

Extraction of Long-term Structures

from

Southern German Runoff Data

by means of

Linear and Nonlinear Dimensionality Reduction

Miguel Mahecha ⁽¹⁾ – Holger Lange ⁽²⁾ – Gunnar Lischeid ⁽¹⁾

(1)



UNIVERSITÄT
BAYREUTH

(2)

SKOG ▲ FORSK
Norsk institutt for skogforskning

(1) introduction

Sources:



Bayerisches Landesamt
für Wasserwirtschaft



Data:

Resolution: *daily*

Window-length: $N = 51$ years

Channels / Sites: $L = 118$

$$x_{i,l} : i = 1, \dots, N; l = 1, \dots, L$$

$$\mathbf{X}^{N \times L} = (X_1, \dots, X_L)$$

Compromise: many channels ... but rather short window

(1) introduction

Hydrological time-series are apparently...

- Correlated in space
- Partly synchronized on a regional level
- Influenced by long-term climate oscillations

Objectives:

- To investigate spatial patterns of long-term components
- To compare different (nonlinear) multi-channel methods

(2) methods

Overview:

Principal Component Analysis

PCA

Nonlinear Principal Component
Analysis

***nl*PCA**

Multidimensional Scaling

MDS

Isometric Feature Mapping
(„*Nonlinear MDS*“)

ISOMAP

(Multichannel)-Singular System
Analysis

(M)-SSA

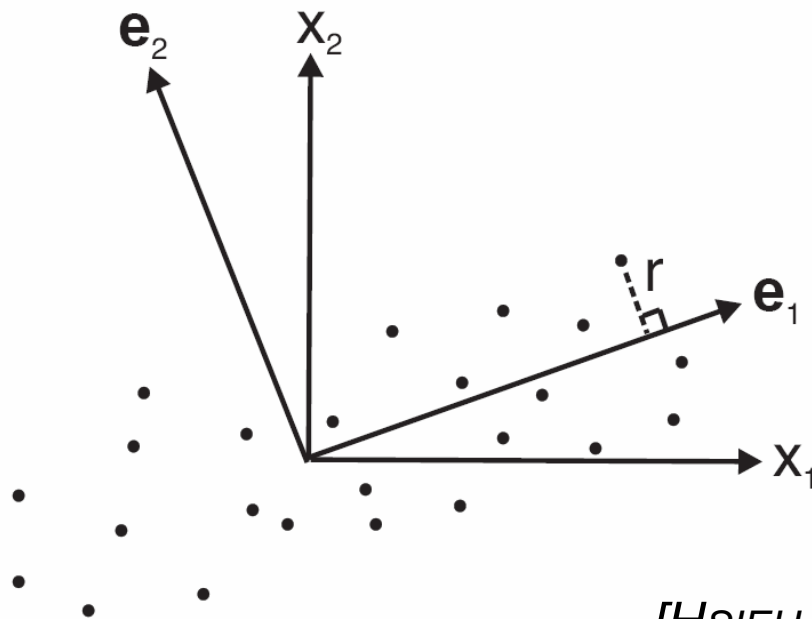
Nonlinear-(Multichannel)-Singular
System Analysis

***nl*(M)-SSA**

(2) methods

Principal Component Analysis (PCA) [PEARSON 1901]:

Find best **linear** approximation for the data-cloud through the centroid of the data-set. Each PC explains the maximum of variance linearly possible.



[HSIEH 2004]

(2) methods

Principal Component Analysis (PCA) [PEARSON 1901]:

- Build Variance-Covariance-Matrix and calculate the eigenvalues and eigenvectors

$$\mathbf{C}_X = \mathbf{X}^T \mathbf{X} = \mathbf{E} \mathbf{\Lambda} \mathbf{E}^T$$

- Allows the estimation of synthetic and representative principal components
the best linear approximation of the data-cloud though the centroid
- Leads to artificial time-series, where the dimensionality of spatial resolution is reduced

$$\mathbf{A} = \mathbf{E} \mathbf{X}$$

$$x_{i,l} : i = 1, \dots, N; l = 1, \dots, L$$

$$a_{i,l} : i = 1, \dots, N; l = 1, \dots, \leq L$$

$$\mathbf{X}^{N \times L} = (X_1, \dots, X_L)$$

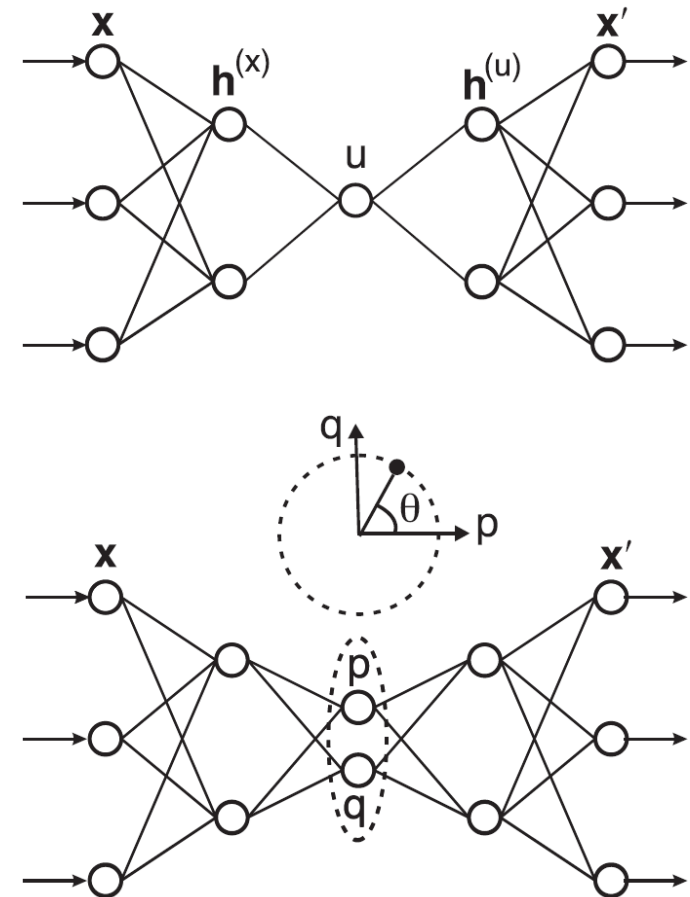
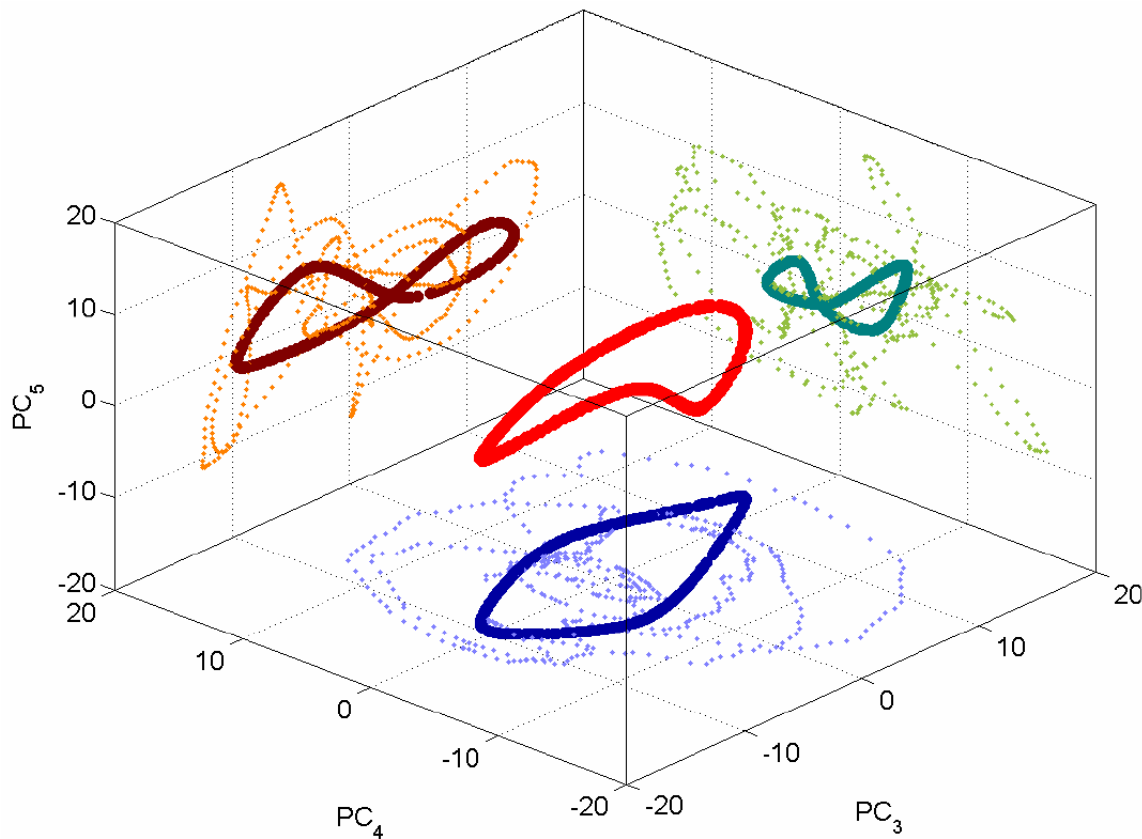
$$\mathbf{A}^{N \times L} = (A_1, \dots, A_L)$$

(2) methods

Non-linear Principal Component Analysis (*n*lPCA) [KRAMER 1991, HSIEH 2004]

PCA: best **linear** approximation for the data-cloud through the centroid $A = EX$

NLPCA: Generalization to open curve solutions $A = f(X)$



(2) methods

Multidimensional Scaling (MDS) [GOWER 1966]:

Find best low-dimensional representation, **trying to preserve the linear inter-point distances**

Mathematics is analogous to PCA, extracting the Eigenvectors and Eigenvalues

$$\mathbf{Z}^{(2)} = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$$

where $\mathbf{D}^{(2)}$ is a matrix of squared inter-point distances, and the entries of \mathbf{J} are $j_{i,j} = \delta_{i,j} - 1/N$

The extracted Principal Coordinates are spatially condensed time series

$$P = EX$$

$$x_{i,l} : i = 1, \dots, N; l = 1, \dots, L$$

$$p_{i,l} : i = 1, \dots, N; l = 1, \dots, \leq L$$

$$\mathbf{X}^{N \times L} = (X_1, \dots, X_L)$$

$$\mathbf{P}^{N \times L} = (P_1, \dots, P_L)$$

(2) methods

Isometric Feature Mapping (ISOMAP) [TENENBAUM ET AL 2000]

MDS: tries to preserve best the linear distances

ISOMAP: tries to preserve the geodesic (non-linear) distances

Technique:

- 1) Calculate the Euclidean distance-matrix \mathbf{D}
- 2) Define a k -nearest neighbor or ε -radius threshold in the Euclidean space
- 3) Connect the chosen neighbors to obtain a connectivity graph
- 4) Calculate the shortest inter-point distances on the Graph: \mathbf{D}_G
- 5) Apply MDS $\mathbf{Z}^{(2)} = -\frac{1}{2}\mathbf{J}\mathbf{D}_G^{(2)}\mathbf{J} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$ to obtain ISOMAP-coordinate $P = EX$

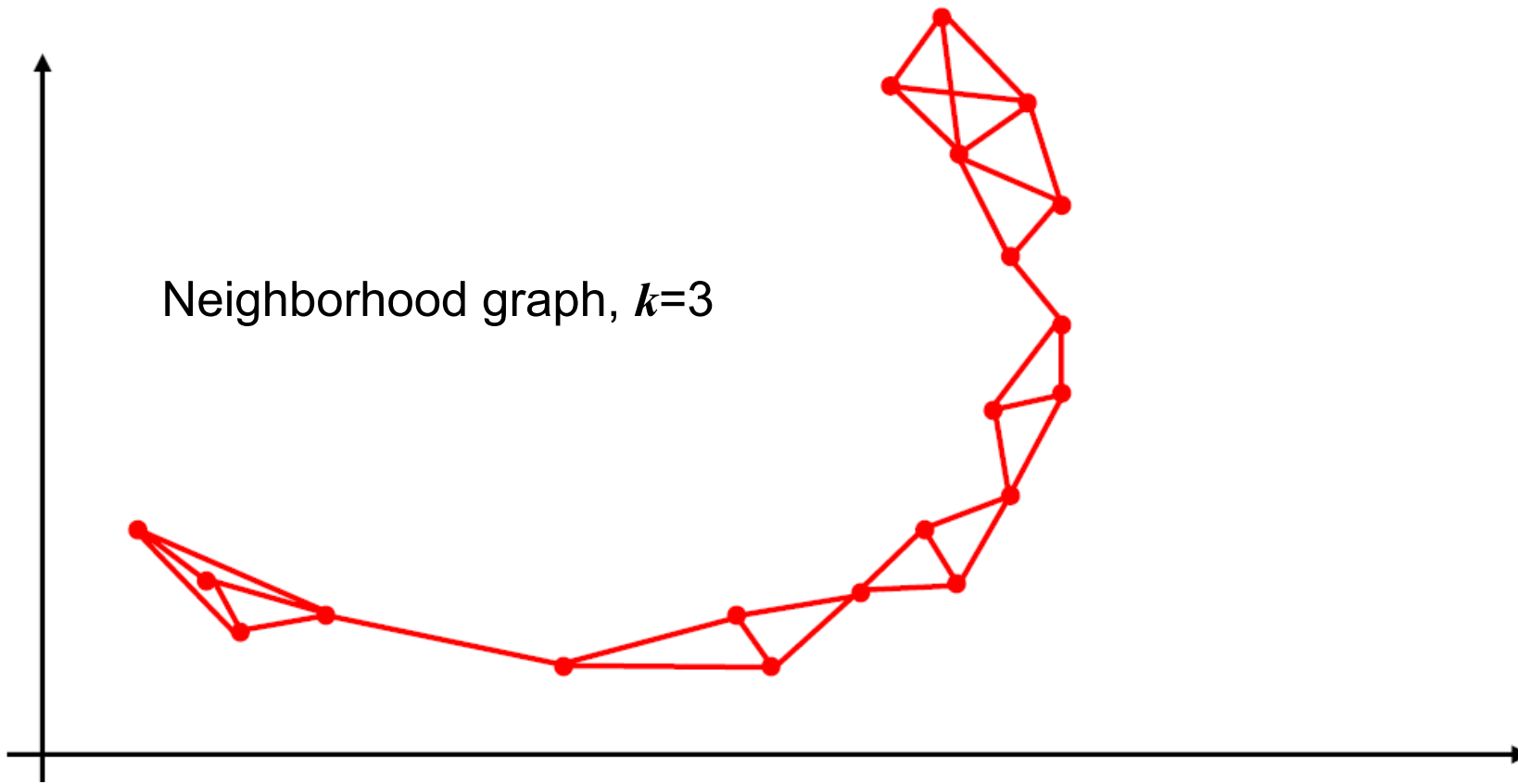
(2) methods

Isometric Feature Mapping (ISOMAP) [*TENENBAUM ET AL 2000*]

MDS: tries to preserve best the linear distances

ISOMAP: tries to preserve the geodesic (non-linear) distances

Technique:



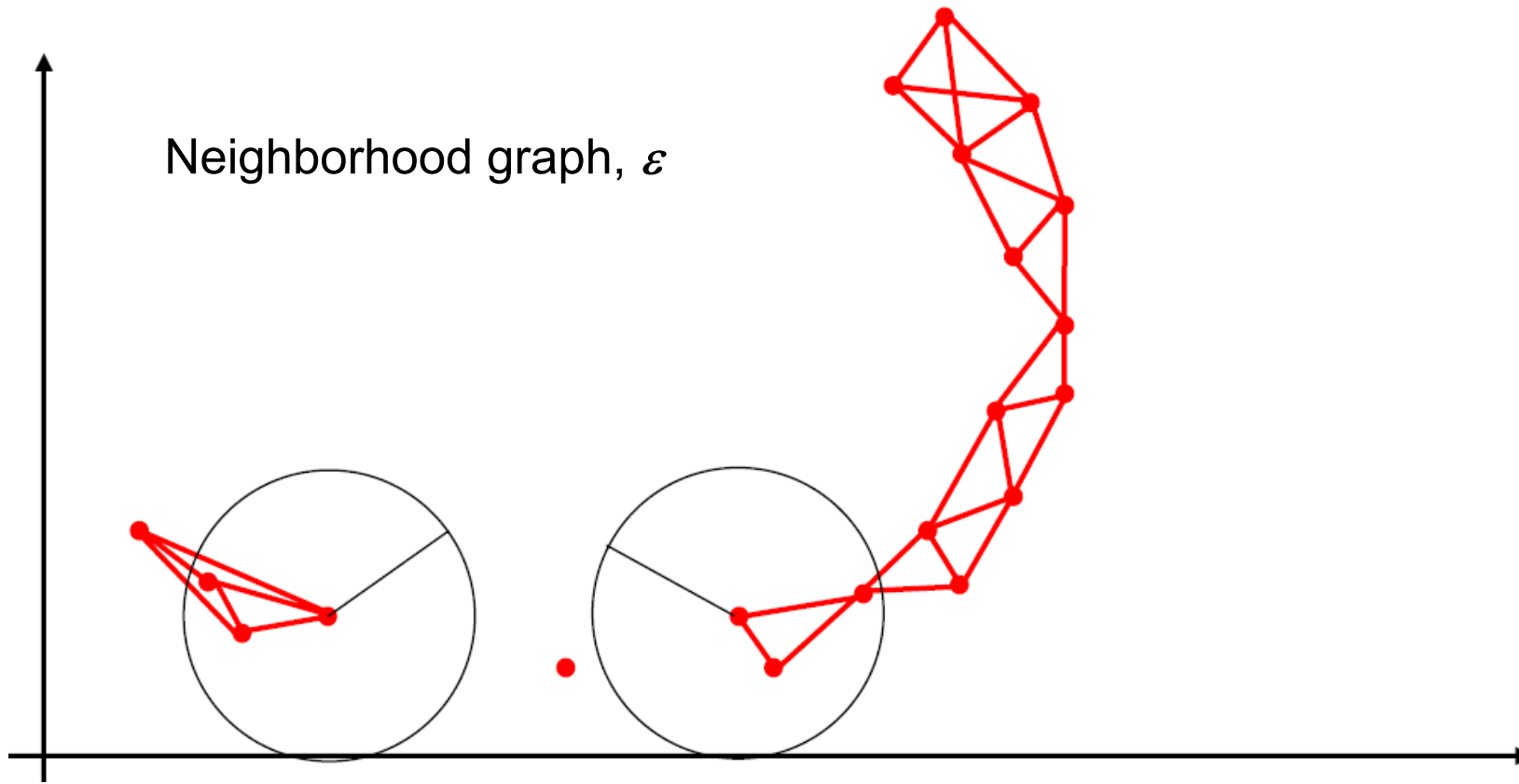
(2) methods

Isometric Feature Mapping (ISOMAP) [TENENBAUM ET AL 2000]

MDS: tries to preserve best the linear distances

ISOMAP: tries to preserve the geodesic (non-linear) distances

Technique:



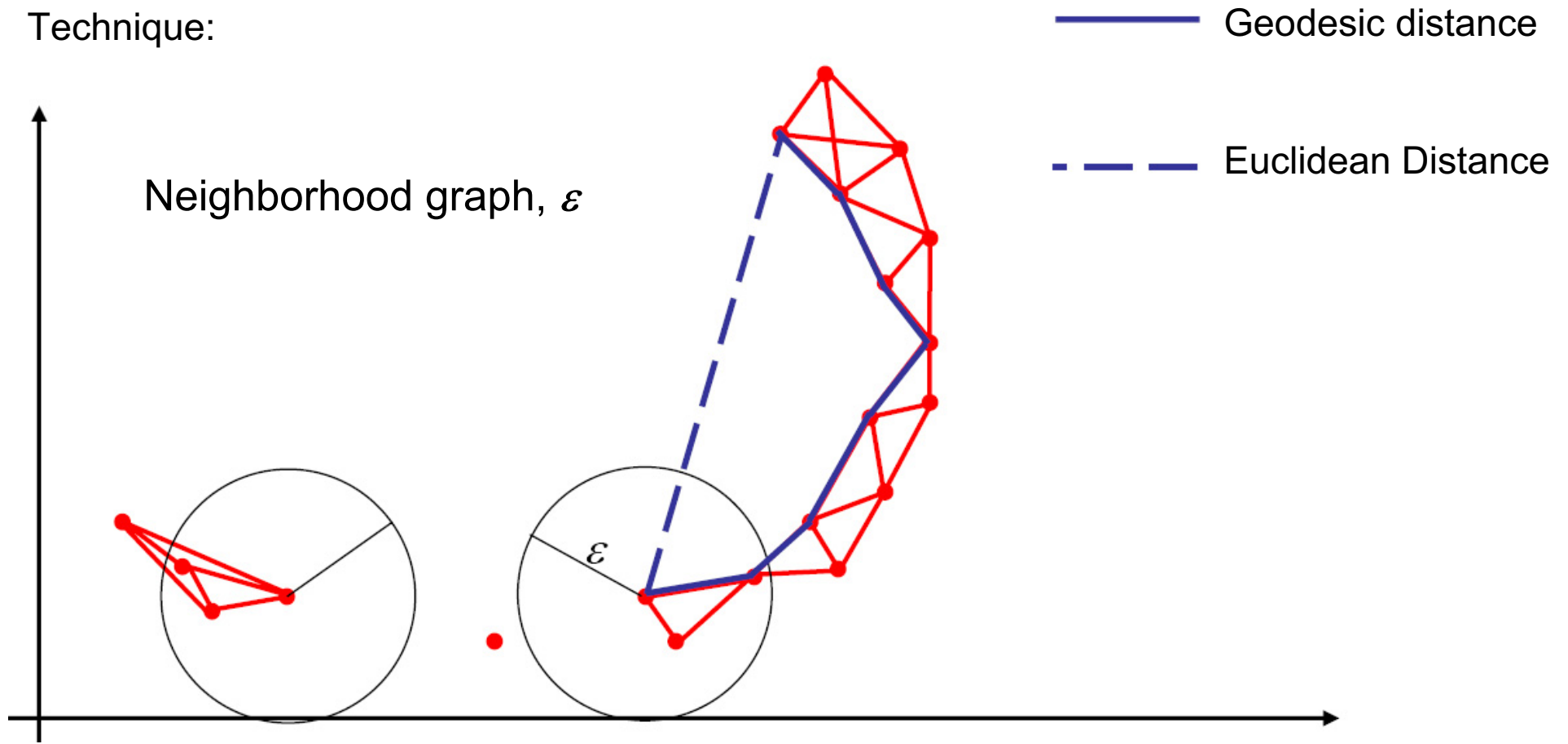
(2) methods

Isometric Feature Mapping (ISOMAP) [TENENBAUM ET AL 2000]

MDS: tries to preserve best the linear distances

ISOMAP: tries to preserve the geodesic (non-linear) distances

Technique:



(2) methods

(Multichannel)-Singular System Analysis (M-SSA)

[GHIL ET AL 2002, GOLYANDINA ET AL 2002]

- Time series embedding for the extraction of components on different time-scales
“PCA for the time-domain”
- Construct a trajectory matrix containing M sub-series of $dim(N - M + 1)$ where the lag is $|i - j|$ to calculate the Toeplitz-(lag-covariance)-matrix...

$$(T_{l,l'})_{ij} = \frac{1}{N - |i - j|} \sum_{t=1}^{N - |i - j|} x_{l,t} x_{l',t+i-j} \quad (1 \leq i; j \leq M)$$

- Keep the Toeplitz-matrix subject to SVD

$$\tilde{\mathbf{T}} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$$

(2) methods

Multichannel SSA (M-SSA), [BROOMHEAD ET KING 1986, GHIL ET AL 2002]

- The projection of the original series on the eigenvectors leads to the “space-time principal components”: → univariate time-series!

$$A = XE, \dim \mathbf{A} = (N - M + 1) \times (L + M)$$

$$a_i^k = \sum_{j=1}^M \sum_{l=1}^L x_{l,i+j-1} e_{l,j}^k$$

- Each PC allows the reconstruction of a part of explained variance at each channel as a sub series called “Reconstructed Component”

$$r_{l,i}^k = \frac{1}{M_i} \sum_{k \in \mathcal{K}} \sum_{j=L_i}^{U_i} a_{i-j+1}^k e_{l,j}^k \quad X = (x_i, \dots, x_N)^T = \sum_{k=1}^M R_k$$

- This allows to identify the effect of each EOF in spatial resolution

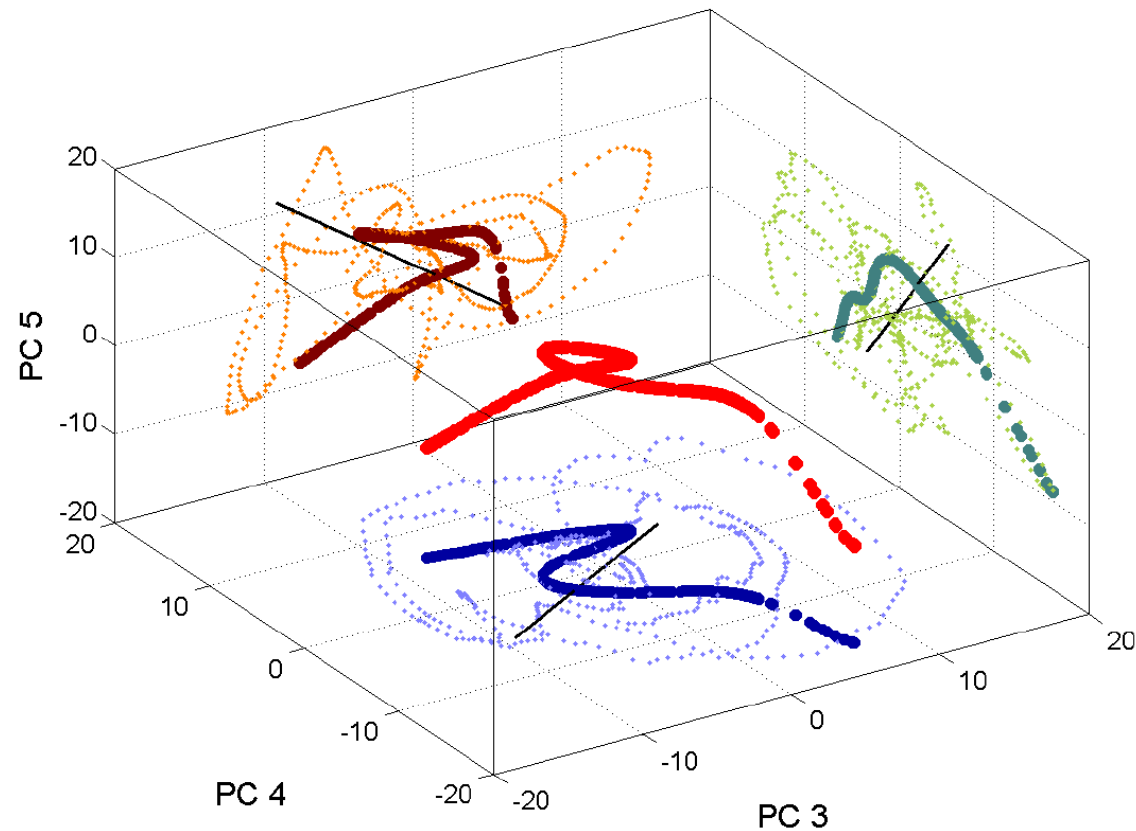
(2) methods

Non-linear M-SSA (*n*lM-SSA) [HSIEH & WU 2002]

Generalize the SSA to open- or closed curve EOF's (Analogue to *n*lPCA) ... two ways

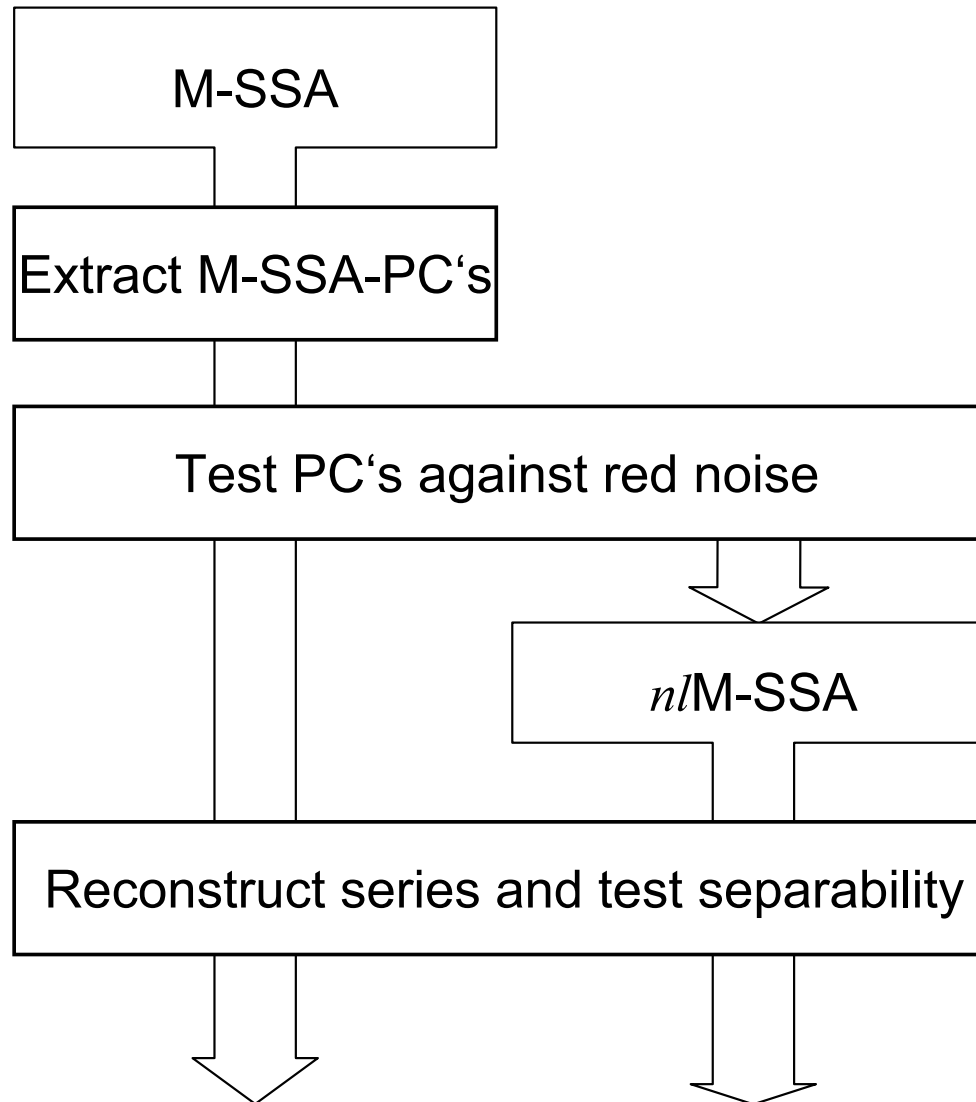
(1) Build trajectory matrix, sequentially *n*lEOF-extraction → Project to series as *n*lPC/*n*lRC

(2) Generalize SSA-PC's with *n*lPCA → *n*lRC's



(2) methods

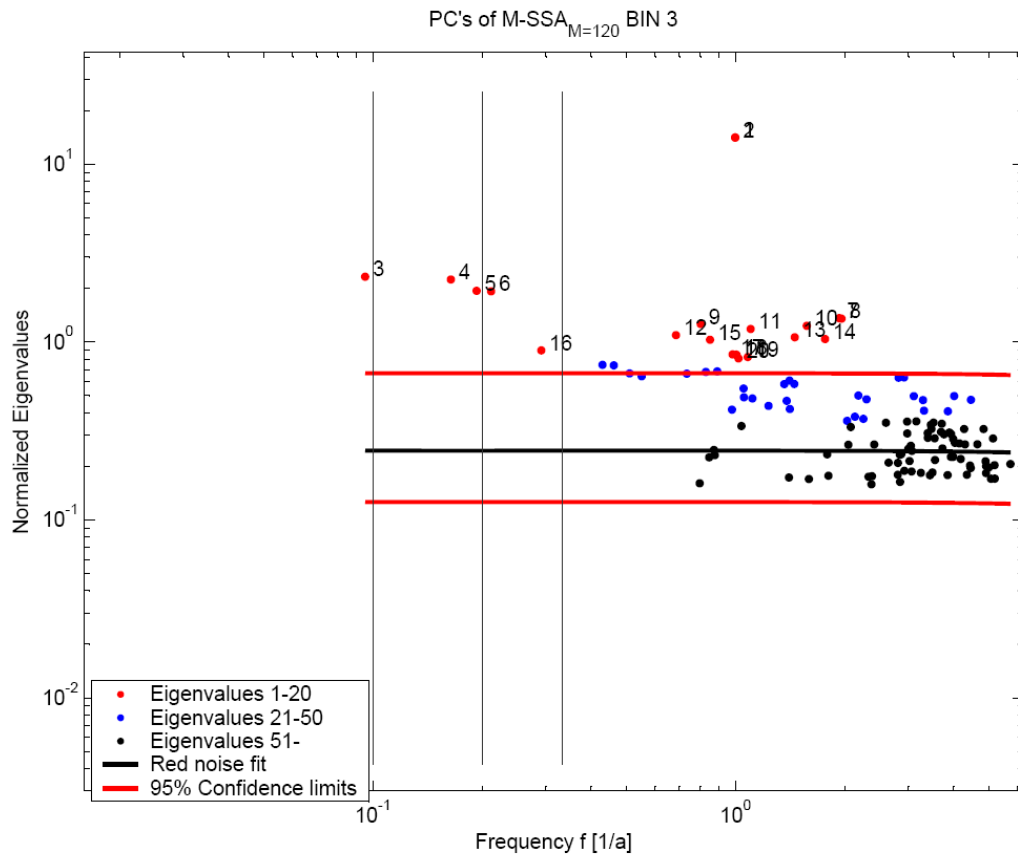
M-SSA: Flow chart for an M-SSA procedure



(3) results

M-SSA & n /M-SSA: Aggregation: monthly-means of 7 to 22 channels

M-SSA_(N/M=5)

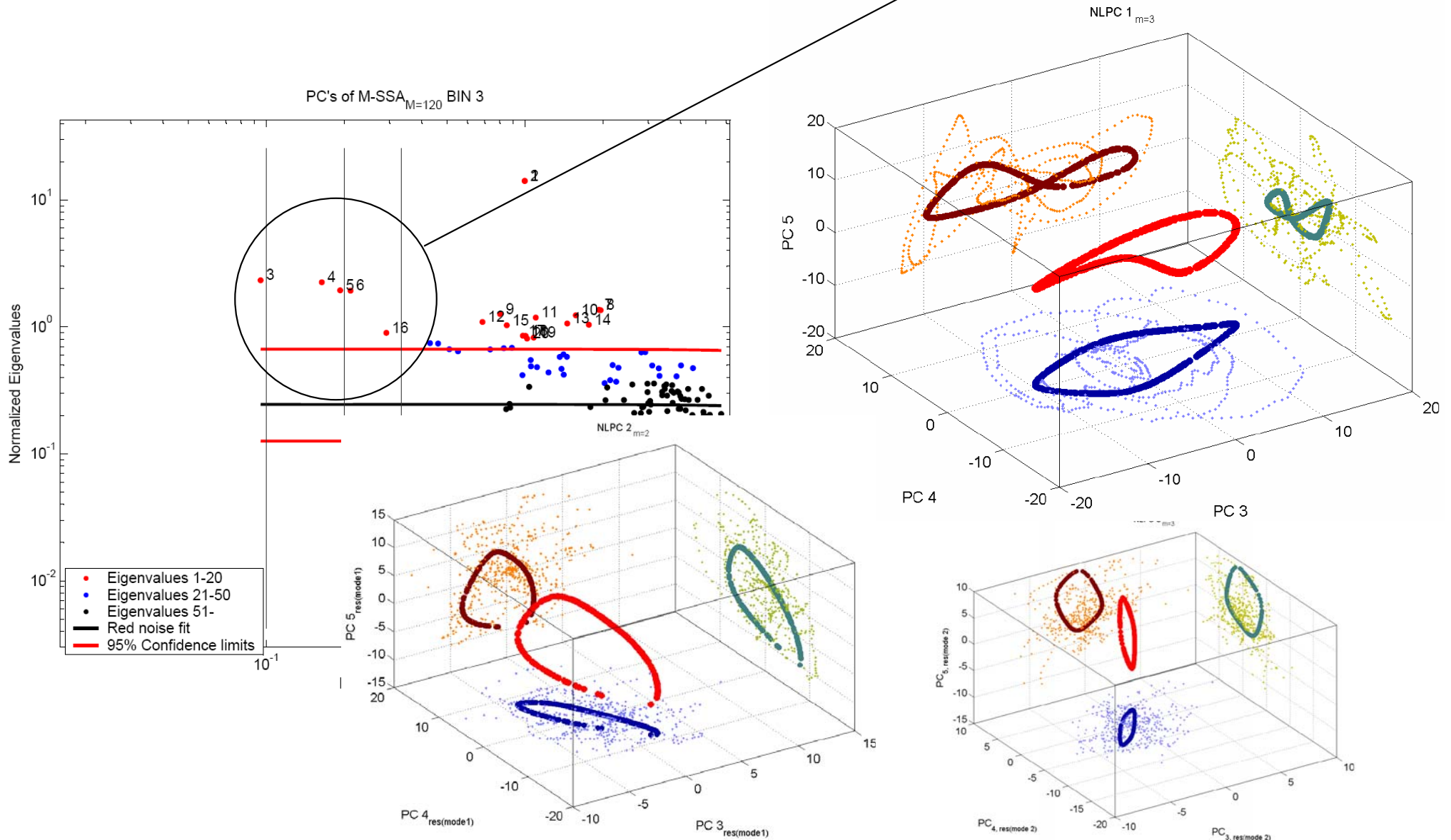


(3) results

M-SSA & *n*/M-SSA: Aggregation: monthly-means & 7 to 22 channels

M-SSA_(N/M=5)

n/M-SSA

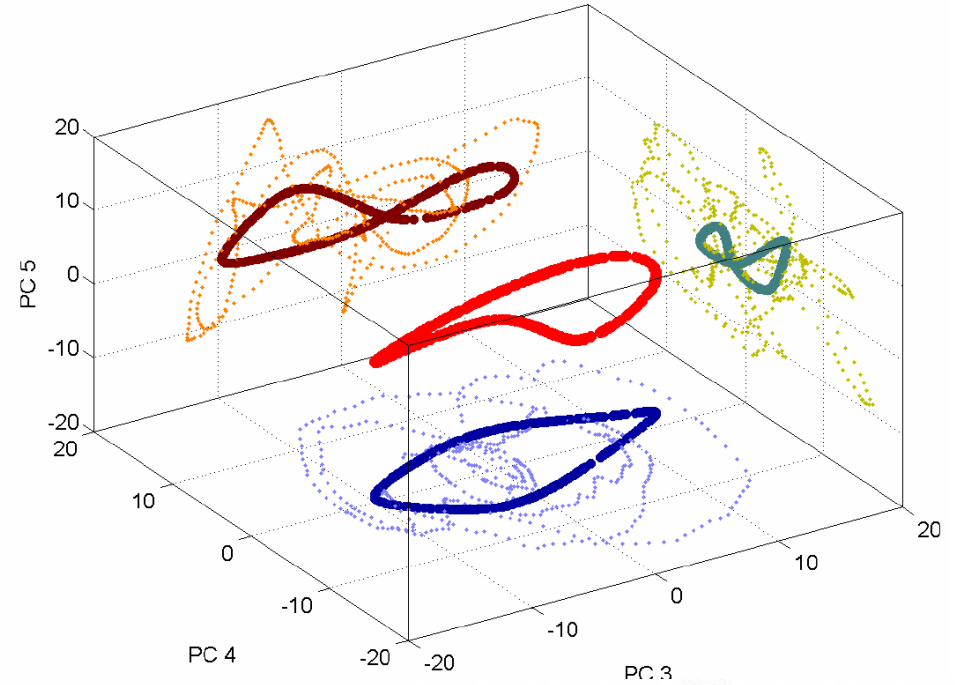
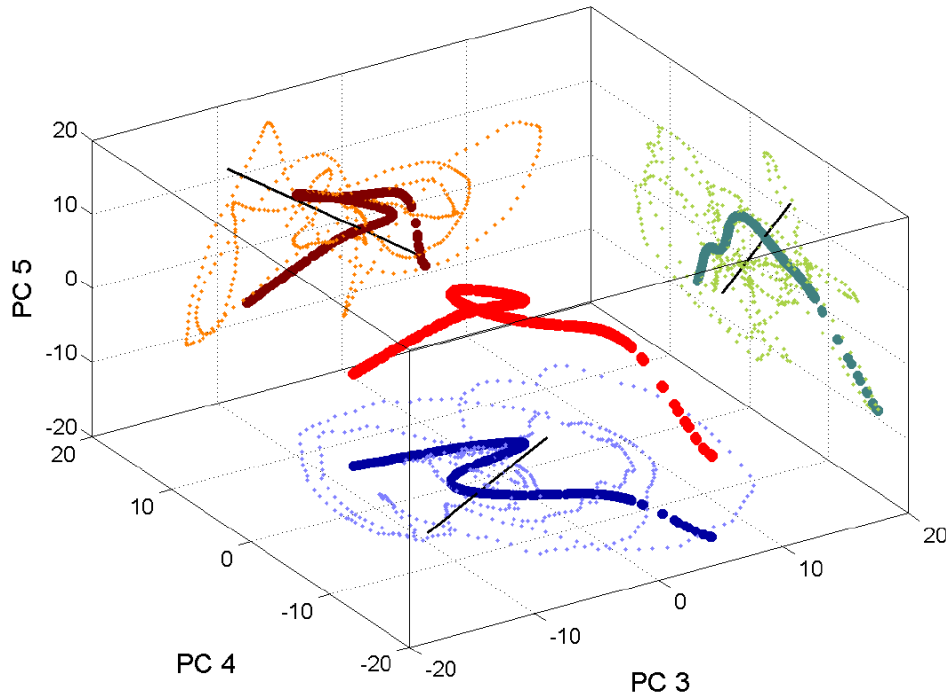


(3) results

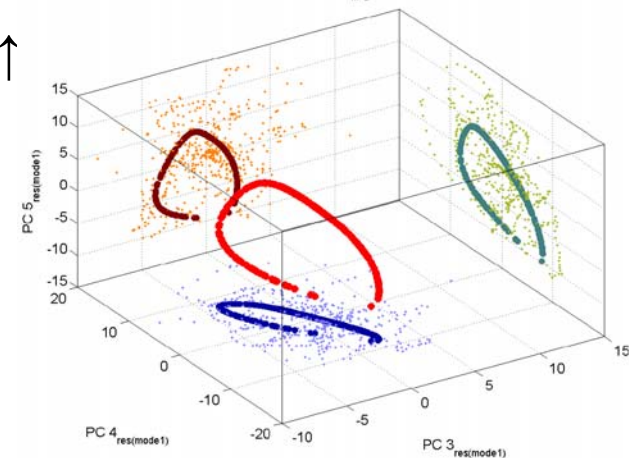
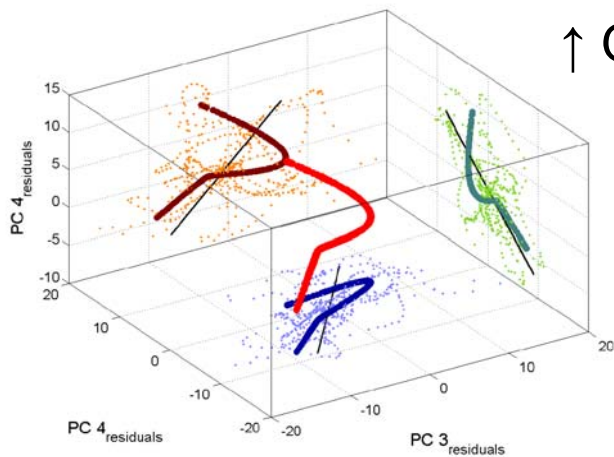
M-SSA & *n*M-SSA: Aggregation: monthly-means & 7 to 22 channels

***n*M-SSA**

NLPC 1_{m=3}

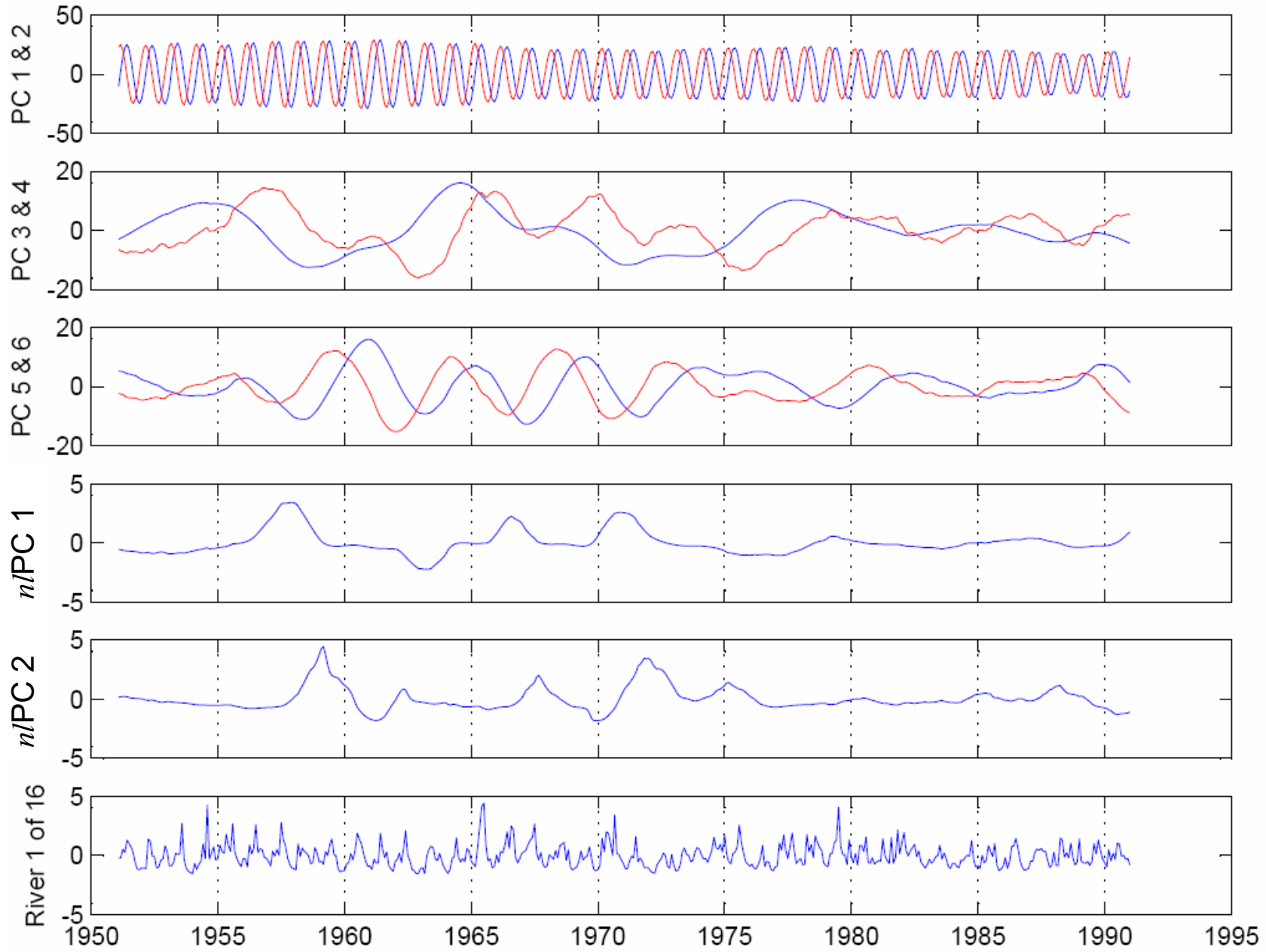


↑ Open versus closed curve solution ↑



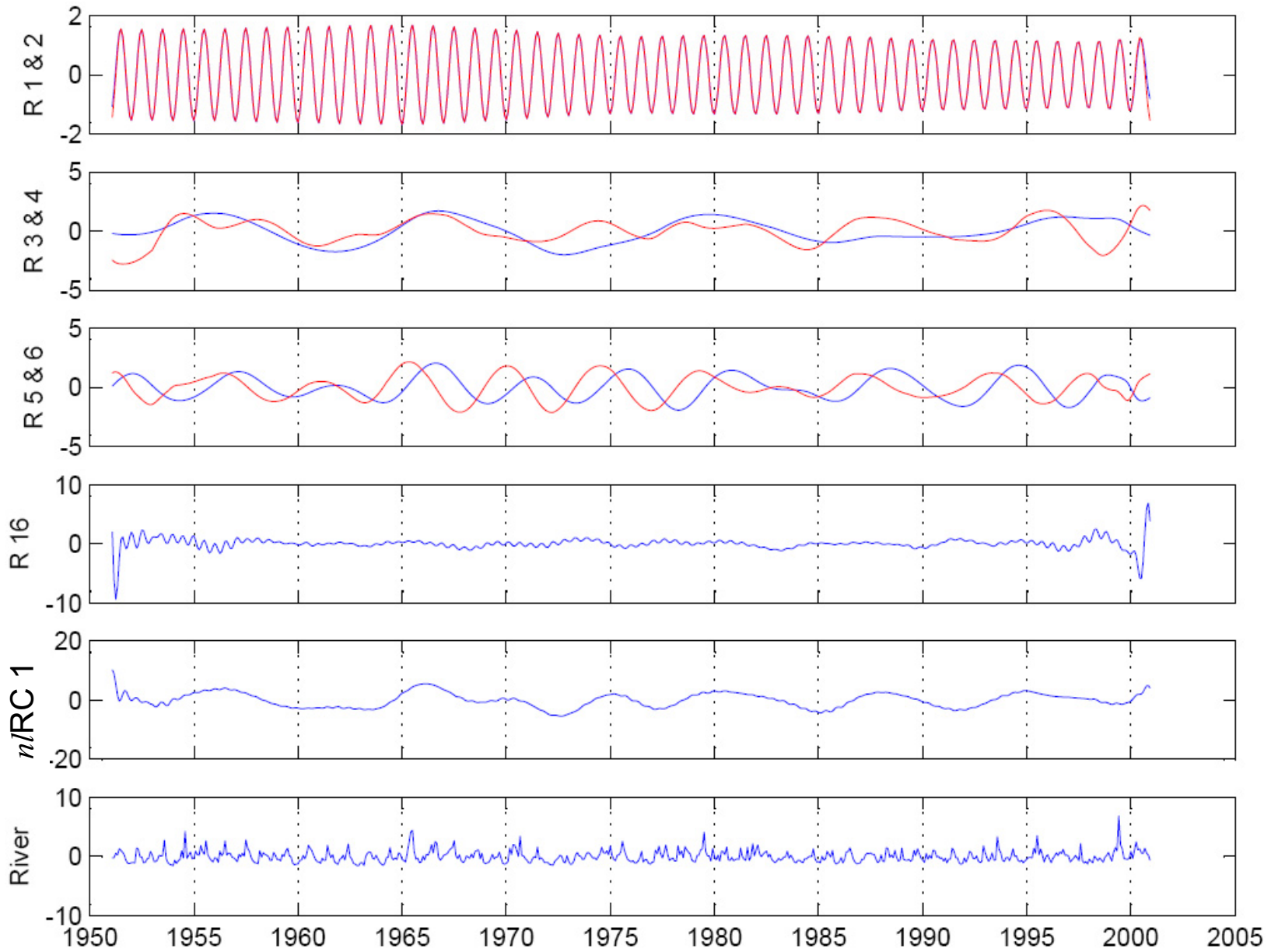
(3) results

M-SSA & n /M-SSA:



(3) results

M-SSA & n M-SSA:



Reconstruction for river 1 „Donau Berg“

(3) results

M-SSA: Week-separability of the reconstructed components RC 's:

- The inner product of two reconstructed components defined as

$$(R_l, R_{l'})_w \stackrel{def}{=} \sum_{i=0}^{N-1} w_i r_{i,l} r_{i,l'}$$

- is called w -orthogonal if

$$(R_l, R_{l'})_w = 0$$

- where the natural measure of deviation from w -orthogonality is

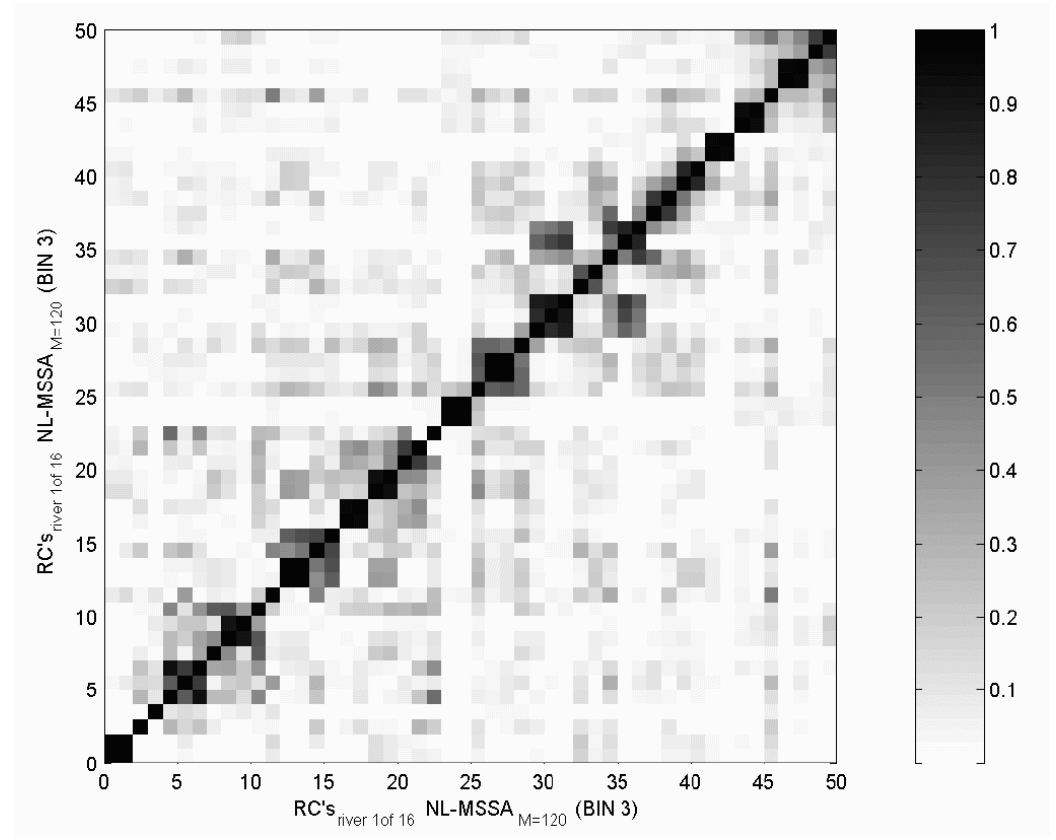
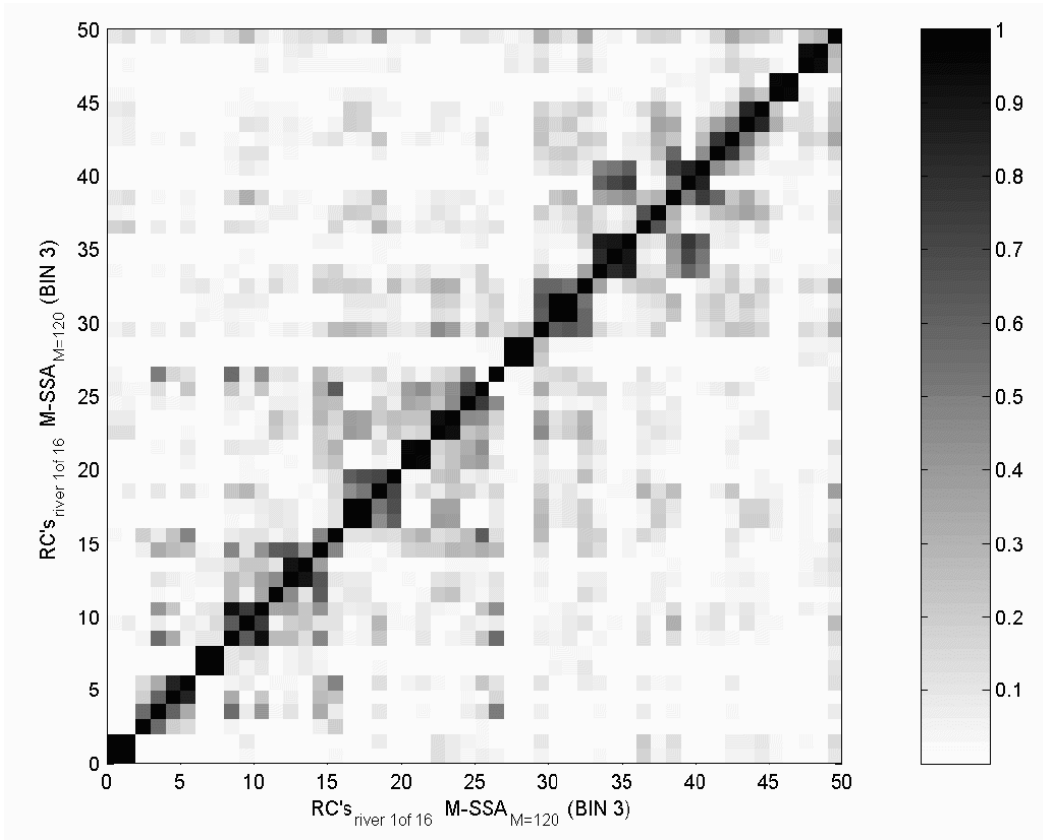
$$\rho_{l,l'} = \frac{(R_l, R_{l'})_w}{\|R_l\|_w \|R_{l'}\|_w} \quad \text{and} \quad \|R_l\|_w = \sqrt{(R_l, R_l)_w}$$

(3) results

M-SSA & n M-SSA: w -correlations for reconstructed components of river 1

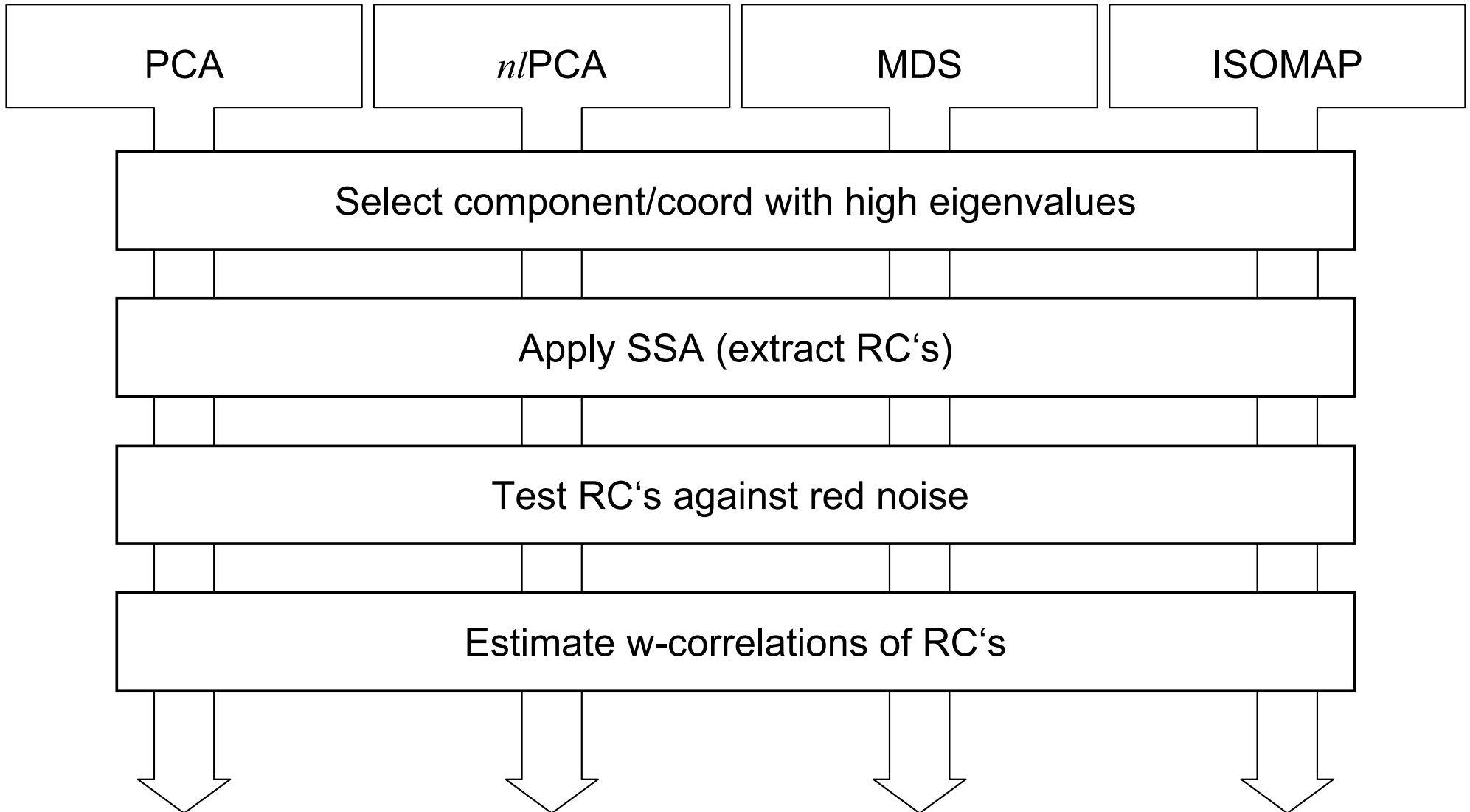
M-SSA_(N/M=5)

n M-SSA



(3) results

Spatial dimensionality reduction of data-set, followed by SSA

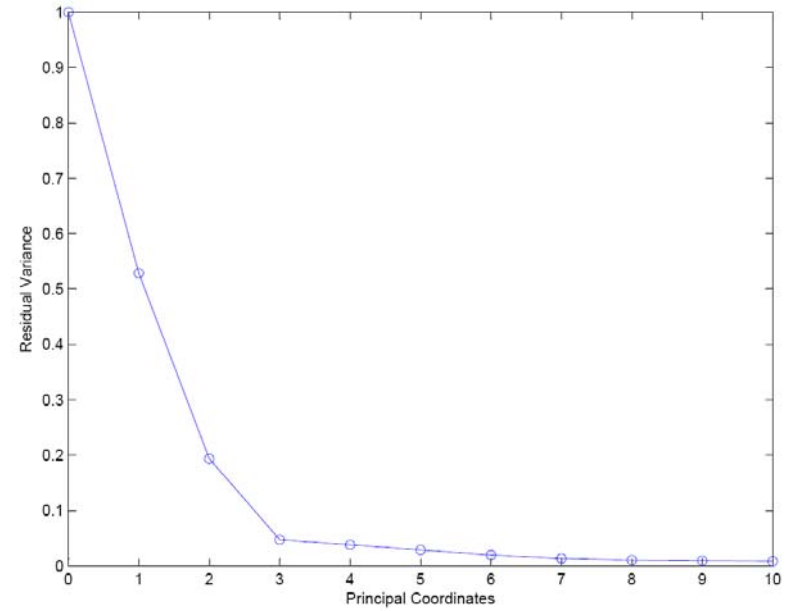
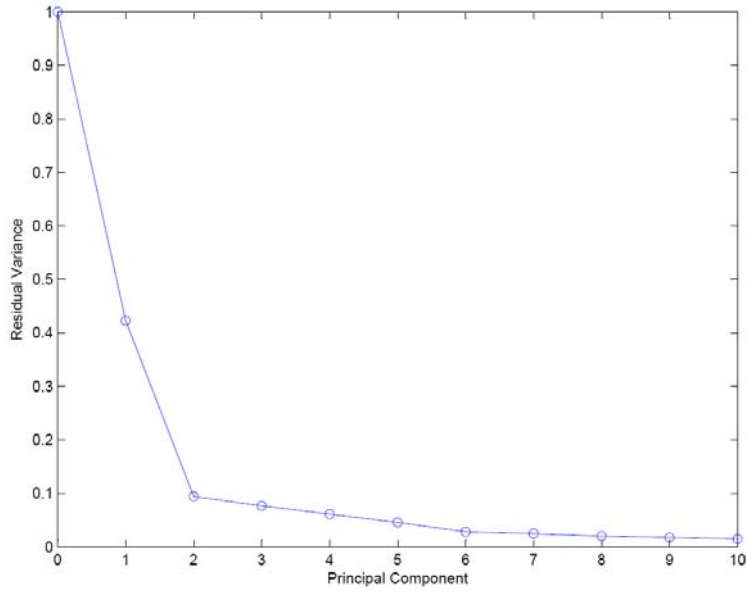


(3) results

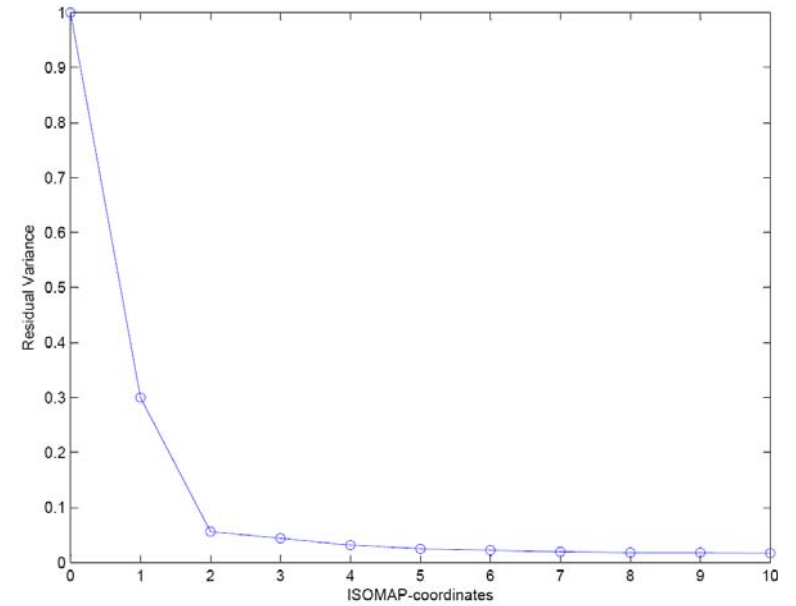
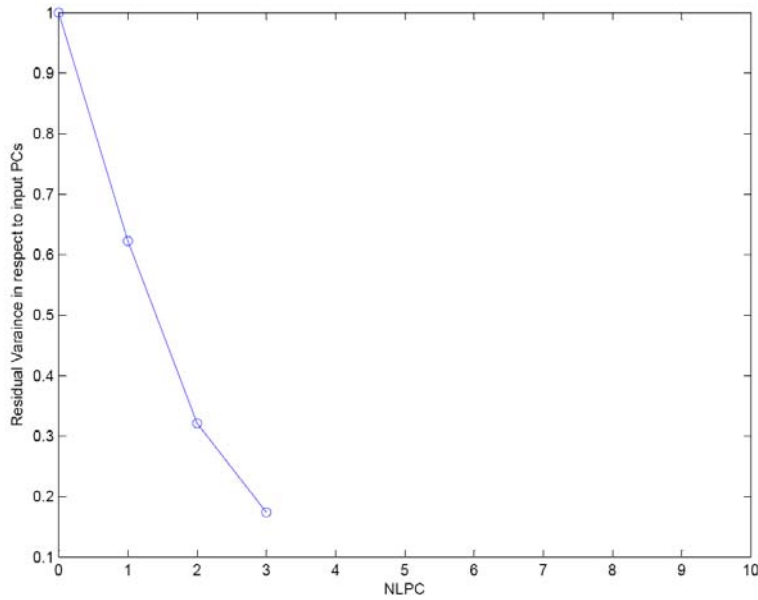
PCA

MDS

linear



Non-linear



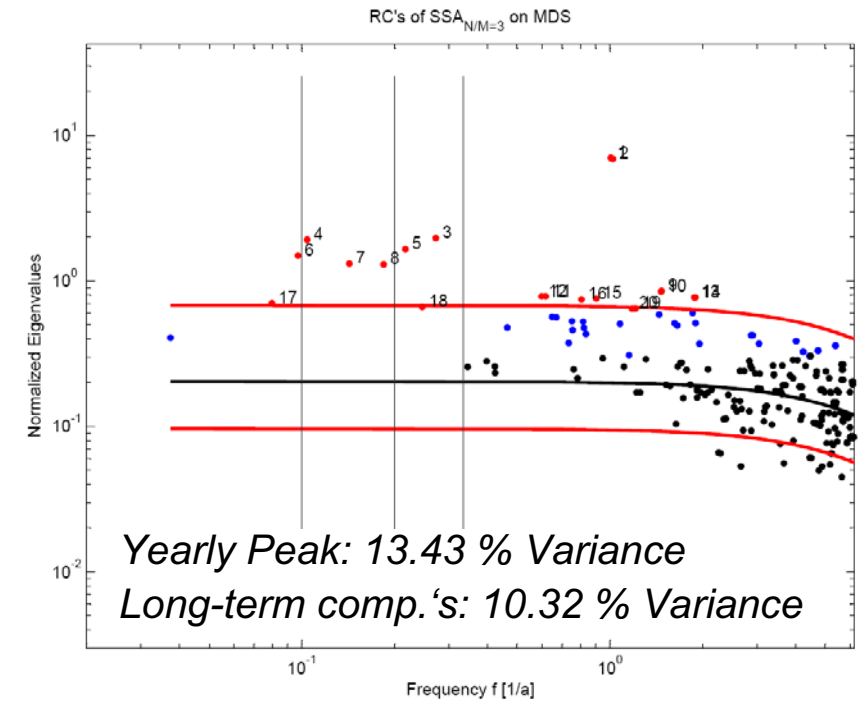
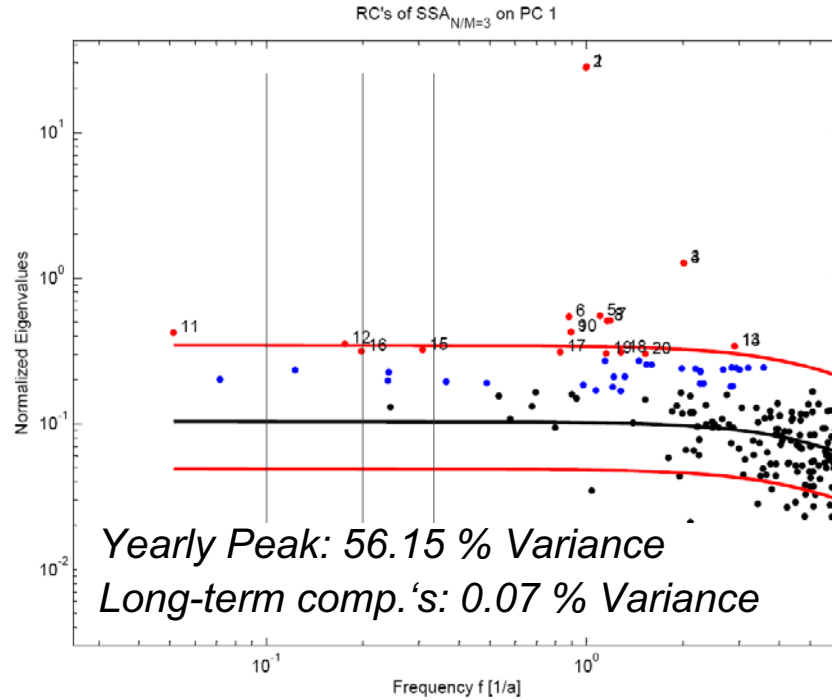
(3) results

PCA

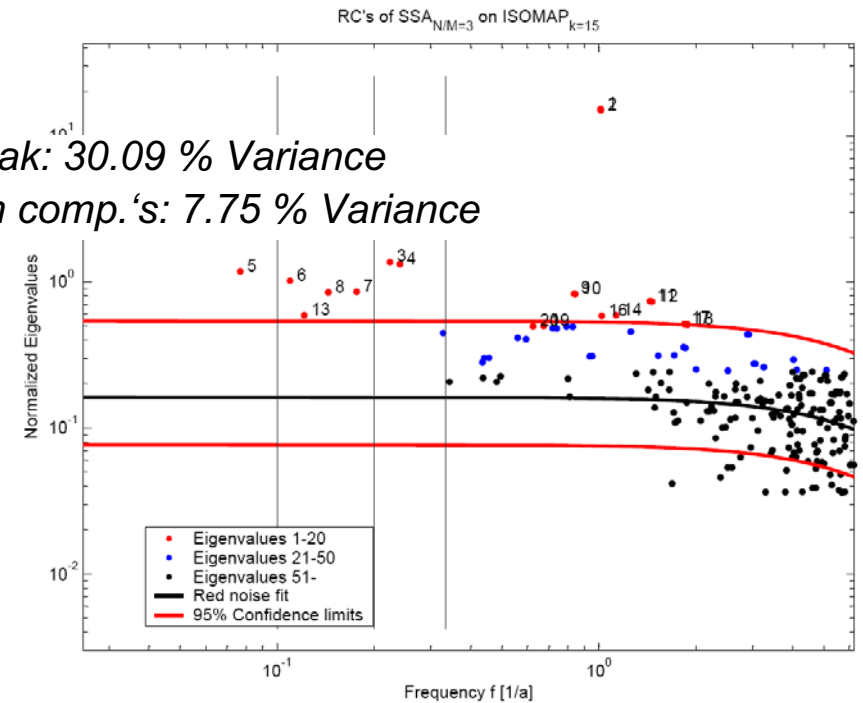
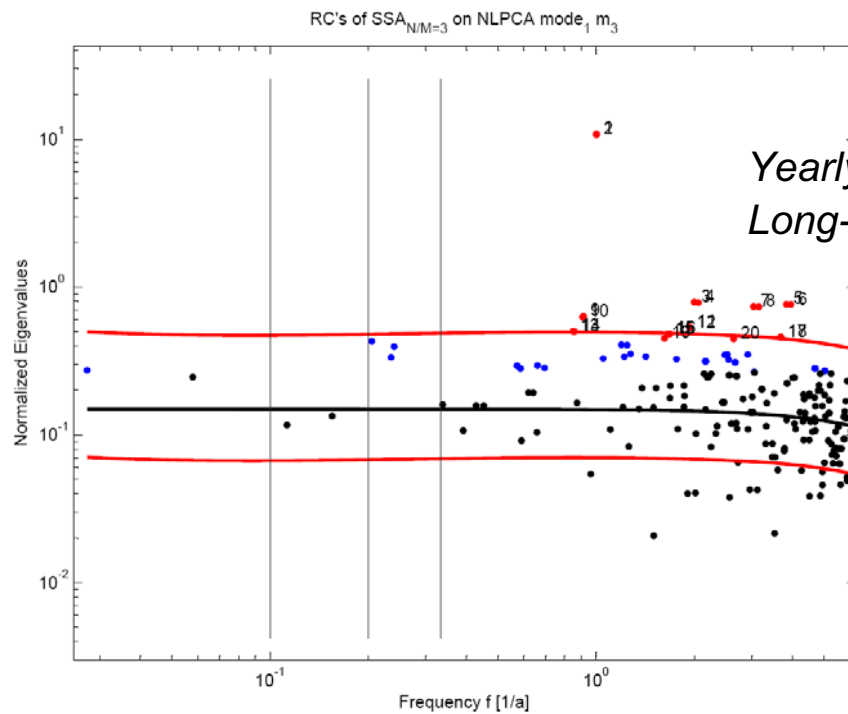
MDS

Application of SSA to 1st components/coordinates

linear



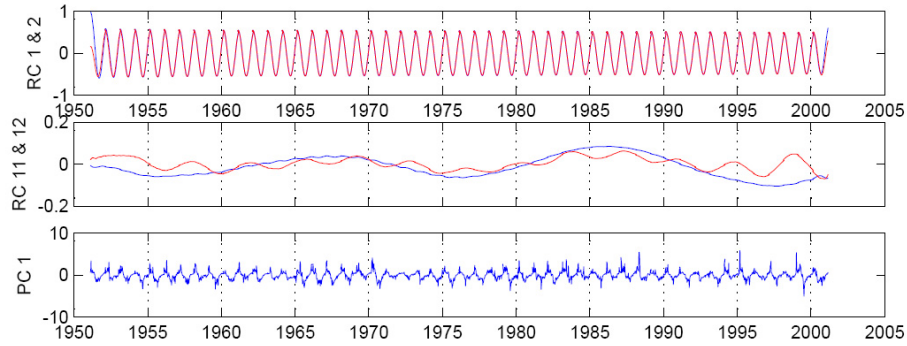
Non-linear



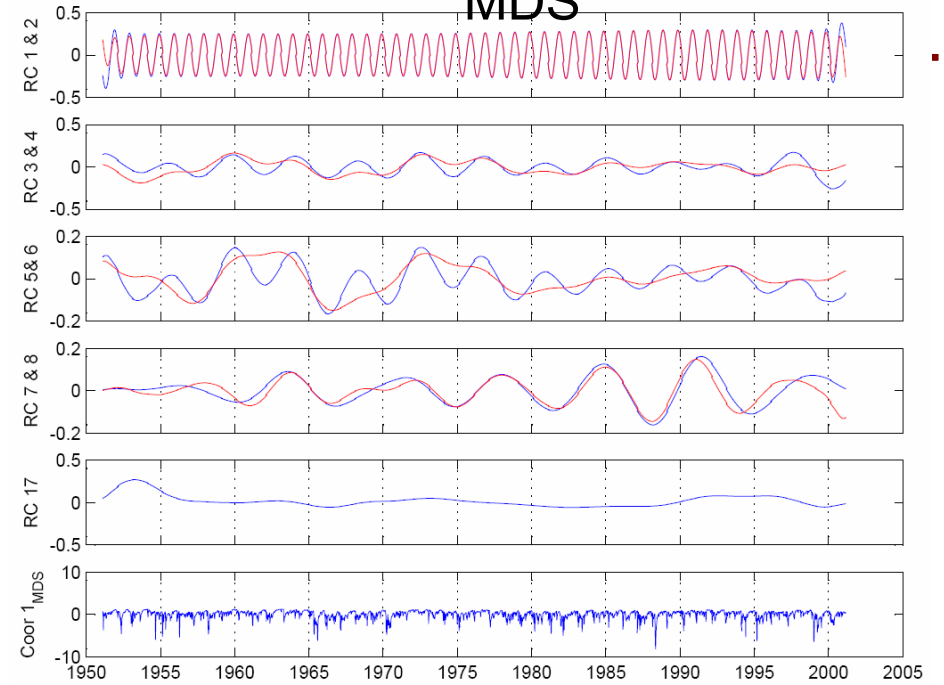
(3) results

PCA

linear

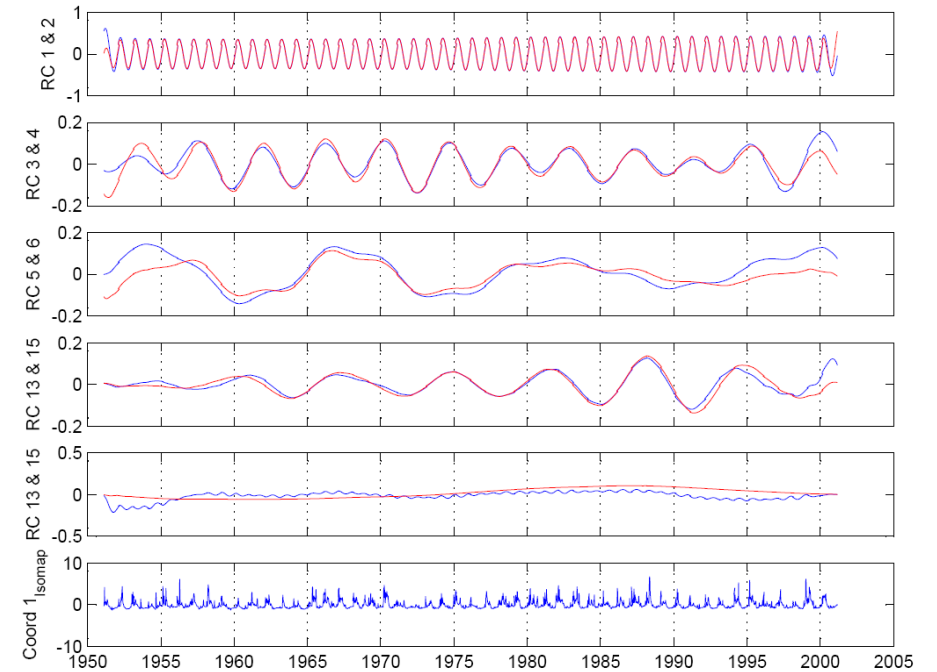
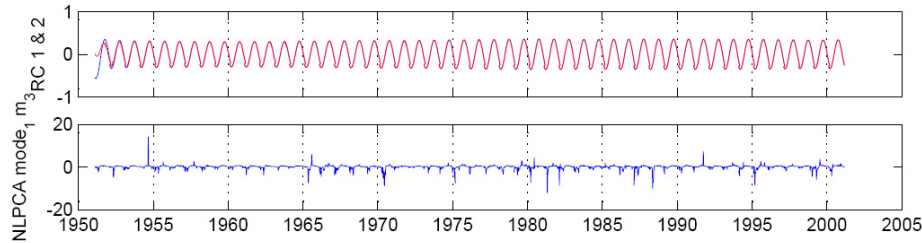


MDS



	<i>Periods (yr)</i>								
PCA	5.68	19.47							
MDS	3.67	9.60	4.61	10.29	7.01	5.42	12.51		
ISOMAP	4.46	4.15	13.01	9.11	5.66	6.92	8.23	48.35	

Non-linear

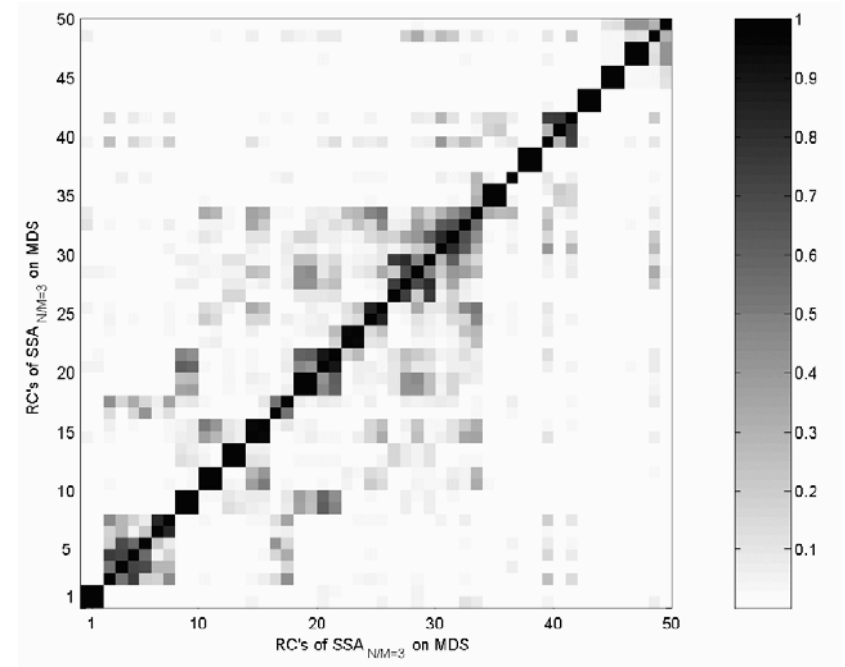
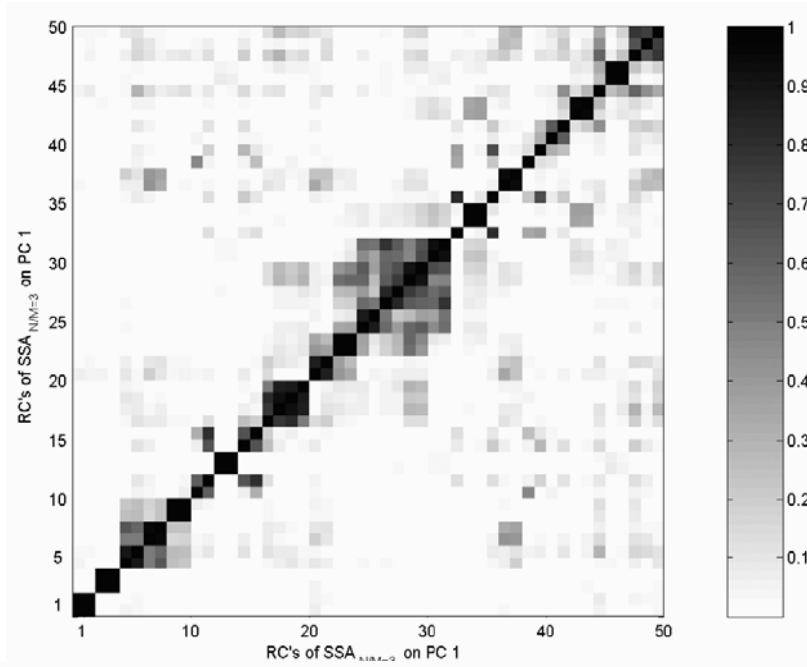


(3) results

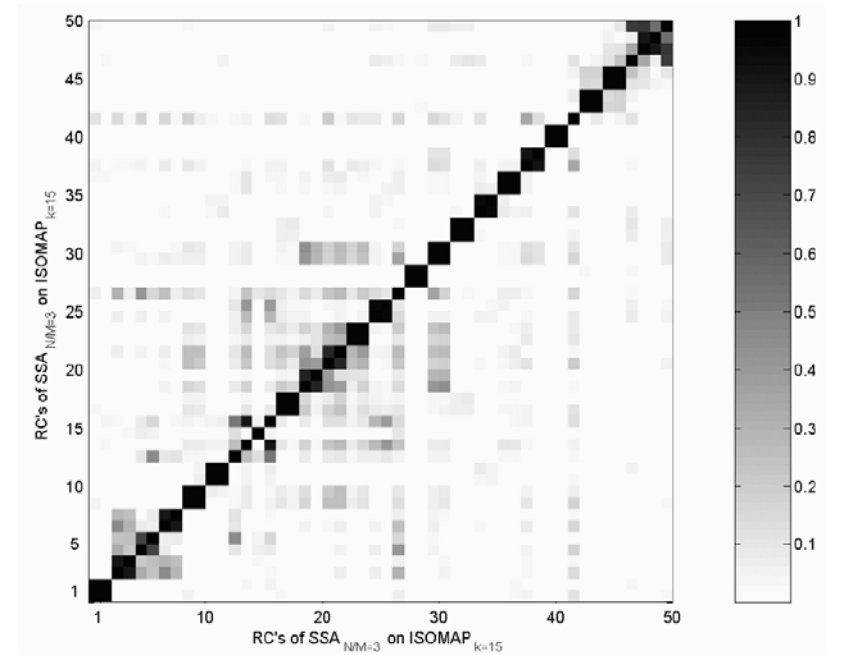
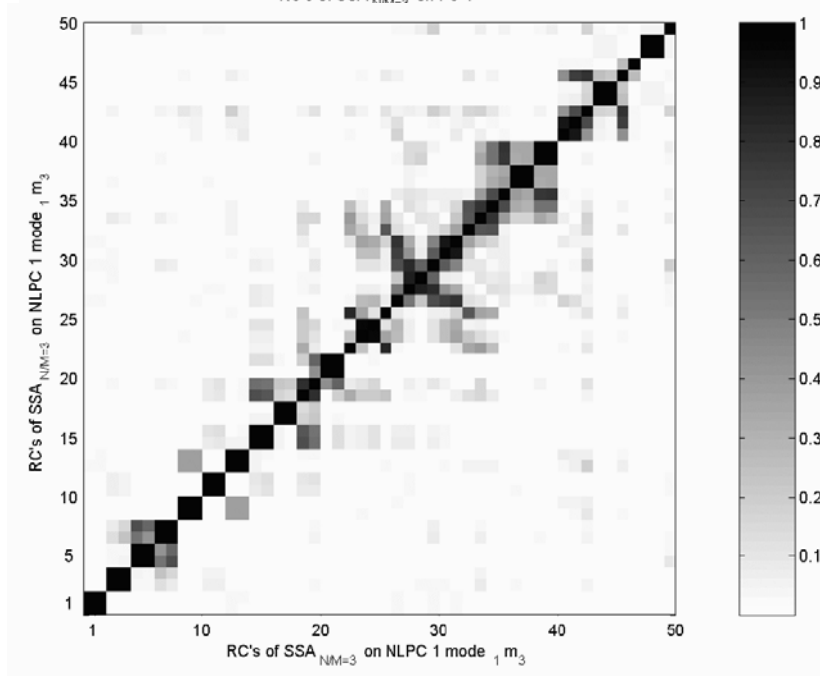
PCA

MDS

linear



Non-linear



(4) discussion

Dimensionality reduction (PCA, *nl*PCA, MDS, ISOMAP) followed by SSA →

- Suitability for the extraction of long-term structures:
ISOMAP > MDS >> PCA > *nl*PCA,
distance-based approaches lead to a better extraction of long-term structures
- Excellent computational efficiency, allows to explore varying aggregation levels
- Lower aggregation-levels lead to a better extraction of long-term structures.
→ to be investigated
- No information of spatial behavior

(4) discussion

Multivariate time series analysis (M-SSA, n /M-SSA) →

- Linear M-SSA performs better than the non-linear variant
- Low computational performance (requires high temporal and low spatial aggregation)
- Information is interpretable in space

(5) Conclusion

Recall the objective and aim of the study:

- To investigate spatial patterns of long-term components
- To compare different (nonlinear) multi-channel methods

Conclusions:

- (1) Long-term components account for approx 10% of the variance
- (2) Spatial (non-linear) dimensionality reduction followed by SSA lead to best results ...
 - highest values of explained variance,
 - best signal- to noise enhancement,
 - most significant long-term components
- (3) No spatial information can be extracted

The next steps to do are:

Explore varying aggregation levels on the NLDR

Find geographical patterns of the significance of the identified long-term-structures