various variables for single years and the whole period. Only 2 (out of 68) insurance days were not detected by the hailpads. On 30 June 1977 only two farmers reported damages, one having less than 10%. On 26 July 1978, one community in the whole test area was reporting damages. The two

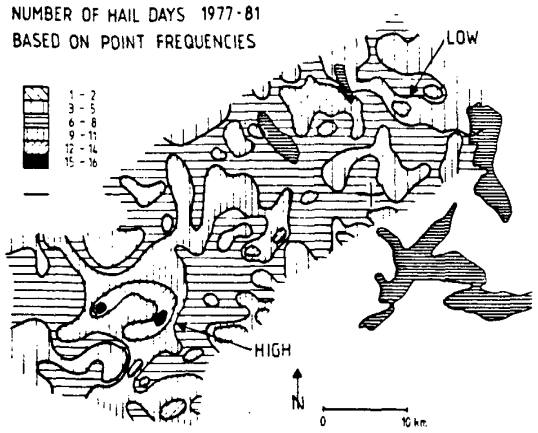


Fig. 5 Small-scale picture of the variation of the haildays in the test area based on point frequencies which are derived from the hailpads.

Days	Number of haildays in the test area					
	1977	1978	1979	1980	1981	1977-81
ID_3	19	22	12	5	10	13.6
PD	31	29	20	12	17	21.8
ID_3^*	1	1	-	-	-	0.4
TD	32	30	20	12	17	22.2
% ID_3 of TD	59.4	73.3	60.0	41.7	58.8	61.3

Table 3 Number of haildays in the test area for the various hailday variables.

days produced probably only graupels (hail < 5 mm in diameter), which were measured by hailpads on one day but have been excluded in this analysis. Thus, one can say that the pads registered almost 100% of the ID_3 whereas the insurance counted only 62% of the actual haildays (TD).

As the PD incorporates almost all ID_3 of the test area a comparison between PD with (identical ID_3) and PD without damage reports (not identical ID_3) can be made, getting an estimate to which extent the two groups differ in the amount of global hail kinetic energy (E_G). It is expected that on days with damages the kinetic energy, measured by the hailpad network, should be much higher than in the other case. Fig. 6 displays a back-to-back stem and leaf plot of the daily E_G in Gigajoule (GJ). The values for days without reported damages are on the left side and for days with damage reports on the right side of the stem. Note that the scale of E_G on the stem changes at 1 GJ.

Above 1 GJ damages are reported in all cases. Between 0.3 and 1 GJ cases occured on both sides but all the events on the left side (marked by a *) were measured in the SW where the insurance coverage is much lower than in the NE. Those cases produced probably damages on the ground but were not detected by the insurance because of the lack of insured crops. Most of the days without assessments obtained energies below 0.3 GJ

$PD \neq ID_3$	E_G (GJ)	$PD = ID_3$
28 2222221111111111 9975555444444	0.0	111333566688 12
7 9965200	0.1	023456 6
2 65	0.2	468 3
2* 70	0.3	355 3
1* 4	0.4	556 3
1* 0	0.5	28 2
1 0	0.6	23 2
	0.7	1446 4
	0.8	
1* 8	0.9	5 (36) 1
	1	000122234567888 15
	2	25679 5
	3	468 3
	4	5 1
	5	29 2
	6	2 1
	11	4 1
	12	22 2
43		66

Fig. 6 Stem and leaf diagram of the daily global hail kinetic energy for two groups of paddays: days with and without reported damages. (*) all cases in the southwestern part of the test area.

but in that range cases with reported damages have been registered too. This fact is of special interest for further research.

Table 4 shows the percentage of days with and without damages and their E_G. On the average (1977-81) 60% of the PD reported damages and produced 95% of the hail energy whereas the remaining days (40%) of the PD (without damages) produced only 5%. The yearly amount varies between 2.9 (78/79) and 17.9% (1981) for the latter category. In 1981 one single day (6 August) was responsible for 0.98 GJ or 10% of the year's total amount. That hailstorm was measured in the SW and would have produced certainly some damages if there were any insured crops.

Year	Total number	% of total number		Total E_G (GJ)	% of total E_G	
		$PD = ID_3$	$PD \neq ID_3$		$PD = ID_3$	$PD \neq ID_3$
1977	31	58.1	41.9	43.6	96.9	4.1
1978	29	72.4	27.6	37.7	97.1	2.9
1979	20	60.0	40.0	20.7	97.1	2.9
1980	12	41.7	58.3	6.8	92.6	7.4
1981	17	58.8	41.2	9.5	82.1	17.9
77-81	21.8	60.6	39.4	23.7	95.2	4.8

Table 4 Percentage of days with ($PD = ID_3$) and without ($PD \neq ID_3$) reported damages and the corresponding global kinetic energy E_G.

4.3 Haildays per region

In an area with homogeneous climatological and topographic conditions the frequency of haildays depends on the size of the investigated area (community). The larger the surface the more hailfalls are observed. In the test area the size of the communities varies between 1 and 62 km^2 and the frequency of haildays between 0 and 9.2 per year. Thus regional differences as are supposed from Fig. 5 cannot be revealed on

the basis of single communities. To overcome that size problem the test area is subdivided in 11 quasi-equal regions in surface and number of hailpads (see Table 5). Fig. 7 exhibits the design of the 11 regions within the test area and the arrangement of the pads. The average is 70.6 km^2 in surface and 19 for the number of pads. The total frequencies (1977-81) of PD, ID_3 and ID_3* which are shown in Table 5 indicate regional differencies which can be discribed as follows.

The regions 1, 2, 4 and 5 belong to the SW; 8, 9, 10 and 11 to the NE of the test area and 3, 6 and 7 form a transition zone between. As already shown in Fig. 5 for the point frequency the total number of PD in the SW-regions are higher than in the NE whereas the ID_3 are higher in the NE, indicating a better insurance coverage there. The insurance coverage in this paper is a qualitative and not a quantitative term. To give an exact answer of the percentage of fields insured in a certain area one has to go through the single insurance contracts of

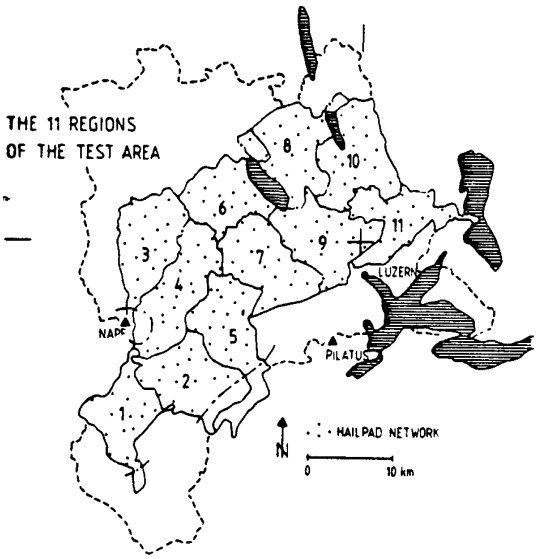


Fig. 7 Design of the 11 regions with quasi-equal surface and number of hailpads within the test area.

all farmers. This will be done in the future for some communities.

The days with reported damages but not detected by the pads (ID_3*) range from 1 day in the whole SW to 5 till 10 (region 8) in the NE. The assumption is made that the more surface the area of insured crops in a region has the more small hailfalls can be detected by the insurance which, on the contrary, might fall through the mesh of a hailpad network. The damages, however, compared with the total amount, are certainly minor. An example of a small storm passing over the region 8 (damage reported in one community), recorded by radar, is depicted in Fig. 8. The 45 and 55 dBZ reflectivity radar contours are shown. 45 dBZ is always heavy rain whereas 55 dBZ is the boundary

Region	Surface (km^2)	Number of hailpads	Total number of days (1977-81) PD	ID_3	ID_3*	
1	62	16	45	9	1	SW
2	78	16	45	11	1	
4	82	20	44	11	1	
5	83	20	41	7	1	
3	65	19	36	15	1	
6	64	19	38	22	4	
7	74	19	37	16	2	
8	68	19	31	27	10	NE
9	72	20	32	17	5	
10	68	20	32	25	8	
11	61	19	27	22	7	

Table 5 The surface, number of hailpads and the number of PD, ID_3 and ID_3* for the 11 regions. SW = southwest and NE = northeast regions of the test area.

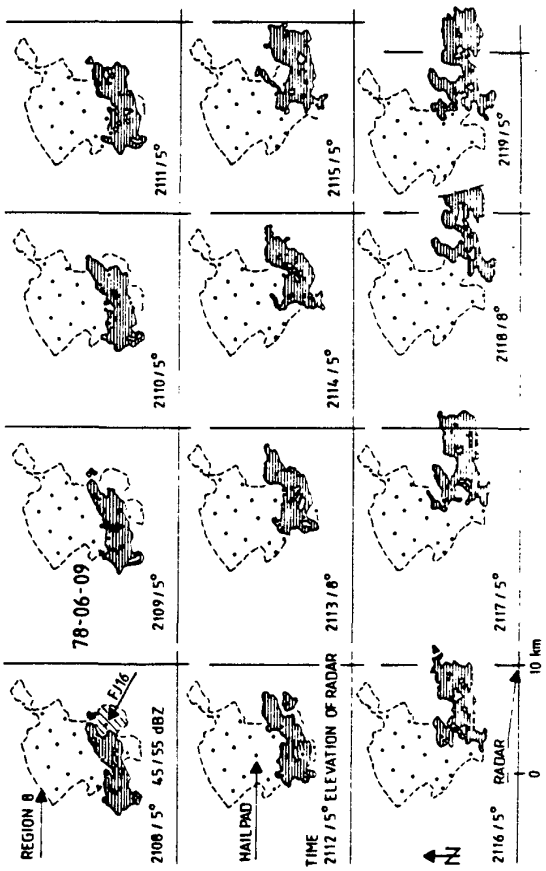


Fig. 8 Example of a small hailstorm passing through the mesh of the hailpad network without touching one hailpad. In the area marked by an arrow (FJ16) damage has been reported by the insurance but the hailpad did not detect hail. The radar contours of each minute in the time-interval 2108-2119 are 45 and 55 dBZ reflectivities.

between rain and hail on the ground (Waldvogel and Federer, 1976). This reflectivity, especially if the covered area is small, can be produced by large drops as well as by graupels or small hail. If it is little hailfall single fields might be affected but

the damages are not significant to the total of all hailfalls. The ID_3* with small hailfalls are of further interest and a quantitative value about the portion of E_G produced by them is not established yet but can be done with the help of the radar information.

Table 6 gives an idea about the percentage of PD with and without reported damages and their E_G-values for each region. In the SW on 80% of the PD no damage claims have been recorded on the ground but more energy is contributed to them (17.2-40.9%) than in the other regions with better insurance coverage. One can suppose, as shown in Fig. 6 that the 6 hailfalls between 0.3 and 1 GJ would have produced some local damage on crops if there were any. The contrary in the NE: the percentage of E_G measured on the average on 50% of the days with reported damage claims are as high as 94% of the total, leaving only 6% to the days without damage claims. Those numbers correspond very well to the ones obtained for the whole test area not making a distinction between SW and NE.

Region	5-year averages (period 1977-81)						
	PD			E_G			
	Total number	% of total number		Total E_G (GJ)	% of total E_G		
		PD ≡ ID_3	PD ≠ ID_3		PD ≡ ID_3	PD ≠ ID_3	
1	9.0	17.8	82.2	3.6	63.9	36.1	SW
2	9.0	22.2	77.8	3.5	77.1	22.9	SW
4	8.8	22.7	77.3	2.9	82.8	17.2	SW
5	8.2	14.6	85.4	2.2	59.1	40.9	SW
3	7.2	38.9	61.1	1.9	89.5	10.5	
6	7.6	47.4	52.6	1.1	90.9	9.1	
7	7.4	37.8	62.2	1.5	80.0	20.0	
8	6.2	54.8	45.2	1.5	93.3	6.7	NE
9	6.4	37.5	62.5	1.1	90.0	10.0	NE
10	6.4	53.1	46.9	1.5	98.6	1.4	NE
11	5.4	55.6	44.4	1.3	92.3	7.7	NE

Table 6 The percentage of days with (PD ≡ ID_3) and without (PD ≠ ID_3) reported damages and the corresponding global hail kinetic energy E_G for the 11 regions. SW = southwest and NE = northeast of the test area.

5. SUMMARY AND CONCLUSION

The comparison of haildays derived from a dense network of hailpads used in the Grossversuch IV experiment and the corresponding hail insurance days with reported damages yielded the following results:

1) About 10% of the insurance days reported from communities had to be judged as wrong because of a false date of the damaging hailfall.

2) The pad network detected on the average 90% of all haildays per community (average size 16 km^2) whereas the insurance days recorded only about 40%, less than half of the best estimate.

3) For the whole test area (~800 km^2) the pads detected almost all insurance days (97%) except 2 with very little damage. From the best hailday estimate the pads recorded 98% and the insurance about 60% of the days.

4) On 60% of the paddays with damage reports 95% of the global hail kinetic energy was measured, leaving only 5% to the remaining 40% of the days without damage claims.

5) A comparison with equal-sized regions in the test area revealed regional differences in hailday frequency and insurance coverage. In the area with a qualitatively good coverage, the insurance days registered more than 90% of the hail kinetic energy whereas in the area with poorer coverage only 64% was measured on the average.

The following conclusion, regarding the use of insurance data for hail climatological purposes, can be drawn:

If the focal point of interest is the "true" frequency of haildays, the insurance data provide only 20-60% of the best estimate, depending on the size and the insurance coverage of the investigated area. If one is interested in the hail kinetic energy which is an objective measure for damage on the ground, the insurance days count at least for 90% of the total energy. That is for a well insured area.

6. REFERENCES

Changnon, S.A., 1977: The scales of hail. J.Appl.Meteor., 16, 626-648.

Federer, B., A.Waldvogel, W.Schmid, F.Hampel, E.Rosini, D.Vento, P.Admirat and J.F. Mezeix, 1978/79: Plan for the Swiss randomized hail suppression experiment. Design of Grossversuch IV. Pure Appl.Geophys., 117, 548-571.

Federer, B., A.Waldvogel, W.Schmid, H.H. Schiesser, F.Hampel, M.Schweingruber, W.Stahel, J.Bader, J.F.Mezeix, N.Doras, G.d'Aubigny, G.DerMegreditchian, D.Vento, 1986: Main results of Grossversuch IV. J.Climate Appl.Meteor., in print.

Long, A.B., 1978: Design of hail measurement networks. Atmosphere-Ocean, 16, 35-48.

Mezeix, J.F., 1978: The measurement of hail at the ground level. Atmosphere-Ocean, 16, 61-68.

Mezeix, J.F. and N.Doras, 1981: Various kinetic energy characteristics of hailpattern in the Grossversuch IV experiment. J.Appl Meteor., 20, 377-385.

Waldvogel, A. and B.Federer, 1976: Large raindrops and the boundary between rain and hail. Preprint Vol.: 17th Conf.on Radar Meteorology, Seattle, 167-172.

Waldvogel, A. B.Federer and P.Grimm, 1979: Criteria for the detection of hail cells. J.Appl.Meteor., 18, 1521-1525.

205-27K 01-10756

ENV. CAN. LIBR./BIB. DOWNSVIEW
QC 980 I583 1986 c.1
Third International Conference on Stati.
0 0002 05142711 1

QC
980
I583
1986

LIBRARY
ATMOSPHERIC ENVIRONMENT SERVICE
4905 DUFFERIN ST.
DOWNSVIEW, ONTARIO
CANADA M3H 5T4

Date Due			IN LIB USE
NOV 26 1987			
88/02/11			
FEB FEV - 6 1991			
NLR 174			