Hauptkomponentenanalyse

Geometrische Interpretation:

Rotation des hochdimensionalen, orthogonalen Koordinatensystems so, dass einige Achsen möglichst durch die Datenwolken verlaufen

(*vorher Zentrierung des Koordinatensystems auf den Schwerpunkt der Daten*)

=> Multiplikation der Datenmatrix mit einer Rotationsmatrix

$$\mathbf{A} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

- die neuen Achsen: ***Eigenvektoren*** (orthogonal)
- Verteilung der Varianz auf die Achsen: ***Eigenwerte***

Prinzip der SSA

Ziel: Identifizierung unregelmäßiger Periodizitäten in einer Zeitreihe

Ansatz: Durchführung einer Hauptkomponentenanalyse anhand der Toeplitz-Matrix (= Zeitreihe *k*-fach gegen sich selbst verschoben)

Prinzip: Periodizitäten daran erkennbar, dass die entsprechend gegeneinander verschobenen Zeitreihen korreliert sind

$$\begin{bmatrix} x_1 & x_{1+1} & \cdots & x_{1+k} \\ x_2 & x_{2+1} & \cdots & x_{2+k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(n-k)} & x_{(n-k)+1} & \cdots & x_{(n-k)+k} \end{bmatrix}$$

Mathematisch

Aus der Zeitreihe $x = x(t)$ werden $(m+1)$ - Einbettungsvektoren

$$\vec{x}(t) = (x_t - \bar{x}, x_{t+\Delta t} - \bar{x}, ..., x_{t+m\cdot\Delta t} - \bar{x})$$

gebildet ($n-m+1$ verschiedene). Diese bilden die Zeilen der $(n-m+1) \times (m+1)$–Matrix.

$C_{i,j} = x_{i+j} - \bar{x}$ und $T = \langle C^T C \rangle$ ist die **Lag-Kovarianz-Matrix:**

$$T_{j,j'} = \frac{1}{n - |j - j'|} \sum_{i=1}^{n-|j-j'|} \tilde{x}_i \tilde{x}_{i+j-j'}$$ (Toeplitz-Matrix der Zeitreihe)

Mathematisch

- Die Toeplitz-Matrix wird diagonalisiert:

  $T = E^T \Lambda E \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_m)$

  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_m > 0$

- Exakte Darstellung mit den Eigenfunktionen E (empirische orthogonale Funktionen = EOF)

  $x_{i+j} = \sum_{k=1}^{m} A_i^k E_j^k$, wobei $A = CE$

SSA und Spektralanalyse

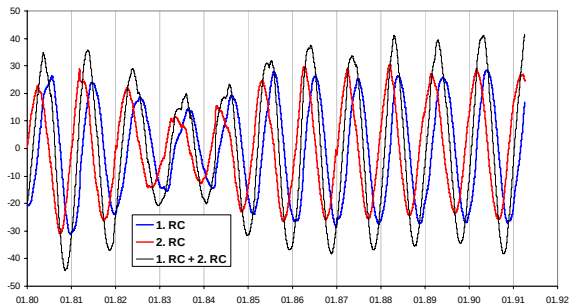
- Das Spektrum einzelner EOFs ist in der Regel einfach (ein oder zwei Peaks)
- Bzgl. Frequenzen kommen alle signifikanten EOFs immer paarweise vor (vergl. Fourieranalyse)
- Signifikante EOFs bestehen den Test auf **rotes Rauschen**:

$$\phi_{RN}(f) \propto (a + bf)^{-2}$$

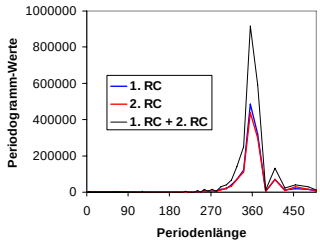
Beispiel

1. RC + 2. RC: 10.5% der Gesamt-Varianz

Zeitreihen



Periodogramm



Parameter der SSA

Time lag τ: erster Nullwert der Autokorrelationsfunktion

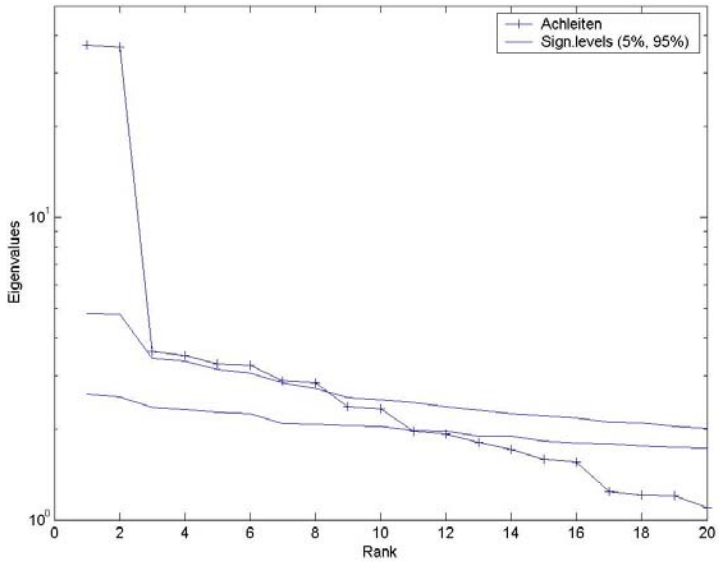
Einbettungsdimension m: Kompromiss zwischen möglichst vielen Datenfenstern und möglichst vielen Basisfunktionen (Periodizitäten). Empfehlung: $0.2 \cdot n \leq m \leq 0.5 \cdot n$

Verarbeitung der Ergebnisse

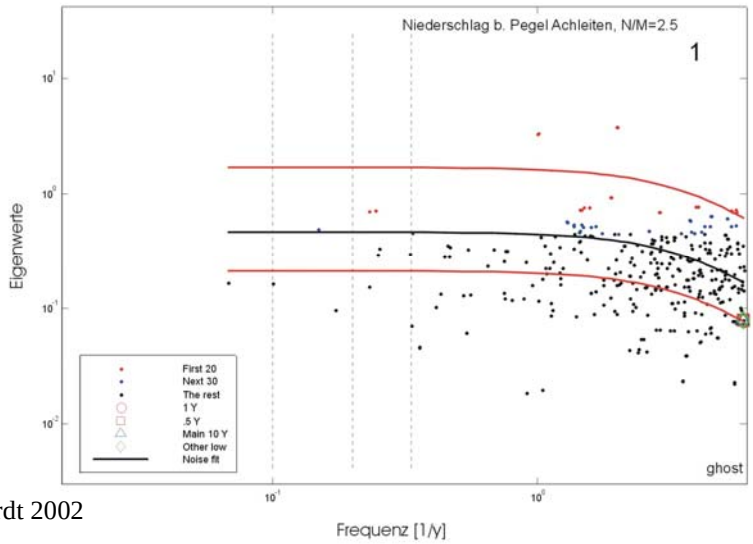
- Quantifizierung des Anteils der Varianz, der auf die einzelnen Komponenten entfällt (proportional zu den Eigenwerten)
- Bestimmung des Spektrums bzw. der vorherrschenden Frequenz der signifikanten Komponenten (in der Regel jeweils zwei Komponenten der gleichen Frequenz)
- Untersuchung des zeitlichen Verlaufs einzelner rekonstruierter Komponenten

$$R_K(t) = \frac{1}{m_t} \sum_{k \in K} \sum_{j=1}^{m} A_k^{t-j} E_k^j$$

Anzahl der signifikanten Komponenten

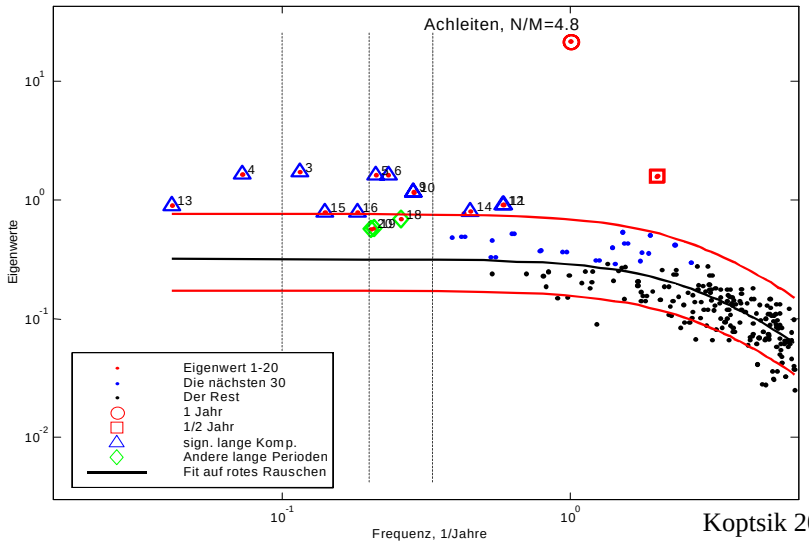


SSA Spektrum und Test gegen rotes Rauschen I: Niederschlag



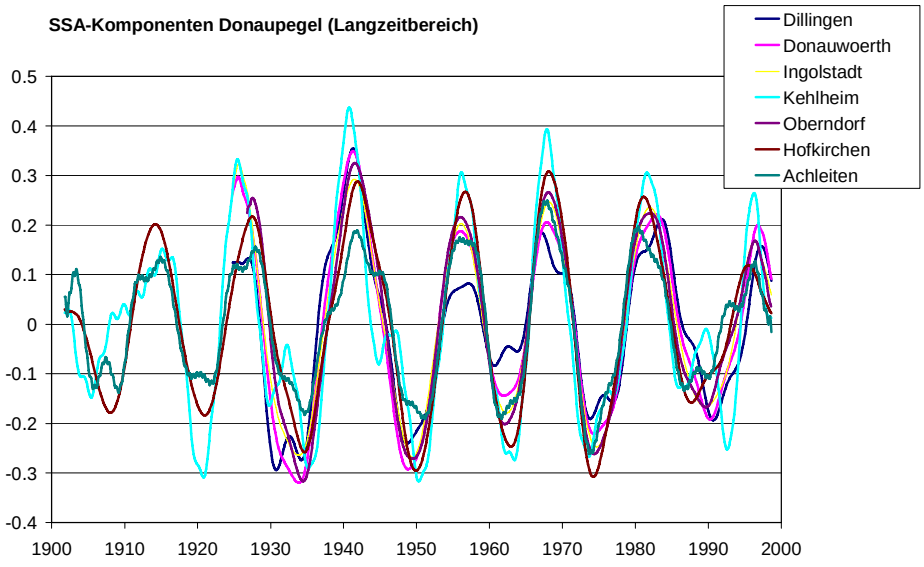
Bernhardt 2002

SSA Spektrum und Test gegen rotes Rauschen II: Abfluss

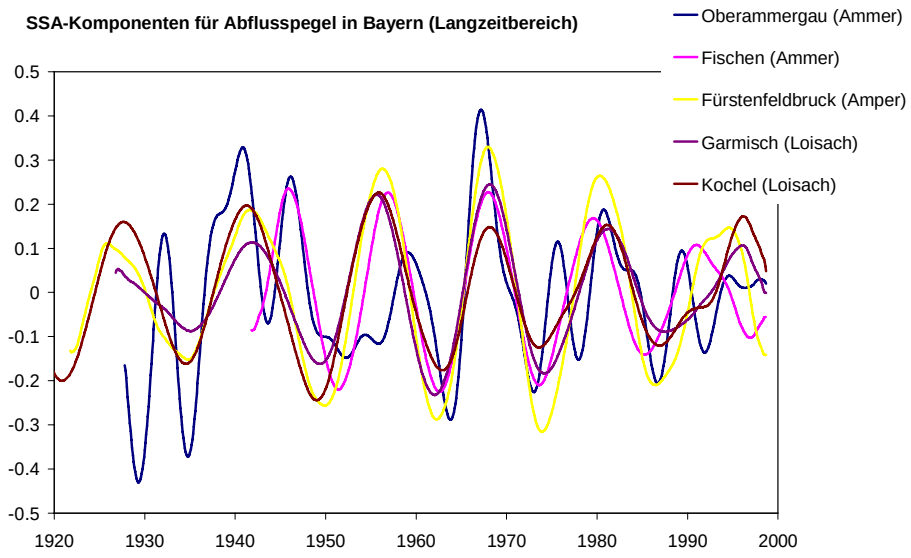


Koptsik 2001

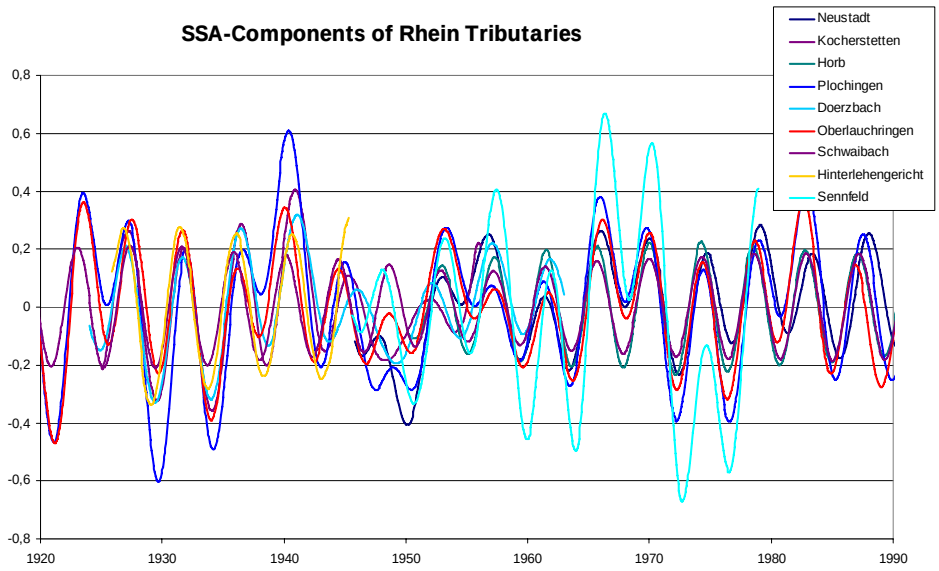
Rekonstruierte Komponenten



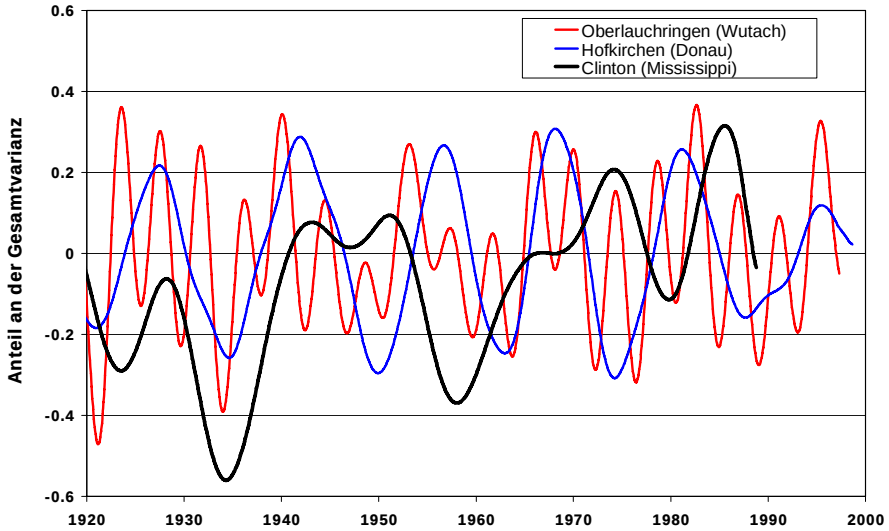
Rekonstruierte Komponenten



Rekonstruierte Komponenten



Rekonstruierte Komponenten



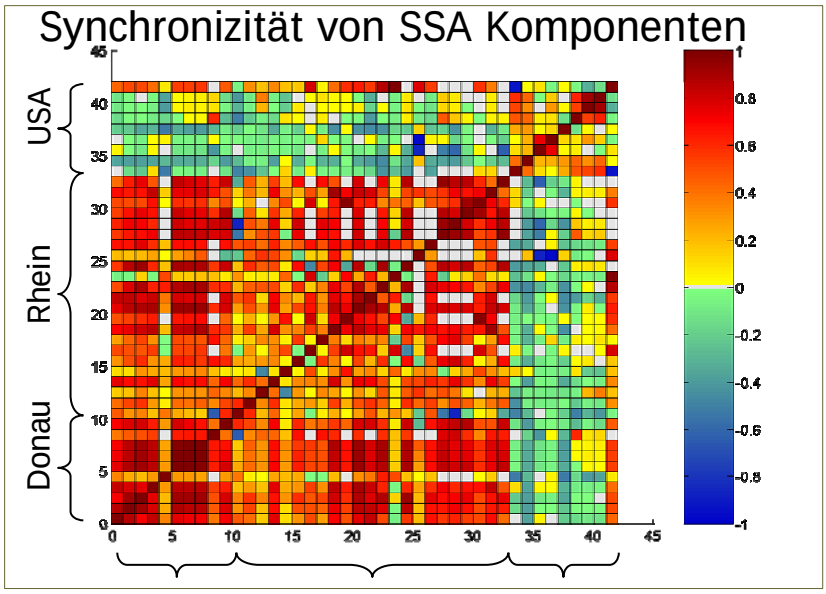
Intuitives Synchronizitätsmaß

Analog zu Kohärenzintegralen in der Signaltheorie:

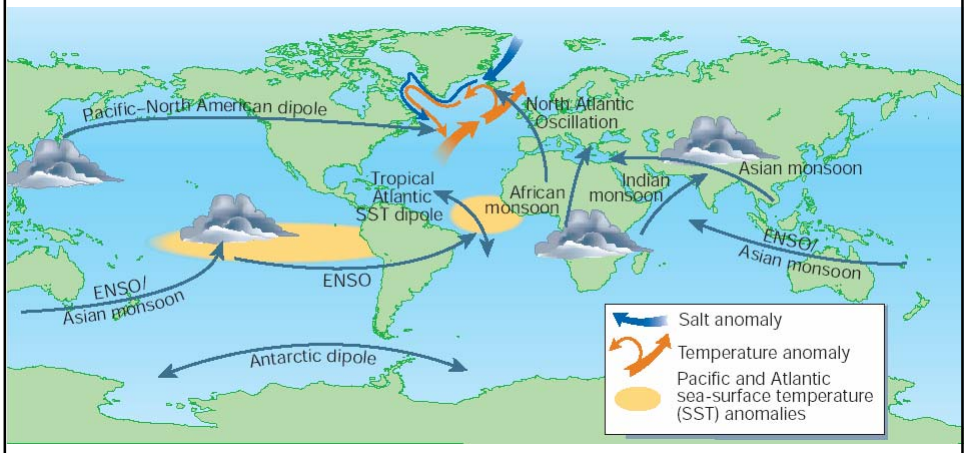
$$S_{coh} = \frac{\int_a^b RC_1(t) RC_2(t) dt}{\sqrt{\int_a^b (RC_1(t))^2 (RC_2(t))^2 dt}}$$

$$-1 \leq S_{coh} \leq 1$$

$S_{coh} = 1$: perfekte Synchronizität



Überall globale Verbindungen...



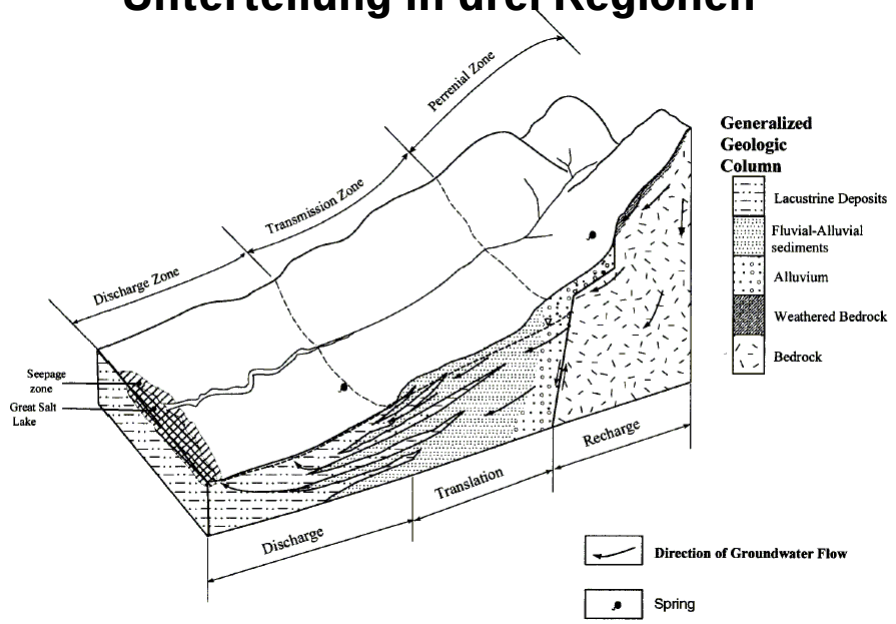
R. Zahn, Nature 421, 324-325 (Jan. 23rd, 2003)

Vielkanal-SSA: Beispiel

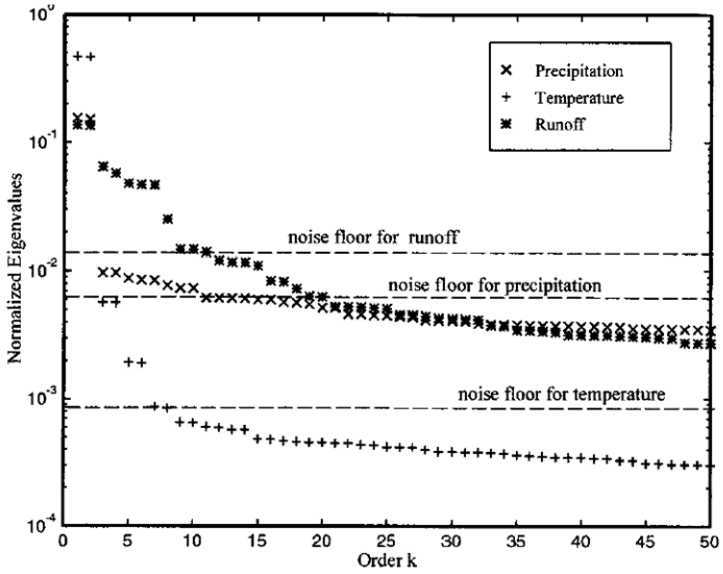
Shun und Duffy, Water Resources Research 35, 191-201 (1999):

- Wasatch-Gebirge, Utah/USA
- Messungen von Temperatur, Niederschlag und Abfluss in 9 verschiedenen Höhenstufen (1287 – 2760 m ü.NN)
- Datensatzlänge zwischen 522 und 1008 Monaten

Unterteilung in drei Regionen



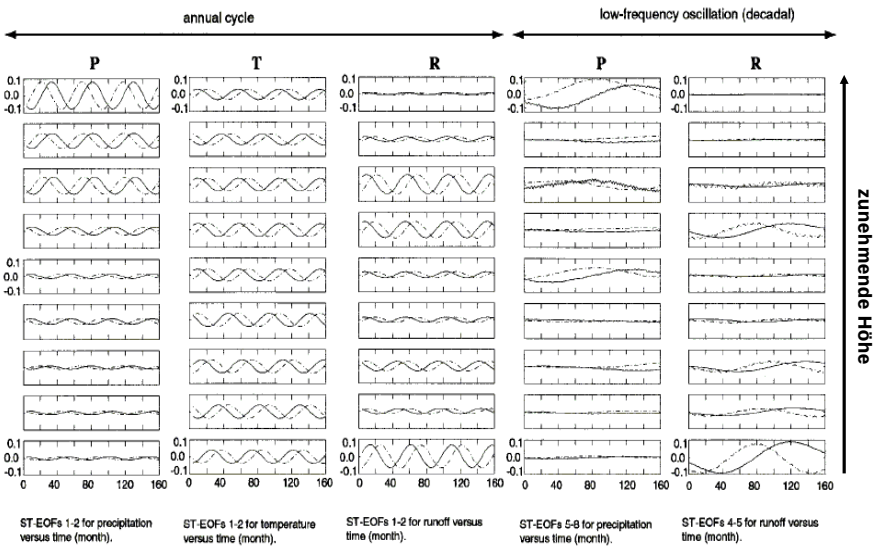
Eigenwerte



Wichtigste Periodizitäten

Experiment	Oscillatory Pair	Period, years	Variance, %
Precipitation	1~2	1	30.8
	3~4	0.33	1.9
	6~7	0.25	1.7
	5~8	11	1.7
	9~10	0.5	1.5
	17~18	3	1.1
Runoff	1~2	1	27.3
	4~5	11	10.5
	6~7	0.5	9.3
	9~10	0.33	2.9
	11~12	7	2.6
	17~18	3	1.1
Temperature	1~2	1	93.4
	3~4	0.5	1.1
	11~12	3	0.1

Rekonstruierte Komponenten



Bedeutung einzelner Komponenten je nach topografischer Höhe

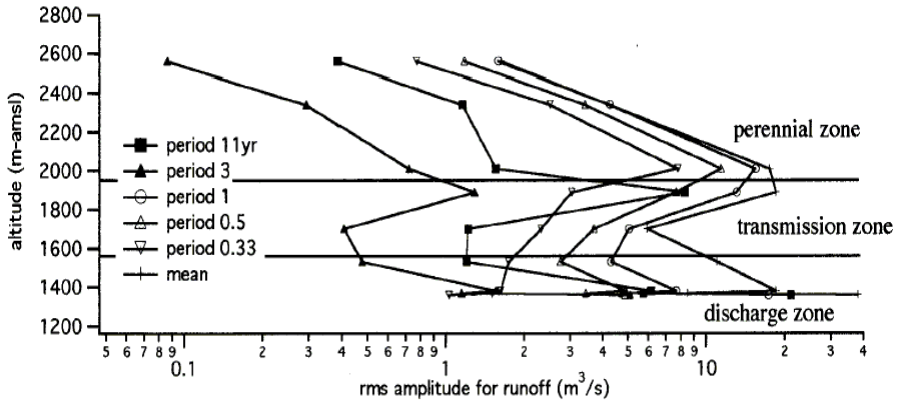


Figure 5. Root-mean-square (rms) amplitude and mean (time average) versus elevation for temperature, precipitation, and runoff.

Aufgabe

1. Erstellen Sie für die Tageswerte von Niederschlag, Temperatur oder Abfluss die Töplitz-Matrix für die maximal mögliche Einbettungsdimension und einem sinnvollen Zeitversatz.
2. Führen Sie damit in Statistica eine Hauptkomponentenanalyse durch.
3. Identifizieren Sie von den Komponenten mit Eigenwert >1 jeweils die beiden zusammengehörigen Komponenten, addieren Sie die Werte, und bestimmen Sie die Periodenlänge.
4. Plotten Sie die Zeitreihen der wichtigsten Komponenten und vergleichen Sie sie mit der ursprünglichen Zeitreihe.