Unsupervised machine learning techniques for studying climate variability

Alexander Ilin, Harri Valpola, Erkki Oja

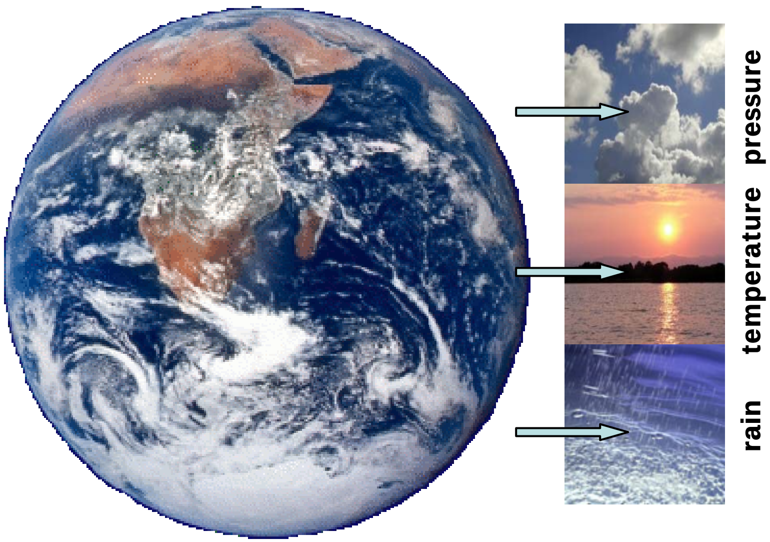
Helsinki University of Technology
Adaptive Informatics Research Center

Adaptive Informatics Research Center

- Develops new methods of information science and *machine learning* (extraction of rules and patterns out of massive data sets)
- Popular techniques developed in the center:
  - Self-organizing (Kohonen’s) maps
  - Algorithms for independent component analysis (ICA)
- *Unsupervised learning*: fitting a model to observed data, no explicit input-output relations

Application in climatology

- Exploratory analysis of global climate data



Important regularities
(structured signals)

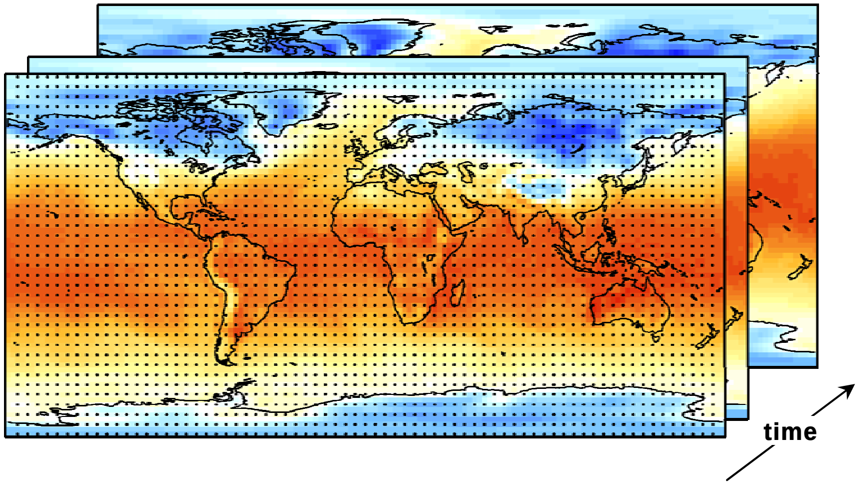
Significant climate
phenomena

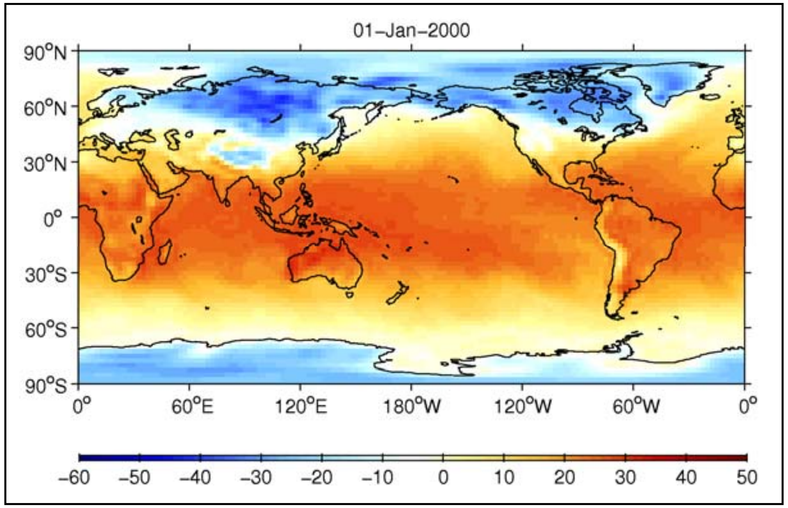
- Tool: *adjustable* technique resembling ICA, tuned to find specific types of signal structures

NCEP/NCAR reanalysis data

- Reconstructed daily averages for several weather variables, global regular grid, 56 years
- Can be summarized in high-dimensional matrix

$$\mathbf{X} = [\ \dots \quad \mathbf{x}(t) \quad \dots\ ]$$





Method of analysis

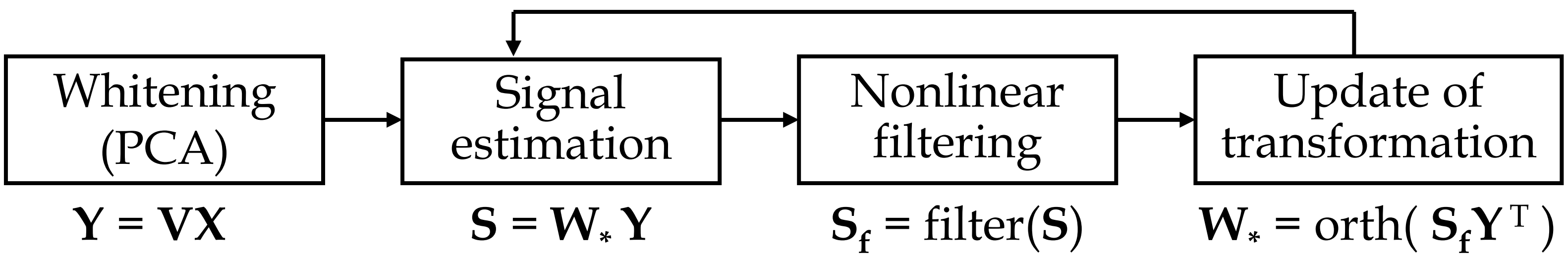
- Linear transformation of data

$$\mathbf{S} = \mathbf{W}\mathbf{X} \quad \text{or} \quad \mathbf{s}(t) = \mathbf{W}\mathbf{x}(t)$$

- Principal component analysis (PCA) is used for preprocessing
- Principal components are further *rotated* such that *a specific type* of signal structure is most prominent in $\mathbf{S}$
- Motivation may come from expert knowledge, simple inspection of data etc

Algorithmic structure

- The method unifies different rotation approaches under one algorithmic framework

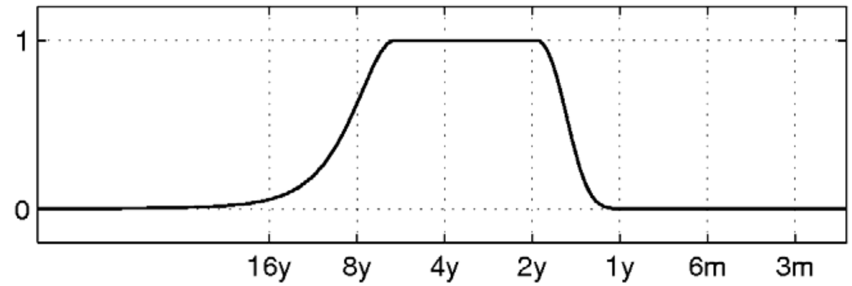


- A proper choice of (non)linear filtering tunes the procedure for a specific task
- Filtering retains desired properties and removes irrelevant structures

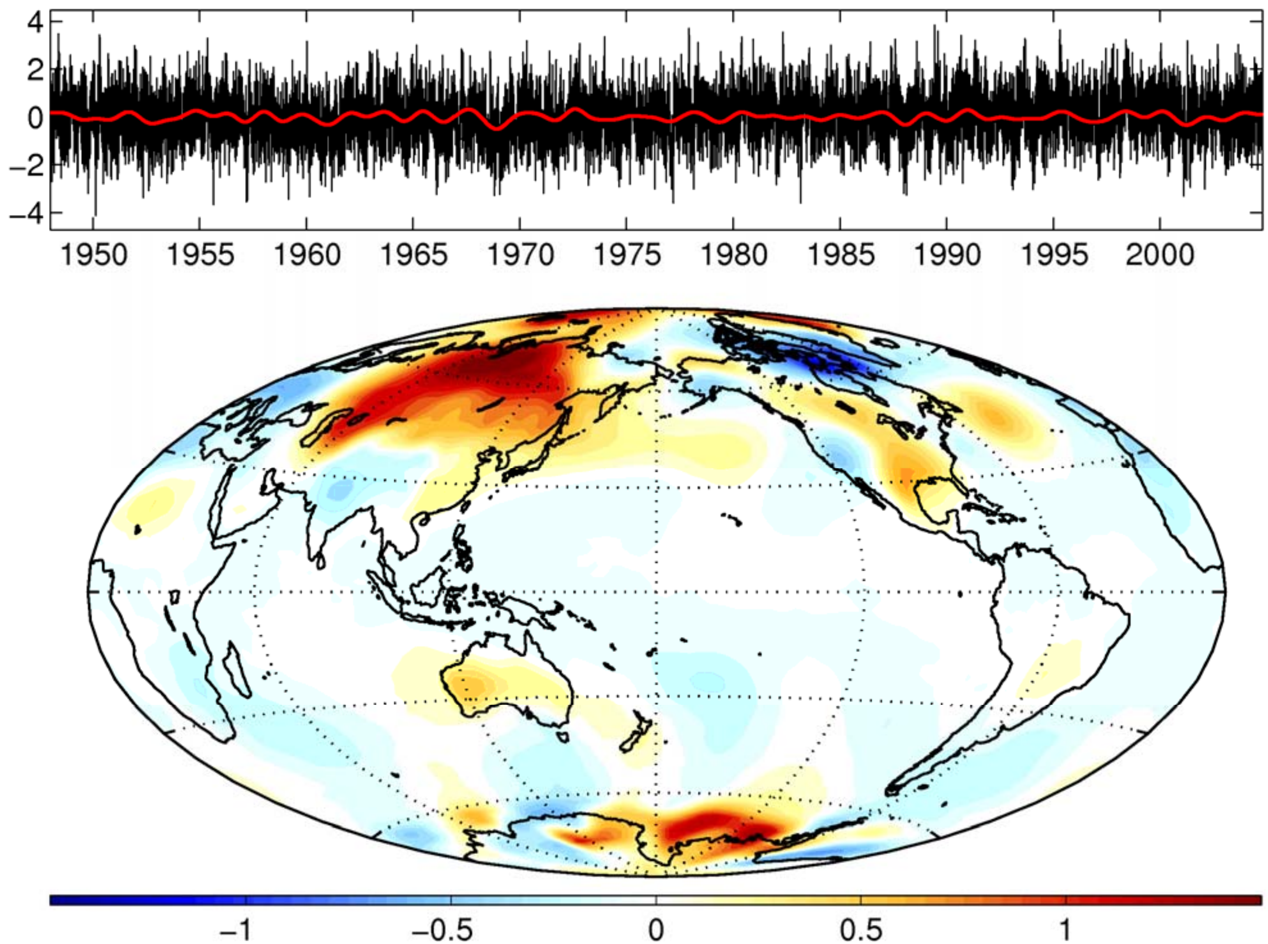
Example: Clarity-based analysis

- The signals are expected to have prominent (clean) variability in a specific timescale:
- The energy contained in the interesting frequencies is large compared to the full energy
- Band-pass filter (which retains only interesting frequencies) can be used in the loop

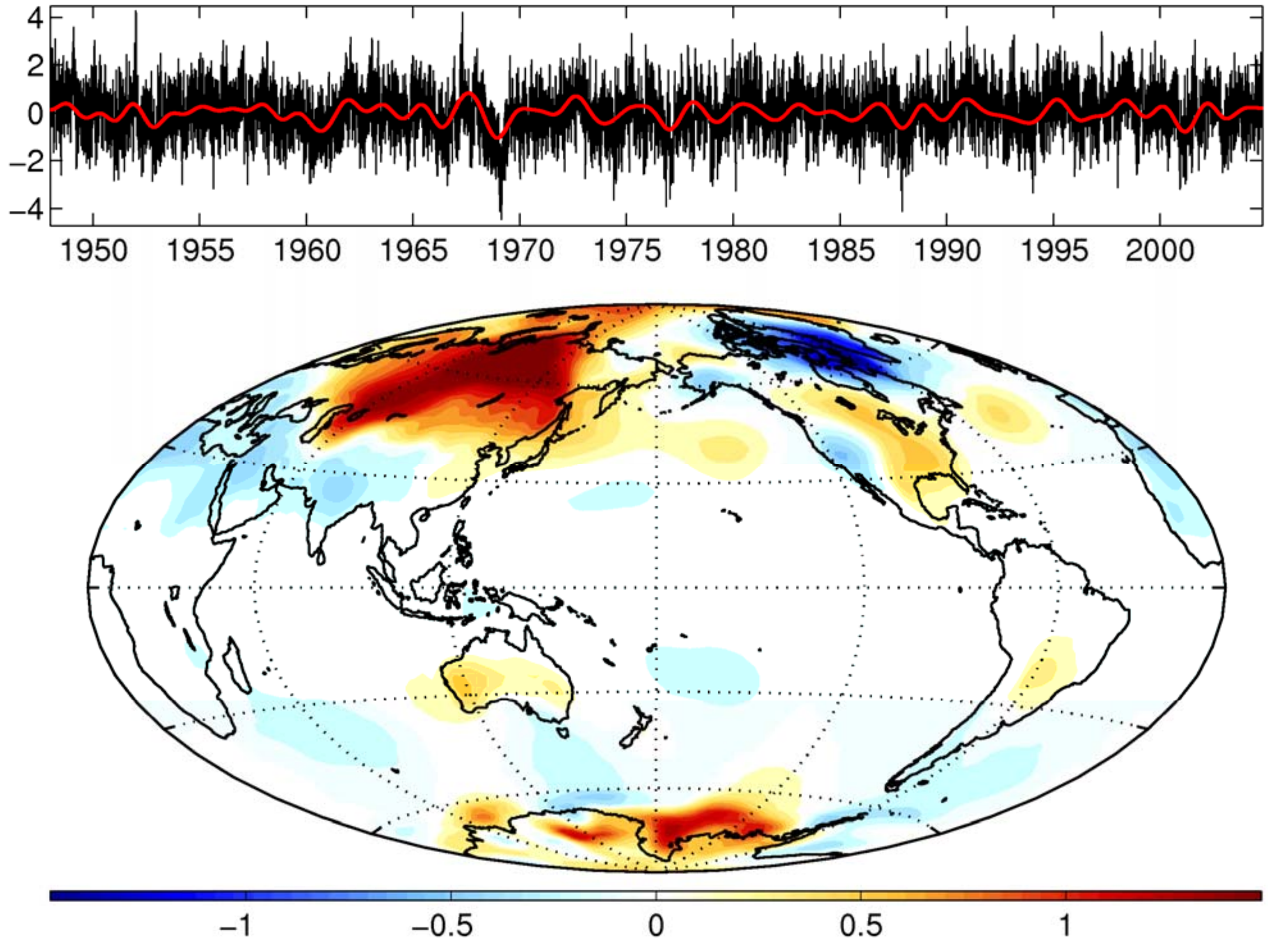
Interannual phenomena:



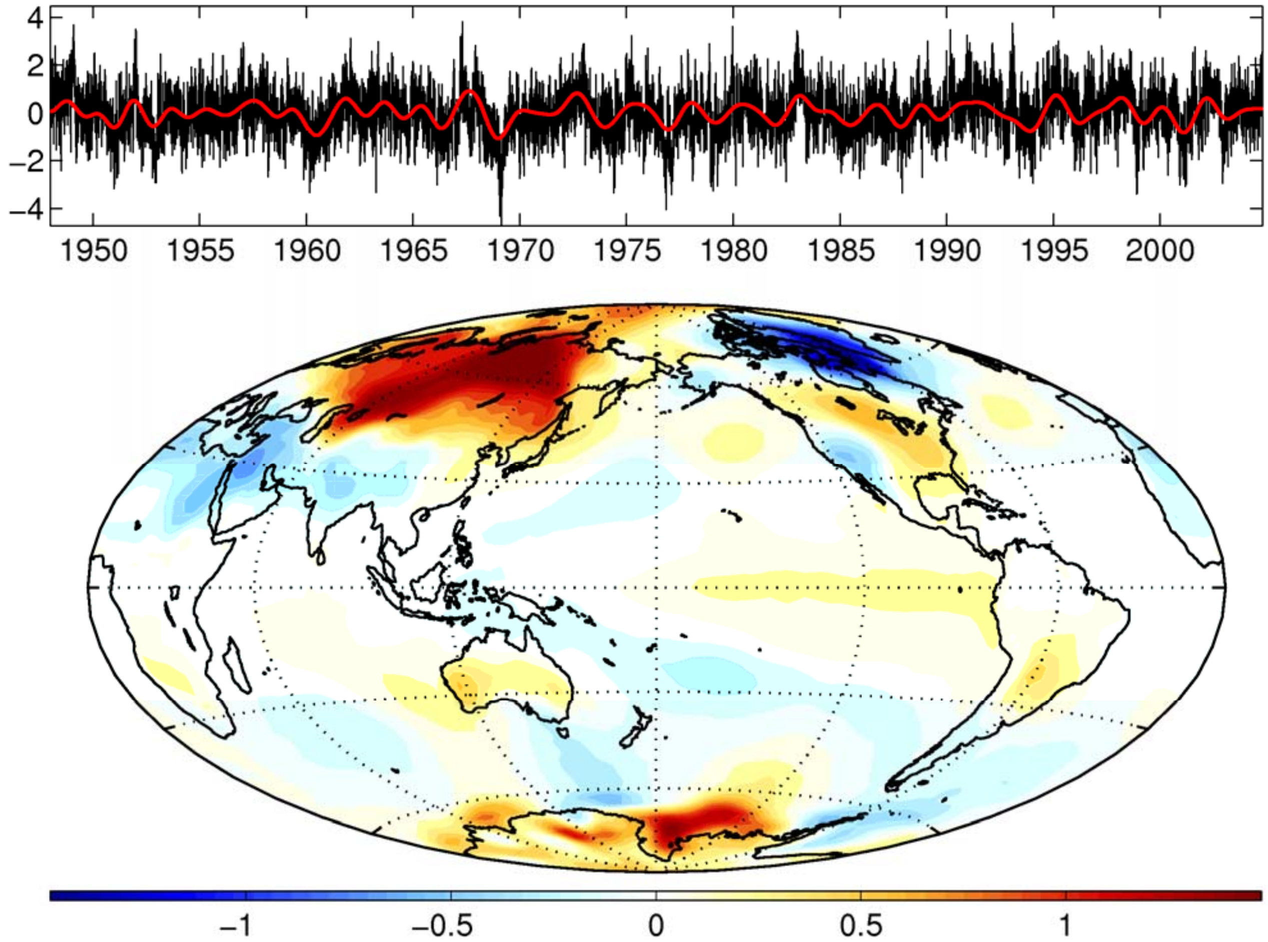
Example: Clarity-based analysis



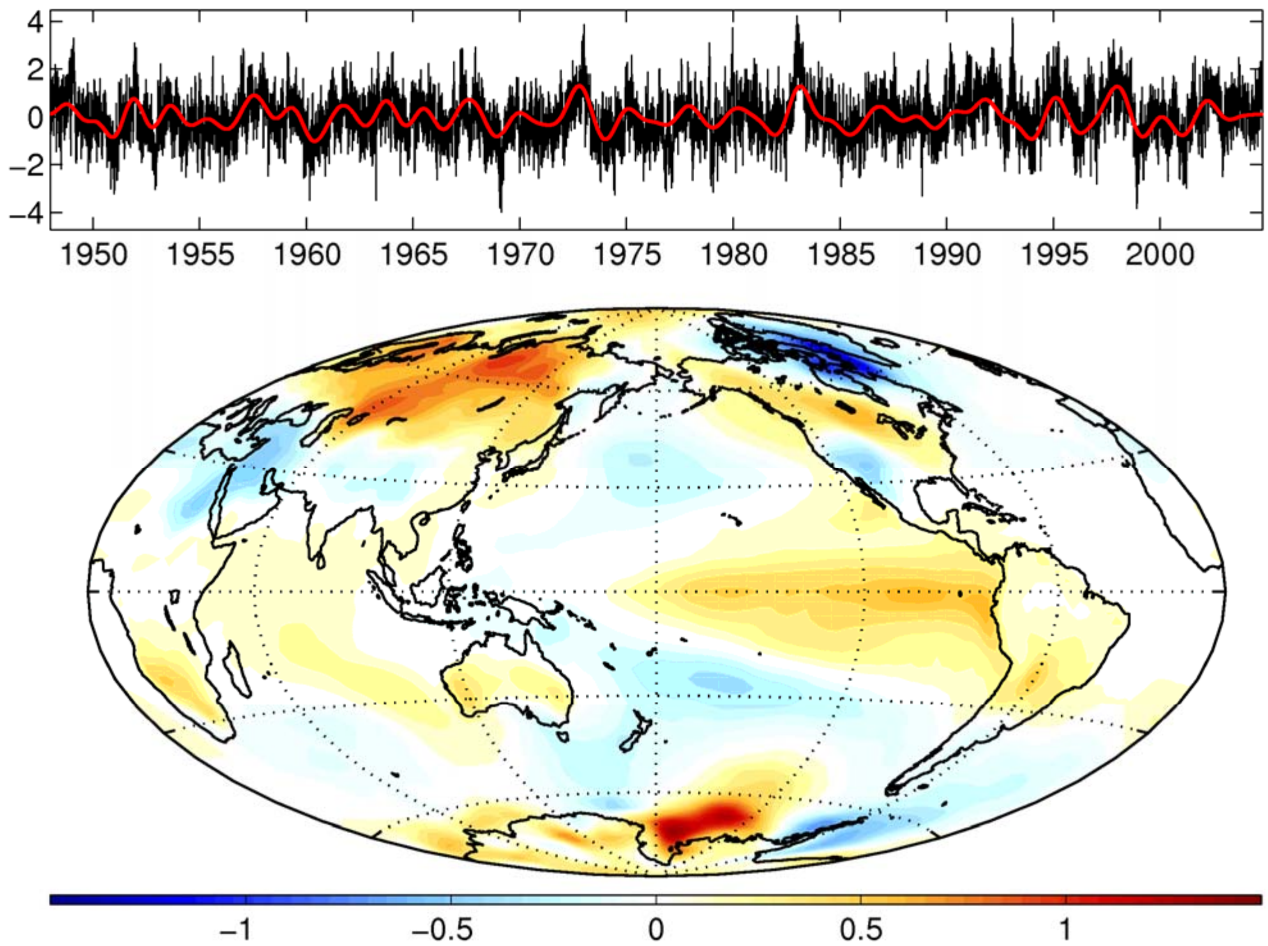
Example: Clarity-based analysis



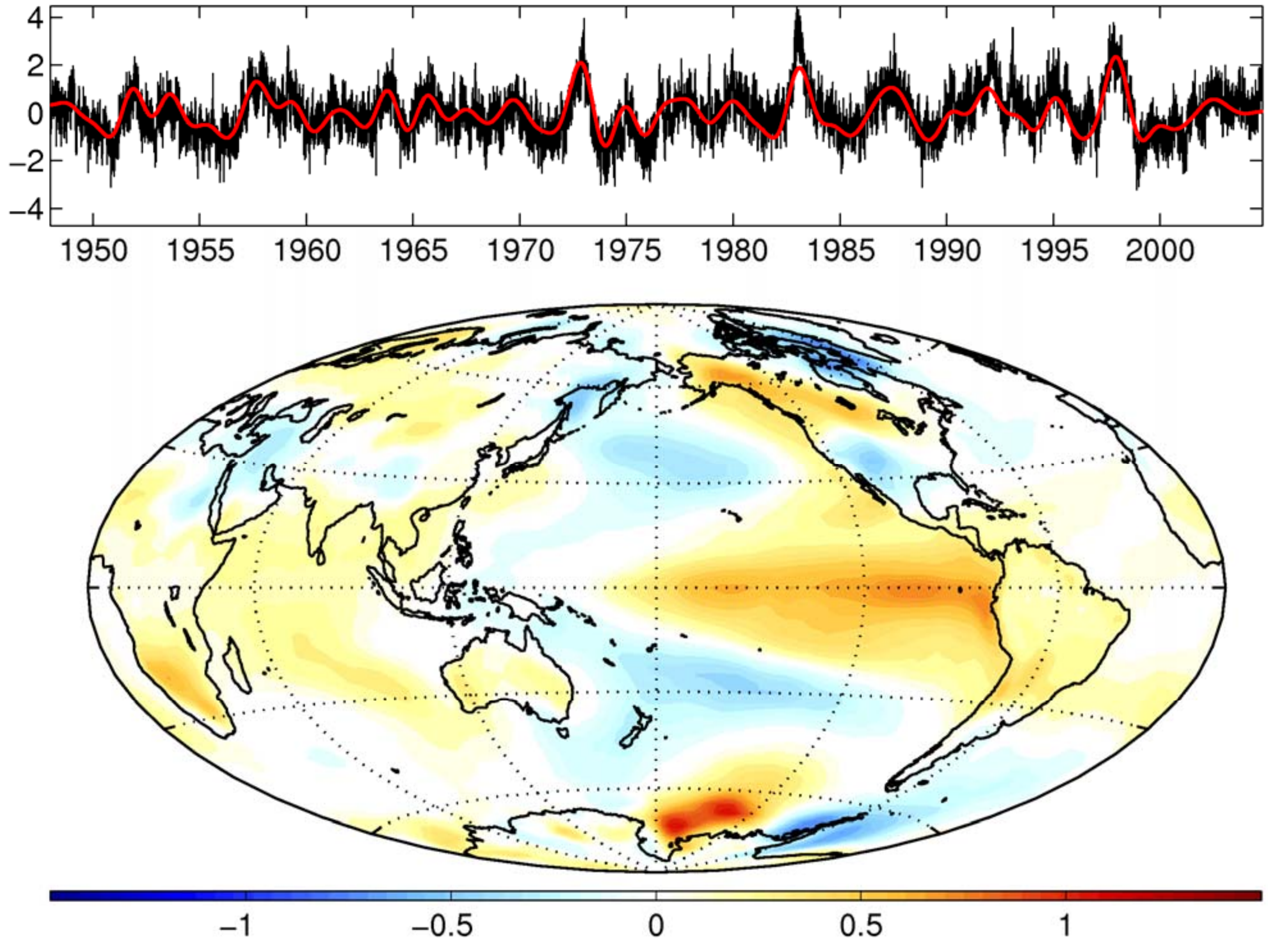
Example: Clarity-based analysis



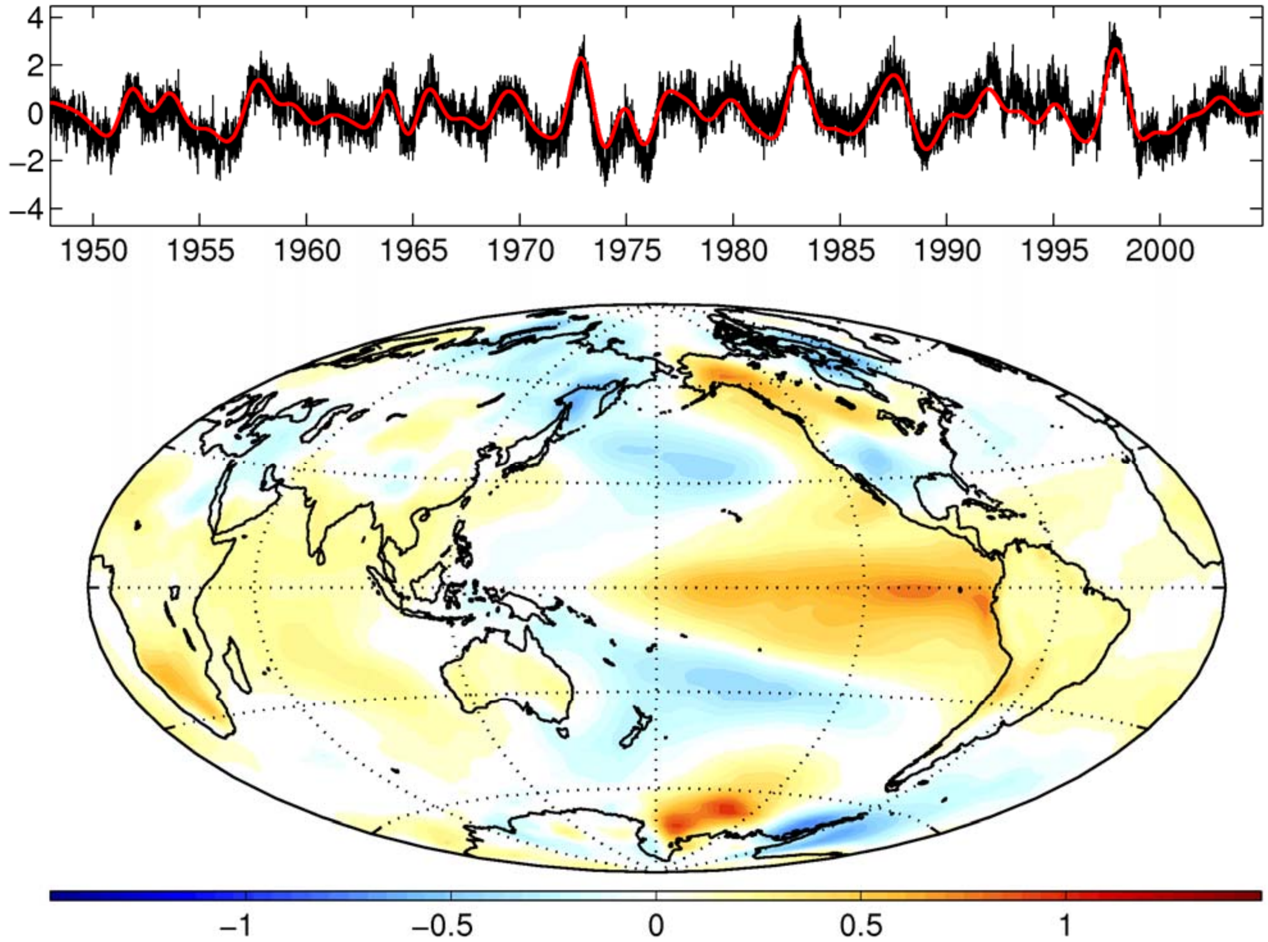
Example: Clarity-based analysis



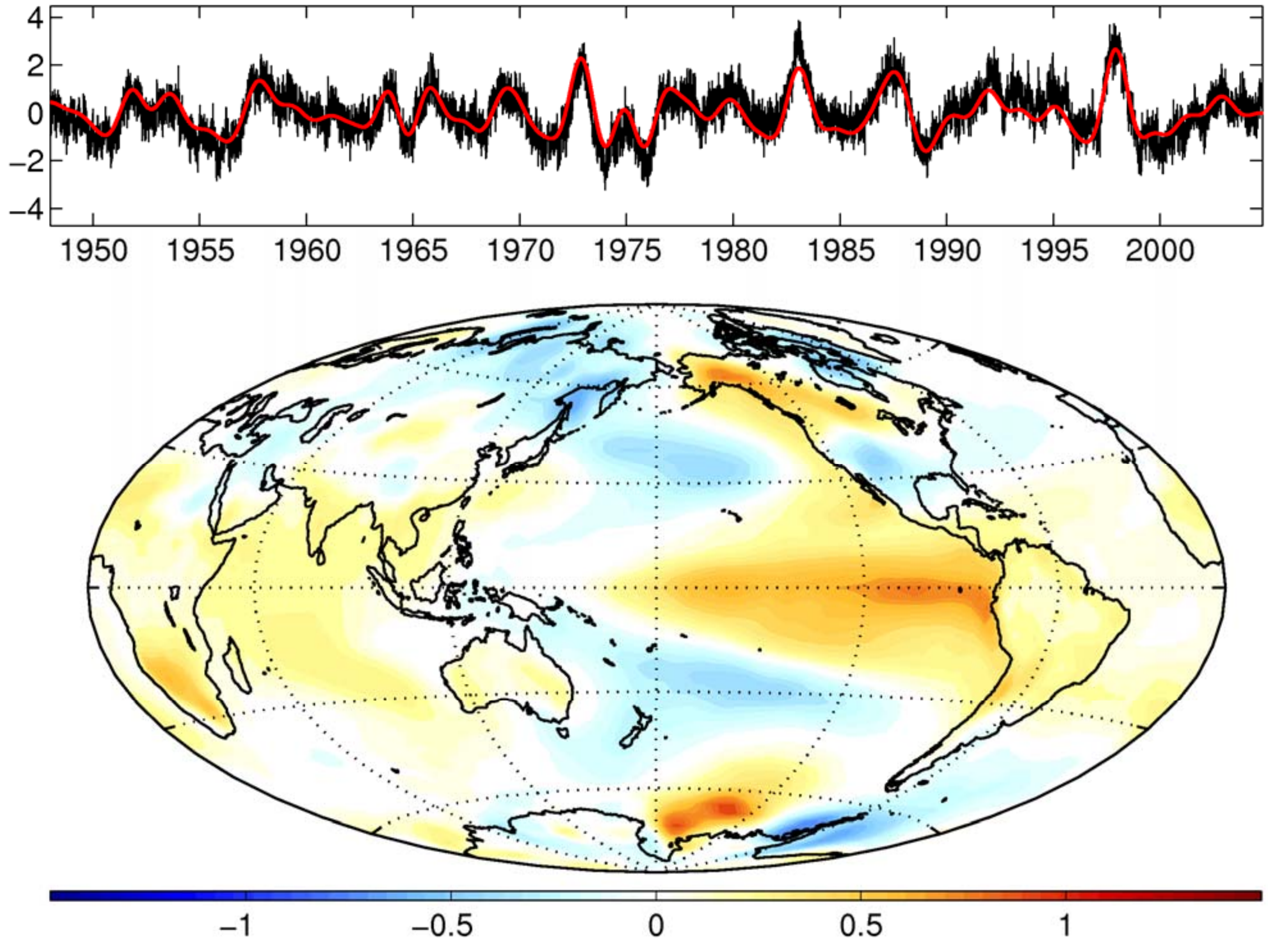
Example: Clarity-based analysis



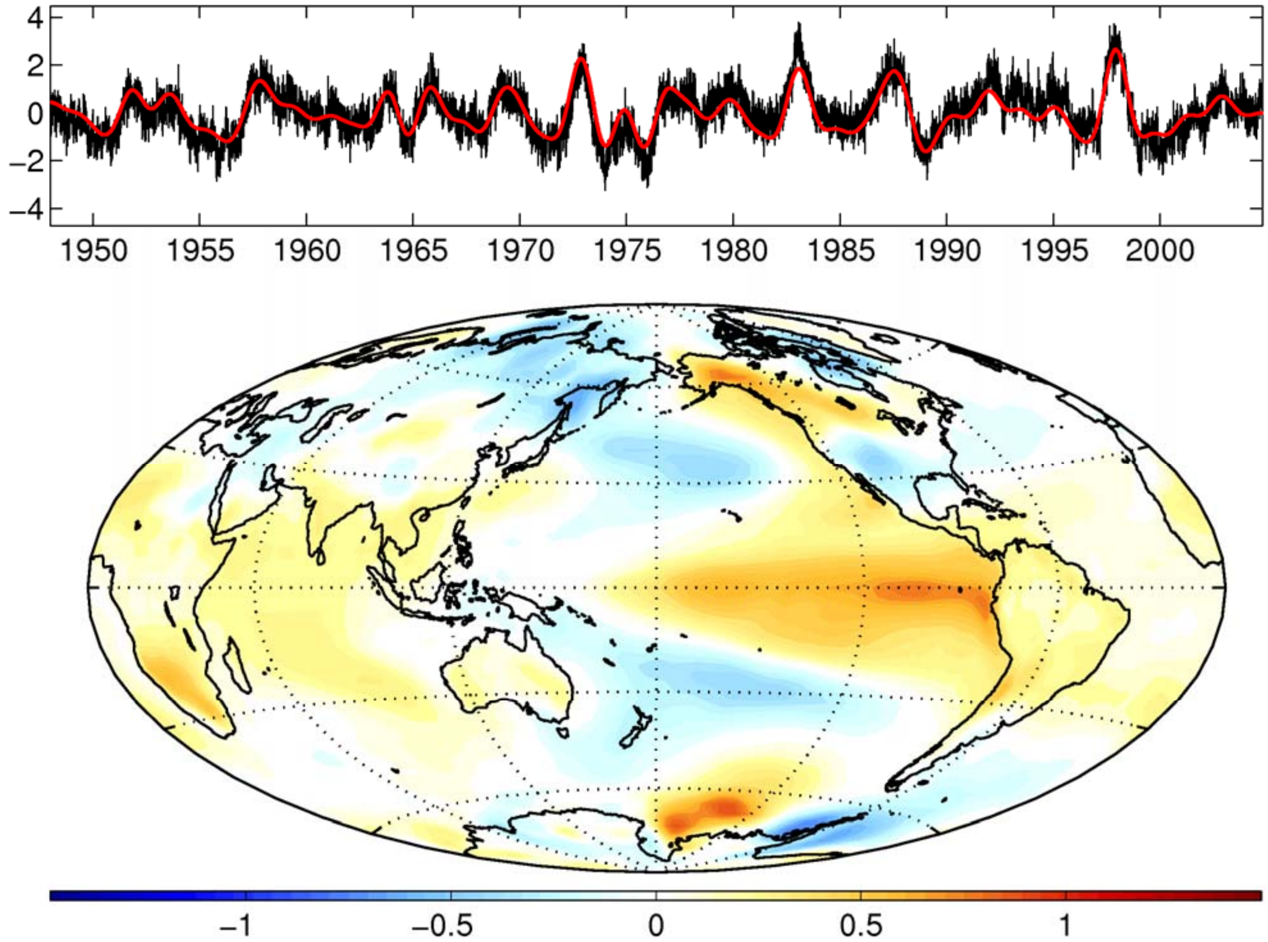
Example: Clarity-based analysis



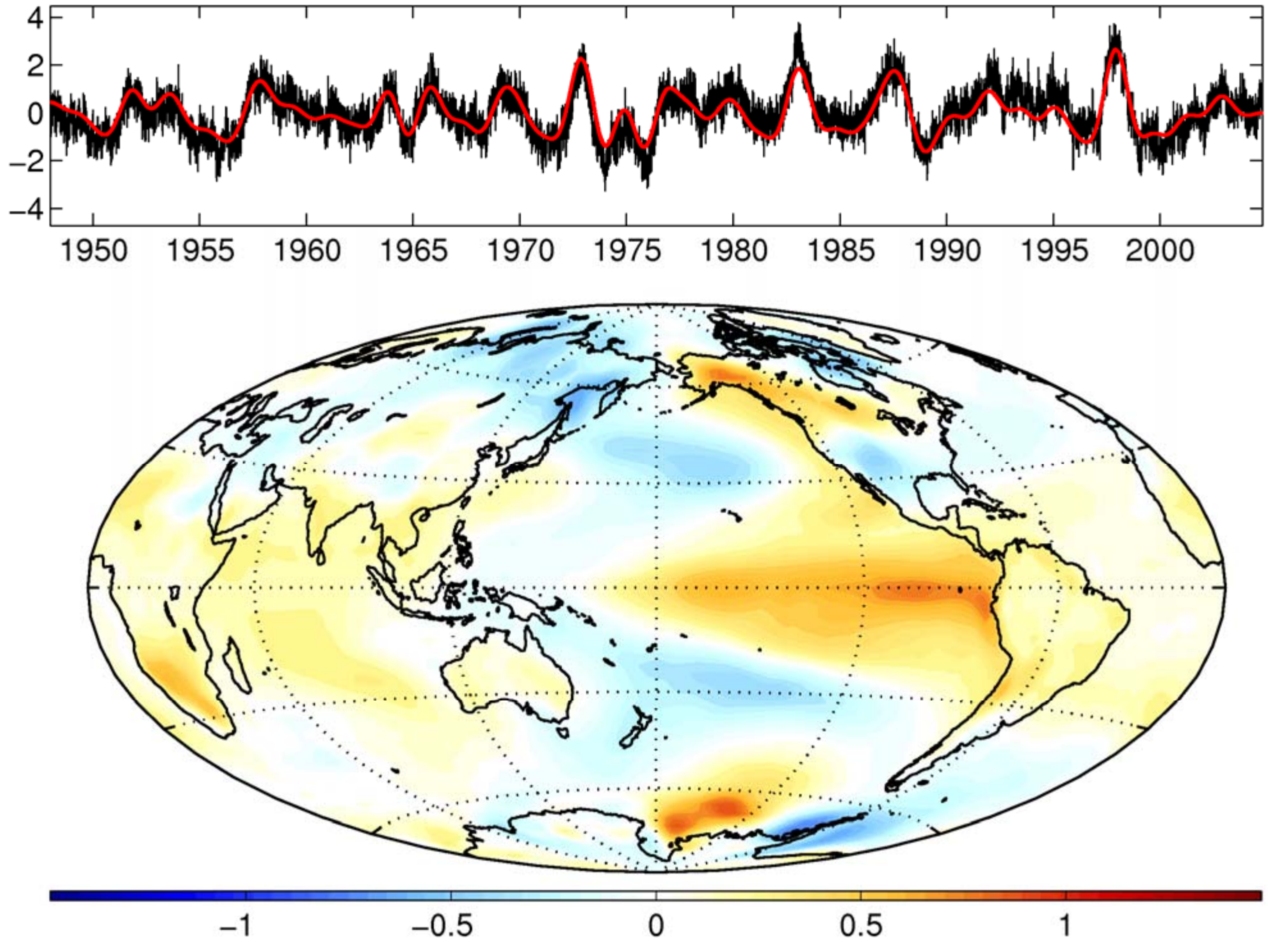
Example: Clarity-based analysis



Example: Clarity-based analysis

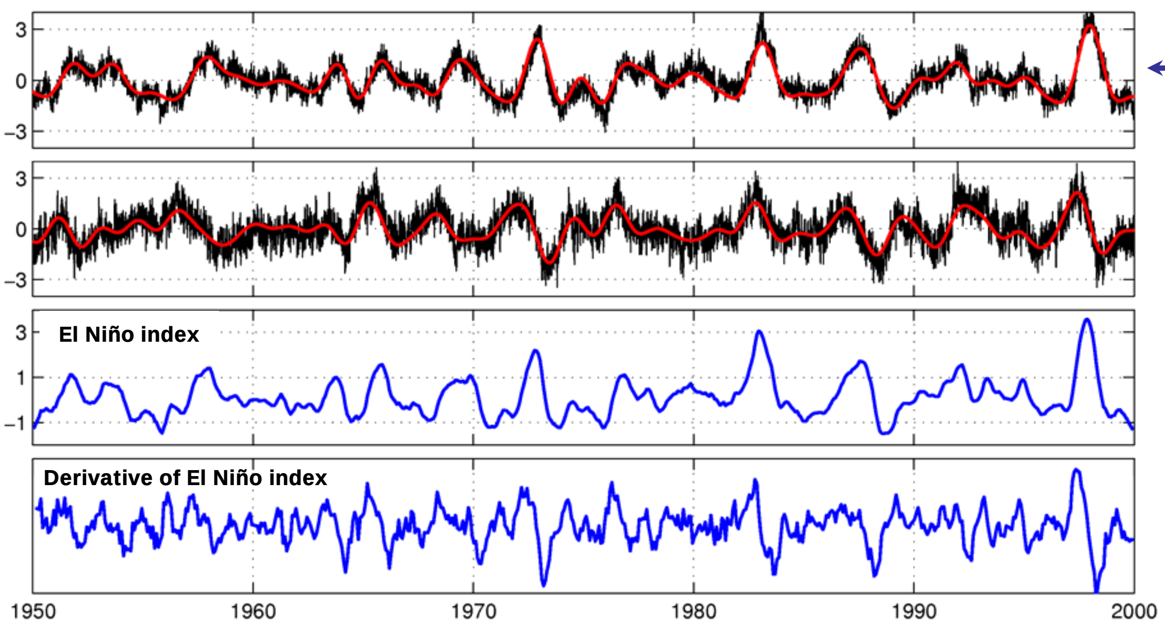


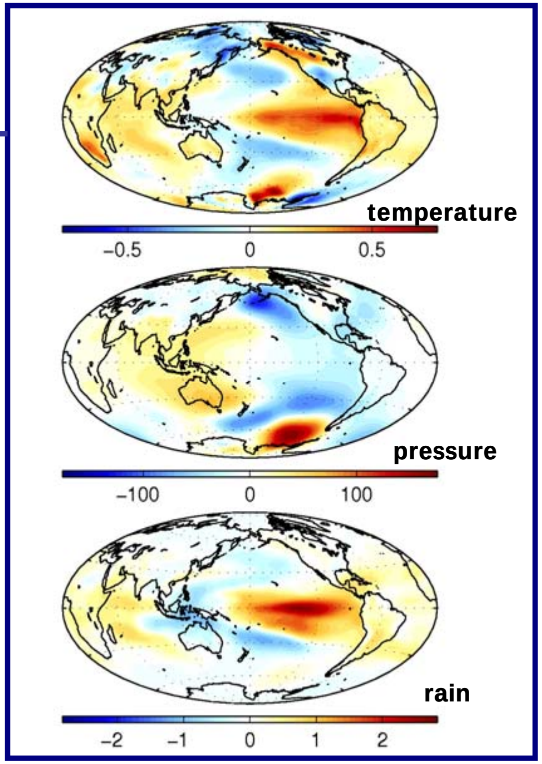
Example: Clarity-based analysis



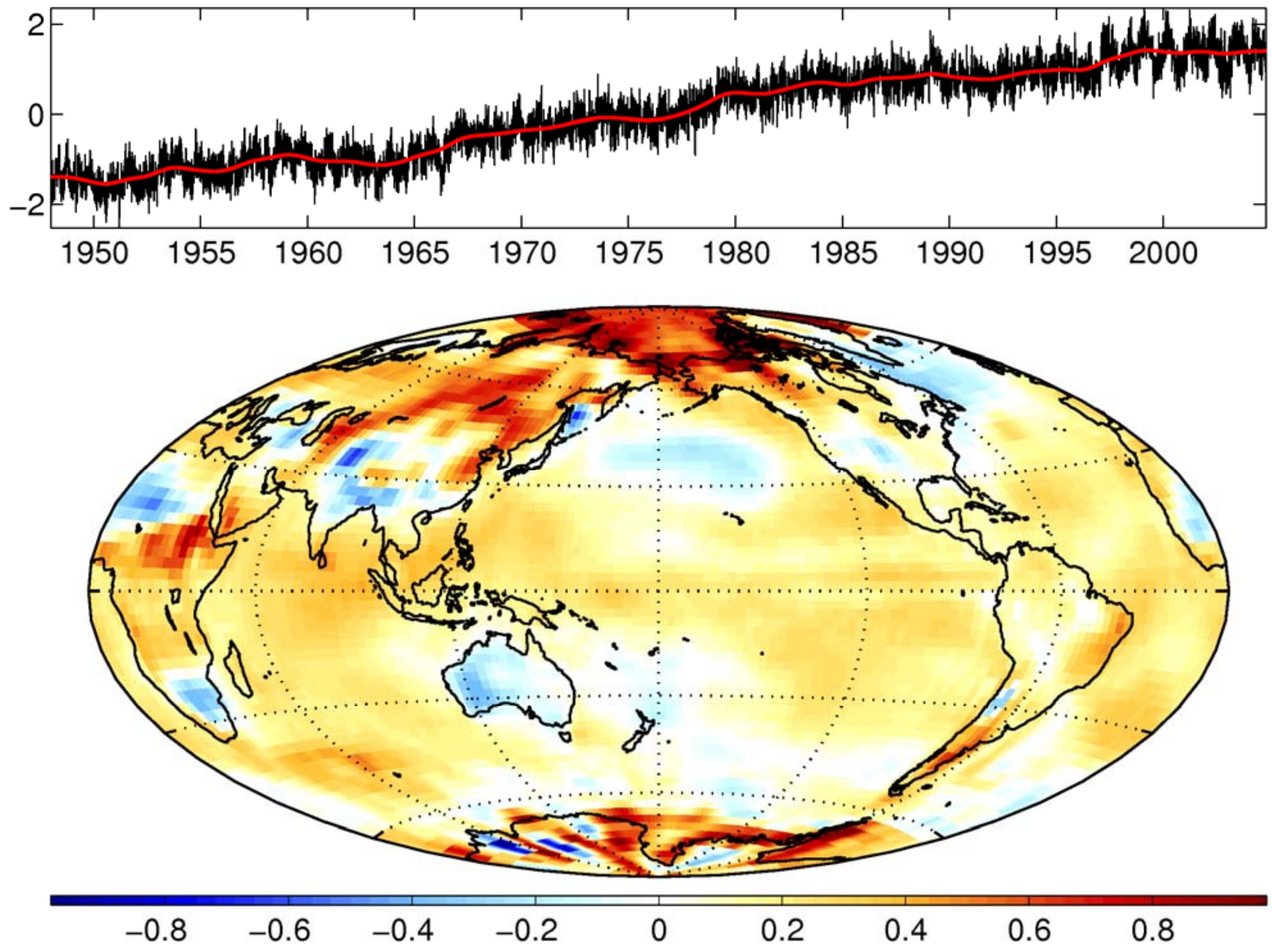
El Niño as cleanest component

- El Niño as the component with the most prominent variability in the interannual timescale





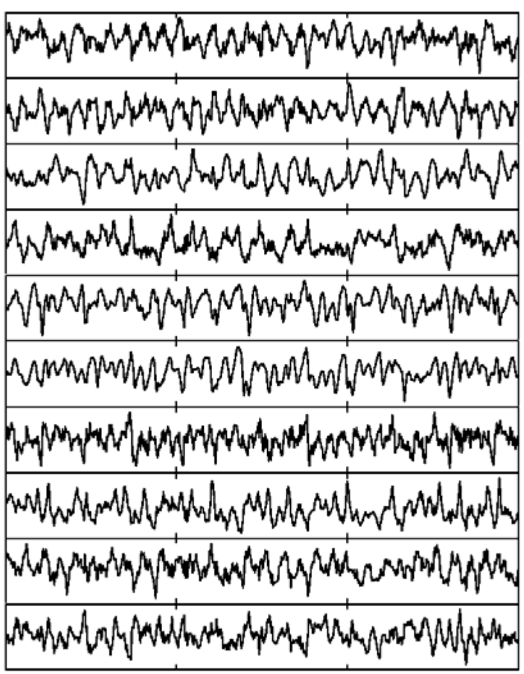
Global warming component



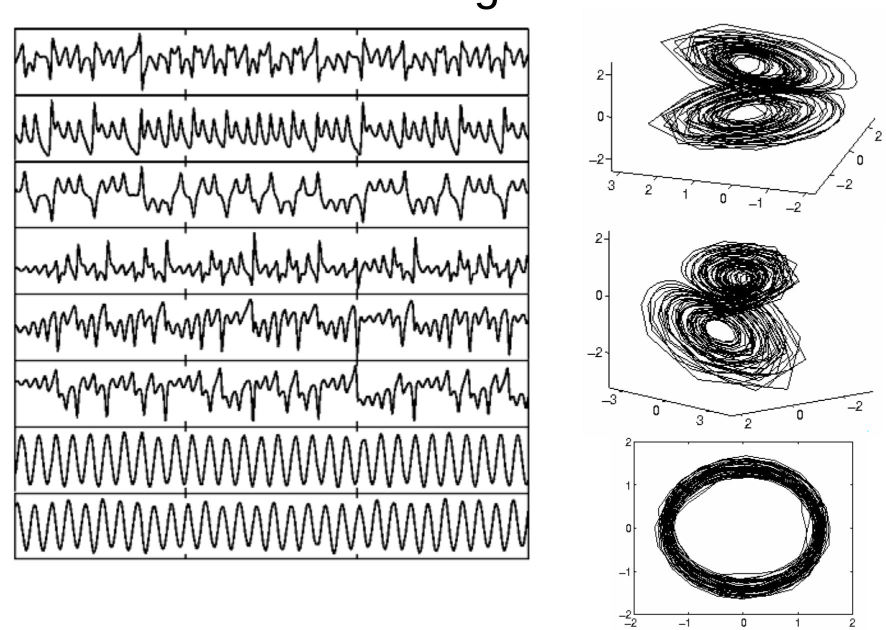
Predictable dynamics

- Modeling assumption: $\boldsymbol{s}_k(t) = \boldsymbol{g}_k(\boldsymbol{s}_k(t\text{-}1)) + \boldsymbol{m}_k(t)$
- Nonlinear function g_k is learned from data

Generated data



Extracted signals



Directions of research

- Developed techniques:
  - Prominent variability in a specific time scale
  - Signals with prominent and *distinct spectral contents*
  - Signals with prominent *variance (activation) structures* in a specific time scale
  - Groups of signals with *predictable dynamics*
- Similar techniques can be used to analyze:
  - climate change, nonlinear effects (e.g. interaction of ENSO with annual cycle), effects of climate (e.g. ecology, economy) etc

Future plans

- Methods and results reported in machine learning literature
- We are open for collaborations with climate experts (a joint research project)
- Contact us:

alexander.ilin@hut.fi

http://www.cis.hut.fi/projects/climate/