PIK

POTSDAM INSTITUTE FOR CLIMATE IMPACT RESEARCH

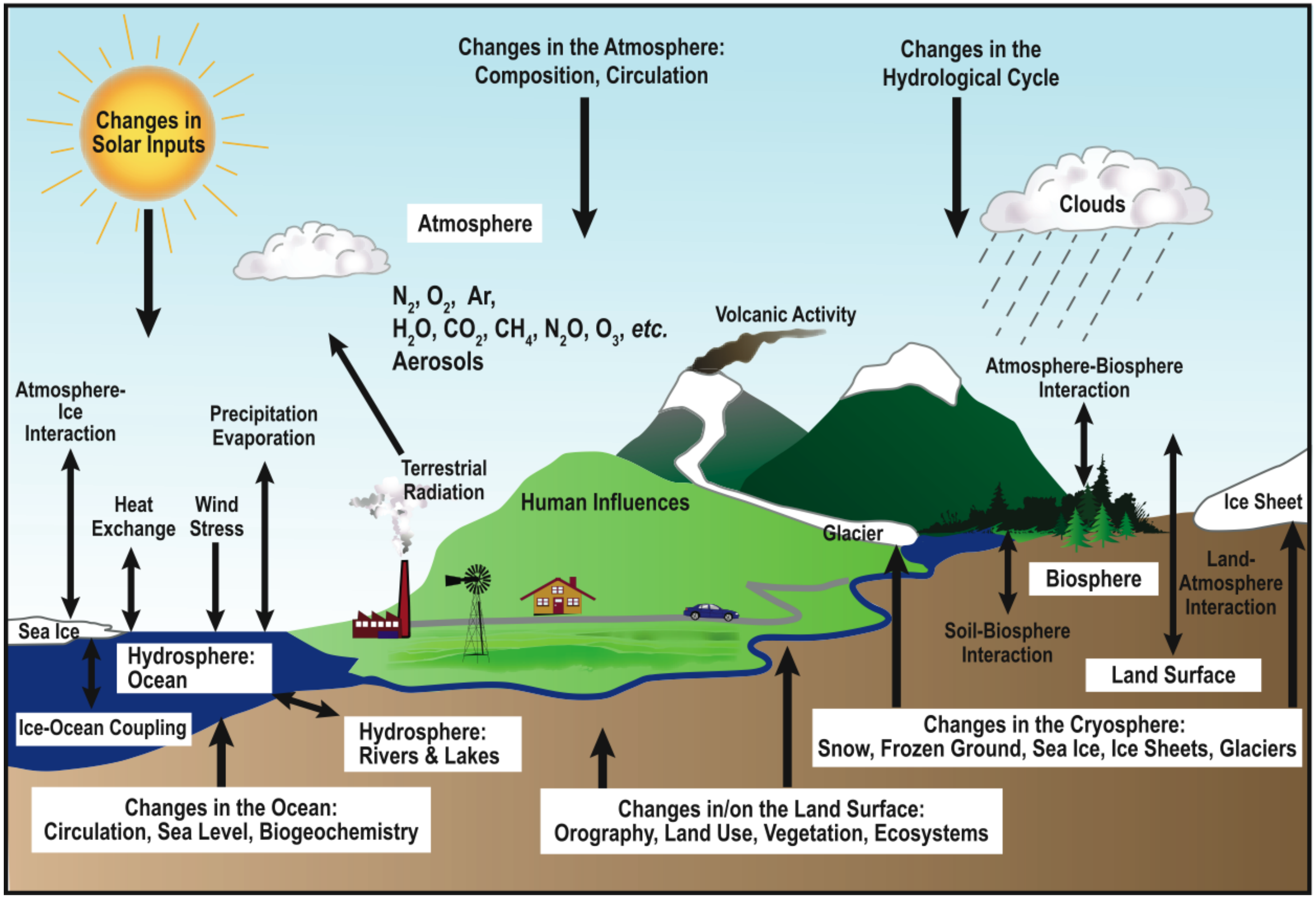
Complex systems approaches for climate data analysis

Reik V. Donner

Leibniz Gemeinschaft

COSY CC²

Climate: A conceptual view



Climate data: Major issues

non-stationarity: multiple regimes (e.g., glacial vs. interglacial climate), tipping elements

nonlinearity: complex interrelationships between subsystems, variables, regions, scales; feedbacks

multi-scale variability: different climate subsystems with distinct time scales of variability

internal dynamics vs. external forcings (solar/orbital variations, volcanism, anthropogenic greenhouse gas emissions)

Why time series analysis?

- generalization of simple one-point statistics to sequences captures information about dynamics of the observed system
- interdependences between observations made at different times
  - existence?
  - type (linear, nonlinear)?
  - strength and direction?
  - characteristic time scales?
  - continuous or variable?

$\Rightarrow$ characterization/identification of underlying processes

Different approaches to time series analysis:

- statistics (linear correlation analysis)
- econometrics (stochastic models based on regression analysis)
- theoretical physics (dynamical systems concepts)

Exploring the complex systems nexus

Applied statistics

Linear time series analysis, moments, trends, extremes, EOF analysis,...

Dynamical systems

Bifurcations, tipping points, phase space concept,...

Data mining

Automated learning, pattern recognition,...

Information theory

Binary data representation, entropy and information, statistical mechanics

Exploring the complex systems nexus

Applied statistics

Linear time series analysis, moments, trends, extremes, EOF analysis,...

Data mining

Automated learning, pattern recognition,...

Dynamical systems

Bifurcations, tipping points, phase space concept,...

Nonlinear methods

Information theory

Binary data representation, entropy and information, statistical mechanics

Testing for nonlinearity in time series

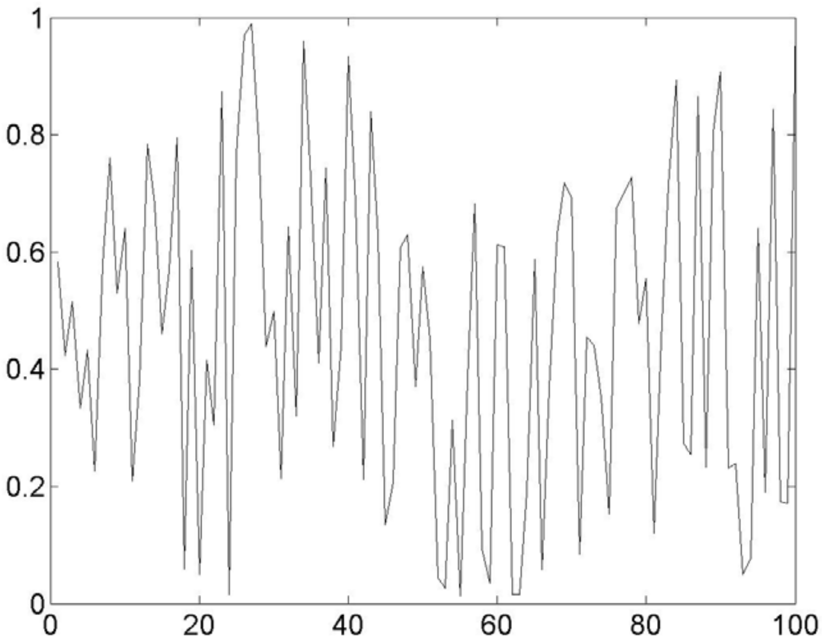
Classical (linear) time series analysis captures many important (statistical) properties of data (correlations, power spectra), but:

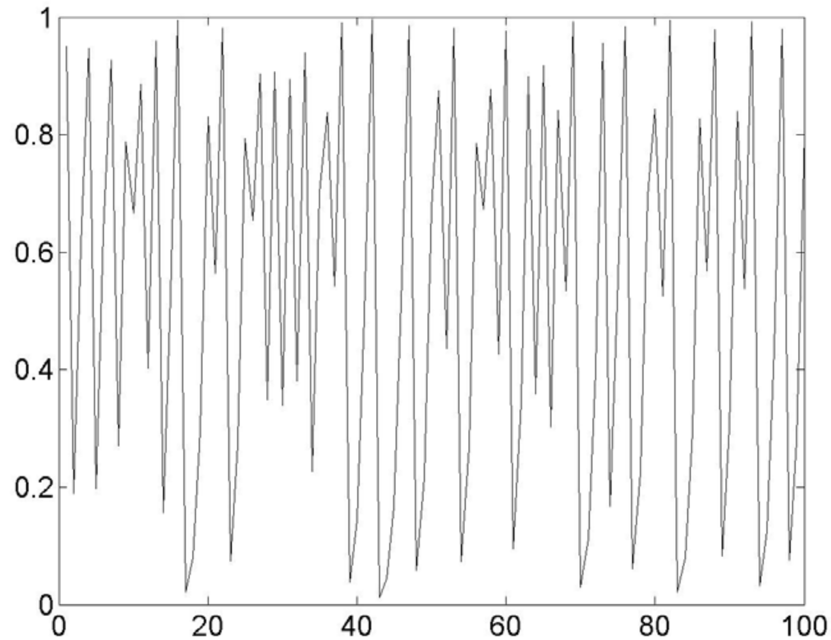
- **relies on the assumption of linear-stochastic processes**
- **whereas nature is most often nonlinear**

Linear statistics cannot distinguish between qualitatively different types of dynamics and is therefore not suitable for fully characterizing nonlinear dynamics!

Testing for nonlinearity in time series

Example: Which process is linear, which one nonlinear?

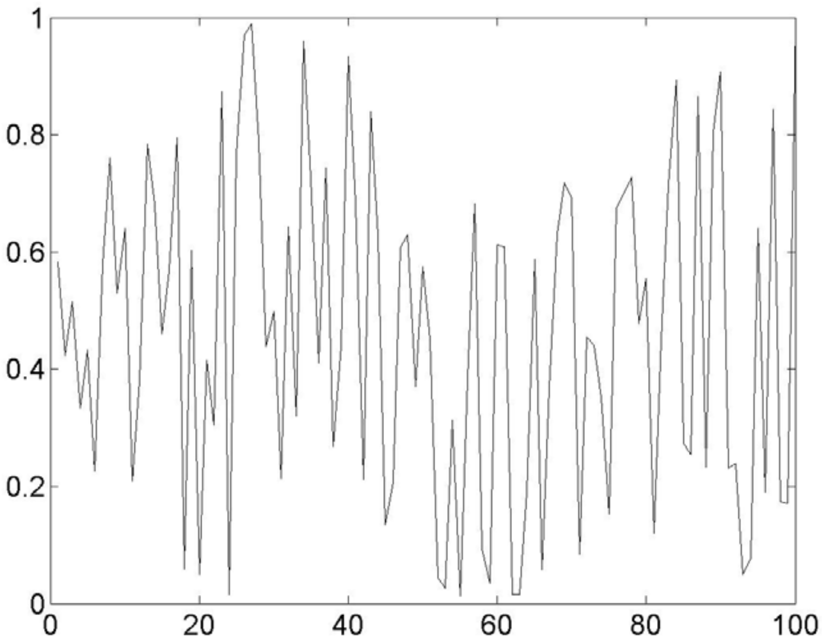




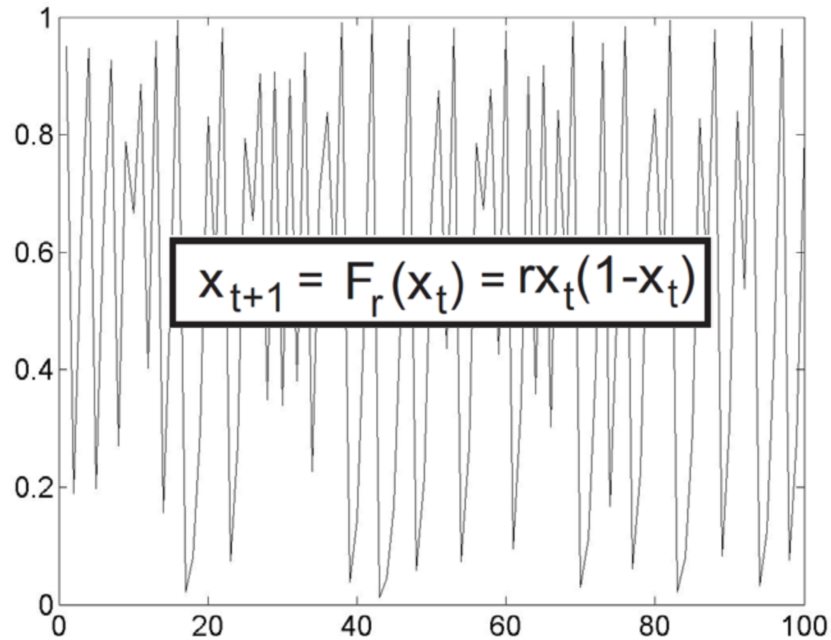
Testing for nonlinearity in time series

Example: Which process is linear, which one nonlinear?

Uncorrelated random process



Logistic map



Assessing nonlinear interrelationships

Information-theoretic methods

Entropy **2013**, *15*, 4844-4888; doi:10.3390/e15114844

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Review

Statistical Mechanics and Information-Theoretic Perspectives on Complexity in the Earth System

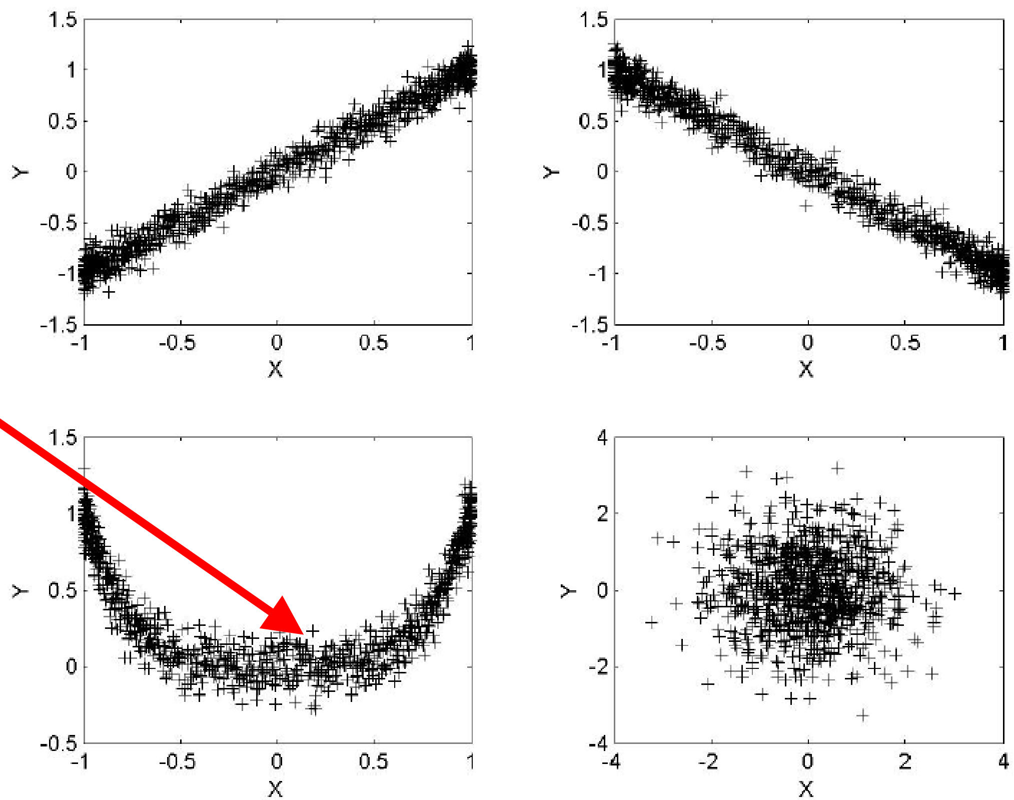
Georgios Balasis [1,*], Reik V. Donner [2,3], Stelios M. Potirakis [4], Jakob Runge [2,5], Constantinos Papadimitriou [1,6], Ioannis A. Daglis [1,6,†], Konstantinos Eftaxias [7] and Jürgen Kurths [2,5,8]

Nonlinear statistical interdependencies

linear Pearson correlation: no information on general statistical associations!

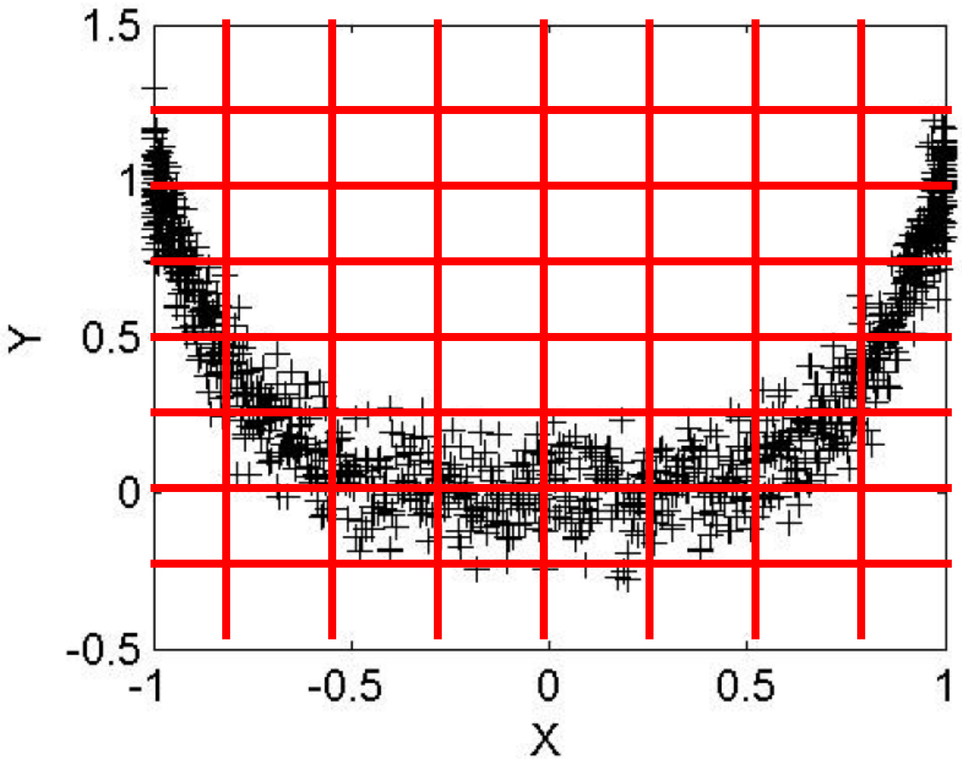
low correlation coefficient,

high mutual information



Mutual information

Basic idea: quantification of contingency table of symbol (joint vs. marginal probabilities)



Symbolic time series analysis

More general: transform time series into discretized representation using abstract symbols *a* from discrete (finite) alphabet *A*

$\Rightarrow$ allows computation of different information-theoretic quantities:

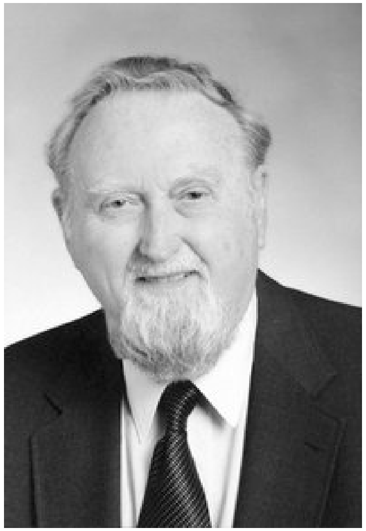
symbolic correlation function

$$\boxed{C_{XY}(\tau) = \sum_{a \in A} P_{aa}^{XY}(\tau)}$$

mutual information (measure for general statistical association)

$$\boxed{I_{XY}(\tau) = \sum_{a,b \in A} P_{ab}^{XY}(\tau) \log_2 \frac{P_{ab}^{XY}(\tau)}{P_a^X P_b^Y}}$$

Identification of causality from time series



Classical approach: linear Granger causality

1. **Build linear regression models (bivariate AR models)**

$$X_1(t) = \sum_{j=1}^{p} A_{11,j} X_1(t-j) + \sum_{j=1}^{p} A_{12,j} X_2(t-j) + E_1(t)$$

$$X_2(t) = \sum_{j=1}^{p} A_{21,j} X_1(t-j) + \sum_{j=1}^{p} A_{22,j} X_2(t-j) + E_2(t)$$

2. **Compare variance of error term E_1 (E_2) with and without inclusion of X_2 (X_1) in the first (second) equation**
   - **If additional term for X_2 in equation for X_1 reduces error: X_2 Granger-causes X_1**
   - **If additional term for X_1 in equation for X_2 reduces error: X_1 Granger-causes X_2**
   - **Practical: are $A_{12,j}$ ($A_{21,j}$) significantly different from 0 (e.g., via F test)?**

Identification of causality from time series

Multivariate problem: given *n>2* variables X_k,

X_i Granger-causes X_j if lagged observations of X_i help in prediction X_j when lagged observations of all other variables X_k are taken into account (cf. partial correlations)

$\Rightarrow$ **Multivariate (conditional) Granger causality**

Problems: linearity, stationarity, unobserved variables are not considered

Many extensions:

- **Nonlinear Granger causality**
- **Spectral Granger causality (fraction of total spectral power of X_1 at a given frequency f that is provided by X_2)**
- **Transfer entropy: conditional mutual information taking lags into account**

Assessing multi-scale dynamics

EOF analysis

Original motivation: extract dominating variations from spatio-temporal fields of climate observations records

<u>Linear PCA:</u>

Diagonalization of lag-zero covariance matrix C of multivariate time series (matrix *X*)

$$C = X^T X \quad \textbf{\textit{with}} \quad C = U^T \Sigma U \quad \textbf{\textit{and}} \quad \Sigma = diag(\sigma_1^2, \ldots, \sigma_N^2)$$

- **Compute correlation matrix of all variables**
- **Estimate eigenvalues and eigenvectors**
- **Eigenvectors: additive decomposition into principal components (weighted superpositions of original variables) with individual variances corresponding to associated eigenvalues**

$\Rightarrow$ **spatial EOF patterns + index/score time series describing magnitude and sign of individual EOF modes**

Singular spectrum analysis

Replace individual data "channels" by time-lagged versions of the same time series
$\Rightarrow$ Identification of distinct temporal variability patterns

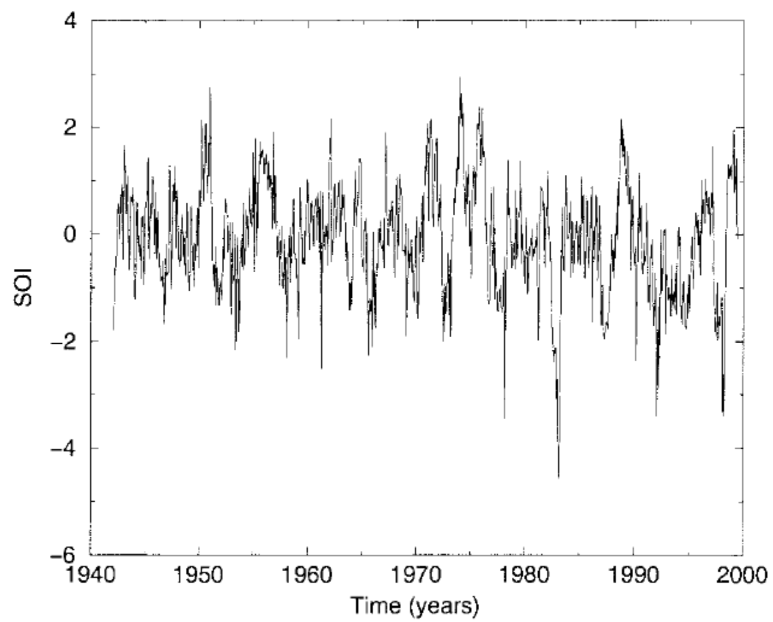
Mathematical roots: time-delay embedding
$\Rightarrow$ Attractors of dissipative deterministic dynamical systems can be approximated from univariate time series by using "independent coordinates" in terms of time-shifted replications of the original data
$\Rightarrow$ Standard tool in nonlinear time series analysis

Relevant information:

- **How many "dimensions" (variables) need to be used to model the observed dynamics? (number of relevant singular values)**
- **What are the associated relevant variability patterns (associated singular vectors)**

Singular spectrum analysis

Example: Relevant dimensions of ENSO activity (Ghil et al., 2001)



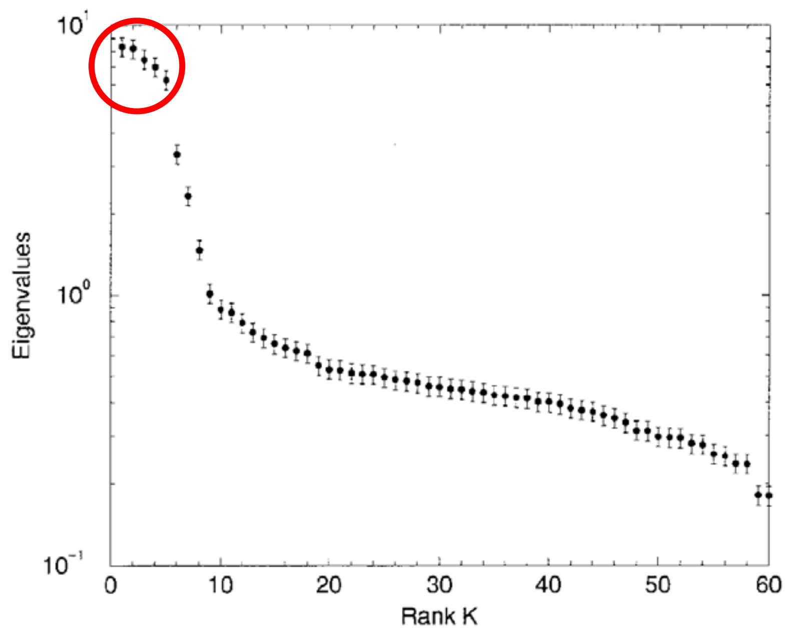
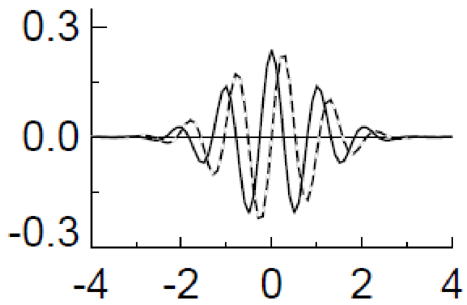


Figure 2. Variations of the Southern Oscillation Index (SOI) between January 1942 and June 1999. Time on the abscissa is in calendar years, and SOI on the ordinate is normalized by its standard deviation.

Wavelet analysis

Widely applicable tool: wavelet analysis

- **Straightforward generalization of classical spectral analysis: convolution of signal with scalable localized oscillatory function (mother wavelet)**
  - **Example: Morlet wavelet – sinusoidal with Gaussian envelope**

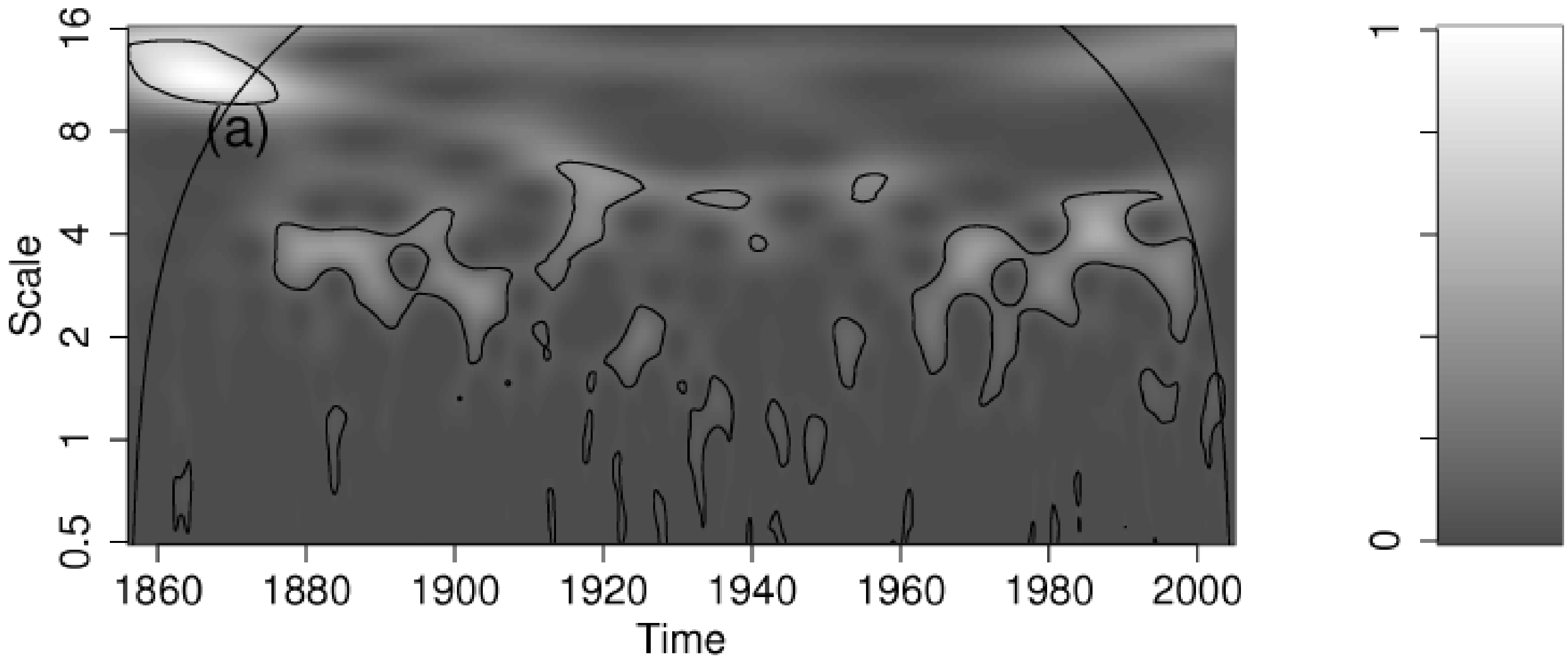


$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2}$$

- **Discrete wavelet analysis: additive decomposition of time series into components in logarithmically spaced frequency bands**
- **Continuous wavelet analysis: filter with central frequency at any desired periodicity**

Wavelet analysis

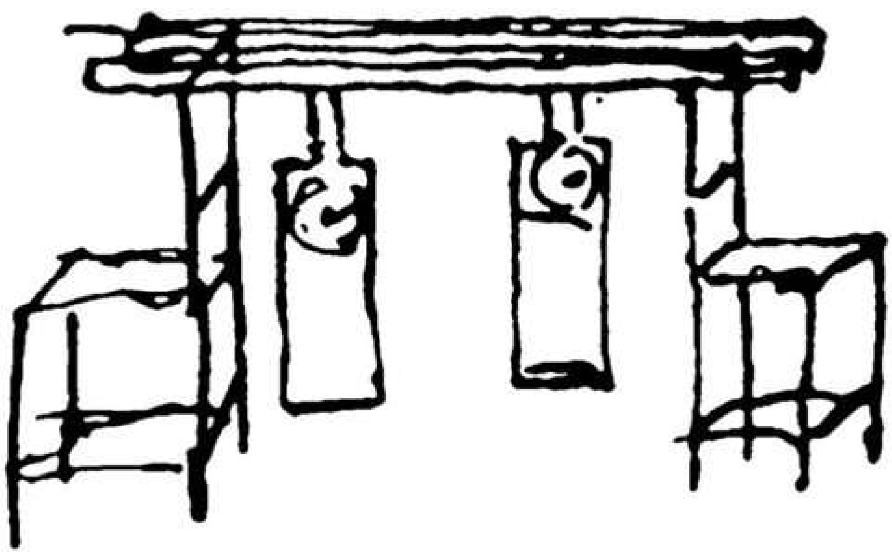
Example: wavelet spectrogram of Nino3.4 index *(Maraun, PhD thesis, 2006)*



Synchronization analysis

Mutual alignment of two or more (not necessarily regular) oscillatory signals

First described by Huygens in 1673 (synchronization of two pendulum clocks hanging on the same wooden beam)



Various types: complete, generalized, phase, lag, event synchronization

Phase synchronization

Classical approach for detecting phase synchronization:

1. **Define phase variable from data (Hilbert or wavelet transformation, stroboscopic mapping,…)**

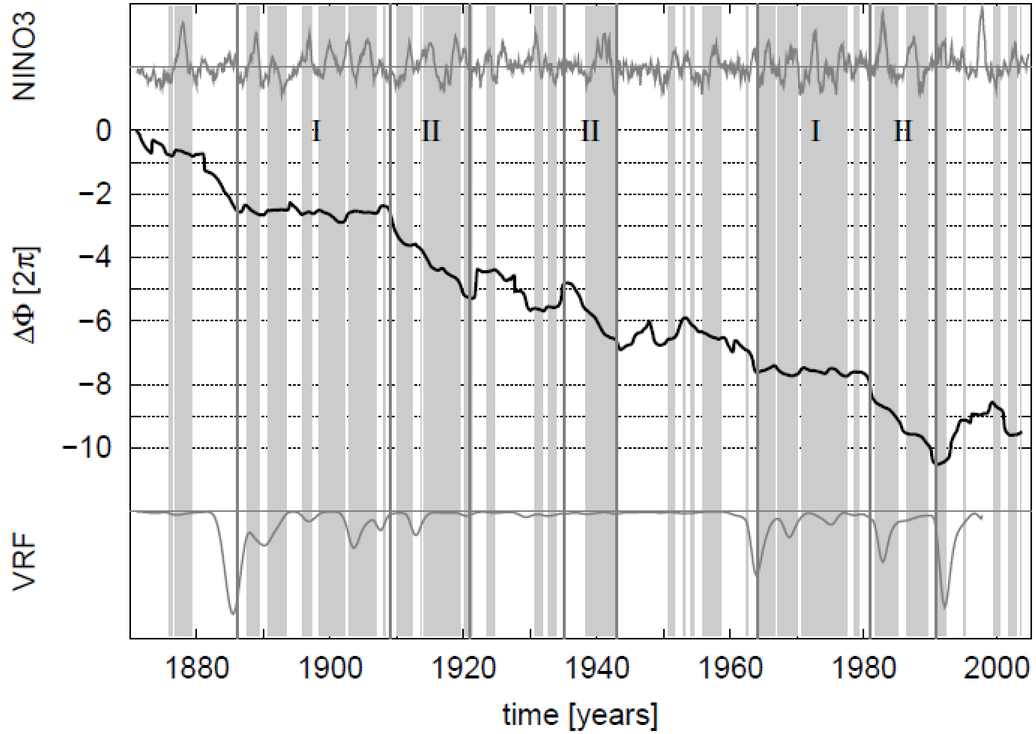
$$\psi(t) = s(t) + j\tilde{s}(t) = A(t)e^{j\phi(t)} \qquad \tilde{s}(t) = \pi^{-1}\mathrm{P.V.}\int_{-\infty}^{\infty} \frac{s(\tau)}{t-\tau}\, d\tau$$

2. **Compute time series of phase differences between two phase time series**
3. **Wrap the phase differences to the interval [0,2π]**
4. **Compute suitable statistics on wrapped phase differences (Rayleigh measure, standard deviation, Shannon entropy)**

Alternative to 2-4: compute mutual information between phase time series (Palus 1997)

Phase synchronization

Example. Phase coherence between ENSO and Indian monsoon *(Maraun & Kurths, GRL, 2005)*



Cross-scale interactions

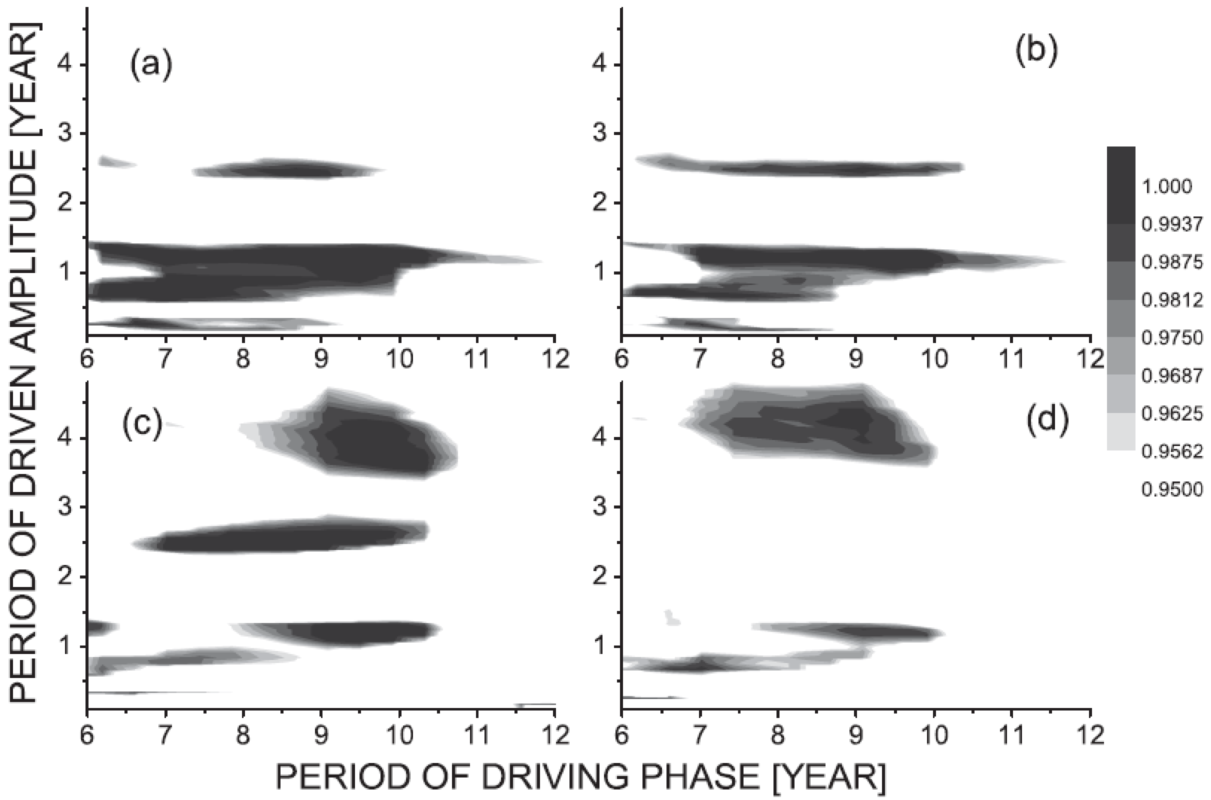
Idea (Palus, PRL, 2014): combine

- **continuous wavelet analysis,**
- **phase synchronization analysis (decomposition into amplitude and phase variables) and**
- **Nonlinear interdependencies (mutual information)**

$\Rightarrow$ **Unveil cross-scale phase-amplitude (phase-phase, amplitude-amplitude) coupling in observed time series**

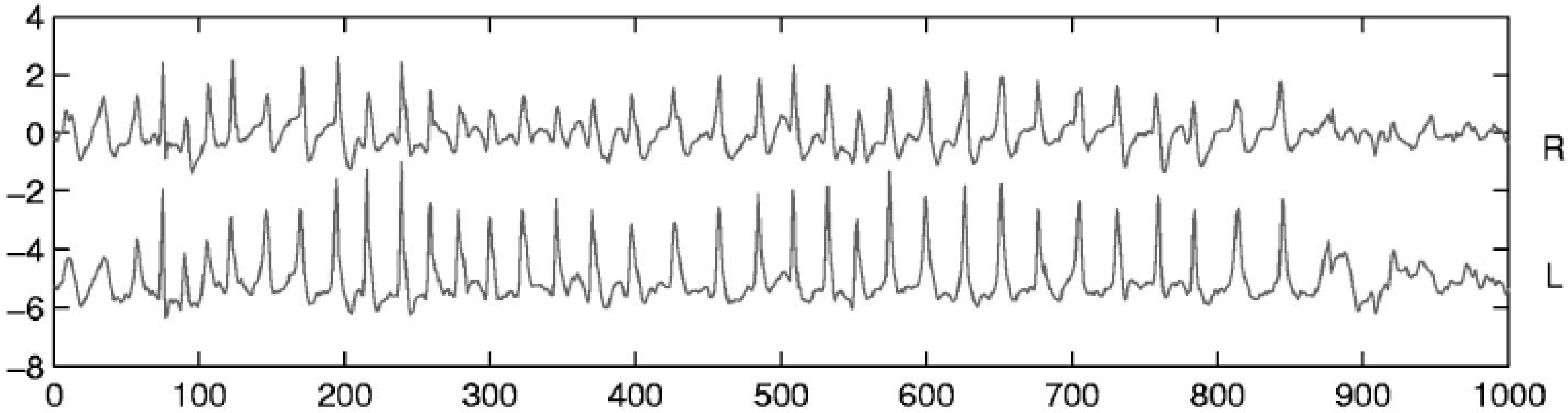
Cross-scale interactions

Example: four long-term air temperature records from Central Europe *(Palus, PRL, 2014)*



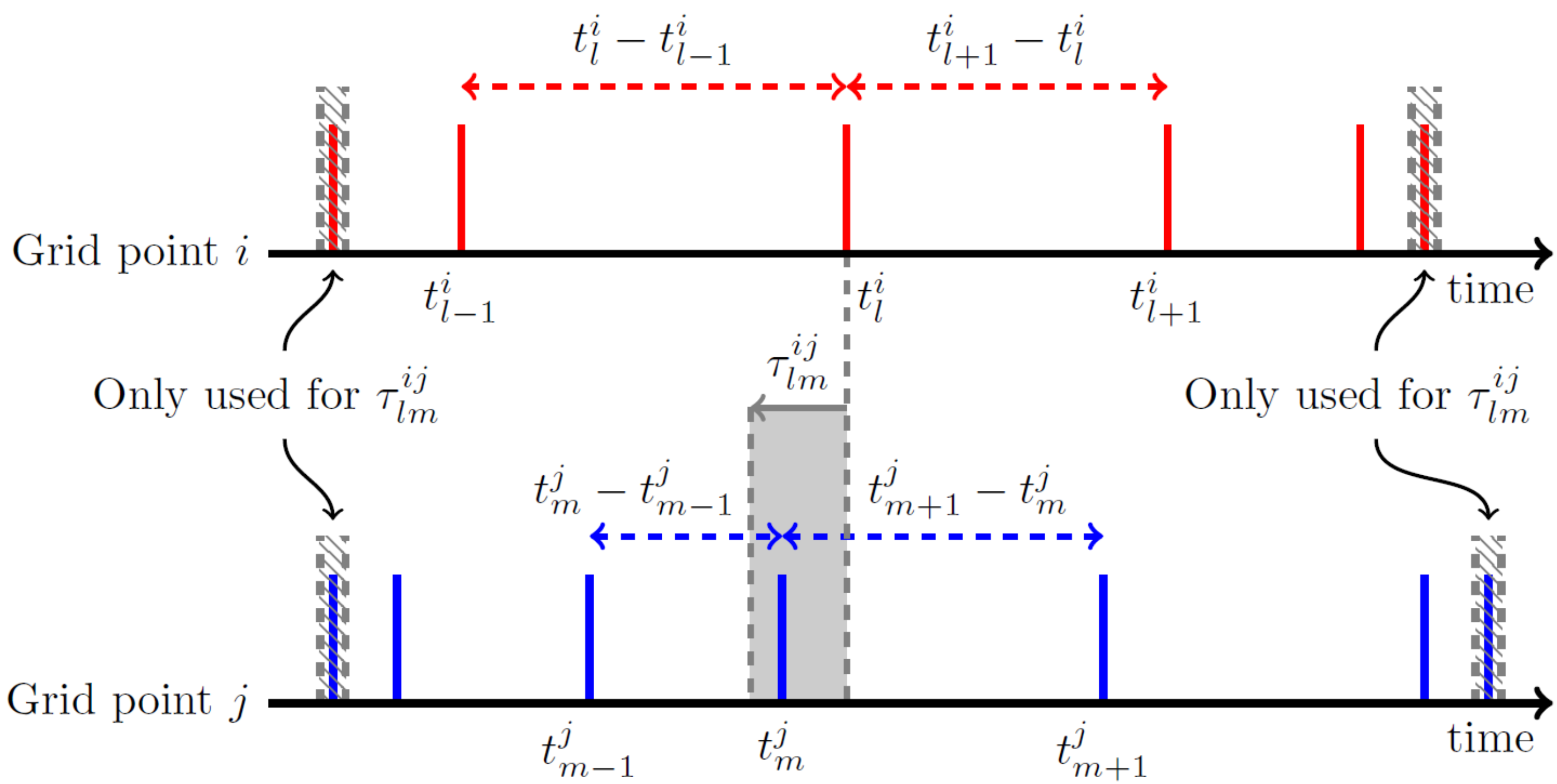
Event synchronization

Common approach from neurosciences (spike train analysis from EEG recordings, Quian Quiroga et al., PRE, 2002) - normalized fraction of temporally close extremes observed at different spatial locations (EEG electrodes, climate data grid point)



Various (even more concepts) further extending this approach (e.g., spike train synchrony), commonly tailored to specific features of neuronal spike trains

Event synchronization



Event synchronization: Definition

$$\theta_{lm}^{jk} = \min\left\{t_{l+1}^{j} - t_{l}^{j}, t_{l}^{j} - t_{l-1}^{j}, t_{m+1}^{k} - t_{m}^{k}, t_{m}^{k} - t_{m-1}^{k}\right\}/2$$

$$c(j|k) = \sum_{l=1}^{s_j}\sum_{m=1}^{s_k} J_{lm}^{jk}$$

with

$$J_{lm}^{jk} = \begin{cases} 1 & \text{if} \quad \tau'^{jk} < t_l^j - t_m^k < \tau^{jk} \\ 1/2 & \text{if} \quad t_l^j - t_m^k = \tau'^{jk} \\ 0 & \text{else,} \end{cases}$$

$$\boxed{\mathbf{Q}_{jk} = \frac{c(j|k) + c(k|j)}{\sqrt{(s_j s_k)}}}$$

Disadvantages:

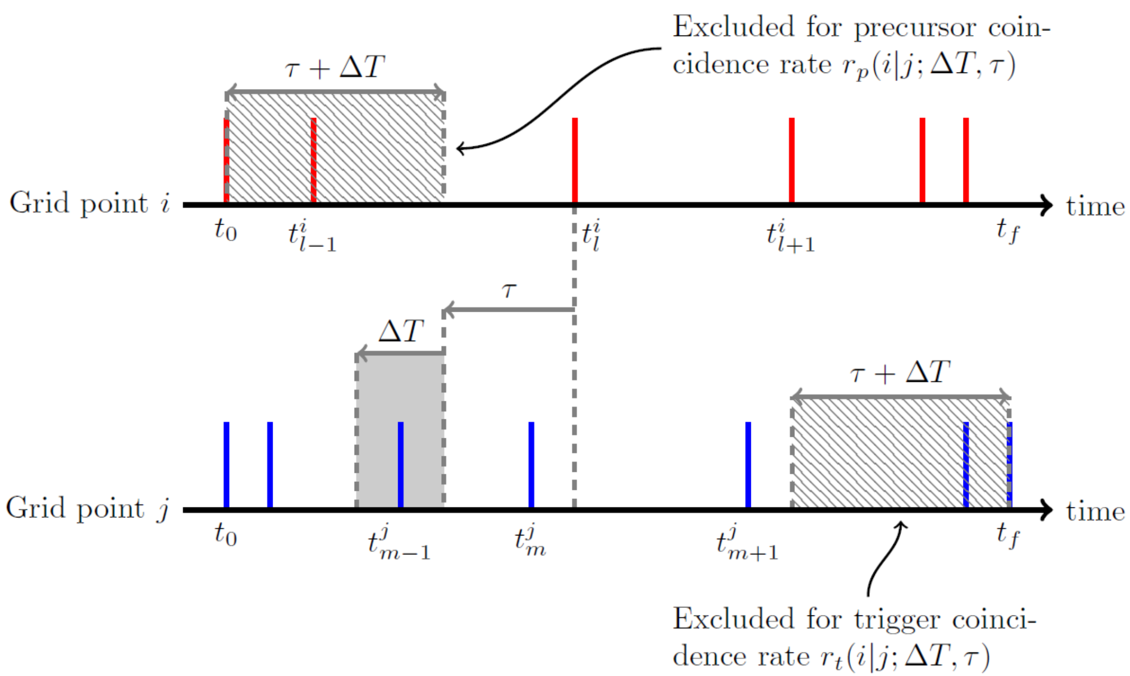
- **adaptive "local" definition of event proximity, tailored to events with approximately constant spacing in time - requires de-clustering in case of temporally close events**
- **no analytics, significance testing at most via Monte Carlo methods**

Event coincidence analysis

Take one of the series as reference and count number of cases in which at least one event in the other series occurs within in given time window relative to the timing of the reference event

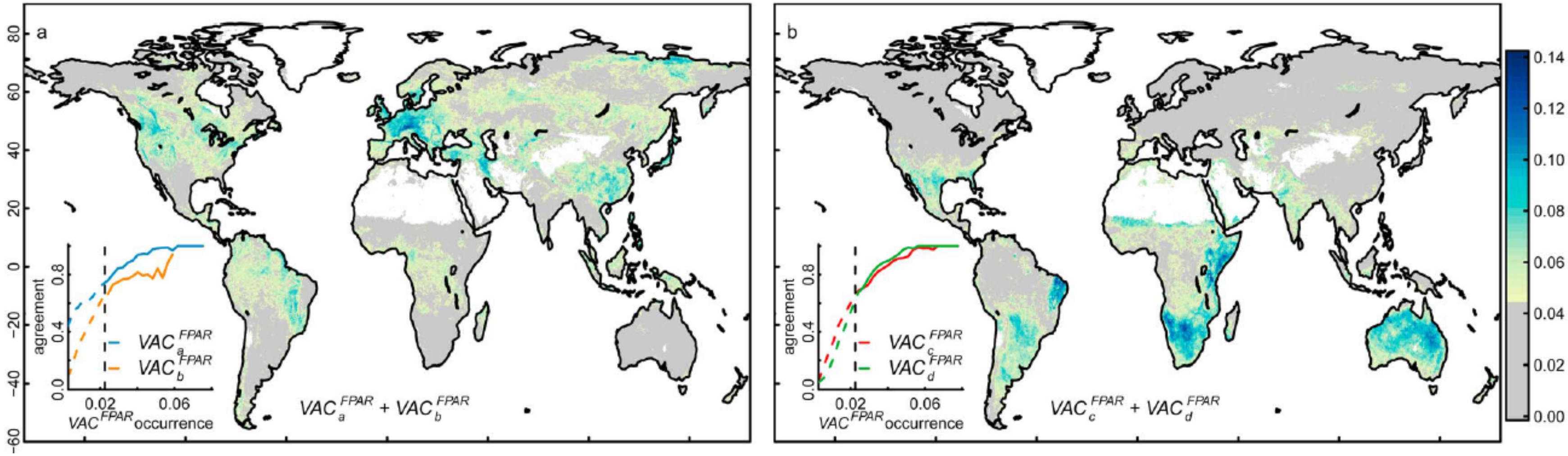
⇒ Asymmetric property (potential for establishing directionality statements)

⇒ Distinction between “trigger” and “precursor” tests, simple significance tests



Example: Remote sensing – FPAR vs. ET extremes

Sign of Anomalies		Index	SM Regime	Status
$T-$	FPAR/ET–	VAC_a	energy-limited	wettening, atmospheric control
$T+$	FPAR/ET+	VAC_b	energy-limited	drying, atmospheric control
$T+$	FPAR/ET–	VAC_c	transitional	drying, land/vegetation control
$T-$	FPAR/ET+	VAC_d	transitional	wettening, land/vegetation control



Zscheischler et al., GRL, 2015

Take home messages

- Climate data exhibit complex variability features
- Classical time series analysis methods are not capable of unveiling these features
- Nonlinearity: interdependencies maybe overlooked by linear analyses, use nonlinear similarity measures like mutual information instead
- Multi-scale variability: time-scale decomposition to identify relevant scales, study interdependence between (oscillatory) components at different scales
- Event based interdependency measures: relevance for extreme event and (discrete) impact studies