Data driven modeling using high resolution water quality monitoring data:

An appetizer.

Gunnar Lischeid

Leibniz Center for Agricultural Landscape Research, Müncheberg (Germany)

and University of Potsdam (Germany)

lischeid@zalf.de

Motivation

Automatic water quality monitoring yields

- Large data sets (continuous time series);
- High temporal resolution data.

These data can be used in a new, unprecedented way:

- New approaches to be applied.
- New questions to be asked.

Some suggestions ...

Basic Assumptions

Observed time series of a ***single variable*** usually reflect effects of a ***variety of different processes***.

But:

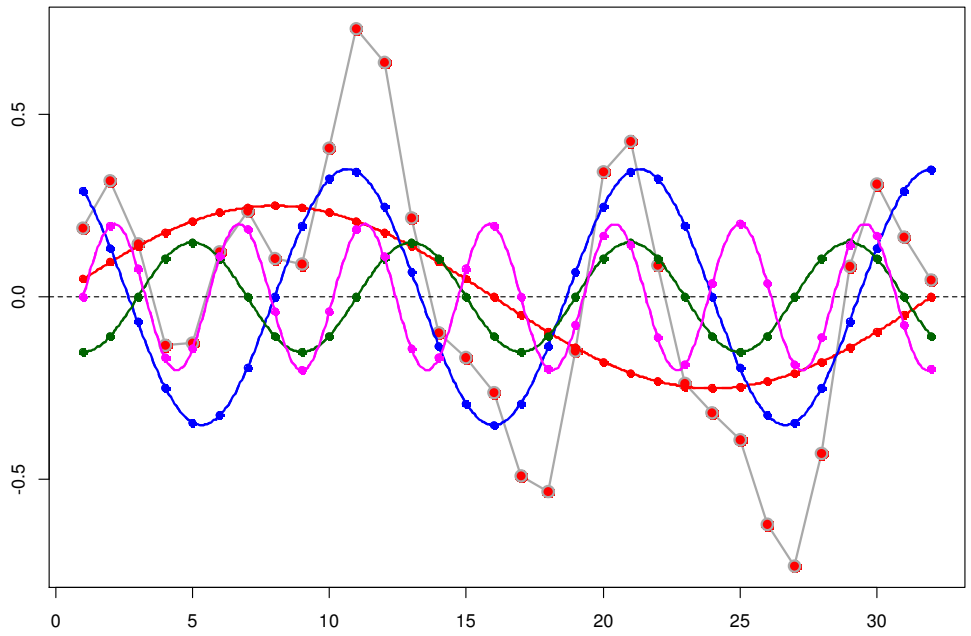
1. Every single process imprints a typical pattern (“signal”) on the observed dynamics.
2. Processes (in most cases) simply superimpose their respective signals.
3. Only few processes prevail.
4. Contribution of single processes to the observed dynamics varies in time and space, even if the basic processes are the same everywhere.

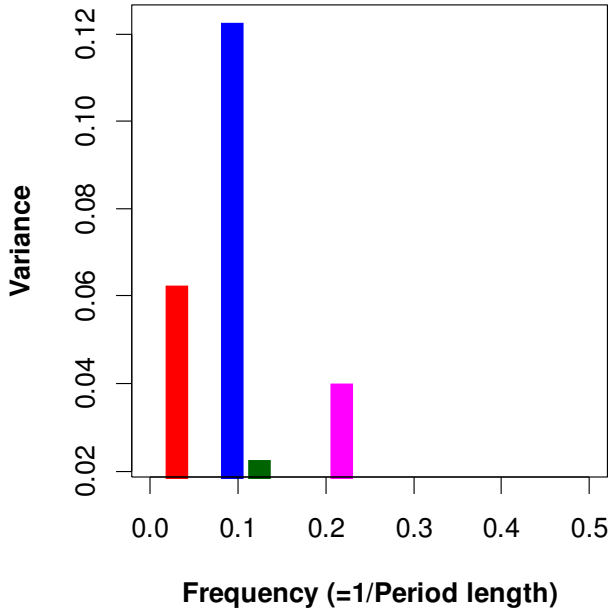
Questions and Approaches

1. **Typical time scales / periodicities**
2. Damping behaviour
3. Visualizing the dynamics
4. Intrinsic dimensionality
5. Process identification

Basic Idea

- Most time series exhibit periodicities (diurnal, seasonal, El Nino, sun spot cycle, ….)
- Even time series, that do not show clear periodicities, can be decomposed into a series of sine and cosine functions via Fourier analysis => “hidden” periodicities can be detected.

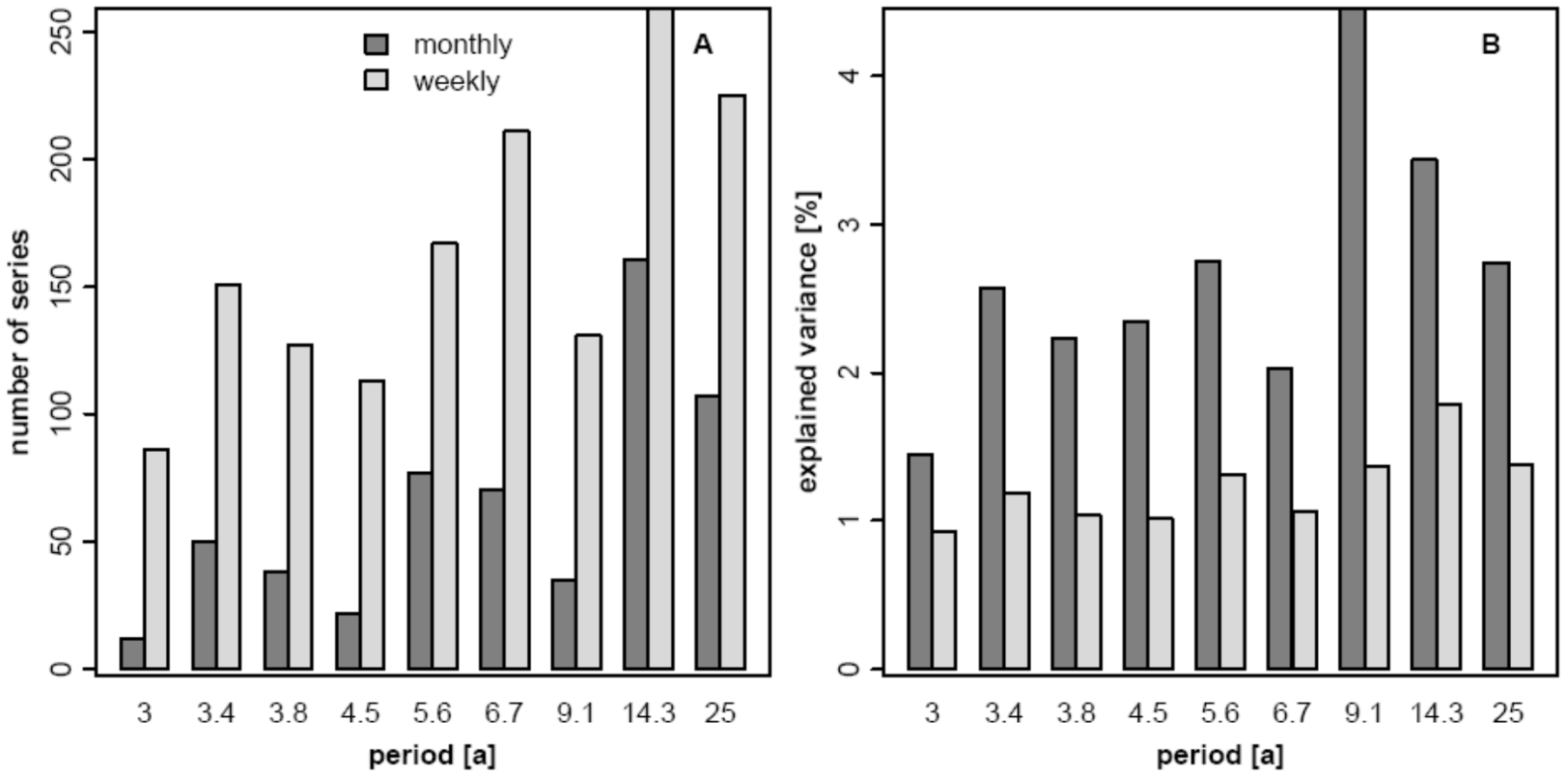




Example: Detected Periodicities

SSA of 387 time series of discharge (1948-2006; USA)

(*Gudmundsson et al. 2007*)

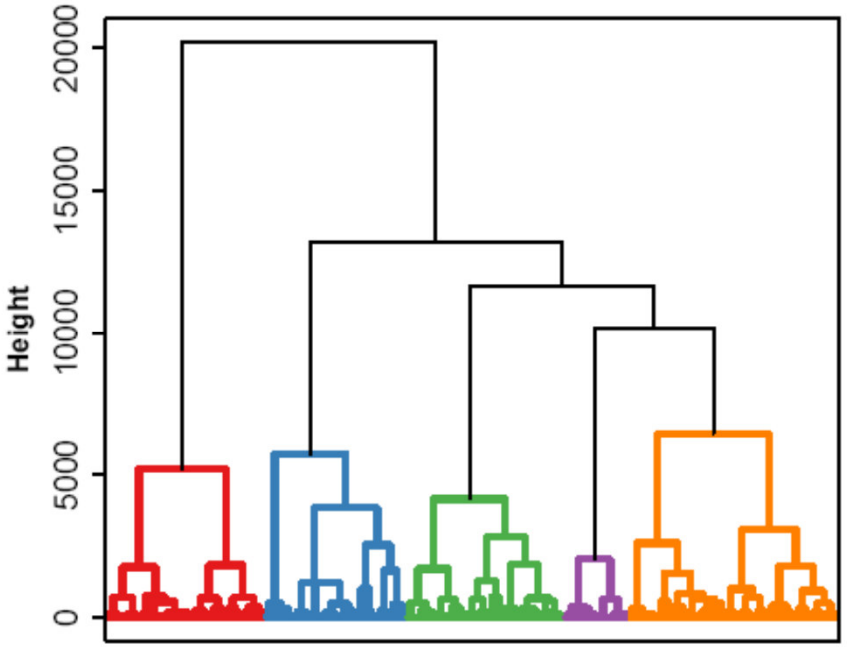


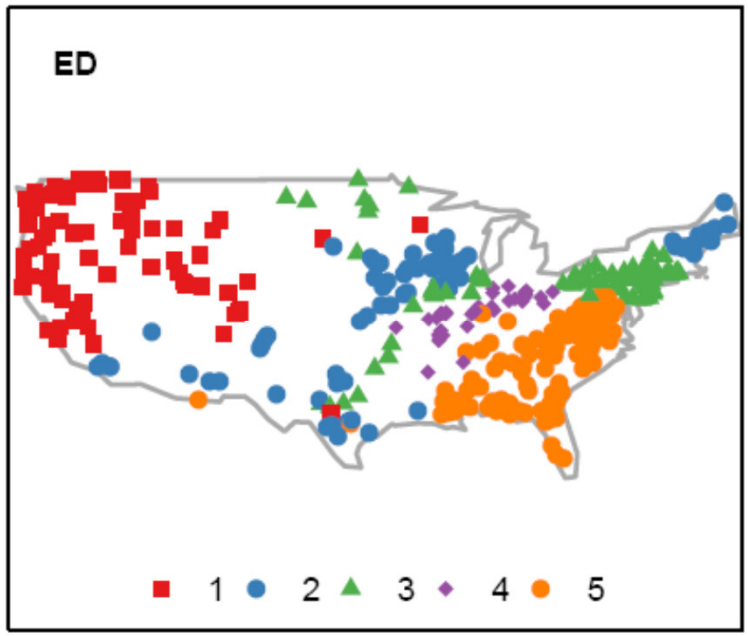
Example: Detected Periodicities

SSA of 387 time series of discharge (1948-2006; USA)

(*Gudmundsson et al. 2007*)

Subsequent cluster analysis:

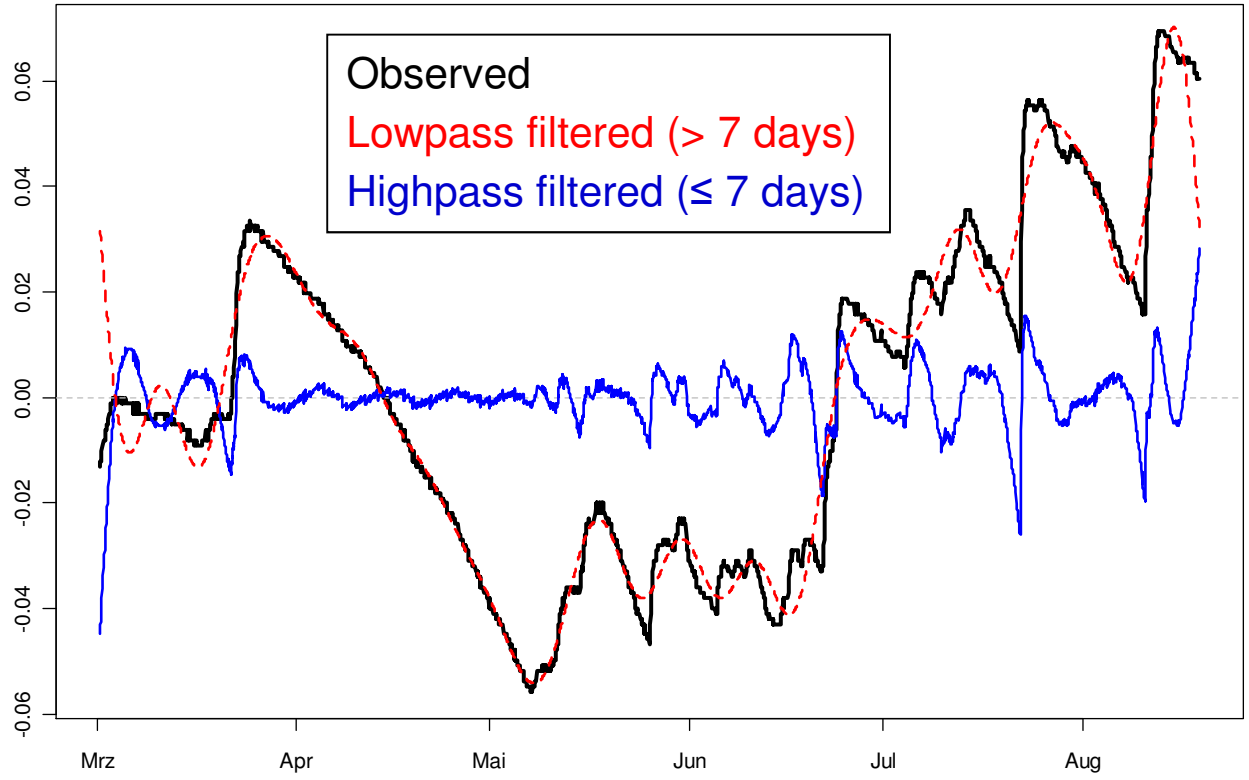




Spectral Filtering: Example

Typical time scales of groundwater head data (pressure transducer):

- Recharge events: Weeks
- Air pressure fluctuations: Hours - days



(*Lischeid et al. 2010*)

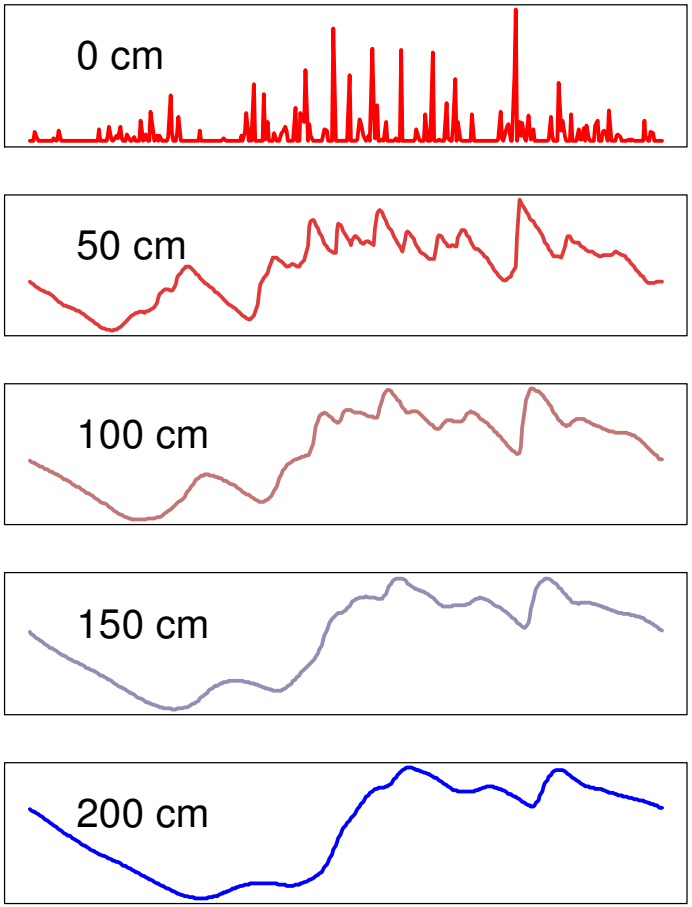
Questions and Approaches

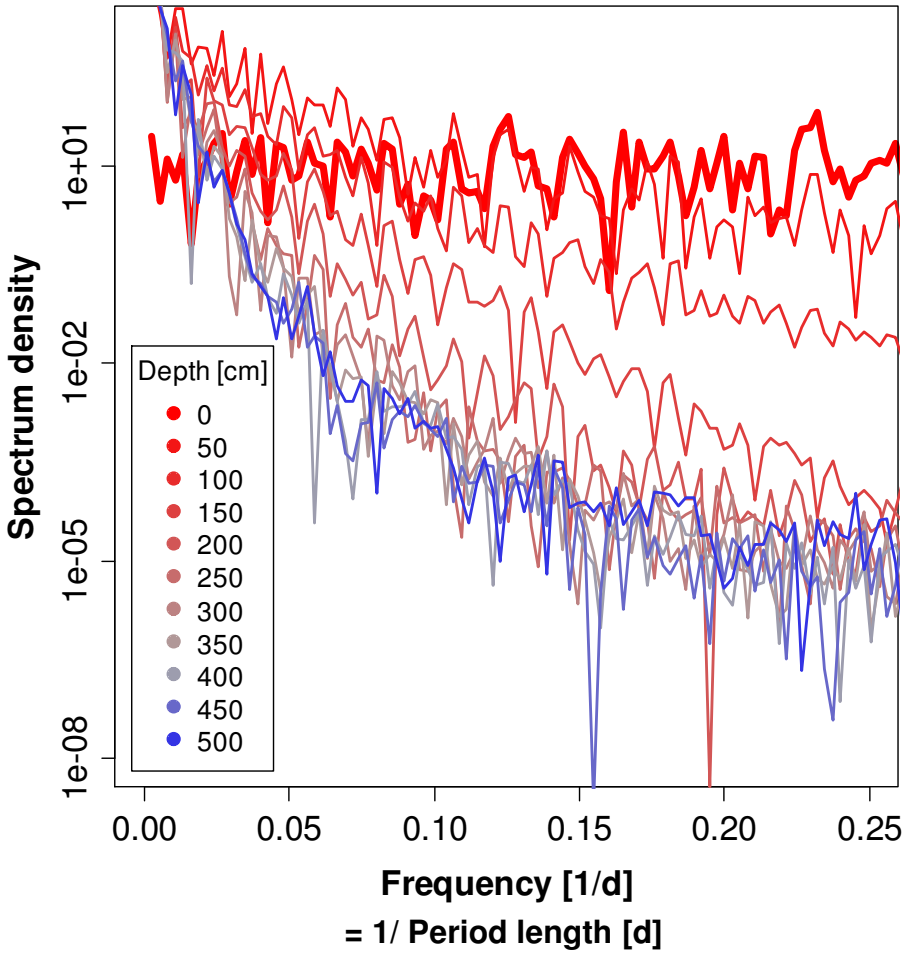
1. Typical time scales / periodicities
2. **Damping behaviour**
3. Visualizing the dynamics
4. Intrinsic dimensionality
5. Process identification

Basic Idea: Example

Time series of modelled soil water content at different depths

(Hydrus 1D; Tobias Hohenbrink)

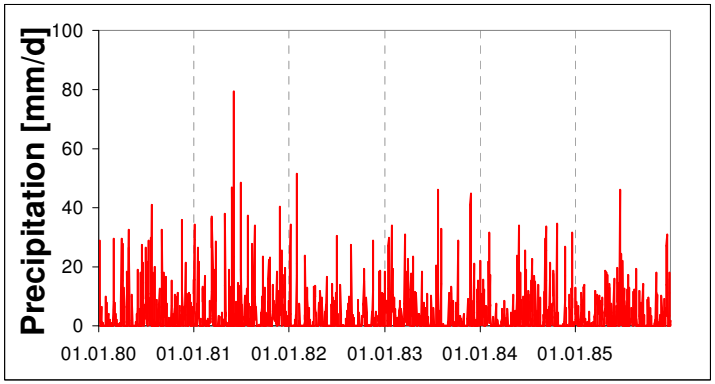


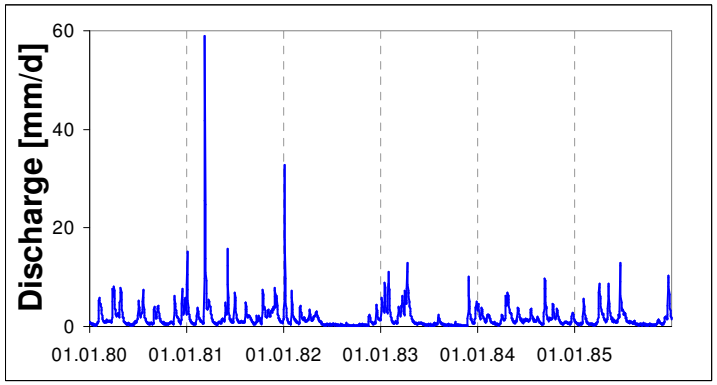


Catchments as Low-pass Filters

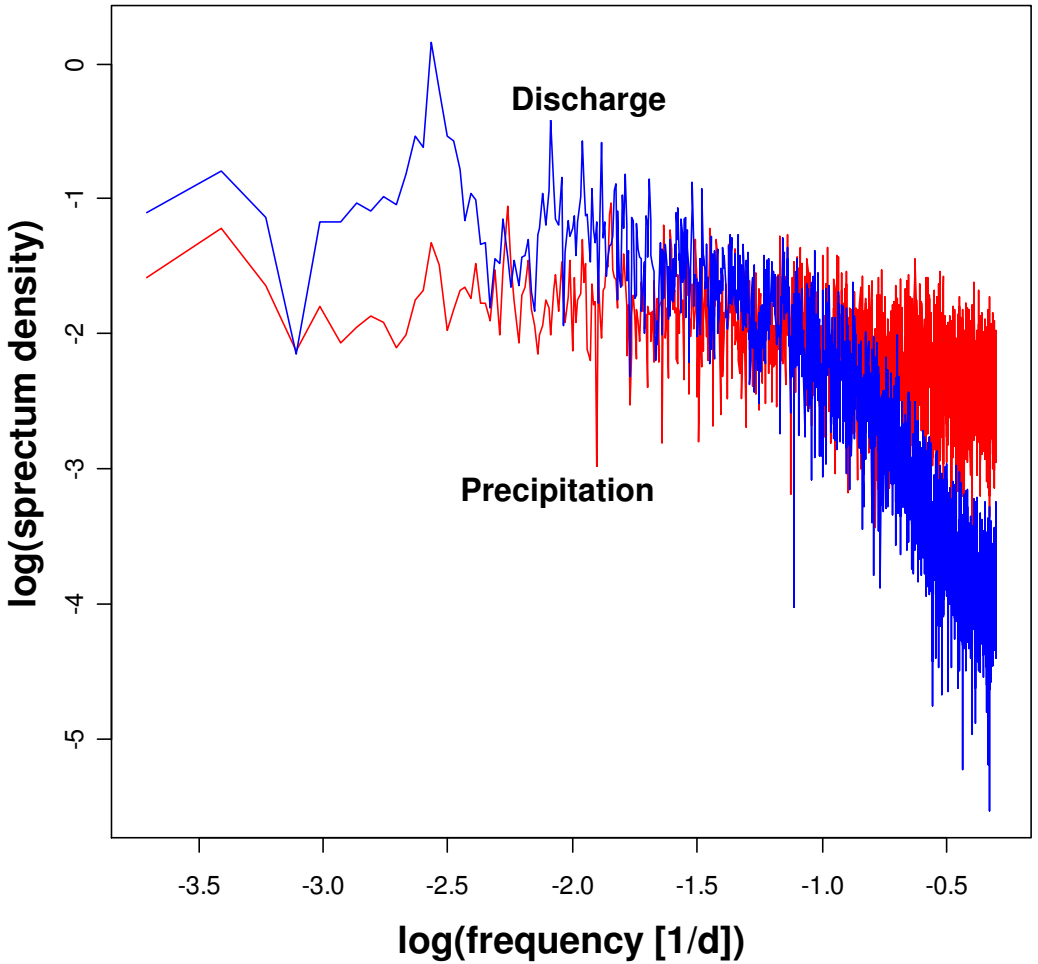
Lange Bramke Catchment (Harz Mountains)

Time series

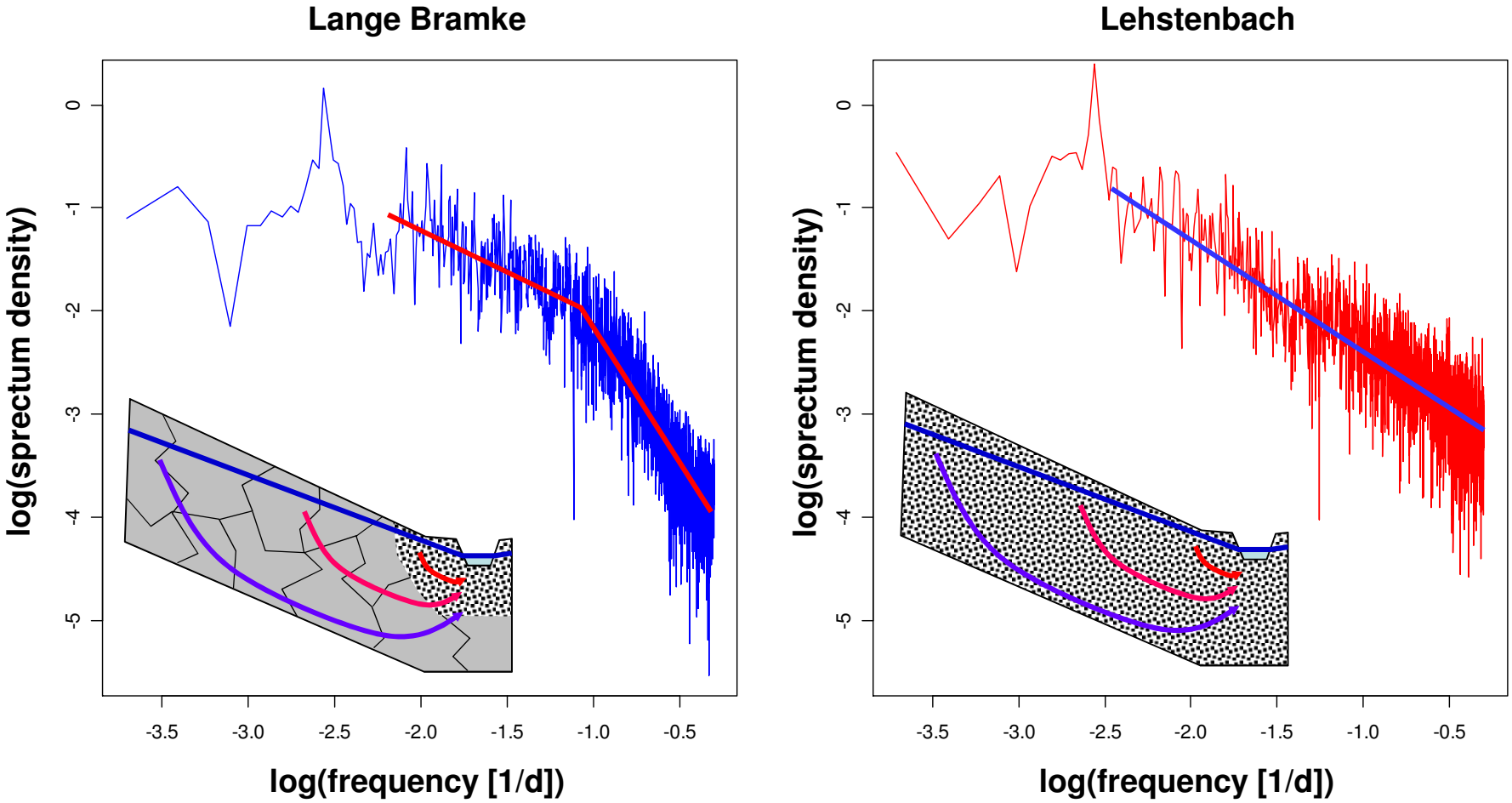




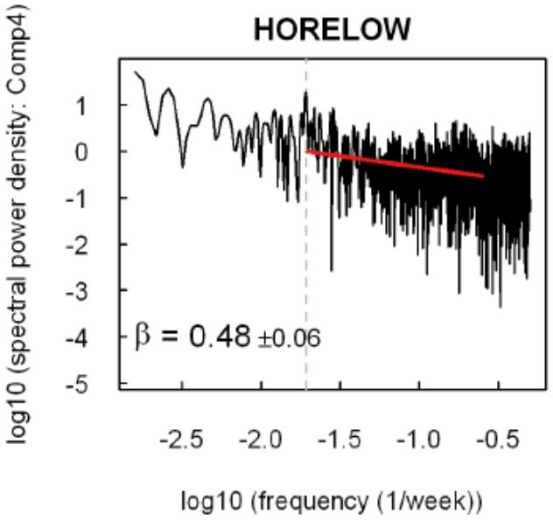
Power spectrum

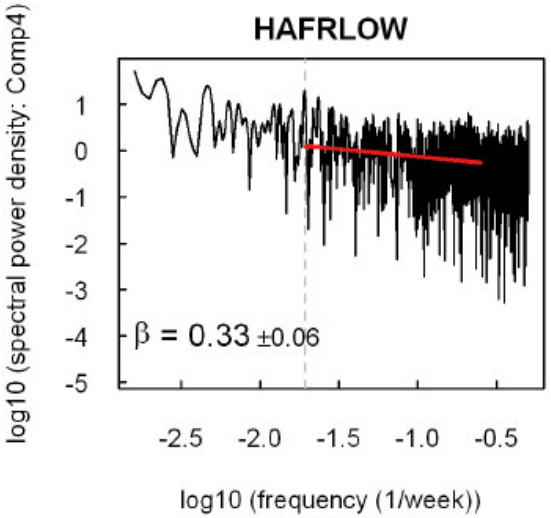


Comparing Different Catchments



Processing Chemical Signals

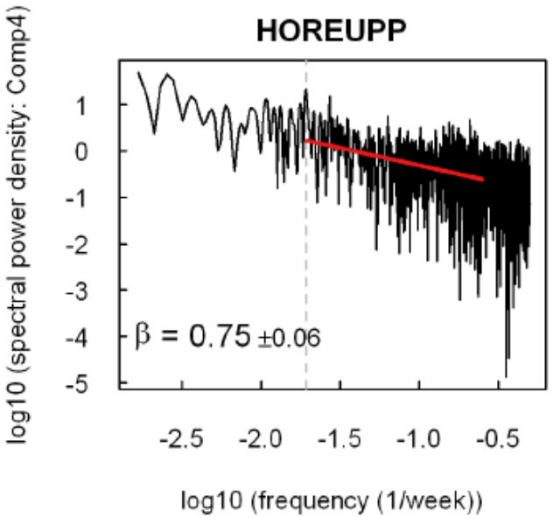


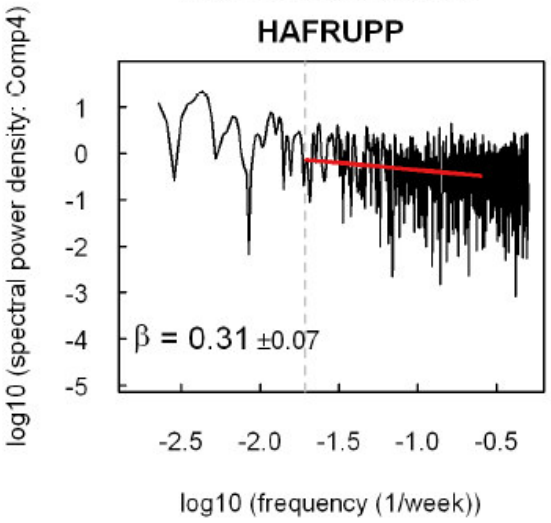


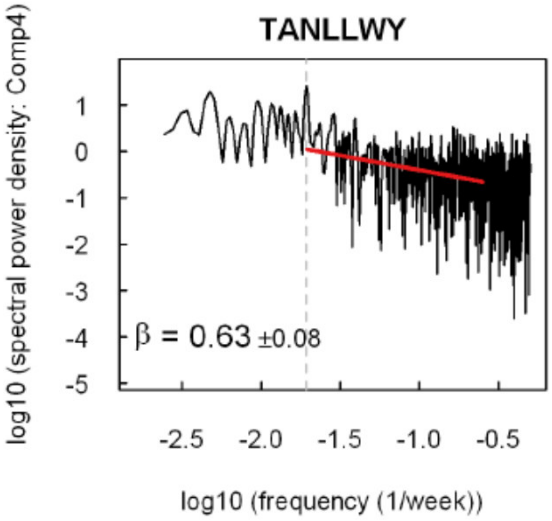
Plynlimon catchment runoff

4. Isomap dimension: Anthropogenic deposition

(*Köck 2008*)







Questions and Approaches

1. Typical time scales / periodicities
2. Damping behaviour
3. **Visualizing the dynamics**
4. Intrinsic dimensionality
5. Process identification

Visualization of Large Data Sets

=> Optimizing the interface between computer (*data*) and human brain (*researcher, decision maker*)

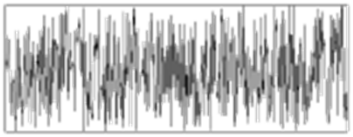
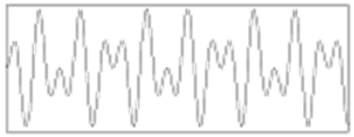
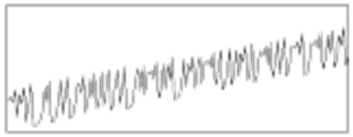
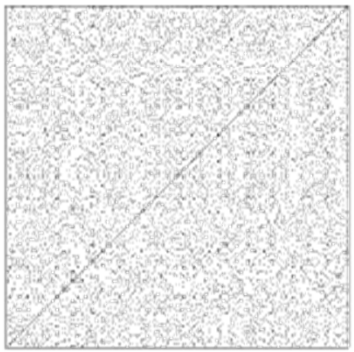
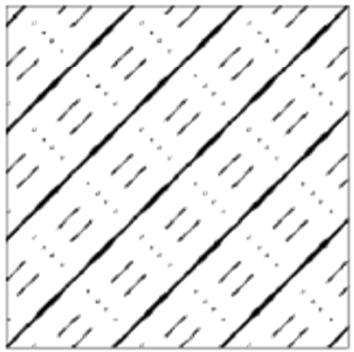
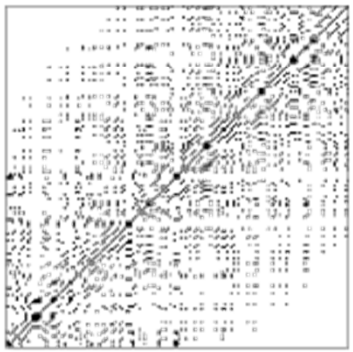
The most powerful interface: **Visualization**

- Human brain focusses on processing visual data (80% of the total information income = 75 MB s^{-1})
- Efficient extraction of information:
  $250 \cdot 10^6$ sensory cells → $2 \cdot 10^6$ ganglia → $2.5 \cdot 10^0$ objects realized

In contrast to modelling approaches:

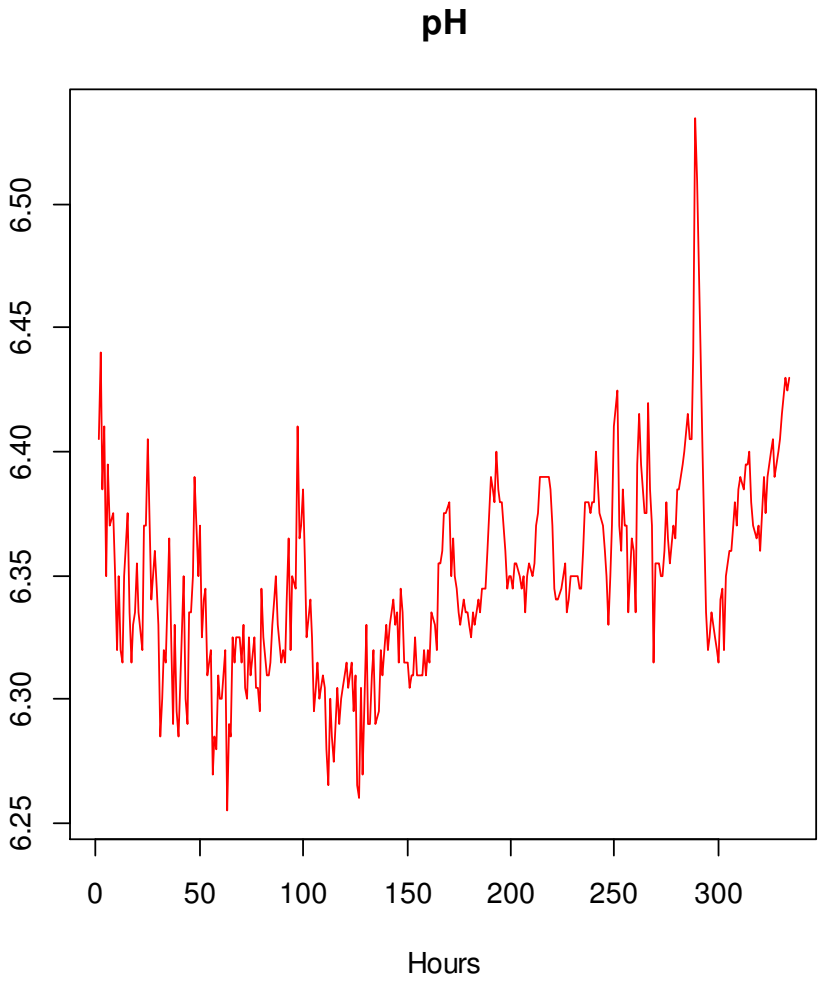
→ Filter information as late as possible in order to capture unexpected information!

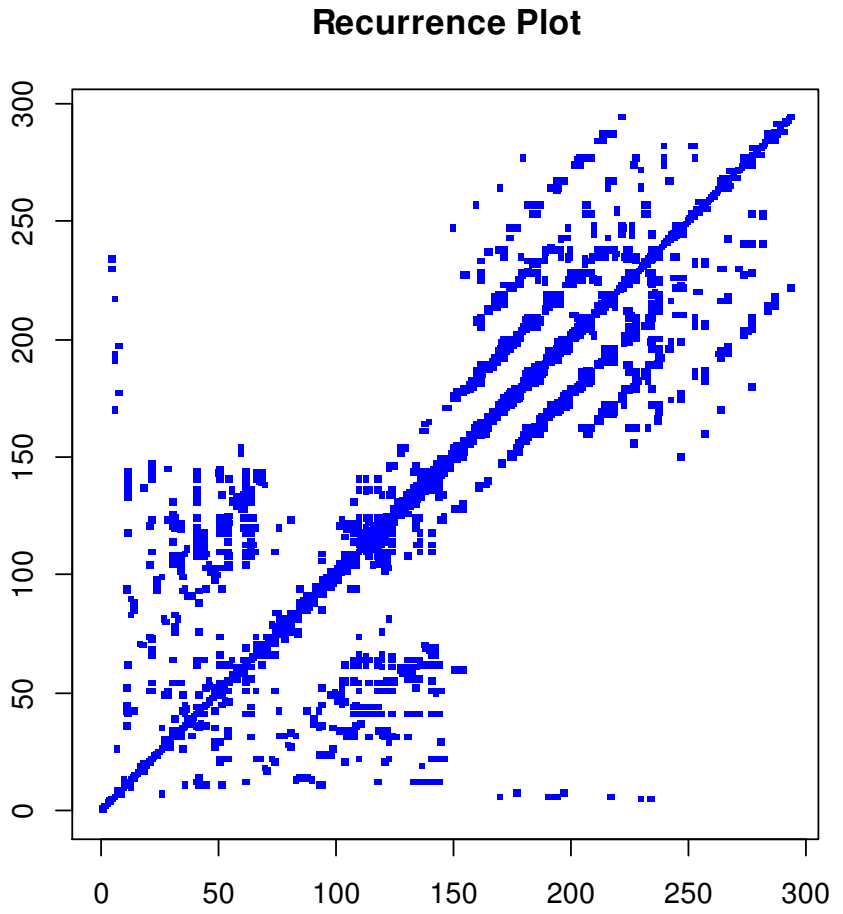
Recurrence Plots

White noise	Harmonic oscillation	Linear trend	Change of dynamics
			
			

Norbert Marwan (http://en.wikipedia.org/wiki/Recurrence_plot)

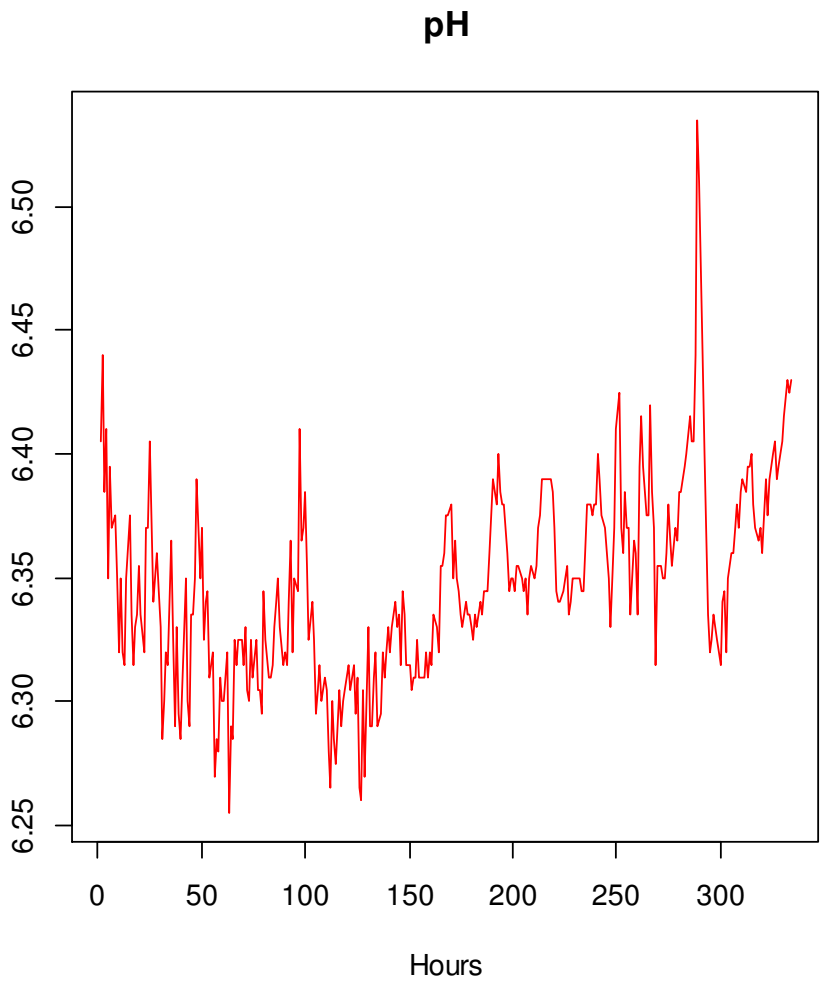
pH in Catchment Runoff

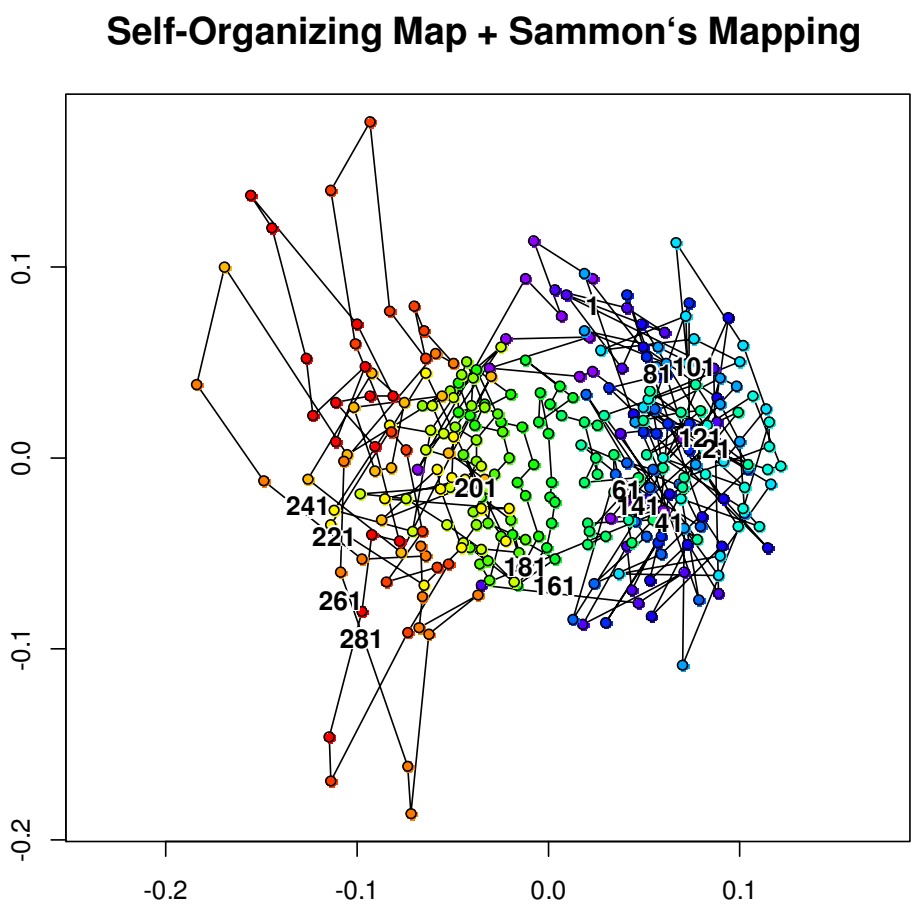




(Windows of 5 subsequent data points, lag width = 10)

pH in Catchment Runoff





(Windows of 5 subsequent data points, lag width = 10)

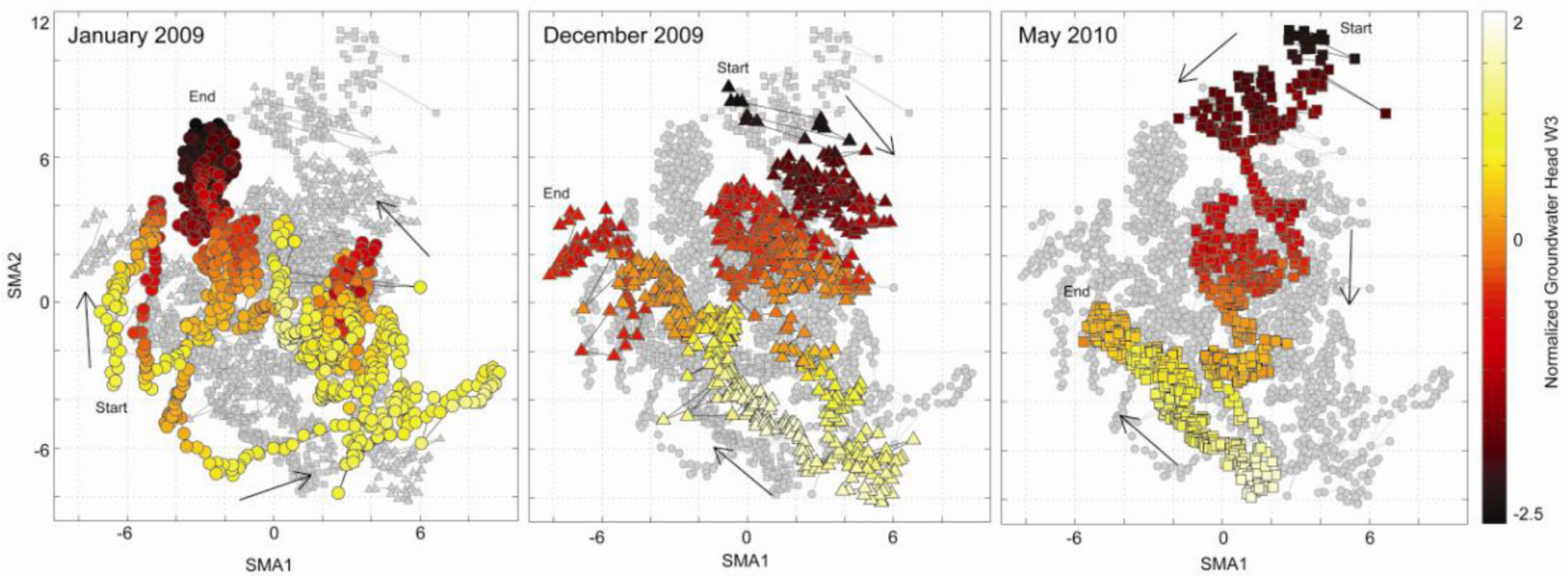
Multivariate Time series

Self-Organizing Map + Sammon's Mapping

Riverine groundwater extraction: Crucial events

(Risk of contamination of a well field by river water during high flow)

(*Page et al., submitted*)



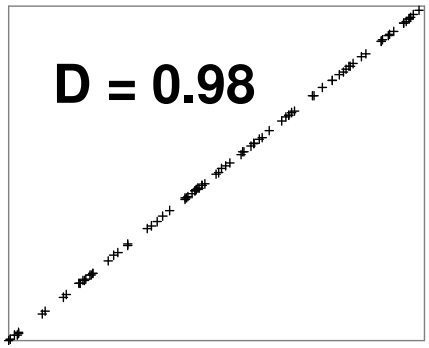
Questions and Approaches

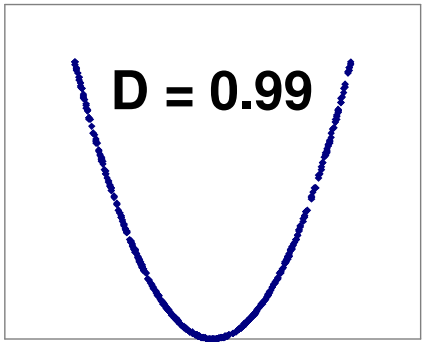
1. Typical time scales / periodicities
2. Damping behaviour
3. Visualizing the dynamics
4. **Intrinsic dimensionality**
5. Process identification

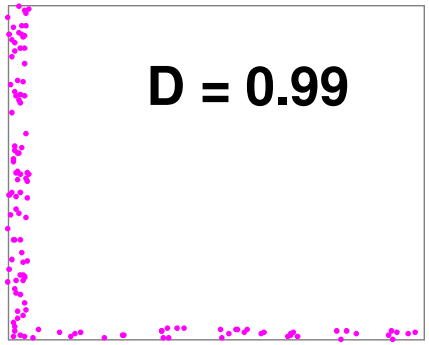
Correlation dimension: Examples

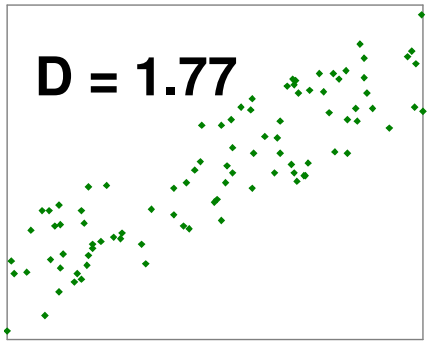
= a measure of to what degree “data fill the available space“

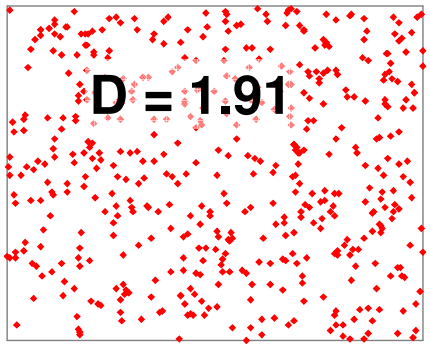
(cf., Principal Component Analysis)











Dimensionality of Runoff Data

(*Sivakumar 2004*)

Table 1
Chaos studies in geophysics: correlation dimension estimates of geophysical time series

Data	Dimension
1. Daily runoff in southwest Idaho, USA	No low dimension
2. Daily flow in Hong Kong (two stations)	0.455, 0.460
3. Daily flow in Po River, Dora Baltea, Italy	<4
4. Daily flow in Canadian Prairies (six rivers)	7–9
5. Flow from the basin Uhlirska in the Jizera Mts, Czech Republic (daily and 30-min)	No low dimension, 2.89
6. Daily flow in Scandinavian region	Low dimension
7. Daily flow in Western Run, MD, USA	No low dimension
8. Daily discharge of Chao Phraya River, Thailand (at Nakhon Sawan)	2.9
9. Daily discharge of Mekong River in Thailand (at Nong Khai and Pakse)	1.69, 1.58
10. Daily discharge of spring Almyros, Greece	3–4
11. Monthly flow in Göta River, Sweden	5.5
12. Daily flow in Adige River, Trento, Italy	2.8
13. Monthly flow in Coaracy/Nunes, Brazil	3.62
14. Daily flow in Little River and Reed Creek, VA, USA	1.19, 1.07
15. Daily flow in English River, Canada	2.4
16. Daily flow in Lindenborg, Denmark	3.76
17. Daily flow in Tryggevaelde, Denmark	1.4
18. Daily flow in Altamaha River, USA	0.85
19. Daily flow in Mississippi River, MO, USA	2.32
20. Annual flood series in Huaihe River Basin in China	4.66

Questions and Approaches

1. Typical time scales / periodicities
2. Damping behaviour
3. Visualizing the dynamics
4. Intrinsic dimensionality
5. **Process identification**

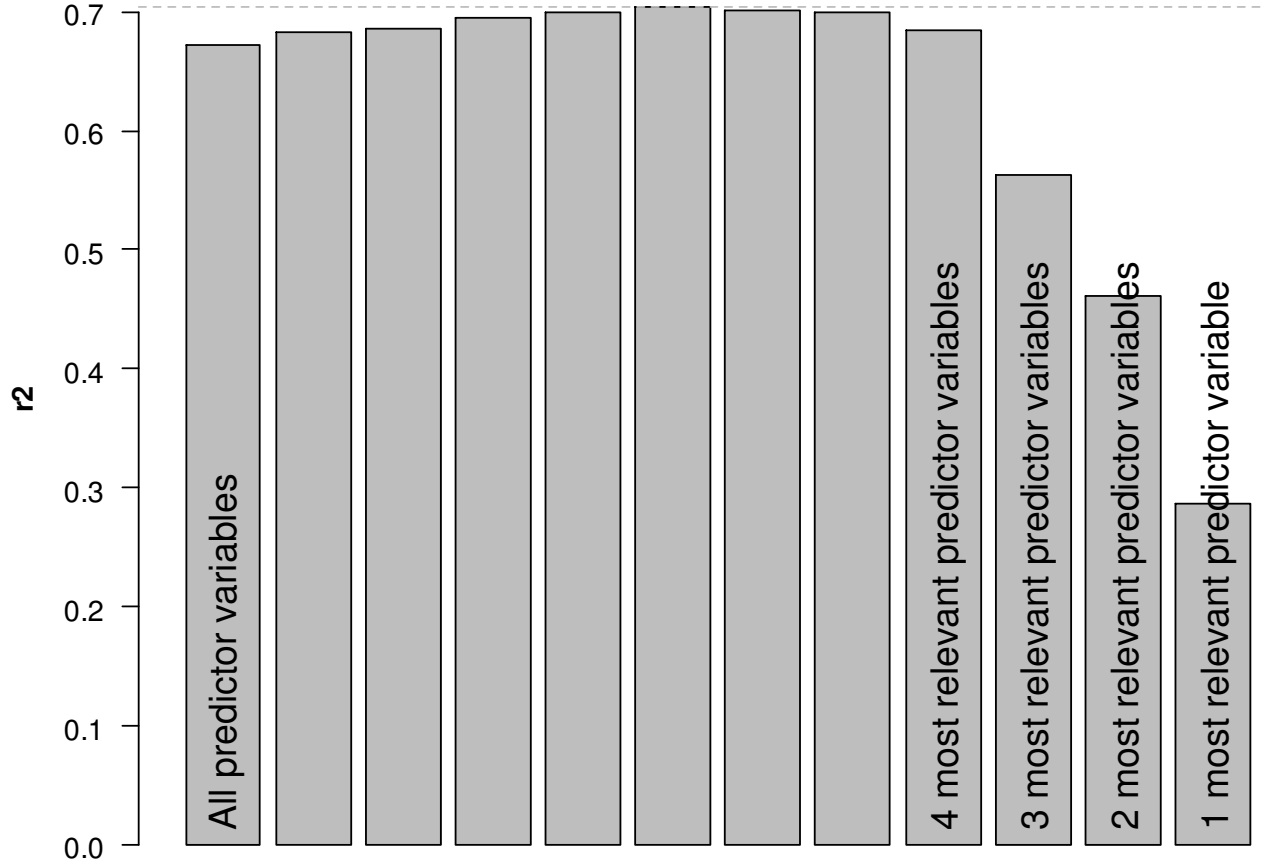
Cause-Effect Relationships

... to be identified via Machine Learning:

- = “Highly flexible, nonlinear, multivariate regression”
- Key variables out of a large number of candidate predictors can be identified (via “pruning”)
- Needs large data sets (>> 100 data points)
- Approaches:
  - Artificial Neural Networks
  - Genetic Algorithms
  - CART / Random Forests
  - Support Vector Machines
  - …

Pruning

→ Subsequently leaving out the respective least relevant predictor variable



Comparative Analysis of Time Series

Working hypotheses:

1. Observed time series of a single variable reflect effects different processes.
2. Every single process imprints a typical pattern (“signal”) on the observed dynamics (*e.g., surface runoff, WWTP efflux, in-stream nitrogen uptake,* …).
3. Processes simply superimpose their respective signals.

Approach:

1. Perform principal component analysis to determine „basic patterns“ that are characteristic for single processes.
2. Use loadings on single components as a quantitative measure for the strength of the respective process.
3. Study spatial patterns of loadings to identify the respective processes.

Conclusions

Automatic on-site measurements yield large and high-resolution data sets.

This opens new options for deepening our understanding about respective processes => a wealth of approaches are available (*but might require some time to get used to*).

Dare to ask strange questions - and to get strange answers!

Questions, remarks, criticism: lischeid@zalf.de