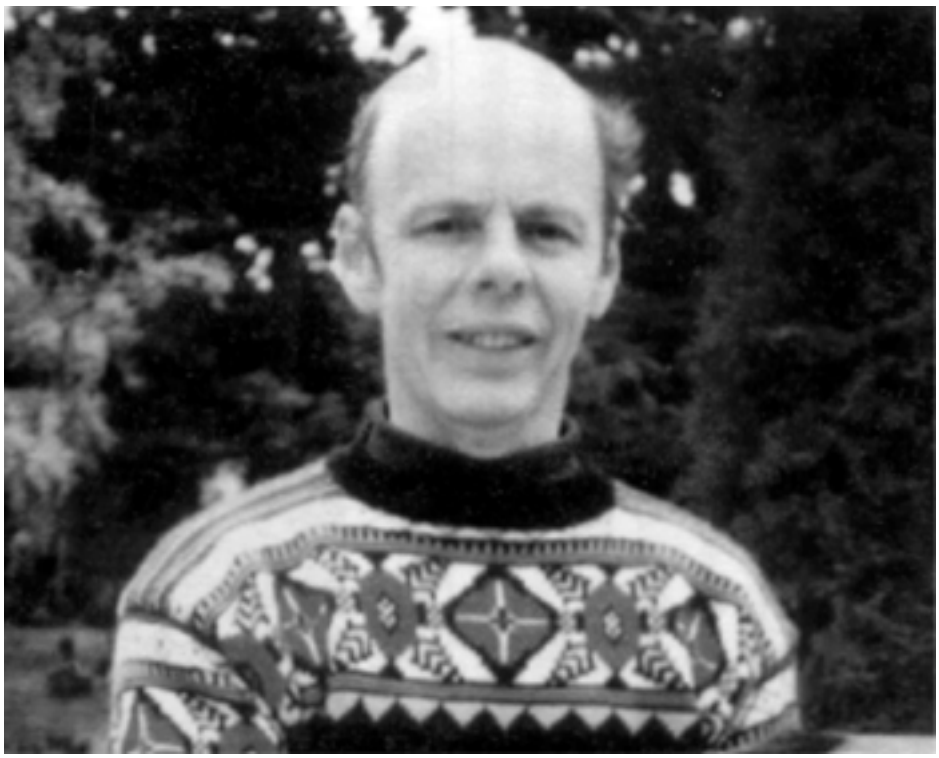
Handout at

Lüneburg,
Germany, 12 – 16 March 2001



Composed by Martin A.J. van Montfort, Eykmanstraat 16, 6706 JX Wageningen; the Netherlands.

Druck GKSS FORSCHUNGSZENTRUM Geesthacht GmbH Layout *Gardeike*

John Revfeim's contributions to IMSC

At the start of the seventh International Meeting on Statistical Climatology (7IMSC) in Vancouver (Canada) in 1998 special attention was paid to the late Allan Murphy who contributed extensively to IMSC, among others by organising 2IMSC in Lisbon (Portugal) in 1983 and being the chairman of the Steering Committee of IMSC in the period 1987–1992.

Nearly at the same time the IMSC-community lost another person with huge merits for IMSC: John Revfeim got a fatal accident by falling from a ladder while helping his neighbour in gathering the very last avocado.

At that time John was preparing his move from Cockayne Road 278, Ngaio, Wellington 6004 (New Zealand) to the place of the roots of the Revfeim family: 33 Parkvale Road, Tauranga (New Zealand), which is still the relevant address of his family. The Cockayne Road facility will be in the remembrance of many visitors interested in statistical climatology: the hospitality of John and his wife Hilary was especially available for scientists from else developed countries.

John only missed the first IMSC in Japan (1979). He was heavily interested in the continuity of the IMSC-series. He used the presence of equal minded persons at the 10th AMS Conference in Probability and Statistics in Atmospheric Sciences at Edmonton (Canada) in 1987, in order to organise a meeting to create a free standing Steering Committee (SC) to oversee the IMSC-series (Murphy & Zwiers, 1993). This meeting supported 4IMSC to be held in New Zealand (March 1989, Rotorua) with John being the chairman of the organising committee (with Steven Goulter, Craig Thompson and John Sansom as co-workers). At the same 10th AMS-CPSAS video tapes in the registration area continuously showed the attractivity of NZ between Cape Reinga and Stewart Island. John introduced the IMSC-logo. He chose Quality Inn in Rotorua as the venue; a picture of (a part of) the participants in front

of that venue is added. The choice of Rotorua will have been based on showing the variation and beauty of nature down under.

John intended to bring together more countries at the IMSC's, not only Europe and North America, but also else developed countries essentially lacking facilities to join such conferences. He found institutions for sponsoring, see the foreword of Pre-prints of 4IMSC, which enlarges the list of sponsors with 'three private sponsors from New Zealand'; personal communication revealed that these had to be found in the local organising committee! Complementary to the scientific program John promoted a broad social program during and after the congress. The after-4IMSC-program brought participants through National parks etc to Wellington; some participants were that motivated that they walked the famous Milford track on the South Island. John's energy in helping people in all walks of life was well beyond the norm. After 4IMSC John sent a detailed scenario of all steps in organising an IMSC to his successor.

In 1989 John took early retirement from the NZ Meteorological Service at Kelburn, which does not mean that he stopped looking for scientifically based solutions for really existing problems in climatology at large. His model choice combined his physical insight and mathematical skills; think about his harmonic approach of the seasonality in parameters due to the changing position of the globe relative to the sun. Only death could stop his scientific work, leaving some papers unfinished, some with relevance for hydrology and climatology. Here three unpublished papers are added, dealing with (1) presenting data to see as much as possible in one view (Modified quantiles…) and with (2, 3) modelling, say, precipitation amounts over a fixed time period as the sum of a random number of events of random size where the events can happen on points (2) on the time scale or on intervals (3) with random duration (entitled resp. Extremes of observations… and An extreme value distribution…). Van Montfort (1997) dealt with ML-estimation of the parameters in the last mentioned paper.

John's death was not only a shock for the statistical climatologists but also for a large group of New Zealand scientists knowing John over different periods of his scientific life with a broader range of interesting topics e.g. solar radiation and wind gusts. Jorgenson (1999) collected information from John's friends all over and wrote an article on John's life. This article contains a nearly complete list of his publications (Appendix 1) including his contributions to 2IMSC till 6IMSC. Here a remark is added concerning his PhD-thesis submitted in support of the application for the degree of Doctor of Philosophy at the University of Manchester (1969), 112 p, entitled 'Iterative techniques for the estimation of

parameters in time series models', with supervisor Dr. M.B. Priestley (and stimulating conversations with Dr. G. Gudmundsson and Dr. T. Subba Rao).

The merits of Kristian John Alasdair Revfeim to science and scientists mainly in statistical climatology are warmly recognised by a broad group all over the world and his ideas may be favourable after his death too. The last picture shows John presenting an IMSC-paper like he remains in our memory. His life was marked by objectivity and independence of mind, filled with achievement and lived with cheerfulness and generosity.

REFERENCES

Murphy, A.H. and F.W. Zwiers, 1993: International Meetings on Statistical Climatology. Bulletin of the American Meteorological Society, Vol. 74, No 9, September, 1721-1727.

Jorgenson, M., 1999: John Revfeim, 1938–1998*. Newsletter of the New Zealand Statistical Association, No 49, May, 6–10.

Revfeim, K.J.A.: unpublished papers*

* Modified quantiles and graded descriptors of 'average' (10 p).

* Extremes of observations bounded by limits on the size or number of events (11 p).

* An extreme value distribution with three physically meaningful parameters (CPEE), (13 p).

Van Montfort, M.A.J., 1997: Some contributions to Quantitative Hydrology. Technical Note (TN 97-10) of Dept of Mathematics (Statistics division) of Wageningen Agricultural University, with chapters on

- ML-estimation of the CPEE-parameters.

- CPEE related to GEV by L-skewness and L-kurtosis.

* Please find attached Jorgenson's obituary of John Revfeim and three unpublished papers of John he was focused on at the time of his sudden death.

John Revfeim, 1938-1998

by Murray Jorgensen

Writing an obituary is always a sad thing to do, but what makes this sadder is the knowledge that to many of those who read this what I write will be an *introduction* to John. This is not because the contributions John made to the scientific applications of mathematics and statistics have been small, but rather that John believed strongly in entering into the disciplines his statistical work served to the point where many people came to think of him as an agricultural scientist or as a meteorologist. This would not displease him, but from time to time I wish that New Zealand mathematical scientists themselves formed a bit more of a community. But John's service to the NZSA was not in the least insignificant: he was Secretary-Treasurer in 1966 and Editor of the *NZ Statistician* in 1970 and 1971. (He may have been on the executive before 1966, but I don't have information about this to hand.)

John grew up in Tauranga, spending much time on the Revfeim family orchard property there. His later education was at Nelson College and Auckland University College, from where he received his BSc (N.Z.) in 1959. There followed a period of teaching at Tauranga College and Auckland's Seddon Memorial Technical College before a return to do an MSc in Mathematics at Auckland which he received in 1962.

After a period at the Ministry of Works Auckland Laboratory John joined the Biometrics Section of the Department of Agriculture in Wellington in March 1964. He showed his musical side at that time by joining the university choir where he met Greg Arnold, who was later to join him at the Biometrics Section.

In September 1966 John, with his wife Hilary, left for the University of Manchester (Institute of Science and Technology) to undertake PhD study under Morris Walker. He returned to the Biometrics Section in later 1969. In those days computer requirements forced a centralization of statistical consulting in government science. "Biometricians" like John were based mainly in Wellington and flew off at regular intervals for several days intensive consulting at regional research stations. One of John's responsibilities was Invermay Agricultural Research Centre at Mosgiel, near Dunedin. During 1971 John suggested to Geoff Jowett, then Professor of Statistics at the University of Otago, that he consider moving to Invermay as a resident Biometrician. Whether or not John's advice was the deciding factor, Geoff Jowett did, in fact move to Invermay in 1972, where he remained for ten years, of necessity doing much pioneering work in the computer analysis of experimental data.

John himself did much work in this area. He also worked with Dr Bob Jordan, an engineer within the Biometrics Section, on 'data loggers' for automatic capture of climatic variables in the field. Bob is now with Hort Research at Ruakura. John was a great admirer of Hewlett-Packard programmable calculators, which in those days extended to large 'desktop' versions with cassette tape drives. I joined Biometrics Section myself in early 1974 and was quickly convinced of the virtues of reverse-Polish command sequence in calculators. Program memory was very scarce in these machines and John showed considerable flair for compressing seemingly complex statistical operations into very few program steps. This is but one example of John's legendary frugality which many who knew him recount. This frugality was, however, coupled with generosity.

> *"In August 1977 he and Hilary packed the Arnold family into their Edinburgh house when we arrived some days before our arranged accommodation became available. In the few weeks before John again left for New Zealand he and Hilary showed us around Edinburgh, and John demonstrated using the university reducing photocopier to squeeze the maximum amount of information onto an A4 sheet. I generally disagreed with John's firmly expressed opinions and was dubious about his economising schemes, but these little eccentricities made up a unique individual who showed me immense generosity."*
>
> - Greg Arnold

Another Research Station on John's visiting circuit was Wallaceville Animal Research Centre in Upper Hutt. Here John had Dr Ken McNatty as a client and did much work on the analysis of hormone concentrations in sheep.

In those days Biometrics Section was made up of about 8 scientists supported by about 8 computer data preparation technicians in two adjacent large rooms. John and his colleague Don Wright contributed to staff moral by organising a version of indoor cricket played with rulers and loose rubber door-stops. They also used to organise pranks with the telephones by dialing two numbers on different phones and holding the handsets together so both persons dialled would assume that the other was the caller.

In May 1976 John and his family departed for the University of Edinburgh, where John had been awarded a post-doctoral fellowship under David Finney. This was just the time when the statistical analysis computer programs and the DSIR Elliot 503 for which they were written had come to the end of their working life. John was not happy with the decision made in 1977 by both MAF Biometrics and Applied Maths DSIR to adopt the Rothamstead-written statistical package GENSTAT to replace the Elliot programs. I remember receiving a letter from him shortly after he returned from Edinburgh in 1978 in which he complained about that large amount of core memory used by GENSTAT. What he would feel about some of today's overblown "Office" suites can only be imagined! In 1979 when John moved to the Meteorological service he found an old calculator (Olivetti or WANG) which was just on the point of being discarded. John 'saved' it and wrote many programs for it. The computations for his papers in the early 80s were carried out on this machine.

In April 1979 John left MAF to join the Meteorological Service, working first in the climate section headed at the time by Jim Hessell who contributes the following reminiscences:

John had no background in synoptic meteorology and did not express any desire or interest in the subject unless it impinged directly on his work. For example, he found from extensive analysis of sunshine cards that many places experienced a greater "afternoon sunshine advantage" (his phrase) over the morning (i.e. pre solar-noon) and would inquire of the synoptic meteorologists why they though this should be so. This independence of thought was typical of his approach and proved advantageous to both parties; The synoptic meteorologists had their impressions of what occurred confirmed by unarguable statistics and he found that there was a real physical basis for his results

As above he had a great interest in sunshine statistics on both a daily and seasonal basis and was also interested in the application of the Poisson distribution to monthly rainfall figures. Many observation stations had good records of monthly rainfalls but little was known about the intra-monthly distributions. Did the rain at any place fall uniformly throughout the month or was it spasmodic? He characterised the rainfall at locales within New Zealand and overseas (notable Singapore) using the Poisson distribution which he found to be as applicable to the tropics as the middle and high latitudes. He visited the Singapore Meteorological Service on one or two occasions and had developed a good working rapport with them.

As well as producing a good volume of work in the Meteorological Service, as his published papers attest, he was very cooperative in teaching other research members of the Climate Section. People he worked with frequently were Craig Thompson, Steve Goulter and John Sansom. Steve Goulter, a well qualified statistician is now in Australia but Craig Thompson and John Sansom are still in Wellington - at NIWA which absorbed the Climate Section on a reorganisation which took place in 1989/90. Craig Thompson produced several publications using Poisson analyses Through his earlier position in the Ministry of Agriculture and Fisheries he was able to help the agricultural meteorologist.

John looked a the methods of archiving climate data used by the Climate Section over many years and pointed out many ways in which they could be simplified and the volume of stored

data drastically reduced. His ideas led to discussions on the "the entropy of information". Most of these discussions were not well appreciated by most of us who had a background in synoptic meteorology rather than statistics.

Although John worked largely on his own he was by no means isolated from the rest of the section and was generally a congenial working partner. Occasionally I would ask him to do some special task such as writing the Introduction to a rainfall statistics publication containing gamma parameters to explain the gamma distribution. Tasks such as these he undertook willingly. He was the instigator of a system of collecting climate data to a daily basis (Dlycli).

John was active on the international scene also through his involvement with the International Statistical Climatology Association. He attended IMSC meetings in Spain and Ireland and organised the 4th Meeting in Rotorua in 1989. Because of these and also because of his papers published in international journals, he was in some ways, better known abroad than at home ; I expect that he will be remembered, and referred to, for many years to come.

Among those he worked with at home he will be remembered by all his colleagues and assistants not only for his productive work but also for his kindness, helpfulness and for his puckish, sometimes devastating, sense of humour.

However John's enquiring mind did not always fit in well with any prevailing official model for how public science should be conducted (and, I am certain, he would have found university red tape no more to his taste.) There is a tale of him spending many months and many letters pursuing a claim through the Met Service bureaucracy for a refund on an anemometer that he purchased during a visit to an Indian Met. research station. John had hoped that NZ would by many more of the devices from India, as they were considerably cheaper than those from other sources. This was just one of many conflicts between John and officialdom. J. T. Steiner, who headed the Research Division at the Meteorological Service in the 1980s recalls:

John was one of the few scientists that I have known who was capable of contributing new ideas. Most of the rest of us merely test the ideas of others , embellish them a bit or apply them to a different problem.

John's capability and work methods thus did not sit well in the model of research proposals, approval and monitoring which I and other research managers were introducing to the science departments of the old Public Service. I think he found my insistence on this model as tedious. Had I been a wiser manager, I might have found some way of extricating him from this model by classing him differently from the others. I left the Met Service before he did I believe that his decision to retire at 50 was at least partly to his disaffection with the increasing requirement for paperwork that was not really relevant to his contribution.

As Jim Hessell points out John was "in some ways, better known abroad than at home". Perhaps it was a desire for better recognition in New Zealand that led him to submit a collection of 23 of his papers, published for the most part in Agricultural and Meteorological journals, to VUW for the degree of DSc. This application was unsuccessful. David Vere-Jones comments:

I felt at the time that his papers on reflectivity provided a valuable and practically useful basis for agricultural procedures, but that for a D.Sc in statistics, a more substantial statistical component would be necessary. John here was in the typical problem situation of someone who works in a cross-disciplinary field but finds it hard to get their papers fully accepted by either side.

Skimming through the papers I am struck by the blend of Agriculture and/or Meteorology with physical science, classical applied mathematics and statistics. This is very much what I remember of John's interest and approach, although I have to confess that my own

background in pure mathematics did not make me very attracted to John's work at the time. Martin van Montfort of the Mathematics Department of Wageningen Agricultural University in the Netherlands has compiled a bibliography of John's publications (which is incomplete at least in respect of some joint articles arising from his collaborations with Wallaceville scientists.) I attach Martin's bibliography as an appendix, the first 23 papers listed constitute, in a slightly different order, the papers submitted for the DSc. As another appendix I include John's introduction to the papers which serves also as a broad indication of John's interests and philosophy at the time.

John's move to the Met. Service did not lead to him quickly cutting his ties with Agricultural Science. He served as Chairperson of the organising committee of an NZ Institute of Agricultural Science convention in 1983 and was Secretary/Treasurer of NZIAS from 1984 to 1986.

In 1989 John took early retirement and set about the establishment of 278 Cockayne Road, Ngaio as a research centre of some note by working on the completion of numerous partly-written papers in his 'bottom drawer'. However with the departure of his youngest son for Medical School in Otago, he felt free to take up the position of Principal Scientific Officer (Climatology) with the Fiji Met. Service, a position that he held between August 1991 and July 1994.

John was a very keen gardener, especially of fruit and vegetables, which allowed him to express both the frugal and the generous sides of his character. After inheriting the family orchard near Tauranga, John's love of gardening caused him to divide his time between Ngaio and Tauranga. Sadly an ironic full circle was closed when an accidental fall while harvesting avocados in his orchard caused his death in April 1998, leaving his wife Hilary, and three adult sons. His insight and wit will be missed by all those who worked with him be they Statistician, Agricultural Scientist, or Meteorologist.

I wish to thank all the many people who sent me information about John, even if I did not directly quote them. I know that I should have got this completed earlier but I hope that now, appearing on the first anniversary of John's death it will serve as a memorial to him.

I have interrupted my study leave to prepare this note and I like to think that John would have been interested in the research work that I have been undertaking. John had always been scornful of statistics that were quoted to many more significant figures than the precision implied by their standard errors. I have just been studying Minimum Message Length (MML) inference with the aid of its founder Emeritus Professor Chris Wallace of Monash University in Melbourne. MML sees excessive precession in estimates as a form of overfitting. In a 1991 note John refers to a 'magnum opus' on entropy moldering in his 'bottom drawer' (surely as much a part of his mind as a part of his desk!) It turns out that 'message length' and entropy are more or less the same thing, and clearly related to John's love of storing the maximum of information in the least physical space. Perhaps I would eventually have seen my ideas converge with John's in this instance, at least!

Appendix 1

Martin van Montfort's Bibliography of John Revfeim's publications.

List of publications and concepts(*) of (KJAR:) Kristian John Alasdair Revfeim (1938-1998). B.Sc(NZ),Dip.Math.Stats,Ph.D.(Manchester).

- KJAR and R.B.Jordan: Precision of evaporation measurements using the Bowen ratio. Boundary layer meteorology,10, p97-111,1976.
- Solar radiation at a site of known orientation on the earth's surface. Journal of applied meteorology,15(6),p651-656,1976.
- A simple procedure for estimating global daily radiation on any surface. Journal of applied meteorology,17(8),p1126-1131,1978.
- Maximisation of global daily radiation on sloping surfaces. New Zealand Journal of science,22,p293-297,1979
- Reviewing agricultural research. New Zealand Agricultural research,14(3),p123-126,1980.
- KJAR and J.W.D.Hessell. Observations and implications of diurnal asymmetry in "bright" sunshine hours. New Zealand Journal of Science,24,p153-160,1981.
- Estimating solar radiation income from "bright" sunshine records. Quaterly Journal Royal Meteorological Society,107(452)p427-435,1981.
- Those illusive decimal points. Weather and Climate,2,p4-8,1982.
- The feasibility of estimating solar radiation flux distributions from 'bright' sunshine data. New Zealand Journal of Science,25,p1-13,1982.
- Simplified relationship for estimating solar radiation incident on any flat surface. Solar Energy,28(6),p509-517,1982.
- Seasonal patterns in extreme 1-hour rainfalls. Water Resources Research, 18(6),p1741-1744,1982.
- Comments "On the study of a probability distribution for precipitation totals". Journal of Applied Meteorology,21(12),p1942-1945,1982.
  Corrigendum and addendum,ibid.,22,p502,1983.
- An interpretation of the coefficient of the Angstrom equation. Solar Energy,31(4),p415-416,1983.
- Stochastic process analysis of rainfall totals and extremes. Proceedings 2nd International Meeting on Statistical Climatology, Lisbon, p10.2.1-10.2.5,1983.
- On the analysis of extreme rainfalls. Journal of Hydrology,62,p107-117,1983.

- KJAR and H.S.Hughes. Physically meaningful parameters that characterise rainfall totals and rainfall extremes. New Zealand Journal of Science,26,p443-445,1983.

- Generating mechanisms of, and parameter estimators for, the extreme value distribution. Australian Journal of Statistics,26,p151-159,1984.

- The cumulants of an extended family of type I extreme value distributions. Sankya,46,p281-284,1984.

- KJAR and J.W.D.Hessell. More realistic distributions for extreme wind gusts. Quaterly Journal Royal Meteorological Society,110,p505-514,1984.

- Is the "100-year-flood" interpreted correctly ? Journal of Hydrology(NZ),23,p4-9,1984.

- An initial model for the relationship between rainfall events and daily rainfalls. Journal of Hydrology,75,p357-364,1984.

- The analysis of maximum wind gusts by direction. New Zealand Journal of Science,27,p365-367,1984.

- Drought prediction from rainfall records.

- A note on the comparison of theoretical and empirical quantiles for monthly rainfall totals. Atmosphere-ocean 23(4) 1985,414-419.

- C.S.Thompson and KJAR. Comment on " Homogeneity analysis of rainfall series- an application of the use of a realistic rainfall model " Journal of Climatology,vol.5,579-581(1985).

- Proverbial outliers. Proc. Pacific Stats Congress 1986 (Elseviers), p296-298.

- Iterative forms in numerical integration. The New Zealand Mathematics Magazine. (1986?) 38-40

- Extracting information from rainfall data.3rd International Meeting on Statistical Climatology, Vienna, p382-387,1986.

- Daily observations: necessity, ritual or imposition ? International Journal of climatology,vol.9,1-6(1989).

- A framework for interpreting rainfall models. Contributed paper 4th International meeting on Statistical Climatology, Rotorua, 1989.

- Approximation for the cumulative and inverse gamma distribution. Statistica Neerlandica (1991)327-331.

- Annual maxima and totals of seasonally varying processes. Stochastic Hydrol.Haudraul.5(1991) 147-153.

- Dominant events in extreme rainfall records.Journal of Hydrology,134(1992)143-149.

- Confronted with Clicom. Proceedings 5th International Meeting on Statistical Climatology, Toronto, p395-396,1992.
- Significant harmonics in distribution parameters of the seasonal process. Proceedings 6th International Meeting on Statistical Climatology, Galway(Ireland), p171-173,1995.

* Identification and interpretation of three bounds occuring in extreme rainfal records.(18pp, 1983)
* A small sample distribution for the m-th largest value (16pp, 1985) [added as an appendix to : Martin A.J. van Montfort: Inference on the maximum based on top-ith and top-i data. Technical Note TN 93-06,Dept. of Mathematics WAU, 21ppp,1993]
* S.W.Goulter and KJAR. Extreme value parameters estimated from short records of ranked observations (9pp,1986)
* Lower bounds, upper bounds and bounded intervals in the analysis of maxima.(10 pp, 1990?)
* Mean exponential statistics: meaningful measures or trivial pursuit ? (5pp,1990)
* Directional and seasonal analysis of wind gusts (7pp, 1991 ?)
* A theoretically derived distribution for annual rainfall totals.(1991?)
* Improved methods for prediction of extreme wind speeds (8 pp,1991 ?)
* An extreme value description with three physically meaningfull parameters (July 1996). [Partially added as an appendix to: Martin A.J. van Montfort: ML-estimation of the CPEE-parameters. Technical Note TN 97-10, Dept of Mathematics WAU,1997.]
* Extremes of observations bounded by a limit on the size of an event or a limited number of events (6 pp, 1989) Extremes of observations bounded by limits on the size and the number of events (10pp, 1998)
* Modified quantiles and graded decriptors of 'average'.(14pp, 1998).

Appendix 2

Introduction to "Some physical, mathematical and statistical properties of environmental observations" by K. J. A. Revfeim (Collected papers, 1976-1985)

In contrast to laboratory experiments with biological material, trials carried out 'under field conditions' involve more than the treatments or varieties in the experimental design. Under field conditions the natural environment includes not only the physical and chemical properties of material below ground level but also the properties of the atmosphere near the

ground. Responses to the natural environment reflect a sequence of effects which are most rapidly varying in the atmosphere.

Some physical properties below ground level, notably moisture and temperature, are subject to seasonal variation. Observed seasonal variation during the course of long-running field trials are often used to qualify deductions from the trail. Diurnal variation is barely within the sphere where it can be related to biological responses.

Seasonal patterns in the physical environment above ground level show more extreme variation between years than below surface patterns. Hence any attempt to make absolute measures of productivity from field trials must take account of the relevant characteristics of phenomena that are used to represent the atmospheric environment (where such phenomena affect the biological response).

Routine monitoring of phenomena in the atmospheric environment such as temperature, rainfall, wind and radiation, has led to vast collections of data. Such data have traditionally been the subject of empirical analyses which, apart from the more obvious seasonal or diurnal patterns, are mostly characterised by the properties of the numbers collected e.g. mean, variance, skewness. These statistical parameters are weak representations of the actual physical processes or events which give rise to the data.

There is considerable scope for developing quite realistic models recognising the broad physical principles underlying environmental processes or events. Statistical moments, such as average and variance, of point observations or cumulative data will be functions of parameters of these models. Thus while a two-parameter model provides no more characteristics than the number properties average and variance, the estimated parameters using average and variance will have a definite meaning. That is with respect to the model the parameters have a physical interpretation as to their expected effect in some other biological response model.

Hence the characterisation of data in physically meaningful terms is a necessary first step to the profitable use of that data.

This collection of papers represents an attempt to put in context the place of environmental data in agricultural research, and to make some realistic assessments, of the scales on which data are measured and the precision to which derived parameters are estimated. Some particular interpretations of data are made which show

a) how radiation on sloping surfaces can be estimated from observations on horizontal surfaces;

b) how radiation can be estimated from sunshine data;

c) how total and intense rainfalls can be characterised as a recurrence process;

d) how wind gusts can be characterised as a recurrence process;

and,

e) how variable data bounds are of only slightly less value than admissible observations.

Modified quantiles and graded descriptors of 'average'

K.J.A. REVFEIM
278 Cockayne Road, Wellington 6004, New Zealand

Summary. In the indefinite sense, *average* implies some middle part of the scale which may be observed by physical measurement, interview assessment, or visual judgement. For comparative purposes particular observations on the scale may be qualified by expressions such as 'above average', 'well below average', 'just' or 'slightly' above/below (*the*) average. Depending on context the indefinite term *average* can have a quite different meaning from *the average*. For a symmetric distribution function $F(x)$ portions $p < 1/2$ and $q = 1 - p$ are simply related through probabilities $P = p,\ Q = q$ to quantiles $F^{-1}(P),\ F^{-1}(Q)$ equidistant from the average. For asymmetric distributions the essential connection between portion p and the average is retained through a relationship with probability $P' = p + 8p^2q^2(P_m - Q_m)$ where P_m is probability at the average ($Q_m = 1 - P_m$). This quartic (p, P) relationship satisfies points $(0, 0), (1/2, P_m), (1, 1)$ and is tangential at the ends. Modified quantiles $F^{-1}(P')$ centred on the mean are called 'queantiles'. A framework of expressions is proposed for modified quintiles and the 17,33,45,55,67,83 'perceantiles' (percent queantiles). This framework may be useful for consistency in reports issued to the media, or for publication in journals, so that the listener or reader has a clearer understanding of the information given.
Key words: quantile, percentile, queantile, perceantile, near-, about-, above-, below-average.

1 Introduction

The term 'average' is widely used in official, economic, educational and scientific reports but only in scientific reporting is the interpretation reasonably clear because in most cases one particular value is described as '*the* average'. Where an amount is loosely qualified as 'around average' there is as much uncertainty as to the extent of 'around' in the mind of the speaker or writer, as the intended listener or reader. Expressions such as 'above average' may be based on (sub)multiples of standard error separation, or be some undefined departure from average. Qualifiers such as 'slightly above' or 'greatly below' are used to reduce or extend the separation from the average. There is even a hint of contradiction in using great-(large)-ly to refer to what may be a small value. In attention demanding headlines, or publicity seeking interviews even more emotive qualifiers, such as 'vastly above' or 'sadly below', may be used for emphasis but this still does not convey a clear sense of departure from an undefined 'average'.

There is further confusion with terminology in climate statistics where the term *normal* is a particular average *viz* the most recent 30 year average completing a natural decade (currently 1961-90, next update 1971-2000). Following the media release of a monthly weather summary it is common to hear or read statements such as "rainfall was 30% above normal" or "only 50% of normal" , with the clarifying word *the* omitted before "normal". Neither of these situations is rare and the interpretation by the general public of *normal* is probably 'about average', or what occurs most often, so that 30% above or 50% below might appear to be noteworthy differences. Use of this particular definition of *normal* in a public statement can be meaningless considering that for climate elements,

with typical symmetric or positively skewed frequency distributions, 40-50%/50-60% of short term averages or totals will be above/below normal.

There is some margin of error in data due to observer or observed, and *the average* is only an estimate. Hence caution is usually applied when describing summary statistics with respect to *average* although there is no universally accepted threshold for distinguishing from the estimated average by simple terminology. Most professionals use statistical methodology for estimating significant differences from the average. Even greater uncertainty applies to the many qualifiers that are used to give degrees of departure above or below average. In some cases two or more words have shades of meaning with no clear difference e.g. just, barely, narrowly. Other words are more applicable (but not exclusively) to one or other side of average e.g. exceedingly, greatly, grossly, highly; a little, minutely. Yet other qualifiers infer some desirable or undesirable outcome e.g. acutely, disastrously, sufficiently. Of course there is a temptation after extreme events, or for highly unusual observations, to make comparisons with average that command the reader or listener's attention. However in general the more emotive the adverb used the less well the true meaning is understood.

It therefore seems to be a useful exercise, for particular threshold crossings and subintervals of the observation scale, to propose a framework of descriptors

- with a natural gradation of meaning that is applicable above and below average;
- that ties in with standard methods for the statistical description of data;
- with a matching colour code.

2 Distribution asymmetry and queantiles

We assume that the statistical concept of a frequency distribution is understood. To briefly summarize and establish notation, the relative frequency $f(x)$ of observations defined on the range of values of x, takes the cumulative form $0 \leq F(x) \leq 1$ called the (probability) distribution function (d.f.). The value of x when $F(x) = 1/2$ is called the median, usually denoted by $\tilde{x}$ for a sample. More general inverses of the d.f. for probability P are called quantiles $x_P = F^{-1}(P)$. As probabilities are often given as integer valued percentages these particular inverses are called percentiles. Certain subdivisions of $0 < P < 1$ are commonly used in descriptions of data, in particular quartiles, quintiles and deciles for 4, 5 and 10 equiprobable intervals respectively.

For symmetric relative frequencies $f(x)$ the average (or mean value) is the same as the median but this is not the case for asymmetric distributions. Hence if we wish to use the probability scale in word descriptions of 'average' (usually denoted by $\bar{x}$ for a sample) it is necessary to centre the probability classifications on $P_m = F(\bar{x})$. A somewhat crude but simple way to define equivalent thresholds below and above the average is to subdivide the respective probabilities P_m and $Q_m = 1 - P_m$ proportionately i.e. a linear fit

$$F(x_{p1}) = 2pP_m, \qquad p \leq 1/2 \tag{1}$$

$$F(x_{p1}) = 2pQ_m + P_m - Q_m, \qquad p > 1/2 \tag{2}$$

As the relationship between the portion p and $F(x)$ is fixed at three points (0,0), (1/2,P_m) and (1,1) an exact quadratic fit is possible

$$F(x_{p2}) = p + 2pq(P_m - Q_m) \tag{3}$$

where $q = 1 - p$. If we add the condition that the relationship should also be tangential to the quantiles, $F(x_{p0}) = p$, at the extremes then we can fit an exact quartic

$$F(x_{p4}) = p + 8p^2q^2(P_m - Q_m) \tag{4}$$

The mapping of the portion space p onto the probability space P is shown in Fig. 1 for $P_m = 3/4$.

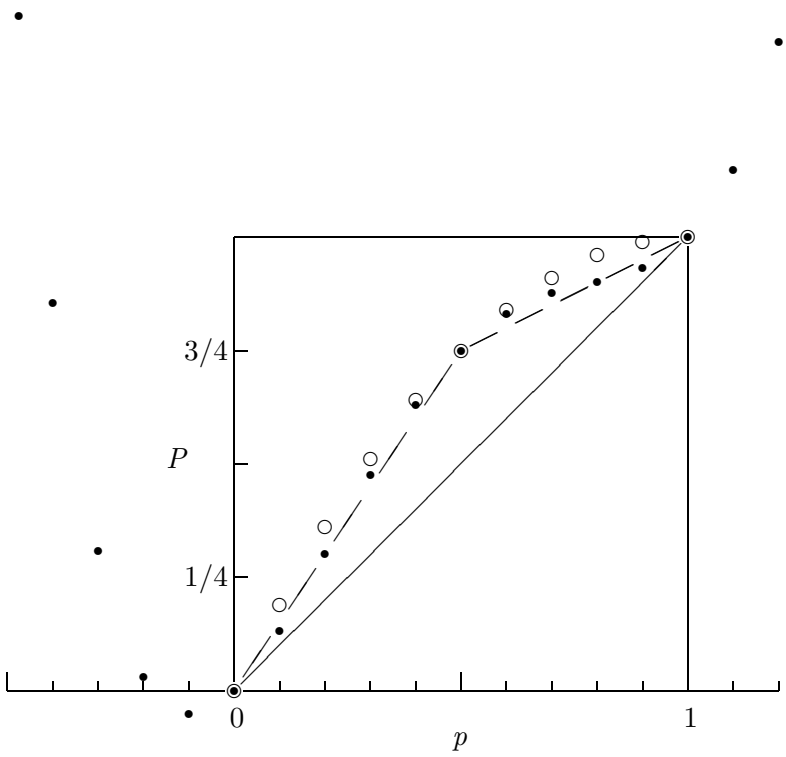


Fig.1. Mapping of portion (p) space onto probability (P) space for probability at the average $P_m = 3/4$. Quantiles (——) and linear (— —)/quadratic(∘ ∘ ∘)/quartic(• • •) relationships.

For the very asymmetric negative exponential distribution $F(x) = 1-\exp(-x)$ the unit mean is beyond the 60 percentile and median value is 0.69. The comparison between equi-probable departures (x_{p0}) from $P = 0.5$ (at the median) and equi-proportional (linear) departures (x_{p1}) around $P_m = 0.63$ (at the mean) are shown in Table 1. Also shown in Table 1 are the inverses of eqns (3) and (4), and the x_{p0}, x_{p4} for the Chi-square distribution with 1 degree of freedom and the arc sine distribution $F(x) = (2/\pi)\sin^{-1}(2x/\pi)$. It is obvious that the x_{p4} are the best modification of quantiles with greatest symmetry about the mean and rapid approach to quantiles near the ends of the range. We propose that the name 'queantiles' be given to these modified quantiles centred on the mean, or more specifically 'quartic queantiles'.

Table 1. Quantile ($p0$) and queantile ($p1, 2, 4$) thresholds for the negative exponential (unit mean), Chi-square (1 d.o.f.) and arcsine distributions.

Portion	p	1/6	1/4	1/3	2/5	9/20	1/2	11/20	3/5	2/3	3/4	5/6
Expl.	x_{p0}	0.18	0.29	0.40	0.51	0.60	0.69	0.80	0.92	1.10	1.39	1.79
	x_{p1}	0.24	0.38	0.55	0.71	0.84	1.00	1.10	1.22	1.41	1.69	2.10
	x_{p2}	0.27	0.43	0.61	0.74	0.86	1.00	1.14	1.29	1.52	1.88	2.36
	x_{p4}	0.23	0.39	0.57	0.73	0.86	1.00	1.13	1.27	1.47	1.73	1.79
χ^2_1	x_{p0}	0.05	0.10	0.19	0.27	0.36	0.46	0.57	0.71	0.94	1.32	1.88
	x_{p4}	0.08	0.21	0.41	0.62	0.80	1.00	1.21	1.43	1.73	2.10	2.56
Arc sine	x_{p0}	0.41	0.60	0.79	0.92	1.02	1.11	1.19	1.27	1.36	1.45	1.52
	x_{p4}	0.36	0.52	0.68	0.81	0.90	1.00	1.09	1.24	1.30	1.42	1.50

For any value of P_m the quartic has turning points at $p_t = (3 \pm \sqrt{3})/6$, or 0.21/0.79, beyond which there is a smooth approach to a line of unit slope. A restriction on the quartic fit is that the mapping must be 1:1 i.e. eqn(4) can only have one maximum or minimum which must lie outside $0 < p < 1$. In other words the cubic differential of (4) with respect to p

$$\partial F/\partial p = 1 - 16pq(p-q)(P_m - Q_m) \tag{5}$$

must have only one real root and two imaginary roots. The boundary of this condition is zero slope of the quartic at the turning points and substituting p_t for p in (5) we get $P_m = (8 \pm 3\sqrt{3})/16$. That is the probability at the mean must lie within $0.175 < P_m < 0.825$ which is of little practical consequence. Alternatively, for a sample of size n, the rank of the value nearest to the mean must lie within $(0.175n, 0.825n)$. A theoretical example of a distribution where the quartic queantile does not give a 1:1 mapping is the beta distribution $f(x) = \alpha(\alpha+1)x^{\alpha-1}(1-x)$, defined on (0,1). If $\alpha = 2^{-k}$ is a binary fraction the mean is $1/(2^{k+1}+1)$, and from $F(x) = x^\alpha(1+\alpha-\alpha x)$ we find that the median tends to $\mathrm{e}^{-1}2^{-2^k}$. As seen in Table 2 P_m exceeds the critical level at $k = 4$.

Table 2. Values of P_m and scaled median for beta(2^{-k},2) distribution.

k	1	2	3	4	5	6
P_m	0.68	0.71	0.78	0.85	0.91	0.94
$2^{2^k}e\tilde{x}$	1.46	1.22	1.07	1.03	1.02	1.01

Consider a ranked sample of n values $x_1, \leq x_2 ..... \leq x_m\ (\leq \bar{x}) \leq x_{m+1} .... \leq x_n$ with m values up to or below the mean (average) $\bar{x}$. For any portion p $(0 < p < 1,\ q = 1-p)$ the sample linear queantile is $x_{i_{p1}}$ where

$$i_{p1} = [2mp] + 1 \qquad p \leq 1/2 \tag{6}$$

$$i_{p1} = [2(mp+nq)] - m \qquad p > 1/2 \tag{7}$$

where [..] represents the integer part of the argument. However because of the obvious superiority of the quartic queantile we use m/n as a distribution free estimate of P_m and from (4) we get the rank queantile for portion p as x_{i_p} where

$$i_p = [pn + 8p^2q^2(2m-n)] + 1 \tag{8}$$

In particular (x_{i_p}, x_{i_q}) are values in the ranked sample that are separated from the mean fairly symmetrically, in comparison with the linear queantiles which are separated in

proportion to the number of x_i below and above $\bar{x}$. For selected values of p these paired thresholds form zones that tie in with words commonly used to describe closeness to or departure from average.

Obviously for symmetric distributions queantiles are equivalent to quantiles. Either from fitting an assumed distribution, or simply from an identified value in the ranked sample, queantiles are easily obtained. The sample properties will be the same as for the equivalent quantiles ζ_P (Cramer, 1946, **28.5**) i.e. the sample queantile x_{i_p} is asymptotically Gaussian distributed $(\zeta_P, \sqrt{PQ/n}/f(\zeta_P))$. To simplify the terminology in the following section we use the same names for quantiles at equal subdivisions of probability and for the equivalently portioned queantiles i.e. terciles, quartiles, quintiles, deciles, bideciles.

3 Statements

With the probability scale appropriately subdivided above and below the *average*, associated statements for degrees of departure from average can be considered. Descriptions with respect to average appear to fall into three classes of intended meaning; low sensitivity including as much around average as seems reasonable (conservative statements), being fairly objective with separation from average (neutral statements), or conscious of small departures from the average (high sensitivity, discriminatory statements). In purely statistical terms the three classes might be seen as applying to large, medium or small coefficients of variation.

The proposal is to use selected queantiles to define thresholds or intervals that might be associated with words or expressions commonly used to describe departure from average to varying degrees.

Neutral statements, equal steps and intervals

The quintiles are often used in graphic or map presentations to show a broad brush picture of a set, or sets, of observations. Changes with time can also be represented with this statistical measure without diverting attention with unnecessary detail. The descriptor 'about average' is suggested for the central quintile or 40-60 perceantile range. This is a smaller probability region than the central tercile, and in the ranking of word pictures 'about' can be more tightly confined than 'around' used in the following section for the 33-67 perceantile range. The inner quintiles (20-40 and 60-80 perceantile ranges) are 'clearly' below and above average (i.e. not in contact with) so that is the natural qualifier. The outer quintiles can simply and unemotively be described as being 'notably' above or below (worthy of being noted) to complete a set of neutral statements with respect to average.

Conservative statements, decreasing steps, inclusive intervals

The word 'average' in the indefinite sense is interpreted as 'around average' as distinct from the alternatives 'above average' or 'below average'. Thus it is natural to use a three part subdivision of the probability scale with a change to above- or below average at the terciles (33 and 67 perceantiles). Emphasis is given to departure from average by adding 'well' (as an adverb, sometimes 'much') as in 'well above average'. The alternative to this state is being 'above', but not 'well above', which naturally divides the probability intervals into two parts at the outer sextiles (17 or 83 perceantiles). In terms of the bell-shaped Gaussian distribution this is very near one standard deviation from the mean

(16 or 84 percentile). Amounts within these intervals might be described as (deficit), insufficient, sufficient, more than sufficient, (surplus). This understanding of sufficiency places the boundaries of deficit/surplus at a realistic distance from the middle region of adaptability.

Further emphasis is occasionally given by adding 'very' to the optional adverb 'much' as in 'very much below' and it is natural that this would again halve the probability at the outer duodeciles (8 or 92 perceantiles). However there is potential confusion because the negative statement 'not very much' means a small amount. At this level of departure from average it is more common to make word comparisons with extremes so we simply recognise the possibility but exclude it from the proposed framework. Anyway in this region the statistical descriptors bottom- or top-deciles (within 0-10 or 90-100 percentiles) are available if required.

The quantile thresholds can be described in terms of the logistic $L = \ln[P/(1-P)]$. The outer terciles, sextiles and duodeciles are approximately the same increment in L of 0.7, 0.9 and 0.8 respectively from the median where $L = 0$. Values of the logistic are shown in Table 3 and there is a near linear increase in L between the thresholds of interest.

Table 3. The logistic L at various quantiles with associated Gaussian deviates as binary (sub)multiples of s.d. σ.

	($\approx \sigma/8$) central	central	($\approx \sigma/4$) c.	($\approx \sigma/2$)		($\approx \sigma$)	outer	outer	($\approx 2\sigma$) outer
Quant.	quadridecile	bidecile	quintile	tercile	quartile	sextile	decile	bidecile	quadridecile
% tiles	(47.5/52.5)	45/55	40/60	33/67	25/75	17/83	10/90	5/95	(2.5/97.5)
L $\pm(\approx)$	0.1	0.2	0.4	0.7	1.1	1.6	2.2	2.9	3.7
Diffce	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	

In Table 3 the logistic given for outer bi- and quadrideciles is calculated at the 5.2/94.8% and 2.4/97.6% subdivisions respectively to maintain the linearly increasing difference. The latter is also near the 2 standard deviation (2.3/97.7%) thresholds for the Gaussian distribution. Parentheses are used for percentages at quadrideciles since integer values only apply to percentiles. The linear trend in differences for the selected quantiles, and symmetry about the 25/75 percentiles, is a satisfying coincidence. This is due to the first three terms in the series expansion $L \approx \ln 3 \pm 16\delta/3 + 32\delta^2/9$ for $P = 3/4 \pm \delta$.

Sensitive statements, increasing steps, exclusive intervals

The preceding classifications do not give a sense of closeness to the average which is sometimes required by the writer or speaker. In this case the further from the average the less important the comparison. The ranking of 'just', 'slightly', 'moderately' and 'considerably' makes them natural to append to above or below the average. Thresholds above the mean with linearly increasing intervals that match these descriptors could be 50,55,65,80,100. However it would be difficult to establish any statistically significant difference between estimated 35/65 perceantiles and the 33/67 terciles. Hence alternate thresholds 50,56,67,83,100, with linear increases in the first three steps, are a better choice also linking with the upper sextile. Setting the 17-33 (25±8) and 67-83 (75±8) perceantiles as 'moderately below/above' the average, and with the link to 'below/above' average (but not 'well below/above') conservative statements, makes this alternative a better choice.

Observing in Table 3 that the logistic at the central bidecile (45/55 percentiles) fits into the pattern of linearly increasing thresholds this is selected as the natural boundary

between 'just-' and 'slightly' below/above the average. From a practical point of view only two additional thresholds to those already specified, the 45 and 55 perceantiles, require tabulation. Thus we have the sensitive classification as shown in Table 4. The central 45-55 perceantile range can easily be described as 'near average' which fits in with the degree of closeness implied by near-, about-, or around average or the indefinite term 'average' by itself.

Table 4. Thresholds between sensitive qualifiers of 'the average'.

above	50		55		67		83		100
		just		slightly		moderately		considerably	
below	50		45		33		17		0

4 Colour coding

For printed or screen display material it is usual to present current or recent observations grouped by colour. By this method monthly weather summaries can show at a glance where regions are well below or above average, perhaps indicating a requirement for administrative action, dampening demand, or some form of social support.

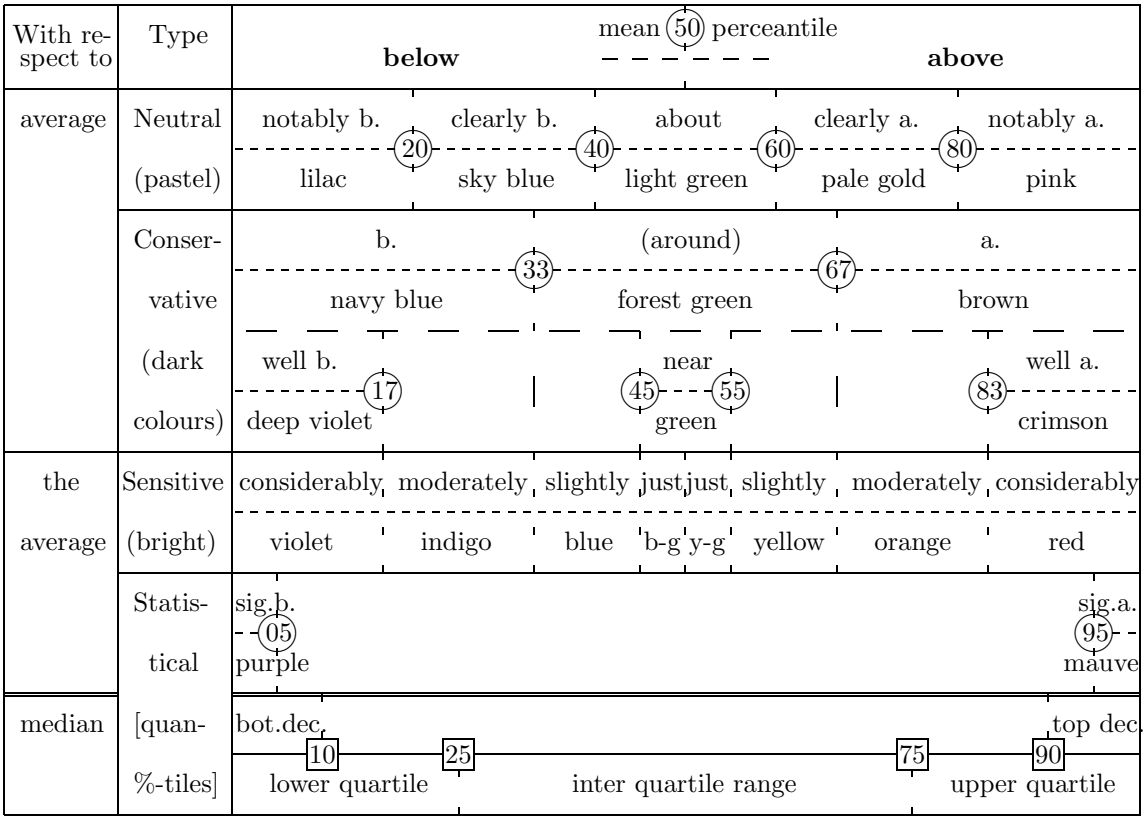
Suggested colours that match the spectral ranking of the descriptors are shown in Table 5 which is also a convenient summary of the statement types. It seems appropriate to equate neutral statements with pastel shades, conservative statements with dark hues, and the seven discriminatory intervals with the full bright colour ROYGBIV spectrum. Standard specifications of colours appropriate for the various screen drivers and printing systems may eventually be produced.

While the order of the colours shown in Table 5 places the red end of the spectrum at the upper end of the observation scale (violet at the lower end) there is no reason why this cannot be reversed. The reverse order is traditionally applied to rainfall data because of the association between red/orange and aridity or dryness (low rainfall).

5 Discussion

Application of the proposed standard descriptors to particular observations or outcomes depends on the availability of tables of queantile thresholds for every variable that is to be described. In general this information is available for its current use in less precise qualification with respect to 'average'. Current use may be based on 'rules of thumb' or less informative statements with respect to 'the average'. There is no difficulty in calculating queantile thresholds from representative data sets.

Table 5. Statement types and descriptors (sig.=significantly) of (the) average (a.=above, b.=below) for queantile intervals with suggested colours (reversible, b-g=blue green, y-g=yellow green) and selected quantiles (bot.=bottom, dec.=decile).



In the educational field assessments of a student are often given by numerical ranking 1 to 5 or alphabetically by A to E. Letters A-E may be further divided into two groups such as B+, B-. These scales can be associated with objective or neutral statements so that it seems natural to set grades 1-5 or A-E+/- by the modified quintiles or deciles respectively. Passes qualified by 'honours' or 'distinction' (*summa cum laude, magna cum laude*) fit into the upper sextiles. This equates 'honours' to above average (but not well-above) and 'distinction' to well above average. The conservative classification with inclusion matches the descriptors (lowest) lower, middle, higher (highest). Though little used now the traditional ranking of bad, very poor, poor, fair, good, very good, excellent fits into the basic sensitive classification. Satisfaction and superiority may be likened to the quantiles. All these classifications are shown in Table 6.

Of course all the above classifications are perceantiles to be translated into actual marks from an assumed distribution or sample rank. In essence this is a method for scaling marks by inverting (4) or (8) to give a portion which can be mapped back onto some prescribed mean.

Actual thresholds for higher passes can be set for a prescribed pass mark, P_m, by applying the modified probability proportionately to the marking scale. For example London Royal Schools of Music (RSM) practical examinations are marked out of 150 with a prescribed 'pass' mark of 100 (i.e. $P_m = 2/3$). From Table 6 and eqn(4) the 'honours' threshold would be $P(\text{honours}) = 194/243$ or 120 marks (which it is) and 'distinction' at $P(\text{distinction}) = 215/243$ or 133 marks (compared with 130 for RSM). By

way of comparison Trinity College (TC) music exams, marked out of 100, require 65 for a pass. In this case the queantile thresholds for 'honours' and 'distinction' would be 78 and 88. The actual thresholds of 75 for 'merit' and 85 for 'distinction' are close to the (neutral) 60 and 80 perceantiles for $P_m = 0.65$. One could say that the TC 'merit' and 'distinction' reflect the upper (modified) quintiles, appropriately with 'distinction' being not(e)ably above average.

Some care needs to be taken in calculating quantile or queantile thresholds for seasonal data. Quarterly or monthly observations within the annual cycle are routinely archived and characterised within quarters or months as distinct data sets. However while the continuity of the underlying process may be reflected in the seasonal pattern of averages this may not be so with the outer thresholds of queantiles or quantiles. Adverse events at particular times can affect the variability by a greater amount than the length of history can accomodate as being representative. This disturbance to the expected pattern of queantiles can be smoothed out by characterising all the seasonal data jointly (Revfeim, 1995). Either by joint fitting of distributions (smoothed seasonal parameters) and derived queantiles, or smoothed fitting of individual queantiles (assuming asymptotic distribution form), unexpected disturbances in the pattern can be avoided.

Table 6. Grades and queantile thresholds of educational achievement.

Neut.		5		:		4		:		3		:		2		:		1							
$\bar{\%}$tile				(20)				(40)				(60)				(80)									
	E-	:	E+	:	D-	:	D+	:	C-	:	C+	:	B-	:	B+	:	A-	:	A+						
Cons.					fail						:					pass									
$\bar{\%}$tile							(33)				(50)			(67)											
			lower				:				middle			:			upper								
	lowest		:														:		highest						
Sens.																									
$\bar{\%}$tile			(17)				(33)		(45)		(55)		(67)			(83)									
		bad	:	v.poor	:	poor	:	fair	:	good	:	v.good	:	excellent											
		inferior		[25]		unsatisfactory		[50]		satisfactory		[75]		superior											

References

Cramer, H.(1946) *Mathematical Methods of Statistics.* Princeton University Press.

Revfeim, K.J.A.(1995) Significant harmonics in distribution parameters of the seasonal process. *Proc. 6th Intl. Mtg. Statist. Climatology* (Ed. I. O'Muirheartaigh), pp. 171-173. Univ. College, Galway.

Extremes of observations bounded by limits on the size or number of events

K.J.A. Revfeim
278 Cockayne Road, Wellington 6004, NZ

Abstract Extreme value distributions are most often expressed in standardised scale/offset form of the argument $y = \alpha(x-u)$. For non-negative measurements it is natural to simply scale x and apply some transform to give shape. Power transforms, x^ν ($\nu > 1$), of measured values may give more realistic frequency distributions over the whole range of observations recognising the non-zero modality of data. Maxima of events occurring as a Poisson process, with beta or Pareto distributed event sizes, have a distribution similar to the so-called Generalised Extreme Value (GEV) distribution. This analogue of GEV leads to the inner exponential limit of the Type I extreme value (EV1) distribution $\mathrm{e}^{-\mathrm{e}^{-y}}$. Binomially distributed event numbers with exponentially distributed event sizes are an alternate generalisation of EV1 which is easily recognised as leading to the outer exponential of $\mathrm{e}^{-\mathrm{e}^{-y}}$. These bounded compound distributions, appropriate for environmental observations, are discriminatory for identifying mixtures, classifying outliers and making more reliable predictions of long term maxima, because the parameters have a physical meaning.

KEYWORDS Upper bounds, Generalised Extreme Value, Compound Poisson Exponential, Probable maximum precipitation.

1 Introduction

Many analyses of extremes are primarily concerned with the upper tail of some statistical distribution. This concentration on the apparent properties of large values, without great concern for the whole frequency distribution, may reduce the possibility of identifying a mixture of processes that could underly observed maxima. That is a mixture may be forced to fit an empirical extreme value model which is not appropriate for either or any component. While this method may give simple engineering and building design criteria a more representative model could reduce extreme risks and possibly save money spent on over-design, especially in situations where maxima come close to their physical limits.

Jenkinson's(1955) extreme value distribution (often called *the* GEV) may be written with scale multiplier α in the form

$$G(x) = \exp\{-[1-\kappa\alpha(x-u)]^{1/\kappa}\} \tag{1}$$

where $-\infty < x < u+1/\kappa\alpha$ for $\kappa > 0$ and $u+1/\kappa\alpha < x < \infty$ for $\kappa < 0$ (parameter nomenclature has been chosen to coincide with Gumbel(1941) in the case $\kappa = 0$). Estimation of the parameters of (1) is not an assured exercise (Otten and van Montfort, 1980; Hosking *et al*, 1985; Buishand, 1986). The mode of the frequency distribution corresponding to (1) is $u + [1-(1-1/\kappa)^\kappa]/\kappa\alpha$, which for small κ is approximated by $u + \kappa/\alpha$, and as κ tends to zero we approach the EV1 distribution $\exp\{-\mathrm{e}^{-\alpha(x-u)}\}$ with mode u.

Many environmental observations are measured on scales with smallest value zero and some physically constrained upper limit which is intuitively understood if not known absolutely. If wind gusts, rainfall amounts within fixed time intervals, river flood levels, earthquakes etc. are assumed to have no upper limit they are conceptually as well as environmentally unattractive. Thus one might infer that eqn(1) with $\kappa > 0$ would be the more appropriate model to represent environmental data maxima. However negative values of κ seem to predominate in published examples of data fitting where the GEV could be considered as an approximation to distributions based on physics.

Table 1: Slopes at origin, upper bounds (u.b.), and parent families of sub-limiting exponential distributions.

Symbol		E_-	E	E_+	
κ	-1/2		0		1/2
$\mu^2 f'(0)$	-3/2		-1		-1/2
u.b.		∞	∞	μ/κ	
Family		Pareto	Exponential	beta	

In the derivation of the EV1 and GEV distributions the observations are assumed to be made at equal discrete intervals and the theoretical distribution is an asymptotic result. Epstein(1949) showed that assuming a Poisson process in continuous space or time, with independent exponentially distributed event sizes, generates a Compound Poisson Exponential (CPE) distribution of maxima that is equivalent to the EV1 distribution. It was emphasized that this distribution was "exact and not merely asymptotic". However assuming an underlying exponential distribution of event sizes is not realistic when considering the relative frequency of wind gust magnitudes or wave heights which have a pronounced non-zero mode. Generalising the exponential form to a Weibull distribution (Revfeim, 1984) retains the simple prediction formula while accommodating the underlying modality.

Pickands(1975) showed that the distribution function

$$F(x) = 1 - (1 - \kappa x/\mu)^{1/\kappa}, \tag{2}$$

was the limiting form for excesses over thresholds as the threshold becomes 'large'. He advocated (2) as being useful for estimating the probabilities of extremes from a sample of highest ranked values. The three distinct shapes of the density

$$f(x) = (1 - \kappa x/\mu)^{1/\kappa - 1}/\mu \tag{3}$$

are convex with respect to the origin ($\kappa < 1/2$), concave ($1/2 < \kappa < 1$), and J-shaped ($\kappa > 1$). For $\kappa > 0$ (3) represents the density of a beta-distribution with support $(0, \mu/\kappa)$. As shown in Fig. 1 all densities start from ordinate $1/\mu$ at $x = 0$ with the distinct shapes separated by straight line decay (triangular) distribution at $\kappa = 1/2$ and the uniform (rectangular) distribution at $\kappa = 1$.

Only convex densities are of practical interest for maxima. For $\kappa < 0$ the Pareto distribution defined on $(0, \infty)$ has mean $\mu/(1+\kappa)$ and variance $\mu^2/(1+\kappa)^2(1+2\kappa)$ which are both finite if $-1/2 < \kappa < 0$. Thus eqn (2) only represents a realistic distribution of event magnitudes over a restricted range $-1/2 < \kappa < 1/2$, as shown in Table 1. These sub-limiting forms of the exponential distribution, belonging to the Pareto and beta families respectively, are designated by E_- and E_+ for $-1/2 < \kappa < 0$ and $0 < \kappa < 1/2$ respectively.

The densities for E_- and E_+ are sometimes referred to as 'fat-tailed' and 'thin-tailed' respectively. Changes in the thickness of the tail are reflected in the probability at the mean $P_m = 1 - (1+\kappa)^{-1/\kappa}$ and the coefficient of skewness $g_1 = 2(1-\kappa)(1+2\kappa)^{1/2}/(1+3\kappa)$ as shown in Table 2.

Table 2: Probability at the mean and skewness of sublimiting exponential distributions.

κ	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
P_m	0.75	0.72	0.70	0.67	0.65	0.63	0.61	0.60	0.58	0.57	0.56
g_1	-	-	16.4	4.6	2.8	2.0	1.5	1.2	0.9	0.7	0.6

In the literature all three forms seem to be included in the name 'generalised Pareto distribution'. For $|\kappa| < 1/2$ a better name would be generalised exponential considering its sub-limiting

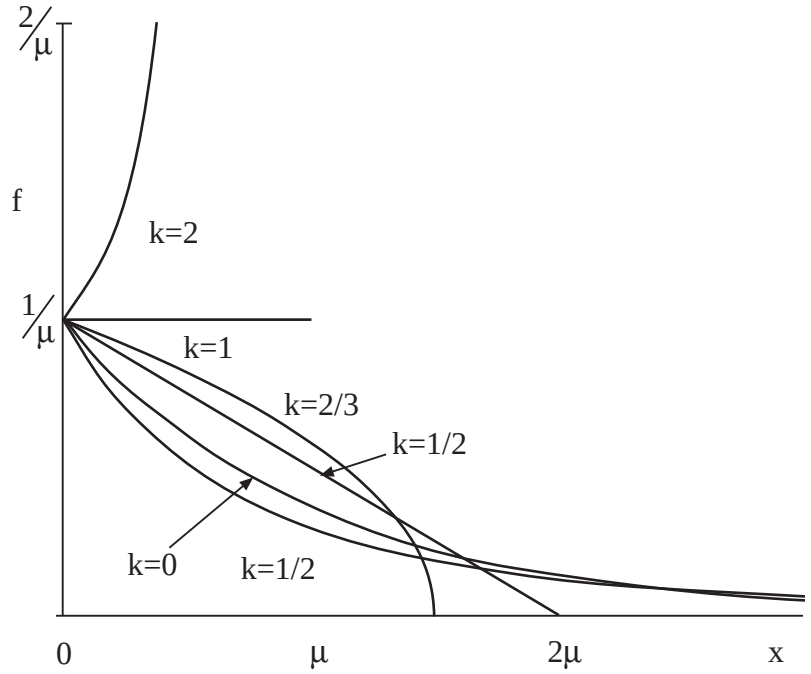


Figure 1: Sub-limiting exponential and related densities $f(x) = (1 - \kappa x/\mu)^{1/\kappa - 1}/\mu$.

exponential form. While the forms of $f(x)$ remain Pareto or (elementary) beta for $|\kappa| \geq 1/2$ neither would seem to have much practical application because of their shapes.

A distribution of event sizes $F(x)$ given by (2) is readily incorporated into Compound Poisson models of extremes. In particular for E_+ we have a finite range that might be appropriate for many environmental measurements. A power transform on x introduces a fourth parameter that is essential if the elementary beta model is to be representative of the whole frequency distribution and not just an approximation to the upper tail.

The Fisher-Tippet (1928) and Gumbel (1941) derivations of EV1 were an asymptotic form for unlimited sample number. The Compound Poisson analogue of the EV also assumes that an unlimited number of events is possible in a finite time interval. A binomial count is the obvious choice of distribution for a limited number of events and this gives an alternate generalisation with or without restriction on event sizes. A double restriction is imposed for binomially distributed event counts and beta distributed event sizes. These models will be developed and estimation methods considered, using re-analysis of data sets previously treated in the literature to measure worthwhile improvements.

2 Models

The maxima of a Poisson process of events at rate ρ per unit time have distribution function

$$G_P(x) = (\exp\{-\rho[1 - F(x)]\} - \mathrm{e}^{-\rho})/(1 - \mathrm{e}^{-\rho}) \tag{4}$$

where $F(x)$ is the underlying distribution of event sizes. Where ρ is moderate to large, say $\rho > 5$, the term $\mathrm{e}^{-\rho}$ can be ignored. In the sequel the distribution of the number of events could change from Binomial over Poisson to Negative Binomial; the distribution of the size of the events could change from beta over exponential to Pareto, where beta and exponential could be expanded to gamma and Weibull resp.

2.1 *CPE$_\pm$ (≈GEV), CPWeibull, CPGamma*

Assuming the distribution of event sizes given by (2) with $-1/2 < \kappa < 1/2$ we get a Compound Poisson Pareto/beta (CPE$_\pm$) d.f.

$$G_{PE_\pm}(x) \approx \exp\{-\rho(1-\kappa x/\mu)^{1/\kappa}\} \tag{5}$$

which is equivalent to (1) with $\mu = 1/\alpha + \kappa u$ and $\rho = (1+\kappa\alpha u)^{1/\kappa}$.

For ρ relatively large it is easy to derive the characteristic function

$$\varphi(s) = \int_0^{\mu/\kappa} \mathrm{e}^{isx} dG(x) \approx \mathrm{e}^{is\mu/\kappa} \sum_{r=0}^{\infty} (-is\mu/\kappa\rho^\kappa)^r \Gamma(1+r\kappa)/r!$$

and from $\ln\varphi(s)$ we get the first four cumulants

$$\begin{aligned} k_1 &= (-\mu/\kappa)[\rho^{-\kappa}\Gamma(1+\kappa) - 1] && (6)\\ k_2 &= C^2 D_2 && (7)\\ k_3 &= C^3(D_3 - 3D_2) && (8)\\ k_4 &= C^4(D_4 - 4D_3 + 6D_2 - 3D_2^2) && (9) \end{aligned}$$

where $C = -\mu/\kappa\rho^\kappa\Gamma(1+\kappa)$ and $D_r = \Gamma(1+r\kappa)/\Gamma^r(1+\kappa) - 1$. From (7)-(9) we get the coefficients of skewness $g_1(\kappa) = (3D_2 - D_3)/D_2^{3/2}$ and kurtosis $g_2(\kappa) = (D_4 - 4D_3 + 6D_2)/D_2^2 - 3$. Substituting (6) in (5) we get the probability at the mean $P_m = \exp[-\Gamma^{1/\kappa}(1+\kappa)]$ which equals $\exp(-\exp(-\gamma)) = 0.5704$ for $\kappa = 0$. Probability weighted moments (PWM's, Hosking *et al*, 1985) are given by

$$\beta_r = \mu\{1 - \Gamma(1+\kappa)/[\rho(1+r)]^\kappa\}/\kappa(r+1)$$

leading to L-skewness and L-kurtosis from definitions of Hosking (1990) as

$$\begin{aligned} \tau_3 &= 2(1-3^{-\kappa})/(1-2^{-\kappa}) - 3\\ \tau_4 &= 5(2^{-\kappa} - \tau_3) - 4 \end{aligned}$$

All these measures of skewness and kurtosis are shown in Table 3.

Table 3: Measures of skewness and kurtosis for CPE$_\pm$ distribution.

κ	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
P_m	0.73	0.69	0.66	0.63	0.60	0.57	0.54	0.52	0.50	0.48	0.46
g_1	-	-	19.2	3.53	1.91	1.14	0.69	0.27	-0.07	-0.36	-0.63
τ_3	0.53	0.45	0.38	0.31	0.24	0.17	0.11	0.05	-0.01	-0.06	-0.11
g_2	-	-	-	45.1	7.8	2.4	0.57	-0.12	-0.29	-0.14	0.24
τ_4	0.42	0.35	0.26	0.19	0.16	0.15	0.12	0.11	0.11	0.10	0.09

It is obvious from Table 3 that the probability at the mean (P_m) reflects the changing asymmetry just as well as the specific measures of skewness, with cross-over from positive to negative skewness near $\kappa = 0.3$. However g_1 is too sensitive for $\kappa < 0$ to be considered a reliable statistic for parameter estimation. The time t to first exceedance of a given threshold, X, has an exponential distribution with mean $T_X = 1/\rho[1 - F(x)] = (1-\kappa X/\mu)^{-1/\kappa}/\rho$.

As stated in the introduction many frequency distributions of event sizes are not realistically represented by (3) and the form $f_{W+}(x) = \nu x^{\nu-1}(1-\kappa x^\nu/\mu)^{1/\kappa-1}/\mu$ gives more flexibility. For integer values of $1/\kappa$ it is not difficult to find values of ν that impart zero skewness to $f(x)$. Obviously $\nu = 1$ for the uniform distribution ($\kappa = 1$) and for $1/\kappa = 2, 3, 4, 5, 10, 20$ we get

$\nu = 1.74, 2.16, 2.43, 2.61, 3.05, 3.31$ respectively. Density $f_{W+}(x)$ is the sub-limiting form of the Weibull density underlying the Compound Poisson Weibull (CPW) model for extremes (Revfeim, 1984). Where $\nu > 1$ the mode moves away from $x = 0$ which is intuitively more suitable for the underlying distribution of events such as wind gusts and wave heights.

Where $\kappa > 0$ the density function (3) is an elementary case of the more general beta distribution $f_{G+}(x) = x^{\eta-1}(1-\kappa x/\mu)^{1/\kappa-1}/\mu^{\eta}B(\eta, 1/\kappa)$. In the limit as κ tends to zero we get the gamma density $f_G(x) = x^{\eta-1}\mathrm{e}^{-x/\mu}/\mu^{\eta}\Gamma(\eta)$ underlying the Compound Poisson Gamma (CPG) distribution for extremes. While the Gamma and Weibull densities are not dissimilar (for $\nu < 4$), statistical properties and parameter estimation are more difficult for the CPG distribution (Revfeim, 1984) than for CPW (a simple power transform on CPE). Hence the bounded density f_{W+} is a preferable generalisation to f_{G+}.

2.2 *CBinE, CBinE$_+$*

For a binomially distributed number of events with unknown upper limit N, at mean rate of occurrence $\rho = Np$, and event sizes with distribution function $F(x)$, the maxima have d.f.

$$\begin{aligned} G_B(x) &= \sum_{r=1}^{N} \binom{N}{r} p^r q^{N-r} F^r/(1-q^N) \\ &= \{(pF+q)^N - q^N\}/(1-q^N) \\ &= \{[1-\rho(1-F)/N]^N - (1-\rho/N)^N\}/\{1-(1-\rho/N)^N\} \end{aligned} \tag{10}$$

where $q = 1-p$. Assuming exponentially distributed event sizes with mean μ, and if $(1-\rho/N)^N$ is negligible, we get the Compound Binomial Exponential (CBE) d.f. (Revfeim, 1989)

$$G_{BE}(x) \approx \{1 - \rho\mathrm{e}^{-x/\mu}/N\}^N \tag{11}$$

which for large N tends to $G_{PE}(x)$. Corresponding to $\kappa < 0$ in (5) we get for $N < 0$ an extreme value distribution generated by the negative binomial distribution (van Montfort & Otten, 1991). Again however it is difficult to perceive many applications for this form since it imposes no restriction on the range.

The inverse of (11) gives $x = \mu(\ln\rho - \ln N + \sum_{t=1}^{\infty} G^{t/N}/t)$ from which we get the mean and variance

$$k_1 = \mu(\ln\rho - \ln N + s_N) \tag{12}$$

$$k_2 = \mu^2\{\sum_{t=1}^{\infty}[N/t^2(N+2t) + 2N(s_{N+2t} - s_t)/t(N+t)] - s_N^2\} \tag{13}$$

$$k_3 = \mu^3\{\int_0^1 (\sum_{t=1}^{\infty} G^{t/N})^3 dG - s_N^3\} - 3\mu s_N k_2$$

where $s_N = \sum_{t=1}^{N} 1/t$. Substituting k_1 in (11) we get the probability at the mean $P_m = (1-\mathrm{e}^{-s_N})^N$. The PWM's are given by

$$\beta_r = \mu\{\ln\rho - \ln N + s_{(r+1)N}\}$$

As L-skewness is dependent on ρ we use a simpler measure (favoured by Greenwood *et al*, 1979) of M-skewness $M_{102}/M_{101} = (\beta_2 - \beta_1)/(\beta_1 - \beta_0) = s_{3N}/s_{2N} - 1$. Some numerical values are shown in Table 4.

If both the number of events and size of event are conceptually constrained by some unknown limits then it is natural to further generalise to the four parameter Compound Binomial beta (CBE$_+$) model

$$G_{BE_+}(x) \approx [1 - \rho(1-\kappa x/\mu)^{1/\kappa}/N]^N, \quad (0 < x < \mu/\kappa)$$

Table 4: M-skewness for the CBE distribution.

N	1	2	3	4	5	6	7	8	16	32
P_m	0.632	0.604	0.593	0.587	0.584	0.582	0.580	0.579	0.575	0.573
M_{102}/M_{101}	0.222	0.176	0.155	0.142	0.133	0.126	0.121	0.117	0.099	0.085

2.3 *Product of independent CPE's*

Many data sets fitted by EV1 or GEV models are not homogeneous but mixtures of seasonal and/or prevailing weather processes with varying rates of occurrence, and magnitudes, of events. An appropriate d.f. for such mixtures can be constructed from the parameters of the component processes (Revfeim, 1991) and the cumulants are

$$\begin{aligned} k_1 &= (\mu/\alpha)[\lambda + C_{\mu\mu}(J_2 + \lambda^2)/2\alpha^2] && (14)\\ k_2 &= (\mu/\alpha)^2[J_2 + C_{\mu\mu}(J_3 + 2\lambda J_2)/\alpha^2] && (15)\\ k_3 &= (\mu/\alpha)^3[J_3 + 3C_{\mu\mu}(J_4 - J_2^2 + 2\lambda J_3)/2\alpha^2] && (16)\\ \text{where } J_r &= (-1)^r \int_0^\infty (\ln\theta + \gamma)^r \mathrm{e}^{-\theta} d\theta \end{aligned}$$

$\lambda = \ln\rho + \gamma$ ($\gamma = 0.5772$ is Eulers constant), and $\alpha = 1 - C_{\rho\mu} + C_{\mu\mu}$ ($C_{\rho\mu}$ is the relative covariation of the ρ's and μ's in the component processes, and $C_{\mu\mu}$ is the squared coefficient of variation of the μ's). These coefficients are assumed small enough for second order terms to be ignored and from (15),(16) it is easy to show that the coefficient of skewness of a CPE mixture is

$$g_{\text{mix}} \approx g_0\{1 + 3C_{\mu\mu}[(J_4 - J_2^2)/J_3 - J_3/J_2]/2\alpha^2\}$$

with g_0 the coefficient of skewness in case of homogeneity where $C_{\mu\mu} = 0$. Using $J_2 = 1.645$, $J_3 = 2.404$, $J_4 = 14.611$ we get $(J_4 - J_2^2)/J_3 - J_3/J_2 = 3.49$. Thus the effect of a mixture is to make the skewness larger than g_0 so that attempts to fit a GEV model to a mixture of true CPE/EV1 processes will reduce the magnitude or may give negative values of κ when the components of the mixture may have positive κ.

It may also be noted that the simplest mixture of two CPE/EV1 processes (Rossi *et al,* 1982) with parameters $\rho \pm \epsilon$, $\mu \pm \delta$ has exact coefficients $C_{\rho\mu} = -\epsilon\delta/\rho\mu$ and $C_{\mu\mu} = \delta^2/\mu^2$.

3 Parameter estimation

3.1 *Moments*

One could consider a two steps procedure, where the first step consists of estimating κ; in the second step μ and ρ are estimated given the result of the first step.

Some estimators for the shape parameter κ could be considered. In the function $P_m = \exp\{-\Gamma^{1/\kappa}(1+\kappa)\}$, P_m could be replaced by an estimate $\hat{P}_m$ in order to get $\hat{\kappa}_0$. An increasingly ordered sample results in an average ($\bar{x}$) and in m with $x_{(m)} < \bar{x} < x_{(m+1)}$. Then P_m could be estimated by m/n or by a minor modification of it: $\{m + (\bar{x} - x_{(m)})/(x_{(m+1)} - x_{(m)}\}$. Note that the support of m is $\{1, 2, \cdots, n-1\}$; so m does not follow a binomial distribution. The ratio R, with $R = \text{var}(m)/\{mP_m(1-P_m)\}$ depends on the distribution of $(x-\mu)/\sigma$; for Student's t_ν var(m) increases with decreasing ν. This ratio R also depends on κ in case of CPE$_\pm(\kappa)$; simulated (10^4 simulations) R-values as a function of κ and of the sample size n are given in Table 5.

The numerical solution of κ as a function of P_m is straight forward. For this purpose the function could be transformed to $\ln\{\Gamma(1+\kappa)\}/\kappa + z = 0$ where $z = -\ln(-\ln(P_m))$.

Table 5: Simulated R-values as a function of κ and sample size n in case of $\mathrm{CPE}_{\pm}(\kappa)$

κ	-0.5	-0.4	-0.3	-0.2	-0.1	0	+0.1	+0.2	+0.3	+0.4	+0.5
$R(n=25)$	0.862	0.735	0.617	0.519	0.443	0.395	0.372	0.357	0.356	0.355	0.354
$R(n=100)$	1.358	1.054	0.752	0.569	0.467	0.410	0.374	0.356	0.348	0.349	0.352

Simplification could be based on approximations, e.g.

$$\begin{aligned} \kappa_0 &= [27 - 10(39P_m - 15)^{1/2}]/19 \qquad (17) \\ P_m &\approx \exp(-e^{-\gamma}) - \frac{\pi^2}{12}\exp\{-(\gamma + e^{-\gamma})\}\kappa \\ &= 0.57038 - 0.26335\kappa \\ \kappa &\approx -3.7972(P_m - 0.57038) \end{aligned}$$

In order to get an idea about the quality of estimating κ by means of P_m with interpolated m and exactly inverting the (P_m, κ)-relation, this estimator is compared to PWM and ML focussing on the standard error (se); for results, see Table 6. This table shows that for common sample sizes PWM and ML are nearly equivalent, and that both are superior compared to P_m-based estimation of κ. The P_m-based estimate of κ will be denoted by κ_0.

Table 6: Comparison of estimation procedures of the shape parameter (κ) (P_m-based, PWM, ML): se-values at $\kappa = -0.2(0.1)0.2$ and $n = 36, 100$.

		$\kappa=-0.2$	-0.1	0	+0.1	+0.2
$n=36$	P_m-based	0.197	0.191	0.191	0.193	0.196
	PWM	0.150	0.140	0.131	0.125	0.123
	ML	0.160	0.154	0.147	0.143	0.139
$n=100$	P_m-based	0.117	0.117	0.115	0.117	0.121
	PWM	0.091	0.083	0.077	0.073	0.073
	ML	0.088	0.083	0.078	0.074	0.070

From the sample L-skewness $t_3 = 2(3b_2 - b_0)/(2b_1 - b_0) - 3$ where $b_0 = \bar{x}$, $b_1 = \sum_{i=2}^{n}(i-1)x_{(i)}/n(n-1)$, $b_2 = \sum_{i=3}^{n}(i-1)(i-2)x_{(i)}/n(n-1)(n-2)$ we get an estimate

$$\kappa_1 = 2 - (47 + 100t_3)^{1/2}/4 \qquad (18)$$

Hosking *et al* (1985) proposed an alternate estimate $\kappa_1 = 7.86y + 2.96y^2$
where $y = 2/(t_3 + 3) - \ln 2/\ln 3$. Using κ_0 or κ_1 we obtain moment estimates of ρ, μ from

$$\begin{aligned} \rho_0 &= [\Gamma(1+\kappa)(1+z^{-1})]^{1/\kappa} \\ \mu_0 &= \kappa\bar{x}(1+z) \end{aligned}$$

where $z = (2b_1/\bar{x} - 1)/(1 - 2^{-\kappa})$ for PWM's and $z = (v/D_2)^{1/2}/\bar{x}$ for ordinary moments (v is sample variance and $D_2 = -1 + \Gamma(1+2\kappa)/\Gamma^2(1+\kappa)$).

3.2 *Maximum likelihood*

The notation of maximum likelihood (ML) procedures is simplified if we use $c = 1/\kappa$. More importantly for numerical purposes c is of similar order of magnitude to ρ and μ. The log-likelihood is

$$L = n(\ln\rho - \ln\mu) + (c-1)\sum_{i=1}^{n}\ln(1 - x_i/c\mu) - \rho S_c$$

where $S_k = \sum_{i=1}^{n}(1 - x_i/c\mu)^k$. If we assume that c (i.e. κ) is known, or estimated by a method other than ML then, noting that $\partial S_k/\partial\mu = k(S_{k-1} - S_k)/\mu$, the ML equations to be equated to zero are

$$\begin{aligned} \partial L/\partial\rho &= n/\rho - S_c \\ \partial L/\partial\mu &= [(c-1)S_{-1} - nc - c\rho(S_{c-1} - S_c)]/\mu \end{aligned}$$

This demonstrates a numerical advantage of the $\mathrm{CPE}_{\pm}(\rho,\mu)$ parametric form over $\mathrm{GEV}(u,\alpha)$ as one parameter ($\hat{\rho} = n/S_c$) can be eliminated by simple substitution and the ML procedure reduces to a single equation to get an estimate of μ:

$$(c-1)S_{-1}S_c - ncS_{c-1} = 0 \tag{19}$$

Based on $1 - \kappa x/\mu > 0$ for all x one gets $\mu > \kappa x_{(n)}$ for $\kappa > 0$ and $\mu < \kappa x_{(1)}$ for $\kappa < 0$ where $x_{(1)}$ and $x_{(n)}$ are the smallest and largest observation resp.

From the expected values of the second order partial differentials we get the (co)variances (c.f. CPE structure in square brackets [..])

$$\begin{aligned} \mathrm{Var}(\rho) &= \rho^2(1 + q^2/d_2)/n \qquad [\rho^2(1 + \lambda^2/\zeta_2)/n] \\ \mathrm{Cov}(\rho,\mu) &= -\kappa\rho\mu q/nd_2 \qquad [-\rho\mu\lambda/n\zeta_2] \\ \mathrm{Var}(\mu) &= \kappa^2\mu^2/nd_2 \qquad [\mu^2/n\zeta_2] \end{aligned}$$

where $p = \rho^{\kappa}\Gamma(1-2\kappa)$, $q = 1 - p$, $\zeta_2 = \pi^2/6$ and $d_2 = p^2 D_2(-\kappa)$ is the essential factor in the determinant of the information matrix.

Full 3-parameter ML parameter estimation has been discussed by several authors (Jenkinson, 1955; Prescott & Walden, 1980; Otten & van Montfort, 1980; Smith, 1985). However as above the $\mathrm{CPE}_{\pm}$ parametric form allows the reduction to a 2-dimensional search. Noting that μ and c are interchangeable in $1 - x/c\mu$ the third ML equation is

$$\partial L/\partial c = (\mu/c)(\partial L/\partial\mu + n/\mu) + T_{01} - \rho T_{c1}$$

where $T_{jk} = \sum_{i=1}^{n}(1 - x_i/c\mu)^j \ln^k(1 - x_i/c\mu)$. Eliminating $\partial L/\partial\mu$ and substituting $\rho = n/S_c$ we get

$$S_c(n + cT_{01}) - ncT_{c1} = 0 \tag{20}$$

which together with (19) can be used for the usual iterative search for $\hat{\mu}, \hat{c}$. This 2-dimensional search should be a more assured process, finding the true maximum, without the possibility of settling on a 'local' maximum that may occur with a 3-dimensional search. Finally with the additional expected values of second order partial differentials with respect to c we get the (co)variance structure

$$\begin{aligned} \mathrm{Var}(\rho) &= \rho^2[(d_2 + q^2)(c\lambda^2 + \zeta_2) - (q\lambda + p\delta)^2]/nd_3 \\ \mathrm{Cov}(\rho,\mu) &= -\rho\mu[c(q\lambda^2 + pq\delta - d_2\lambda) - q\lambda^2 + q\zeta_2 - p\delta\lambda]/cnd_3 \\ \mathrm{Var}(\mu) &= \mu^2[c^2d_2 + (c-1)\lambda^2 + 2cp\delta + \zeta_2]/c^2nd_3 \\ \mathrm{Cov}(\rho,c) &= \rho c(pq\delta - d_2\lambda)/nd_3 \\ \mathrm{Cov}(\mu,c) &= -\mu(cd_2 + p\delta)/nd_3 \\ \mathrm{Var}(c) &= c^2d_2/nd_3 \\ \text{where } d_3 &= (cd_2 - 1)\lambda^2 + \zeta_2 - p^2\delta^2 \end{aligned}$$

and $\lambda = \ln\rho + \gamma - 1$, $\zeta_2 = \pi^2/6$, $\delta = \psi(1-\kappa) - \psi(1) = \psi(1-\kappa) + \gamma$.

(CPW+) The value of ν could be found by direct estimation procedure or might be obtained by a simple search on $\nu > 1$ for the highest sample likelihood already maximised in ρ, μ, c. Then for chosen ν a first estimate of the predicted mean is the ν-th root of $X_T = E\{x^{\nu}\}$ and further corrections can be made using a Taylor series expansion.

Table 7: $CPE_{\pm}$ parameter estimates for rivers Trent (Trent Bridge, 1884-1933), Mississippi (Vicksburg, 1890-1939), and Rhône (Lyon, 1826-1936).

	Trent($n=50$)				Mississippi($n=50$)				Rhône($n=111$)			
	κ	ρ	μ	X_{50}	κ	ρ	μ	X_{50}	κ	ρ	μ	X_{50}
EV1												
moments	-	8.8	187	1136	-	96	264	2235	-	54	546	4308
PWM	-	9.0	185	1131	-	77	276	2273	-	45	570	4388
ML	-	8.41	192	1156	-	70	282	2296	-	30	636	4635
$CPE_{\pm}$												
$\kappa(P_m)$;ML	0.026	7.99	202.8	1175	0.229	15.3	611	3242	0.114	18.3	885	5361
κ(PWM);ML	0.146	6.34	261.3	1260	0.087	35.5	394	2713	0.160	15.4	997	5620
ML	0.179	8.22	202.6	996	0.079	66.2	288	2178	0.192	29.7	655	4020

4 Examples

Three sets of flood discharge data taken from Gumbel (1941, Mississippi & Rhône) and NERC Flood Studies Report (1975, Trent) are used to compare parameter estimates and 50 year return value predictions, see Table 7. The data are annual maxima of daily flood discharges (expressed in m^3/sec) where the year runs from July till June and the day has a fixed starting point. All three datasets result in a positive estimate of κ, so indicate more or less a finite upperbound of the support of the size-distribution. The ML-estimated κ-values for Mississippi, Rhône and Trent are resp. 0.179, 0.192 and 0.079; standardisation to normits (z) results in $z = 1.36$, 3.05 and 0.72. So the Rhône data support E_+ for the size distribution. The tests, described by Van Montfort and Otten (1991) to detect lack of fit w.r.t. Poisson-counts and Exponential sizes, show significant deviations only at the Rhône data supporting Negative Binomial counts ($z = -2.32$) and E_+-sizes ($z = +2.24$). The method of estimating the shape parameter is differentiated by p_m (17), t_3 (18) or ML.

A likelihood ratio test showed for all three datasets that adding a parameter ν (CPW) did not show evidence for $\nu \neq 1$; this is reasonable because ν is especially relevant for the left tail of the size distribution whilst the right tail is important for the extremes.

5 Discussion

A common limiting factor in obtaining sufficiently precise parameter estimates for EV distributions (and confidence in predictions) is the length of the data series. However with many records this situation is compounded by the summarizing to annual maxima from monthly or decadal values. While this data reduction is perceived as a useful filtering process information is being discarded. More importantly the underlying seasonal process makes annual maxima a non-homogeneous mixture for most elements. In effect there is a trade-off between filtering out lower maxima (which are considered to be less representative of extremes, inflating variability with larger s.e.'s of parameter estimates and predictions) and the size of the data set (contributing to larger s.e.'s by decrease in $n^{1/2}$). This author considers that by jointly fitting a seasonal model to monthly data, with harmonic forms of parameters (thus allowing for the pattern of variability, see Revfeim (1982)), more than compensates for any perceived benefits of filtering. Further only 2 or 3 harmonics are likely to be significant so , for example, the 12 fold increase in n for monthly data is a gain on the 5- or 7- fold increase in the number of parameters required.

There is a contradiction between the concept of statistical models for extremes with unbounded range of event magnitudes and the concept of 'probable maximum' of some variable.

For rainfall a methodology has been developed for estimating PMP (the probable maximum precipitation) from a representative column of air (radiosonde sounding) saturated at all levels with some inferred maximum observable dew-point at the surface. The validity of such physically based estimates has been questioned by Reynolds(1978). An alternative procedure is to form an envelope of greatest observed rainfalls over various times intervals T and fit an empirical function of the form $P(\text{mm}) = aT^b$ (Court and Salmela, 1963). Typically b is about 0.5 (square root "response") and with T measured in hours a was estimated as 393 for all-world data up to 100 days by Court and Salmela, and 112 for New Zealand data up to 24hrs by Tomlinson(1980). The upper bound of the beta distribution is a third method of estimating a probable maximum for a homogeneous data set with the advantage of being qualified by standard errors of the parameter estimates. Obviously where some data is detected by the estimation procedure as an outlier of the 'normal' range then there will be two or more probable maxima.

In both $G_{PE_\pm}$ and G_{BE} the traditional form of exponential limit is apparent because $\lim_{\theta\to\infty}(1 - z/\theta)^\theta = \mathrm{e}^{-z}$. However a more general definition (Hardy, 1941) is

$$\lim_{\theta\to\infty}[1 - f(z,\theta)]^\theta = \mathrm{e}^{-z} \text{ if } \lim_{\theta\to\infty}\theta f(z,\theta) = z$$

and there are many choices for $f(z,\theta)$ e.g. $\ln(1+z/\theta)$, $\sinh(z/\theta)$, $\mathrm{I}_1(2z/\theta)$. Of course $f(z,\theta) = 1-\mathrm{e}^{-z/\theta} = z/\theta - z^2/\theta^2 2! + \ldots$ is the identity function. The quotient expression $f_1 = z/\theta(1+z/2\theta)$ is perhaps the simplest form so that $E(f_1) = [(1 - z/2\theta)/(1 + z/2\theta)]^\theta$ and taking the logarithm we get $\ln E(f_1) = -z - z^3/12\theta^2 - z^5/80\theta^4 \ldots$ with residual of order θ^{-2}. In fact $E(f_1)$ is the first of a sequence of polynomial quotient approximations $E(f_k)$ (Lanczos, 1957) which give ever smaller residuals of order θ^{-k}. Thus we may choose in place of (2)

$$F(x) = 1 - [(1 - \kappa x/\mu)/(1 + \kappa x/\mu)]^{1/2\kappa}, \;\; 0 < x < \mu/\kappa$$

which approaches $1-\mathrm{e}^{-x/\mu}$ much faster with decreasing κ. If κ is small we may choose to ignore fourth and higher order terms so that

$$\ln g \approx \ln(\rho/\mu) + (\kappa x/\mu)^2 - x/\mu - \kappa^2 x^3/3\mu^3 - \rho\mathrm{e}^{-x/\mu}(1 - \kappa^2 x^3/3\mu^3)$$

which may be compared with the equivalent form for (2)

$$\ln g \approx \ln(\rho/\mu) + \kappa x/\mu - x/\mu - \kappa x^2/2\mu^2 - \rho\mathrm{e}^{-x/\mu}(1 - \kappa x^2/2\mu^2)$$

6 Summary

The GEV distribution has been put in context of sub-limiting Exponentially distributed event magnitudes with Poisson recurrence. The family of Compound Poisson (or Binomial) recurrence with Exponential (sub-limiting beta or Pareto)/Weibull/Gamma event magnitudes covers a wide range of possible shapes and boundary conditions. Initial estimates of the $\mathrm{CPE}_\pm$ shape parameter can be made using the simple estimate of probability at the mean. This allows comparison of alternate sample statistics, variance or first order PWM, in the estimation of other parameters. The parametric form of the $\mathrm{CPE}_\pm$ analogue of the GEV distribution allows a 2-dimensional search to obtain ML estimates of the 3 parameters. The explicit form of the covariance structure is easy to relate to that for the CPE model.

As the compound model parameters have a physical meaning this can be used to decide whether estimates are sensible or if an alternative compound model would give more realistic values.

References

Buishand, T.A., 1986: Extreme-value analysis of climatological data. *Proceedings 3rd. Intl. Meeting on Statistical Climatology*, Vienna.

Court, A., and Salmela, H.A., 1963: Improbable weather extremes and measurement needs. *Bull. Amer. Met. Soc.,* **44**, 571-575.

Epstein, B., 1949: A modified extreme value problem. *Ann. Math. Stats.,* **20**, 99-103.

Fisher, R.A., and Tippet, L.H.S., 1928: Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.,* **24**, 180-190.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C., and Wallis, J.R., 1979: Probability weighted moments:Definitions and relation to parameters of several distributions expressible in inverse form. *Water Resour. Res.,* **15**, 1049-1054.

Gumbel , E.J., 1941: The return period of floods. *Ann. Math. Stats.,* **12**, 163-189.

Hardy, G.H., 1941: *A Course of Pure Mathematics.* 8 edn, Cambridge University Press, London.

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**, 251-260.

Hosking, J.R.M., 1990: L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J.R. Statist. Soc.* B, **52**, 105-124.

Jenkinson, A.F., 1955: The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quart. J.R. Met. Soc.* **81** , 158-171.

Lanczos, C., 1957: *Applied Analysis.* Pitman, London, 539pp.

NERC, 1975: *Flood Studies Report,***1**. London, Natural Environment Research Council.

Otten, A., and Van Montfort, M.A.J., 1980: Maximum-likelihood estimation of the general extreme-value distribution parameters. *J. Hydrology,* **47**, 187-192.

Pickands, J., 1975: Statistical inference using extreme order statistics. *Annals Stats.,* **3**, 119-131.

Prescott, P., and Walden, A.T., 1980: Maximum likelihood estimation of the generalized extreme value distribution. *Biometrika,* **67**, 723-4.

Revfeim, K.J.A, 1982: Seasonal patterns in extreme 1-hour rainfalls. *Water Resour. Res.,* **18**, 1741-1744.

Revfeim, K.J.A, 1984: Generating mechanisms of and parameter estimators for the extreme value distribution. *Australian J. Statistics,* **26**, 151-159.

Revfeim, K.J.A., 1989: A framework for interpreting rainfall models. *Proc. 4th Intl. Mtg. Stat. Climatology*, NZ Meteorological Service, 233-237.

Revfeim, K.J.A., 1991: Annual maxima and totals of seasonally varying processes. *Stoch. Hydrol, Hydraul.*, **5**, 147-153.

Reynolds, G., 1978: Maximum precipitation in Great Britain. *Weather,* **33**, 162-166.

Rossi, F., Fiorentino, M., and Vesace, P., 1984: Two-component extreme value distributions for flood frequency analysis. *Water Resour. Res.,* **20**, 847-856.

Smith, R.L., 1985: Maximum likelihood estimation in a class of non-regular cases. *Biometrika,* **72**, 67-90.

Tomlinson, A.I., 1980: The frequency of high intensity rainfalls in New Zealand. Tech. Pubn. No. 19, Water and Soil Division, MOWD, Christchurch, NZ.

Van Montfort, M.A.J., and Otten, A., 1991: The first and second e of the extreme value distribution, EV1. *Stochastic Hydrol. Hydraul.,* **5**, 69-76.

AN EXTREME VALUE DISTRIBUTION WITH THREE PHYSICALLY MEANINGFUL PARAMETERS

K.J.A. REVFEIM
278 Cockayne Road, Wellington 6004, New Zealand

ABSTRACT

Selection of "appropriate" extreme value distributions for the same element, over different time intervals at the same location or similar intervals at different locations, has very uncertain outcomes. There is a framework of theoretically based compound event-recurrence/event-magnitude model distributions that seems to cover the same range of statistical properties as the large number of empirical distributions often employed in the selection procedure. The distinct advantage of these compound distributions is that their parameters have a physical meaning that can be tested against reality. The framework has been extended with event duration (or "event" length) added to the model and the additional parameter appears to have little effect on estimates of event reccurrence and magnitude. Applying this model to rainfall data we find that moment estimates of event duration seem reasonable. Further work on robust parameter estimates with some measure of precision is required.

INTRODUCTION

Since it was first advocated by Epstein (1949) the Compound Poisson/Exponential (CPE) distribution has not been used often as a model for extremes. The exact small sample theory of CPE is generally passed over in favour of the Fisher & Tippet(1928) or EV1 model based on asymptotic theory and implied discrete time. Similarly the Frechet(1927) or EV2 distribution is used for maxima in favour of its Compound Poisson/Pareto (CPP_x) analogue with the Pareto distribution having a positive threshold (Revfeim, 1984a), i.e. applicable to extreme rainfall "peaks over threshold (x)" (POT) analysis. Finally EV3 or Weibull is not recognised as a distribution of minima of a Compound Poisson/ J-shaped-beta process (Revfeim, 1984a). Hence the physical interpretation available in the CPE distribution parameters has been under-utilised, particularly in application to extreme rainfalls. Physical contradictions apparent in CPE parameter estimates are not recognised in EV1 e.g. for a rainfall data set of several durations (12,24,48,72hrs) the logical

expectation of decreasing numbers of the longer duration events may not be realised if shorter duration maxima dominate (Revfeim, 1992).

In a review article Bobee and Rasmussen (1995) discuss the dilemma between identifying an appropriate statistical distribution (more parameters, better fit) and estimation of parameters for the selected distribution (fewer parameters, smaller standard errors). Regionalization is seen as the possible means of solving the dilemma by identifying appropriate distribution "shapes" with only a scale parameter being station or site specific. However there is no generally accepted procedure to decide which distribution will be identified as appropriate, nor are there necessarily large differences between the "shapes" of distributions (18 referenced in the review article). At least part of the problem is that parameters of empirical distributions do not have a physical meaning that can be used in the selection procedure. For example in the case of EV1 the parameters in $G(x) = \exp(-\mathrm{e}^{-\alpha(x-u)})$ (Gumbel's 1941 formulation) have no meaning apart from u being the mode and α the scale factor. In the CPE analogue $G(x) = \exp(-\rho \mathrm{e}^{-x/\mu})$ the parameter ρ is the mean rate of occurrence of events which has a strong spatial connotation of regions being affected by the same geophysical processes (i.e. ρ is nearly constant within a region), while mean event size μ is the local scale factor. Bobee and Rasmussen's review ends with a conjecture from Klemes (1993) that "if more light is to be shed on the problems of hydrological extremes, then it will have to come from more information on the physics of the phenomena involved".

In the spirit of this quotation the CPE model is extended by including event duration or length under simplistic assumptions for event intensity or thickness (i.e. magnitude/duration or length). While the extended model has general application this initial study deals with rainfall events for which storm occurrence, magnitude and duration are relatively simple concepts. In this context "regionalization" (as an identifier in distribution selection) benefits from the physical interpretation of mean duration as a possible (sub-)regional factor as well as its effect on the shape of the distribution. It is hoped that this study will help to establish compound occurrence (in time or length or space) and magnitude statistical distributions as a versatile framework for extremes with the added

benefit of their physical interpretation.

MODEL

Like the model constructions of Todorovic & Yevjevich (1969) and Eagleson (1978) for rainfall totals we assume that a rainfall process can be approximated, albeit crudely, by a Poisson process in time of event initiation (rate ρ), an exponential distribution of event amounts (mean μ), and exponential distribution of event duration (mean δ). As with the CPE model the component processes are assumed to be independent. In studies of rainfall totals Eagleson (1978) observed that the exponential distribution gave a reasonable fit to "storm durations" where storms were defined by a minimum dry period separation of at least 2 hours. Using the same definition storm occurrence and duration were found to be uncorrelated ("independent") by Grace and Eagleson (1966), while average intensity (amount/duration) and duration were also deduced to be "independent" by Eagleson (1978). Chandler *et al* (1995) assume a conditional exponential distribution of amount given duration of the form $E\{X|Y\} = Y\mathrm{e}^{-Y/\delta}$. For the purposes of tractability we also make the assumption that the intensity of rainfall events is uniform over its duration; without this assumption the partitioning of amounts presents difficulties with integration. We will refer to this model as Compound Poisson/Exponential/Exponential abbreviated to CPEE.

Thus for a specified accumulation time T the probability that a rainfall event duration exceeds T is $Q = \mathrm{e}^{-T/\delta}$. For $y > T$ the depth of the event amount u within a predetermined time T is Tu/y and this is less than some arbitrary maximum x if $u \leq xy/T$. Given that r events occur we have a binomial probability distribution of $r-s$ events of duration $\leq T$ and s events greater than T denoted by $B(r,s,P) = {}^{r}C_{s}P^{r-s}Q^{s}$ where $P = 1-Q$. Hence we get the conditional probability for the maximum of r events

$$G_r(x|y) = \sum_{s=0}^{r} B(r,s,P)F^{r-s}(x)F^{s}(xy/T)$$

where $F(x) = 1 - \mathrm{e}^{-x/\mu}$. Taking the expectation over event durations $> T$ we get the unconditional probability

$$
\begin{aligned}
G_r(x) &= \sum_{s=0}^{r} B(r,s,P)F^{r-s}(x)\mathrm{e}^{T/\delta}\sum_{t=0}^{s} {}^{s}C_t(-1)^t \int_{T/\delta}^{\infty} \exp[-(1+t\delta x/T\mu)z]\, dz \\
&= \sum_{s=0}^{r} B(r,s,P)F^{r-s}(x)\sum_{t=0}^{s} {}^{s}C_t(-\mathrm{e}^{-x/\mu})^t/(1+t\delta x/\mu T) \\
&= \sum_{s=0}^{r} B(r,s,P)F^{r-s} \\
&\quad -\sum_{s=1}^{r} B(r,s,P)F^{r-s}s(1-F)/(1+\delta x/\mu T) \\
&\quad +\sum_{s=2}^{r} B(r,s,P)F^{r-s}s(s-1)(1-F)^2/2!(1+2\delta x/\mu T) \\
&\quad ..... + (-1)^r(1-F)^r/(1+r\delta x/\mu T) \\
&= B^r - rAB^{r-1}/(1+\delta x/\mu T) + r(r-1)A^2B^{r-2}/(1+2\delta x/\mu T).... \qquad (1)
\end{aligned}
$$

where $A = Q(1-F)$, $B = A + F$ and in the last two lines of (1) $F = F(x)$. Now using the assumed Poisson count of events we finally get the (conditional) distribution function (d.f.) for maximum depths within T as

$$
\begin{aligned}
G(x) &= \mathrm{e}^{-\rho}\sum_{r=1}^{\infty} \rho^r G_r(x)/r! \\
&= \mathrm{e}^{-P\theta} \bullet \sum_{s=0}^{\infty} (-Q\theta)^s/s!(1+sz/\alpha) \qquad (2) \\
&= \mathrm{e}^{-P\theta} \bullet \mathrm{e}^{-Q\theta}\Gamma(1+\alpha/z)\sum_{s=0}^{\infty} (Q\theta)^s/\Gamma(1+s+\alpha/z) \qquad (3) \\
&= \mathrm{e}^{-\theta}[1 + Q\theta z/(z+\alpha) + (Q\theta z)^2/(z+\alpha)(2z+\alpha) + ...] \qquad (4)
\end{aligned}
$$

where $\theta = \rho\mathrm{e}^{-x/\mu}$, $z = x/\mu$, $\alpha = T/\delta$ and the terms after $\bullet$ in (2) and (3) are forms of the confluent hypergeometric function ${}_1\mathrm{F}_1$ (Sneddon 11.3,11.14, 1956). Alternatively ${}_1\mathrm{F}_1$ may be written as an integral in terms of the incomplete gamma function (Abramowitz & Stegun **6.5.4**,1970). From (4) it is seen that the d.f. for the CPEE model takes the form of the CPE (or equivalently EV1/Gumbel) d.f. multiplied by a rapidly converging summation that includes the duration parameter α (relative to T). When x is large $G(x)$ tends to CPE$(\rho Q, \mu)$ and when x is small $G(x)$ tends to CPE(ρ, μ).

Terms in the summations decay at a greater than exponential rate and for numerical work the summations might be approximated by the first few terms only; perhaps justified

by the argument that the smallest maximum in a data set is of order twice the mean event amount i.e. $z_{min} \approx 2$. Similarly mean event duration is less than the prescribed accumulation times of particular interest $T = 24, 48, 72$ hours so that α is likely to be at least 2. Hence the first series term in (4) $Q\theta z/(z+\alpha) = \rho z e^{-z-\alpha}/(z+\alpha)$ is typically of magnitude $\rho e^{-4}/4$ or smaller. Now intuitively the recurrence rate of all rainfall events is less than one event per week, and selected events of longer duration will occur at less than half that rate, so that this first term is at most $\rho/100 = 1/4$. With this "ball park" largest value and the greater than exponential decay the first few terms approximation is seen as being quite reasonable. The above argument may also be seen as a benefit of physically meaningful parameters in the model without which numerical assessment would be difficult, but no similar empirical statistical distribution would have been likely to be developed either!

PARAMETER ESTIMATION

The raw moments are given in the appendix (14)-(16) and can be written as

$$\begin{aligned}
m_1 &= \mu(\lambda - \alpha A_0)/P \\
m_2 &= \mu^2(J_2 + \lambda^2 - 2\alpha B_1 + \alpha^2 A_1)/P \\
m_3 &= \mu^3[J_3 + 3\lambda J_2 + \lambda^3 + 3\alpha(PB_1 - C\lambda + \alpha B_2 - 3\alpha^2 A_2)]/P \\
\text{where } A_k &= \sum_{s=1}^{\infty}(-Q/P)^s R_s/s!s^k \\
B_k &= \sum_{s=1}^{\infty}(-Q/P)^s \gamma(s,\rho P) R_s/s!s^k \\
C &= \sum_{s=1}^{\infty}(-Q/P)^s/s^2 \\
&\approx 2 + Q/3P(1+Q/3P)^{3/20} - 2(PQ)^{1/2}\cos(P^{1/2}) \\
\lambda &= \ln(\rho P) + \gamma \text{ (Eulers constant)}
\end{aligned}$$

and $\gamma(s,y)$ is the incomplete gamma function (see Appendix), $J_2 = 1.645,\ J_3 = 2.404$. If in the first instance we ignore the K_j in R_s (for definitions see Appendix), assume that $e^{-\rho P}$ can be ignored (i.e. $A_k = 0$), and take the first terms only as approximations to B_k, C (i.e. $B_1 = B_2 = C = -Q/P$), then we get expressions for the mean and scaled

cumulants

$$m_1 = \mu\lambda/P \tag{5}$$

$$\kappa_2/m_1^2 = (PJ_2 + 2\alpha Q)/\lambda^2 - Q \tag{6}$$

$$\kappa_3/m_1^3 = P[PJ_3 - 3\alpha Q(P+\alpha)]/\lambda^3 - 3Q[PJ_2 + \alpha(1+Q)]/\lambda^2 + Q(1+Q) \tag{7}$$

Of course (6) is the squared coefficient of variation, v^2. Directly from (5) and (6) we get

$$\lambda^2 = (PJ_2 + 2\alpha Q)/(v^2 + Q) \tag{8}$$

which can be used to eliminate λ from (7). The resulting equation can be expressed in terms of the coefficient of skewness $g_\delta = \kappa_3/\kappa_2^{3/2}$ as

$$D_2[D_2E_3 + 3Q(D_2 + \alpha P)E_2]^2 = (PD_3)^2E_2^3 \tag{9}$$

where $D_2 = PJ_2 + 2\alpha Q$, $D_3 = PJ_3 - 3\alpha Q(P+\alpha)$, $E_2 = v^2 + Q$ and $E_3 = v^3 g_\delta - Q(1+Q)$. By the usual iterative procedures the moment estimate of $\hat{\alpha}$ is obtained from (9), then $\hat{\lambda}$ by back substitution in (8) (from which we get $\hat{\rho}$), and finally $\hat{\mu} = Pm_1/\lambda$. However to get a solution for α from (9) the condition $g_\delta < g_0$ must be met (g_0 is skewness for the CPE model $J_3/J_2^{3/2} \approx 1.14$), and coefficient of variation v must be greater than some value above 0.1 to give positive values of α. The latter is scarcely a constraint for extreme values!

Using the first two terms only from the summation in the density from the derivative of (2)

$$g(x) = -\theta' \mathrm{e}^{-P\theta}\alpha \sum_{s=0}^{\infty}(-Q\theta)^s\{P - s[1 + 1/(sz+\alpha)]/\theta\}/s!(sz+\alpha)$$

the log(likelihood) of a sample $\{x_i\},\ i = 1.....n$ is

$$L = n(\ln\rho - \ln\mu) - n\bar{x}/\mu - \rho P S_0 + \sum_{i=1}^{n}\ln D$$

where $D = P - \alpha Q[P\theta - 1 - 1/(z+\alpha)]/(z+\alpha)$. First and second partial differentials with respect to $\rho,\ \mu,\ \alpha$ give the usual maximum likelihood equations to be solved by Newton-Raphson iteration. However this does not appear to give satisfactory estimates.

Probability weighted moments can be obtained by the same decomposition of the d.f. as used in the Appendix for raw moments.

$$\beta_r = -\mu \int_0^{\rho} z d\{(\sum G_s)^{r+1}\}/(r+1)$$

However it is not possible to separate ρ and P by the approximation used for m_1 in which $\beta_r \approx \mu[\ln(\rho P) + \ln(r+1) + \gamma]/(r+1)P$. We must at least add the first residual integral so that

$$\beta_r \approx \mu[\ln(\rho P) + \ln(r+1) + \gamma - 3\alpha Q K_{1,r+1}]/(r+1)P$$

where $K_{1,j} = \int_0^{j\rho P} \mathrm{e}^{-\xi} d\xi / jP[\ln(j\rho P) + \alpha - \ln \xi]$. As stated at the end of the Appendix there is still some uncertainty about the accuracy of approximations proposed for K_j.

EXAMPLE

Annual extreme rainfalls over 24-72 hours at Kelburn (New Zealand, 41°S) from 1928-1980 that were previously fitted by the CPE model (Revfeim, 1983) have been re-analysed. In the original study a wider range of T from 10 minutes to 12 hours was included in the data set but it seems reasonable at this initial stage to restrict the data to time intervals of similar order as the physical event duration. It is recognised that these are not homogeneous data but a mixture of seasonal and storm directions. However for the purposes of this initial model fitting the results should demonstrate on a published data set any adjustment to CPE values of the recurrence and event size parameters, and give physically reasonable values of the duration parameter.

Table 1. Parameter estimates for CPE and CPEE models for 24,48,72 hour maxima.

		24 hour			48 hour			72 hour		
		ρ	μ	δ	ρ	μ	δ	ρ	μ	δ
		(evts/yr)	(mm)	(hrs)	(evts/yr)	(mm)	(hrs)	(evts/yr)	(mm)	(hrs)
CPE	Mom	26.7	20.3	0.0	20.8	26.6	0.0	22.8	28.4	0.0
	ML	30.0	19.6	0.0	25.7	24.8	0.0	29.0	26.2	0.0
	(SE)	(10.0)	(2.0)	-	(8.4)	(2.6)	-	(9.9)	(2.8)	-
CPEE	Mom	26.4	20.4	3.8	20.8	26.6	6.3	22.7	28.4	10.2

The significant result is that there is virtually no difference between CPE and CPEE in the estimates of event magnitude but there is a small difference in the recurrence parameter. This may be interpreted as a justification for the point process CPE model. It is apparent that the approximations used in the expressions for moments provide estimates of δ that are realistic in terms of increasing order but smaller than expected. PWM parameter estimates have not yet been assessed.

Not surprisingly the CPEE model does not overcome the "dominant events" in longer interval maxima which can give illogical estimates of more underlying events per year than for shorter intervals. Dominant events are defined as occurring when the maximum over the longer interval is the same as for the shorter interval, i.e. while large events of longer duration may have occurred they did not exceed the amount of the shorter interval maximum. However the hydrological effect of a smaller amount accumulated over longer lead times may be more significant in its flood potential in the lower reaches of a catchment system. As discussed previously (Revfeim, 1992) this problem is solved by treating the amount of a dominant event as an upper bound rather than a point observation. Parameter estimation for data as a mixture of upper bounds and point values can only be solved by maximum likelihood which is still unsatisfactory for the CPEE model.

DISCUSSION

The assumption of uniform intensity of storms obviously detracts from the model. It is also obvious that this only affects those storms of duration greater than the specified time interval T, for which the fraction of the storm amount underlying the T-maxima will be larger than given by a uniform intensity. That is, in the model, the size of exponentially distributed amounts is underestimated, but this also lessens the effect of the uniform intensity assumption since some restriction on the upper tail of the distribution of amounts is more realistic. Obviously when values of T fall well below mean storm duration the uniform intensity assumption becomes inappropriate. Finding a mathematically tractable non-uniform intensity that allows analytic integration of the conditional distribution $G_r(x|y)$ over storm durations greater than T is a challenge for future research.

Including event durations in the model may be compared with extending the simple point process CPE model by generalisations of the component Poisson and Exponential distributions. The particular extensions in just one of the component distributions, leading to Compound Poisson Beta (CPβ) and Compound Binomial Exponential (CBE) (Revfeim, 1989), have strong physical meanings of upper bounds on event size and number of events respectively. However lifting the point process restriction and including the duration parameter in CPEE has at least equal priority, in terms of added realism, over the CPβ and CBE generalisations.

It may be noted that the alternate sub-limiting forms of the exponential functions in the CPE d.f., namely Compound Poisson/Pareto (CPP_0^a, above zero threshold c.f. EV2) and Compound Negative- Binomial/Exponential (CB^-E) (van Montfort and Otten, 1991), represent weaker conditions of distribution shape change but no upper bounds. It is also worth noting that Jenkinson's (1955) "Generalised Extreme Value" (GEV) distribution generalises only one of the exponential functions (essentially event magnitude) in EV1 i.e. the CPβ/CPP_0^a sub-limiting forms. The fact that most applications of GEV find a best fit for the CPP_0^a analogue may not be the outcome when the duration parameter of CPEE is included.

Further generalisation of CPEE to include bounds on event size, e.g. CPEβ appear to be less tractable (analytically if not numerically) than bounds on the number of events. In the latter case $G_r(x)$ as in (1) is applicable and we get the 4 parameter form of the d.f. for CBEE

$$G_4(x;\rho,\mu,\alpha,\lambda) = \sum_{r=0}^{\lambda}(-Q\theta)^{r\lambda}C_r(1-P\theta/\lambda)^{\lambda-r}/(1+rz/\alpha) - (1-\rho/\lambda)^{\lambda}$$

Recognition that annual maxima arise from a mixture of seasonal processes, and in many cases modal ("prevailing weather") directions, will give rise to additional terms in $G(x)$ involving the relative covariances of the mixture parameters. Mixtures have an effect on the skewness which may make it impossible to obtain moment estimates of the CPEE parameters. For the CPE model the cumulants of a mixture are

$$\kappa_1 \;=\; (\mu/c)[\lambda + C_{\mu\mu}(J_2+\lambda^2)/2c^2]$$

$$\kappa_2 = (\mu/c)^2[J_2 + C_{\mu\mu}(J_3 + 2\lambda J_2)/c^2] \quad (10)$$

$$\kappa_3 = (\mu/c)^3[J_3 + 3C_{\mu\mu}(J_4 - J_2^2 + 2\lambda J_3)/2c^2] \quad (11)$$

where $c = 1 - C_{\rho\mu} + C_{\mu\mu}$, and $C_{\rho\mu}$, $C_{\mu\mu}$ are the coefficients of (co-)variation of the component values of ρ, μ in the mixture (Revfeim, 1991); these coefficients $C_{..}$ are assumed to be small enough for terms higher than first order to be ignored. From (10) and (11) we can get an approximate value for the skewness of a CPE mixture

$$g_{\text{mix}} \approx g_0\{1 + 3C_{\mu\mu}[(J_4 - J_2^2)/J_3 - J_3/J_2]/2c^2\}$$

in which $(J_4 - J_2^2)/J_3 - J_3/J_2 = 3.49...$ Thus the positive skewness of a CPE mixture is greater than g_0, and a similar effect for a CPEE mixture could put g_δ above the g_0 threshold where a moment estimate of the duration parameter is possible.

A similarly based model assuming independence of exponentially distributed amounts and durations was assumed for rainfall totals (Revfeim, 1984b) from which the mean duration of **all** events could be deduced using the estimated recurrence rate ρ_t (again of all events) of the Compound Poisson Gamma (CPG) distribution and the mean number of raindays ($\bar{n}$) viz. $\hat{\delta}_t = \bar{n}/\rho_t - 1$. In contrast the duration parameter in the current study is for **selected** events underlying the prescribed accumulation time T. That is while the relative parameter $\alpha = T/\delta$ may not differ greatly for different values of T the mean duration δ is expected to increase with T.

SUMMARY

The CPEE model for extremes is a natural extension of the point process CPE model to include a parameter of (rainfall) event duration. Estimates of the duration parameter need to be realistic because of the physical meaning. Moment estimates of the duration parameter are not totally satisfactory which may be due to approximations used in the moment formulae, the assumption of uniform intensity being inadequate, or the unreliability of sample estimates of skewness. However the model development given does provide a platform for testing these deficiencies and making improvements where possible and exploring other methods of parameter estimation. It also sets the suite of small sample compound recurrence/magnitude/extent models in context with their asymptotically based analogues.

Relaxation of the uniform intensity assumption will probably add another parameter but this may not necessarily improve the prediction of long term extremes i.e. design criteria. Feasible extensions to the 4 parameter CBEE model, with its constraint on the number of events in a finite time interval, appear intuitively to be of greater importance than applying constraints on event amounts.

REFERENCES

Abramowitz, M., and Stegun, I., 1964. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. Dover, New York, 1046 pp.

Bobee, B., and Rasmussen, P.F., 1995. Recent advances in flood frequency analysis. Rev. Geophys.(Suppl.), 1111-1116.

Chandler, R.E., Isham, V., Kakou, A., & Northrop, P., 1995. Spatial-temporal rainfall processes: stochastic models and data analysis. Proc. 6th Intl. Mtg. Stat. Climatology, Univ. Galway.

Eagleson, P.S., 1978. Climate , soil and vegetation, 2. The distribution of annual precipitation derived from observed storm sequences. Water Resour. Res., 14: 713-721.

Epstein, B., 1949. A modified extreme value problem. Ann. Math. Stat., 20: 99-103.

Fisher, R.A., and Tippet, L.H.S., 1928: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Phil. Soc., 24: 180-190.

Frechet, M., 1927. Sur la loi de probabilite de l'e cart maximum. Ann. Soc. Polon. Math., 6: 93-116.

Grace, R.A. and Eagleson, P.S., 1966. The synthesis of short-time increment rainfall sequences. Report No. 91, Hydrodyn. Lab., MIT, Cambridge, Mass., USA.

Gumbel, E., 1941: The return period of flood flows. Ann. Math. Statist., 12: 163-190.

Jenkinson, A.F., 1955. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quart. J.R. Met. Soc., 81: 158-171.

Klemes, V., 1993: Probability of extreme hydrometeorological events- a different approach. in Extreme Hydrological Events: Precipitation Floods and Droughts, IAHS, Publ. No. 213: 167-176.

van Montfort, M.A.J., and Otten, A., 1991. The first and second e of the extreme value distribution. Stochastic Hydrol. Hydraul., 5: 69-76.

Revfeim, K.J.A., 1983: On the analysis of extreme rainfalls. J. Hydrol., 62: 107-117.

Revfeim, K.J.A., 1984a. The cumulants of an extended family of Type 1 extreme value distributions. Sankhya, 46 B: 281-284.

Revfeim, K.J.A., 1984b: An initial model of the relationship between rainfall events and daily rainfalls. J. Hydrol., 75: 357-364.

Revfeim, K.J.A., 1989. A framework for interpreting rainfall models. Proc. 4th Intl. Mtg. Stat. Climatology, NZ Meteorological Service.

Revfeim, K.J.A., 1991: Annual maxima and totals of seasonally varying processes. Stoch. Hydrol. Hydraul., 5: 147-153.

Revfeim, K.J.A., 1992: Dominant events in extreme rainfall records. J. Hydrology, 134: 143-149.

Sneddon, I.N., 1956. Special functions of mathematical physics and chemistry. Oliver and Boyd, London, 164pp.

Todorovic, P., and Yevjevich, V., 1969: Stochastic processes of precipitation. Col. State Univ. Hydrol. Papers No. 35, 61pp.

APPENDIX

Using the nomenclature of (4) equation (2) can be written in the form $G(\theta) = \sum_{s=0}^{\infty} G_s(\theta)$ where

$$G_s(\theta) = \alpha(-Q\theta)^s \mathrm{e}^{-P\theta}/s!(sz+\alpha)$$

Thus the raw moments can be expressed as

$$\begin{aligned} E\{x^p\} &= -\mu^p \int_0^{\rho} z^p dG(\theta) \\ &= -\mu^p \sum_{s=0}^{\infty} \int_0^{\rho} z^p dG_s(\theta) \end{aligned} \tag{12}$$

The first term in (12) simply generates the CPE raw moments $\mu^p I_p/P$ (with ρP in the place of ρ) where

$$\begin{aligned} I_p &= \int_0^{\rho} (\ln\rho - \ln\theta)^p \mathrm{e}^{-P\theta} P d\theta \\ &= \int_0^{\rho P} [\ln(\rho P) - \ln\xi]^p \mathrm{e}^{-\xi} d\xi \\ &= \sum_{q=0}^{p} {}^pC_q \lambda^{p-q} J_q \end{aligned}$$

$\lambda = \ln(\rho P) + \gamma$, and when ρ is moderately large the J_q obey a recursive relation in zeta functions $J_{q+1} = q! \sum_{r=1}^{q} \zeta_{r+1} J_{q-r}/(q-r)!$ with $J_0 = 1,\ J_1 = 0$ (Revfeim, 1984a).

For $s > 0$ the remaining terms in (12) can be integrated by parts giving

$$\begin{aligned} \int_0^{\rho} z^p dG_s(\theta) &= p \int_0^{\rho} z^{p-1}\theta^{-1} G_s(\theta) d\theta \\ &= [p\alpha(-Q)^s/s!] \int_0^{\rho} z^{p-1}\theta^{s-1}\mathrm{e}^{-\theta P} d\theta/(sz+\alpha) \\ &= [p\alpha(-Q/P)^s/s!] \int_0^{\rho P} z^{p-1}\xi^{s-1}\mathrm{e}^{-\xi} d\xi/(sz+\alpha) \end{aligned} \tag{13}$$

where z now takes the form $\ln(\rho P) - \ln\xi$. Hence for the second and third moments we can reduce the integral (13) using $z = (sz + \alpha - \alpha)/s$ and $z^2 = [(sz + \alpha)(sz - \alpha) + \alpha^2]/s^2$ and finally get

$$E\{x\} = (\mu/P)\{\lambda - \alpha\sum_{s=1}^{\infty}(-Q/P)^s R_s/s!\} \quad (14)$$

$$E\{x^2\} = (\mu^2/P)\{J_2 + \lambda^2 - 2\alpha\sum_{s=1}^{\infty}(-Q/P)^s[\gamma(s,\rho P) - \alpha R_s]/s!s\} \quad (15)$$

$$E\{x^3\} = (\mu^3/P)\{J_3 + 3\lambda J_2 + \lambda^3 - 3\alpha\sum_{s=1}^{\infty}(-Q/P)^s[s!(\lambda - \sum_{j=1}^{s-1}\gamma(j,P\rho)/j!) \quad (16)$$

$$-\alpha\gamma(s,\rho P) + \alpha^2 R_s]/s!s^2\} \quad (17)$$

where $\gamma(s,y) = \Gamma(s)[1 - \mathrm{e}^{-y}\sum_{j=0}^{s-1} y^j/j!]$ is one form of the incomplete gamma function and the residual integral involving inverse powers of z

$$\begin{aligned} R_s &= \int_0^{\rho P} \xi^{s-1}\mathrm{e}^{-\xi}d\xi/(sz+\alpha) \\ &= \sum_{j=0}^{s-2} j!s^j\Gamma(s-j)\{[\gamma(s-j,\rho P)/\Gamma(s-j) - 1 + \mathrm{e}^{-\rho P}]/\alpha^{j+1} \\ &\qquad + K_{j+1}(s) + (j+1)sK_{j+2}(s)\} \\ \text{with } K_j(s) &= \int_0^{\rho P} \mathrm{e}^{-\xi}d\xi/(sz+\alpha)^j \end{aligned}$$

It is obvious that $K_j(s)$ is smaller than $O(1/s)$ and using a binomial expansion of the denominator (where it is valid) it might be expressed as

$$\begin{aligned} K_j(s) &\approx (s\lambda+\alpha)^{-j}\sum_{k=0}^{\infty} {}^{-j}C_k(-1)^k(\lambda+\alpha/s)^{-k}\int_0^{\rho P}(\ln\xi+\gamma)^k\mathrm{e}^{-\xi}d\xi \\ &= (s\lambda+\alpha)^{-j}\sum_{k=0}^{\infty} {}^{-j}C_k(\lambda+\alpha/s)^k J_k \end{aligned}$$

The binomial expansion is valid for $\xi > Q^{1/s}/\rho P$ so the approximation has excluded

$$\begin{aligned} \int_0^{Q^{1/s}/\rho P} \mathrm{e}^{-\xi}d\xi/s^j(\rho P/Q^{1/s} - \ln\xi) &< (1 - \mathrm{e}^{-Q^{1/s}/\rho P})/2\ln(\rho P/Q^{1/s}) \\ &\approx Q^{1/s}/2\rho P\ln(\rho P/Q^{1/s}) \end{aligned}$$

and included

$$(s\lambda+\alpha)^{-j}\sum_{k=0}^{\infty} {}^{-j}C_k(-1)^k(\lambda+\alpha/s)^{-k}\int_0^{Q^{1/s}/\rho P}(\ln\xi+\gamma)^k\mathrm{e}^{-\xi}d\xi$$

the effect of which has yet to be evaluated.