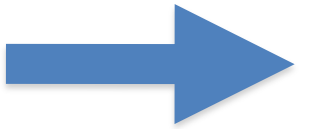
Fabio D'Andrea
LMD – 4e étage
01 44 32 22 31

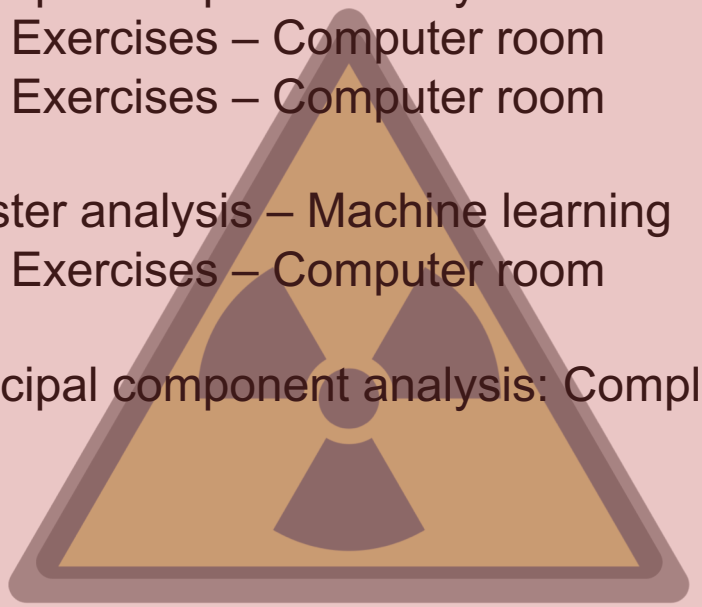
dandrea@lmd.ens.fr

http://www.lmd.ens.fr/dandrea/TEACH

Program

17/1	Elementary statistics – 1
23/1	Elementary statistics - 2
14/2	Exercises – Computer room
14/2	Fourier Analysis -1
3/3	Fourier Analysis -2, stochastic processes
6/3	Exercises – Computer room
13/3	Exercises – Computer room
13/3	Principal component analysis -1
20/3	Principal component analysis -2
27/3	Exercises – Computer room
3/4	Exercises – Computer room
17/4	Cluster analysis – Machine learning
24/4	Exercises – Computer room
15/5	Principal component analysis: Complements
22/5	Exam



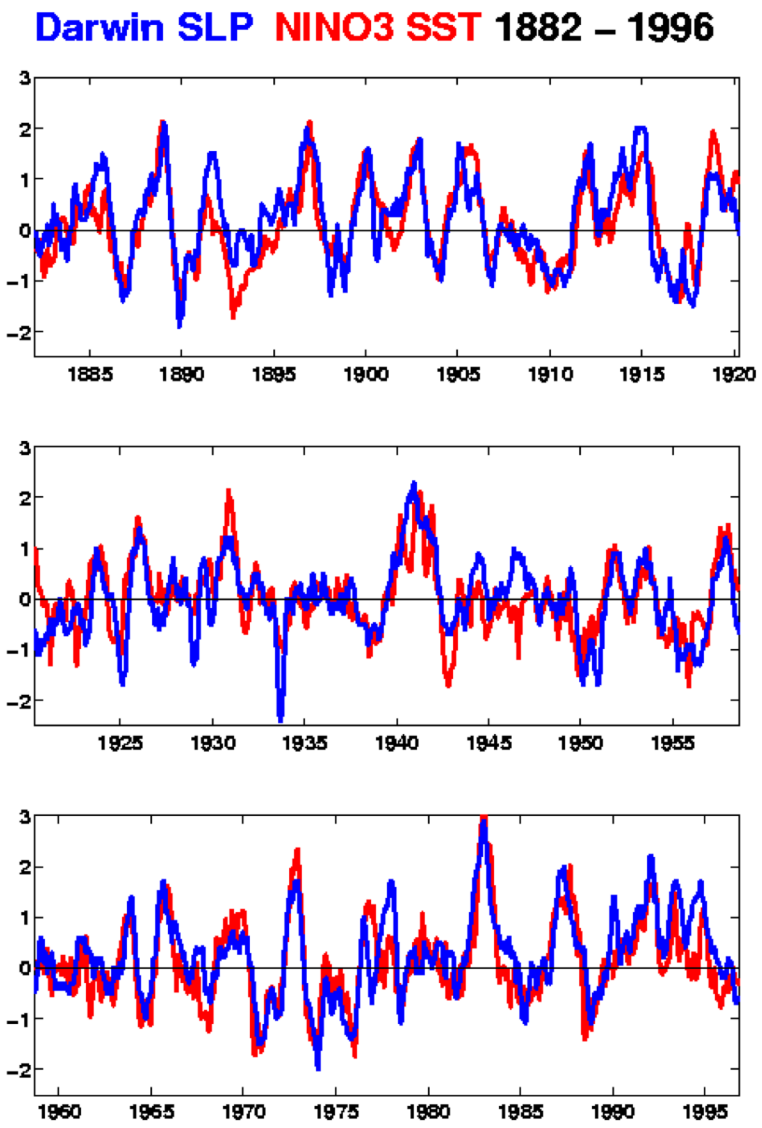


1) Maximum Covariance Analysis:
application of the PCA to the case of two vector time series. In the same way as the EOF generalized the concept of variance to the case of a vector data, this generalizes the concept of covariance of two timeseries.

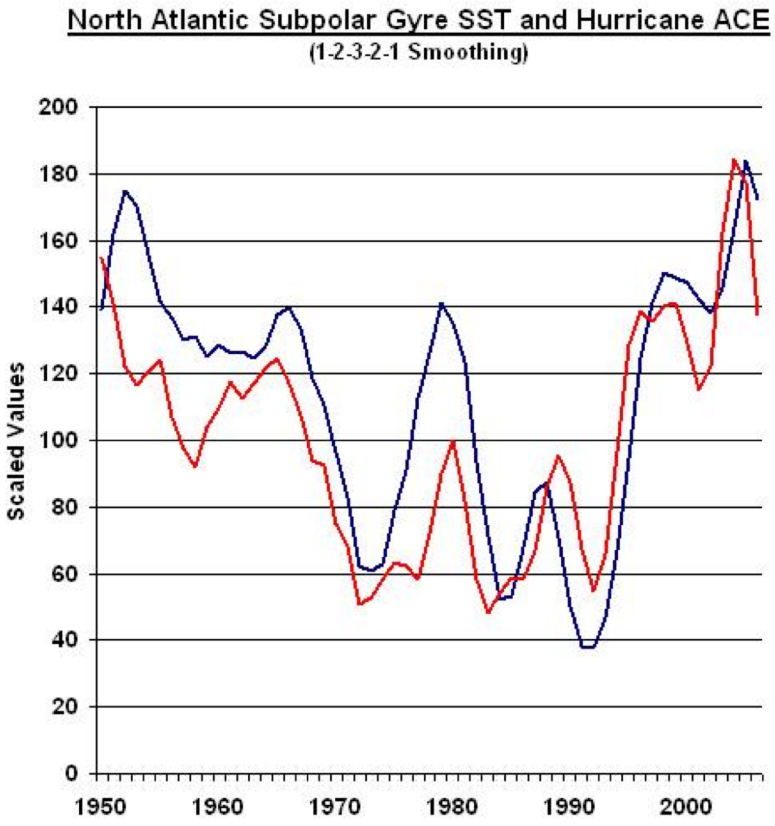
2) Rotation of EOF.
Interpretation of the EOF patterns

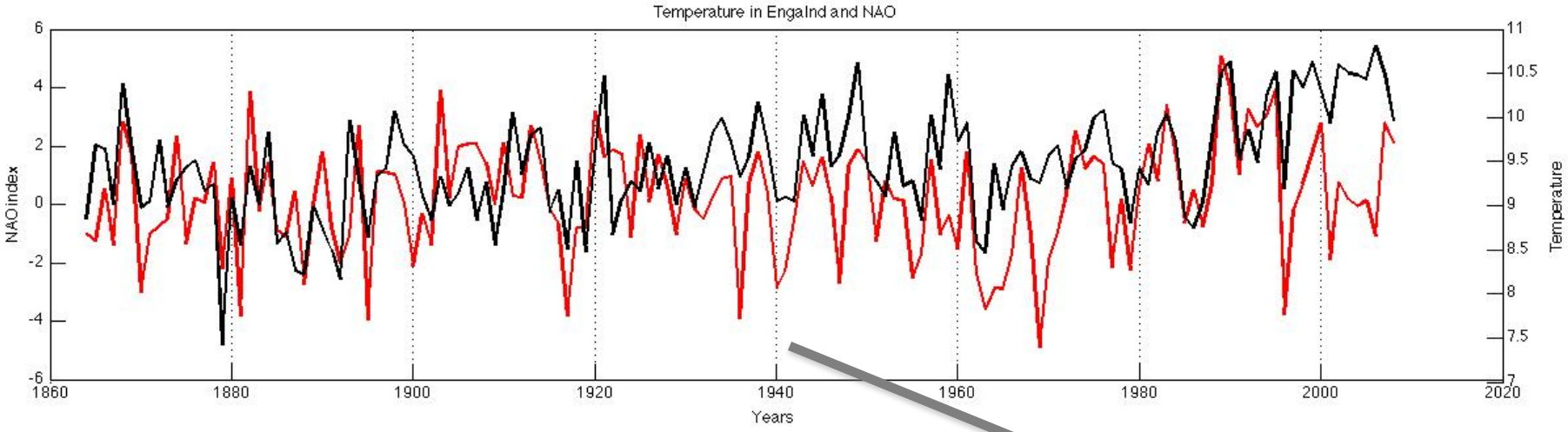
3) PCA in the time domain: SSA and Extended EOFs.

Are two timeseries correlated?



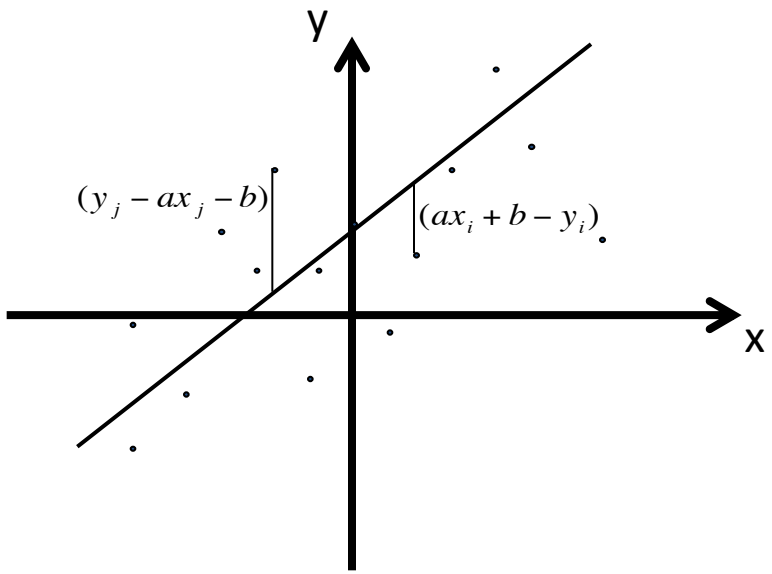
In geophysics, sometimes one wonders whether two timeseries are linked.

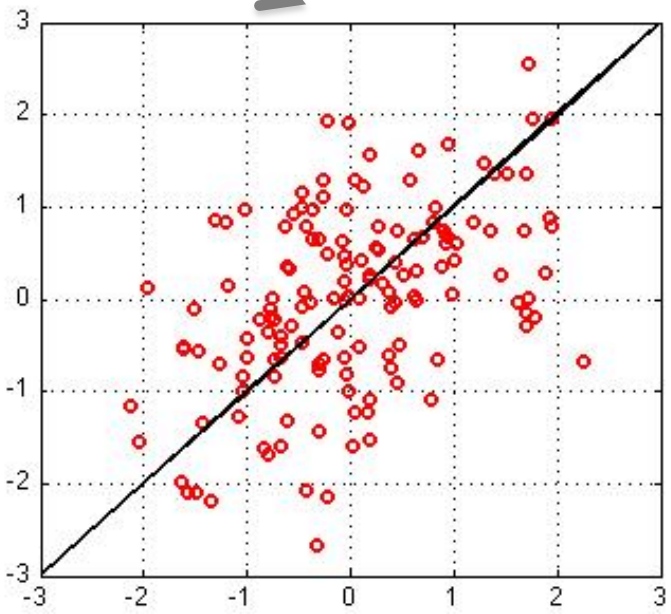


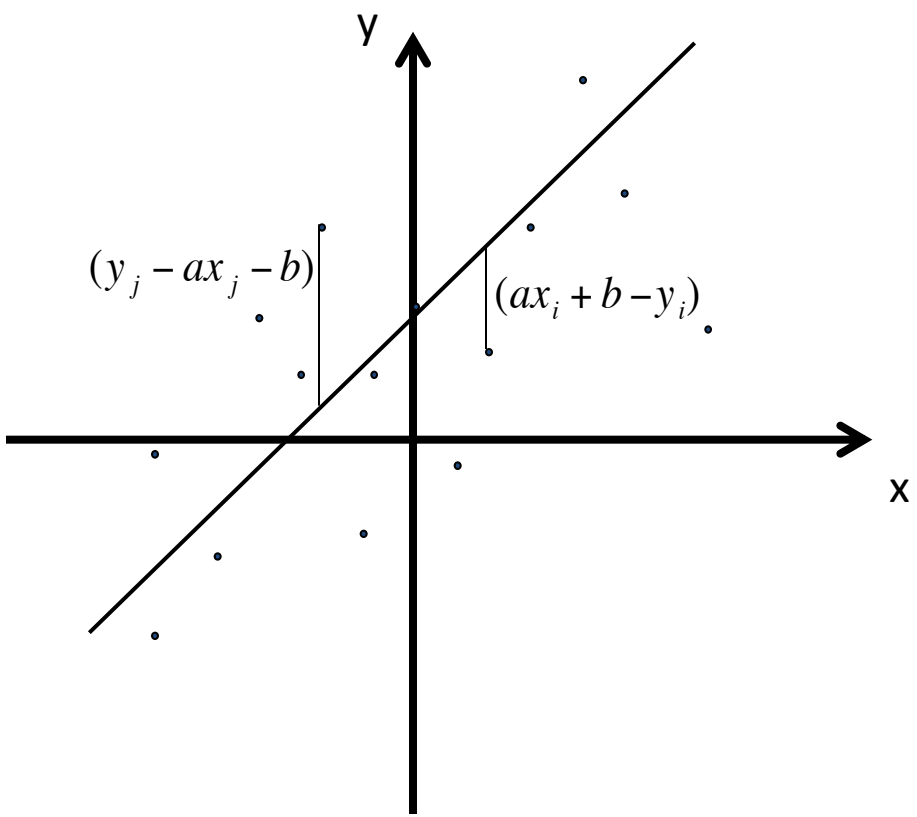


One way to put it is to estimate if one series can be obtained by linear transformation from the other:

$$y_i = ax_i + b$$







We want to minimize:

$$\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2$$

We take the derivative with respect to *a* and *b* and we obtain the two conditions:

$$a)\quad \sum x_i\left(y_i - ax_i - b\right) = 0$$

$$b)\quad \sum\left(y_i - ax_i - b\right) = 0$$

Condition b) gives:

$$b)\quad \sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i - Nb = 0 \Rightarrow \boxed{b = \bar{y} - a\bar{x}}$$

Substituting b) into a) gives:

$$a)\quad \sum_{i=1}^{N}\left(y_i x_i - ax_i^2 - \bar{y}x_i - a\bar{x}x_i\right) = \sum_{i=1}^{N}\left(y_i' x_i' - ax_i'^2\right)$$

Where we have introduced the definitions :

$$x_i = \bar{x} + x',\ y_i = \bar{y} + y'.$$

Hence

$$a = \frac{\sum_{i=1}^{N} x_i' y_i'}{\sum_{i=1}^{N} x_i'^2}$$

$$b = \bar{y} - a\bar{x}$$

Regression

How good is the regression?

The regression is not perfect: $\hat{y}_i = ax_i + b \neq y_i$

Introducing the error $y_i^* = y_i - \hat{y}_i$ We can write $y_i = ax_i + b + y_i^*$

And the variance of y becomes:

$$\overline{y'^2} = \overline{a^2 x'^2} + \overline{y^{*2}} \quad \Rightarrow \quad \boxed{\frac{\overline{a^2 x'^2} + \overline{y^{*2}}}{\overline{y'^2}} = 1}$$

Explained variance + unexplained variance = 1

Substituting the value of *a* found above we find:

$$\frac{\overline{a^2 x_i'^2}}{\overline{y_i'^2}} = \frac{\left(\overline{x'y'}\right)^2}{\overline{x'^2}\,\overline{y'^2}} = r^2, \qquad r = \frac{\overline{x'y'}}{\sigma_x \sigma_y}$$ Is the correlation coefficient

$$\boxed{r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}; \quad 1 - r^2 = \frac{\text{Unexplained Variance}}{\text{Total Variance}}}$$

In PCA one looks for the structure, or vector, that, projected on the data defined a time series with maximum variance.

For generic norm defined by the metric matrix M we maximized:

$$\overline{(\mathbf{e}^T M\mathbf{x})^2} = \mathbf{e}^T M\overline{\mathbf{x}\mathbf{x}^T} M\mathbf{e} = \mathbf{e}^T MCM\mathbf{e}.$$

This is the variance of the scalar time series $\mathbf{e}^T MX$, if X is the vector timeseries of data, expressed as an array, i.e. an array whose columns are the vectors of the timeseries at different times.

And remember that the variance of it was exactly equal to λ, the higher eigenvalue of $\frac{XX^T}{N} M$

We will now generalize the PCA approach.

Suppose we have two timeseries, X and Y, of the measurements of two different physical variables. The two timeseries don't need to have the same spatial location (the grid) but they do have one dimension in common, time. I.e. the measurements are synchronous.

We want to find the two unit vectors that, each one projected on a timeseries, maximize the covariance between the two resulting scalar series. Using the **canonic norm**, that gives the following constrained maximization problem:

$$\max\left(\mathbf{e}^T \frac{XY^T}{N} \mathbf{f} - \lambda(\mathbf{e}^T\mathbf{e} - 1) - \mu(\mathbf{f}^T\mathbf{f} - 1) \right)$$

That uses the lagrange multipliers lambda and mu.

Maximum Covariance Analysis (MCA)

Taking the derivative with respect to **e** and **f** one gets

$$\begin{cases} \frac{XY^T}{N}\mathbf{f} - 2\lambda\mathbf{e} = 0 \\ \frac{YX^T}{N}\mathbf{e} - 2\mu\mathbf{f} = 0, \text{ or} \end{cases}$$

$$\begin{cases} C_{xy}\mathbf{f} - 2\lambda\mathbf{e} = 0 \\ C_{xy}{}^T\mathbf{e} - 2\mu\mathbf{f} = 0, \text{ calling } C_{xy} \text{ the cross-covariance matrix of the two timeseries.} \end{cases}$$

The two equations can be decoupled by substitution, which gives:

$$\begin{cases} C_{xy}C_{xy}{}^T\mathbf{e} - 4\mu\lambda\mathbf{e} = 0 \\ C_{xy}{}^TC_{xy}\mathbf{f} - 4\mu\lambda\mathbf{f} = 0. \end{cases}$$

Thus, the vectors that we are looking for are the eigenvectors of two matrices that are one the transpose of the other, and share the same eigenvalues.

At this point one can solve the two eigenvalues problems and get the vectors.
But there is a more confortable way to get the solution, by **Singular Value Decomposition**

Any rectangular matrix can be decomposed in this way:

$$A = U\Sigma V^T$$

If A is *m* x *n*, U is *m* x *m* and V is *n* x *n*. Σ Is *m* x *n,* and has non-zero elements only on its diagonal.

Furthermore, it is:

$$AA^T U = U\Sigma^2$$

$$A^T AV = V\Sigma^2$$

So **e** and **f** are given by the SVD decomposition of C_{xy} :

$$C_{xy} = E\Sigma F^T$$

Where E and F are the matrices whose columns are all the **e** and **f,** respectively.

What is the meaning of Σ? What is the covariance $\mathrm{cov}(\mathbf{e}^T X, \mathbf{f}^T Y)$?

$$U^T \frac{XY^T}{N} V = U^T U \Sigma V^T V = \Sigma$$

Σ contains the values of the covariances between corresponding projection timeseries.

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & & 0 \\ 0 & \sigma_2 & & & \\ \vdots & & \ddots & & \\ 0 & \cdots & \cdots & \sigma_n & \cdots & 0 \end{pmatrix}$$

Two generalizations and a by-product

1)Weighting by the variance: (Canonical Correlation Analysis)
2)Extension to any norm (the ‘weight matrix’)

3) Computing EOFs by SVD.

1) Canonical Correlation Analysis

In CCA we want to find the two vectors that maximixe the projection timeseries correlation. In this case the maximization problem to solve is:

$$\max\left(\frac{\mathbf{e}^T C_{xy}\mathbf{f}}{\sqrt{\mathbf{e}^T C_x \mathbf{e}\ \ \mathbf{f}^T C_y \mathbf{f}}}\right)$$

It can be shown that its solutions are:

$$\begin{cases} {C_x}^{-1} C_{xy} {C_y}^{-1} {C_{xy}}^T \mathbf{e} - 4\mu\lambda\mathbf{e} = 0 \\ {C_y}^{-1} {C_{xy}}^T {C_x}^{-1} C_{xy}\mathbf{f} - 4\mu\lambda\mathbf{f} = 0. \end{cases}$$

It can also be showed that the same result is obtained applying MCA to data that have been previously projected on the first EOFs, and the PCs are normalized to have variance 1.

2) Computing MCA under a generic norm.

As in the case of PCA we can write the general problem using the metric matrix M.

$$Max\left[\mathbf{e}^T M_x C_{xy} M_y \mathbf{f} - \lambda(\mathbf{e}^T M_x \mathbf{e} - 1) - \mu(\mathbf{f}^T M_y \mathbf{f} - 1)\right]$$

And the solutions are:

$$\begin{cases} M_x C_{xy} M_y C_{xy}{}^T \mathbf{e} - 4\mu\lambda\mathbf{e} = 0 \\ M_y C_{xy}{}^T M_x C_{xy} \mathbf{f} - 4\mu\lambda\mathbf{f} = 0. \end{cases}$$

Note that the result for CCA above is the same as this here, if one takes

$M_x = C_x{}^{-1}$ and $M_y = C_y{}^{-1}$. In other words, CCA is MCA using the Mahalanobis norm in the vector spaces of X and Y.

It can also be shown that for diagonal norms the trick of computing the weights by the change of variable $\mathbf{x}' = \mathbf{m}_x \mathbf{x}$ and $\mathbf{y}' = \mathbf{m}_y \mathbf{y}$ and then applying the inverse of this to the eigenvectors is still valid.

3) Computing the EOFs by SVD

From the property of the SVD decomposition stated above, we can see that if X is the array containing the data, the SVD gives:

$$X = E\Sigma F^T .$$

E are the eigenvectors of XX^T, i.e. the EOFs, and Σ contains the square roots of the eigenvalues of the covariance matrix of X, multiplied by N, the number of data.

The rows of F simply contains the PCs.

So to get the eigenvalues of XX^T :

$$\lambda_i = \frac{\sigma_i^{\,2}}{N}.$$

Example 1

Canonical Correlation Analysis of winter-mean sea level pressure in the North Atlantic and the winter-mean precipitations in the Iberian Peninsula.

Von Storch et al (1993) J. Clim.

The Canonical correlation is 0.75

The modes represent 65% of the total variance of SLP and 40 of the variance of the precipitations.

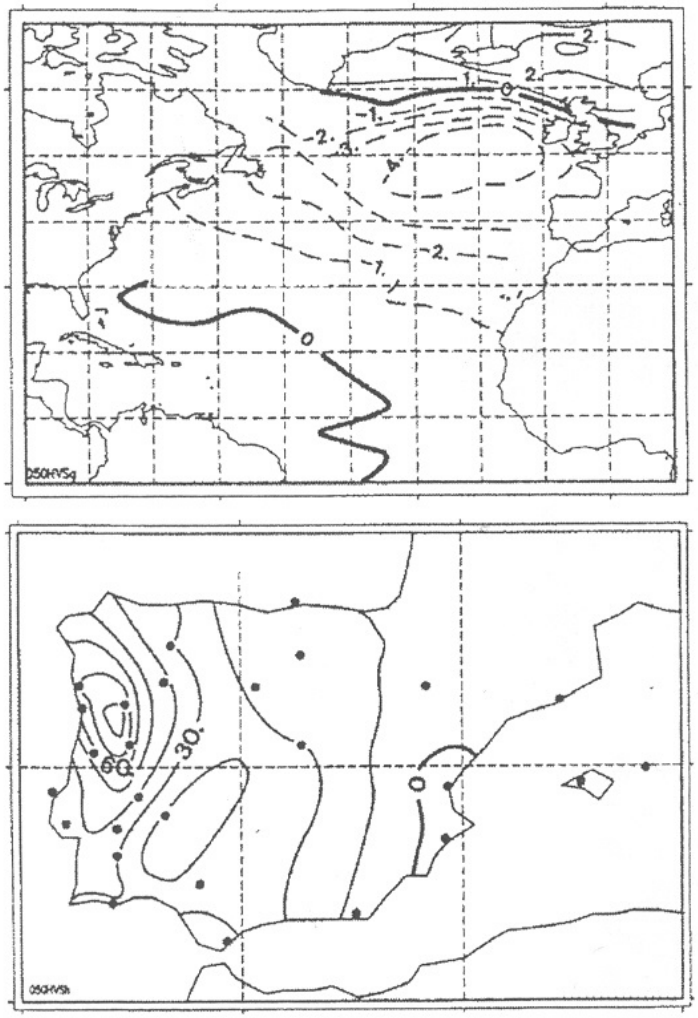
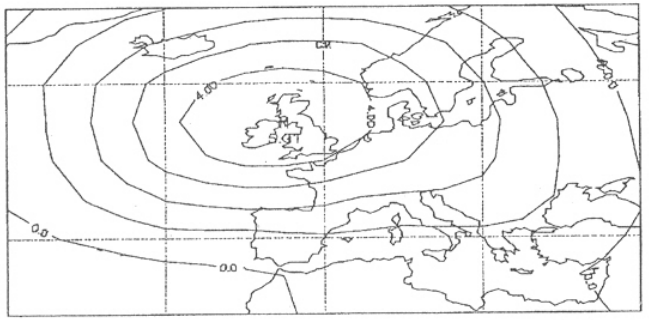


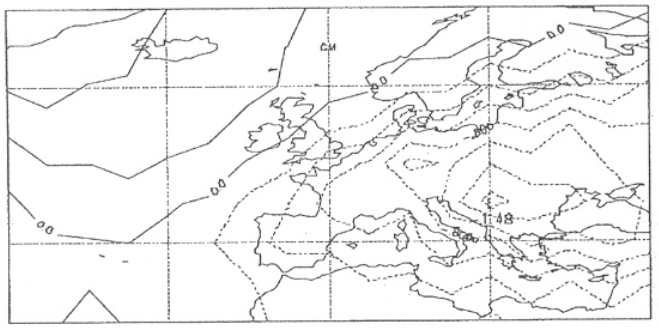
Figure 14.4: *First pair of canonical correlation patterns of the North Atlantic winter mean sea-level pressure* $\vec{\mathbf{Y}}$ *and a vector* $\vec{\mathbf{X}}$ *of seasonal means of precipitation at a number of Iberian locations [403].*

Example 2

Canonical Correlation Analysis of winter-mean sea level pressure and temperature in the North Atlantic and a few local climatic variables in Bern.

Gyalistras et al (1994) Clim. Res.





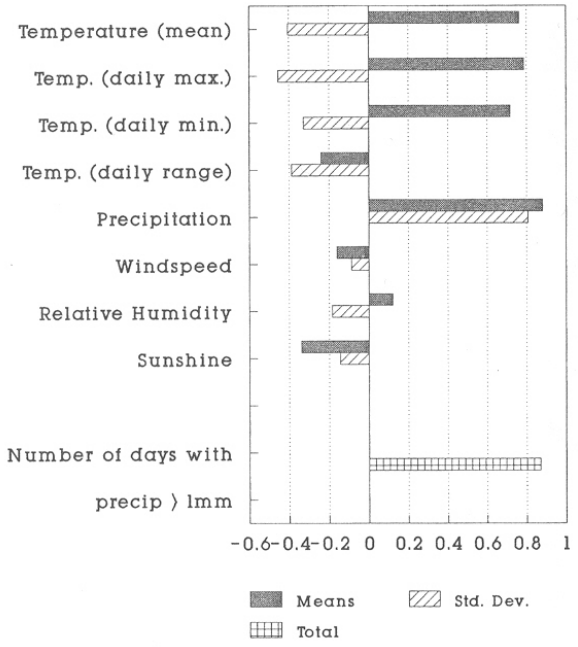
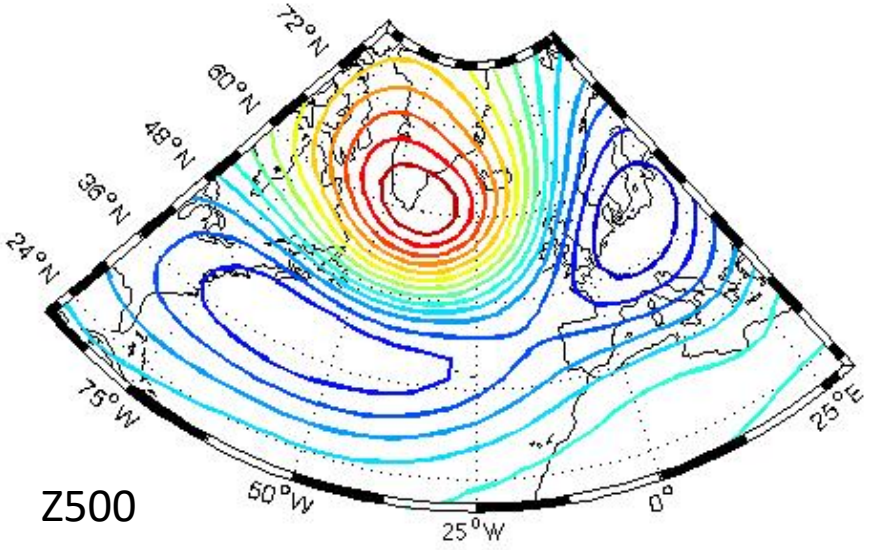


Figure 14.1: *First pair of canonical correlation patterns of* $\vec{\mathbf{Y}}$ *= (DJF mean SLP, DJF mean temperature) and a vector* $\vec{\mathbf{X}}$ *of DJF statistics of local weather elements at Bern (Switzerland).*
Top: The SLP part of the first canonical correlation pattern for $\vec{\mathbf{Y}}$.
Middle: The near-surface temperature part of the first canonical correlation pattern for $\vec{\mathbf{Y}}$.
Bottom: The canonical correlation pattern for the local variable $\vec{\mathbf{X}}$.
Note that the correlation between the corresponding pattern coefficients is negative.
From Gyalistras et al. [152].

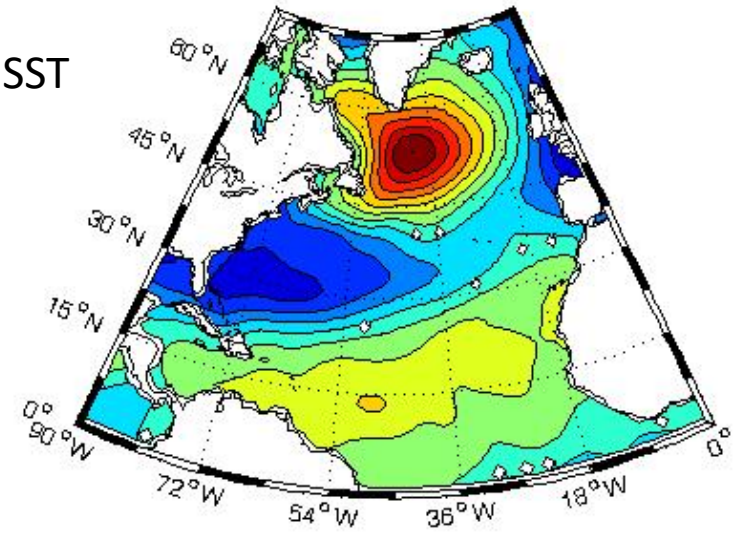
EXAMPLE 3

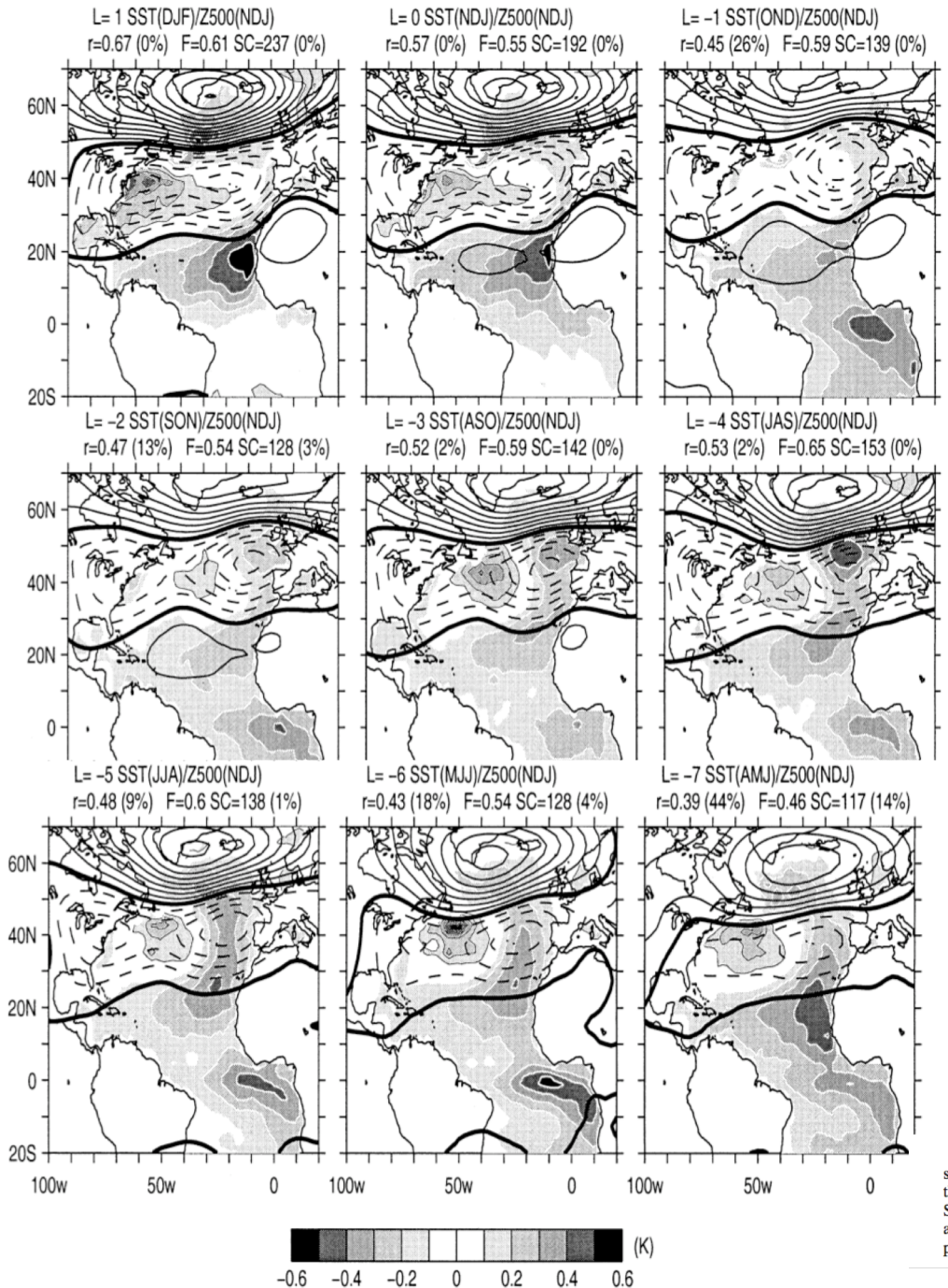
MCA analysis of the 500mb geopotential height monthly means in the Euro-Atlanctic region and the Sea Surface Temperatures of the Atlantic.



Z500

SST





Czaja and Frankignoul 2001, J. of Climate.

FIG. 2. Heterogeneous Z_{500} (thick contours, CI = 5 m with negative values dashed) and homogeneous SST (gray shading in K with white contours for positive values and black contours for negative values) covariance maps for the first MCA mode between Pan-Atlantic 500-mb height and SST anomalies in early winter (Z_{500} fixed in NDJ, SST lagged as indicated). The results are shown from lag -7 to $+1$. The correlation coefficient r between the SST and Z_{500} MCA time series, the SC fraction F, and the SC of the mode are given for each lag. The percentages in parentheses for r and SC give their estimated significance level.

Rotation of EOFs

The EOFs are a statistical construct, so they cannot *a priori* be linked to a given physical mechanism. They are the signature of the dynamics of a given physical system. The physical interpretation is done *a posteriori* by the user. Sometimes it is evident, sometimes not.

Rotation is a way to:
- alleviate the strong constraints of EOFs, namely orthogonality/uncorrelation of EOFs/PCs, domain dependence of EOF patterns
- obtain simple structures,
- be able to physically interpret the patterns.

Here one looks for a rotation matrix Q to construct the rotated EOFs U according to:

$$U = EQ$$

Where $E = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_m]$ is the matrix of the leading m EOFs.
The criterion for choosing the m x m rotation matrix Q is what constitutes the rotation algorithm, which is generally expressed by a minimization or maximization problem:

$$\max f(EQ)$$

With the constraint, for an orthogonal rotation, that Q is orthogonal:

$$QQ^T = Q^T Q = I$$

over all choices of rotation matrices Q. The functional f represents the rotation criterion.

An example, the VARIMAX criterion:

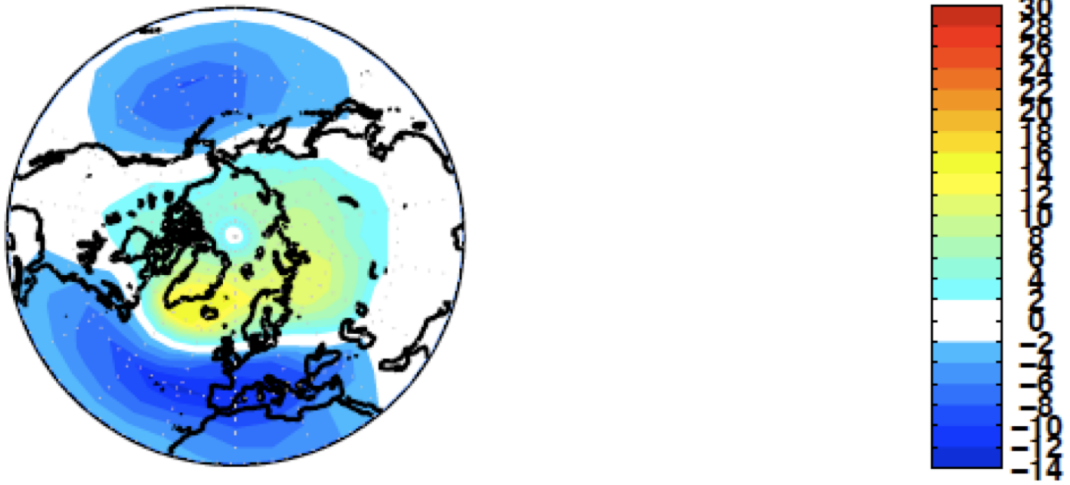
$$\max\left(\sum_{k=1}^{m}\left[p\sum_{j=1}^{p}u_{jk}^{4}-\left(\sum_{j=1}^{p}u_{jk}\right)^{2}\right]\right)$$

Where $U=[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_m]$.
This criterion tends to maximize the spatial variance of each mode, i.e. the values of the EOFs on the gridpoints tend to approach zero or ± 1.

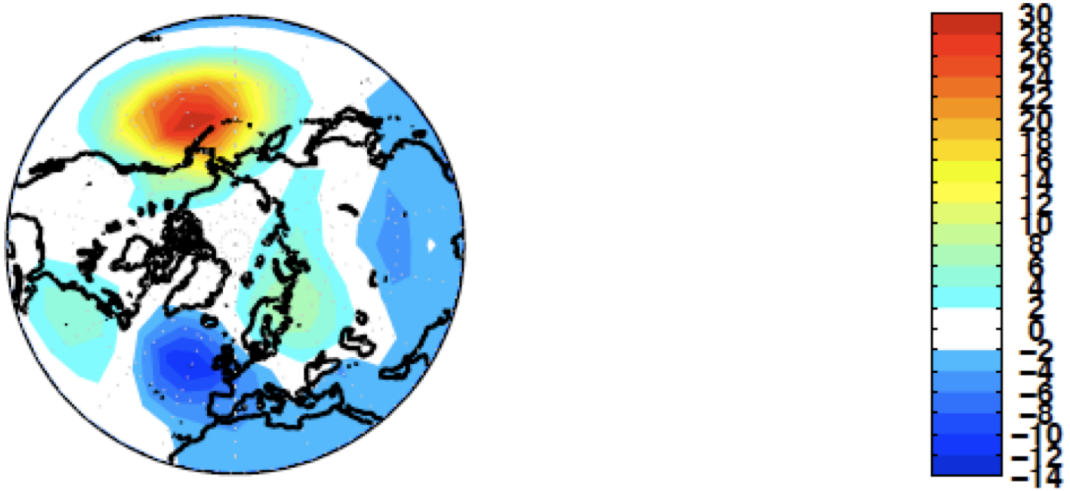
Exmaple of VARIMAX rotation.

EOFs of surface pressure for DJF in the Northern Hemisphere.

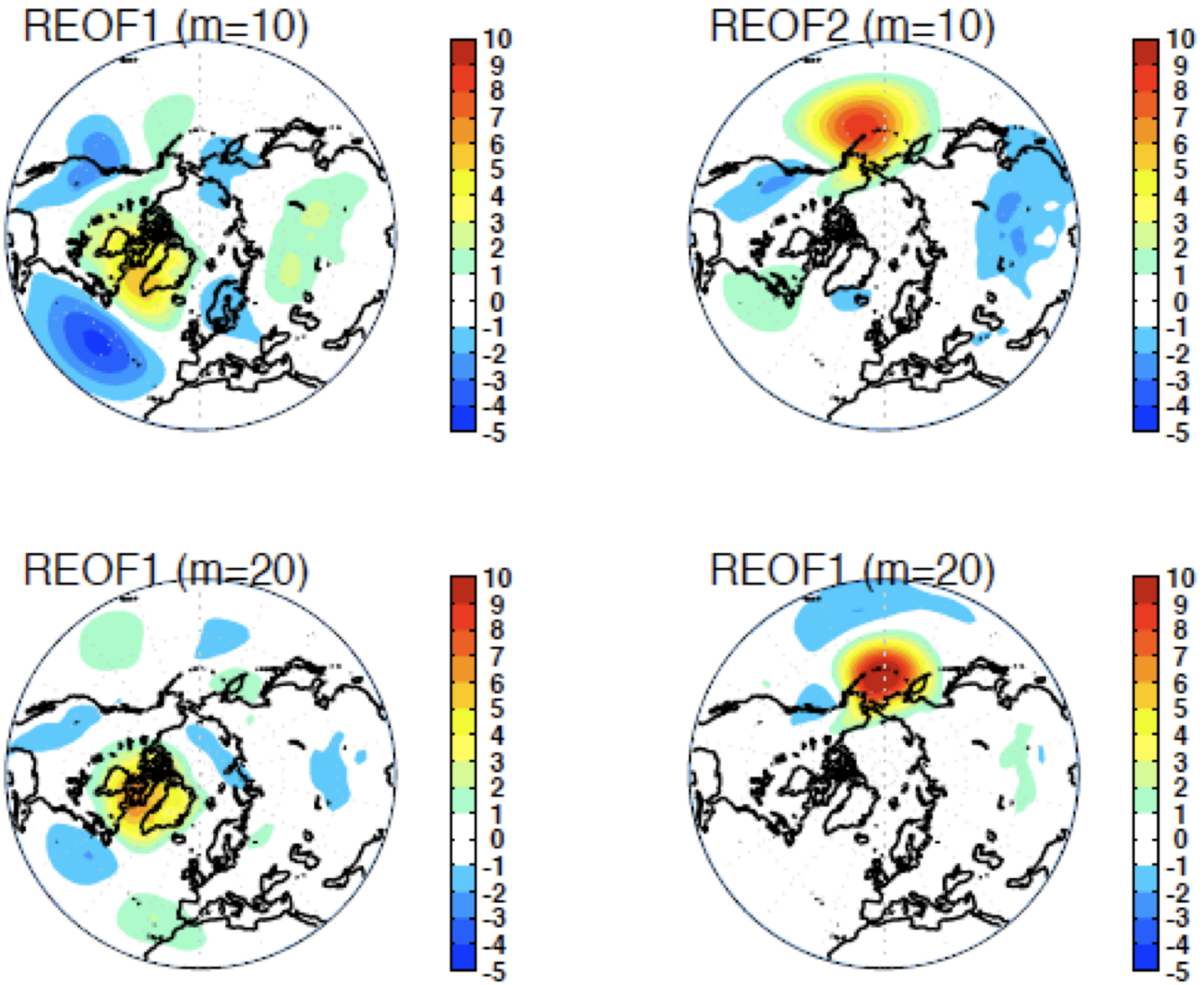
DJFSLP EOF1 (21%)



DJFSLP EOF2 (13%)



VARIMAX rotated EOFs



Credits: Abdel Hannachi (http://www.met.rdg.ac.uk/~han/Monitor/eofprimer.pdf)

PCA in the time domain: Singular Spectrum Analysis.

SSA is a method to extract periodicities from very noisy timeseries, or to detect intermittent spells of periodicities.

Suppose you have a scalar timeseries X, SSA consists in performing Principal Component Analysis on the vector timeseries Y defined like this.

$$\mathbf{y}_i = \left(X_{i-k}, X_{i-k+1}, \ldots, X_i, \ldots, X_{i+k-1}, X_{i+k}\right)^T$$

Y is called the "delay" timeseries, with window 2k+1.

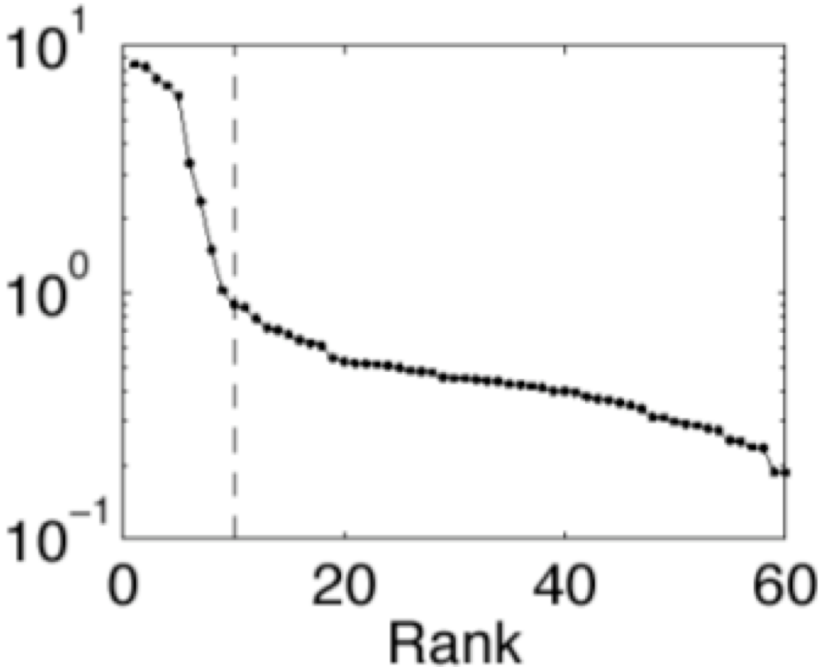
The covariance matrix $\frac{1}{N} YY^T$ (the Toeplitz matrix) is hence diagonalized to obtain time-EOFs.
The time-EOFs can be seen as typical chunks of time evolutions of length 2k+1.

The delay timeseries Y can be reconstructed by expansion on a few time-EOFs in the usual way:

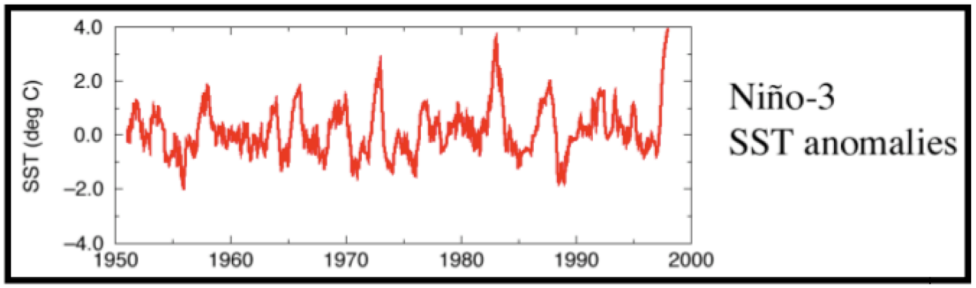
$\mathbf{y}_i = \sum_{n=1}^{T} a_{n,i} \mathbf{e}_n$ where the $a's$ are given by a scalar product in the delay coordinates:

$$a_{n,i} = \mathbf{y}_n{}^T \mathbf{e}_i$$

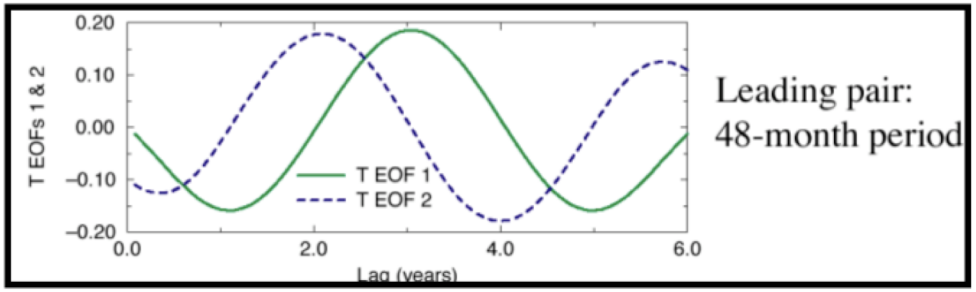
SSA analysis of SOI timeseries



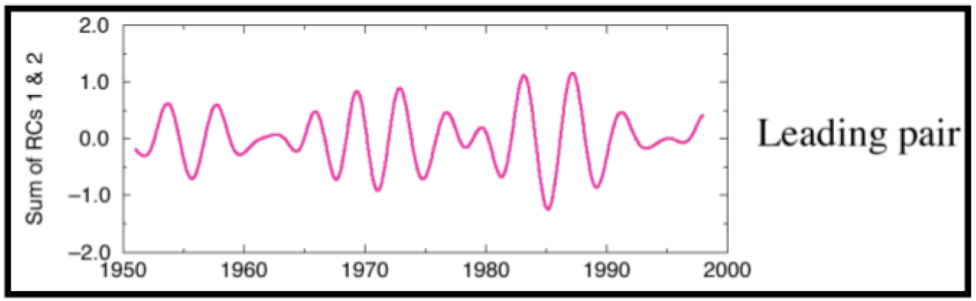
Time series

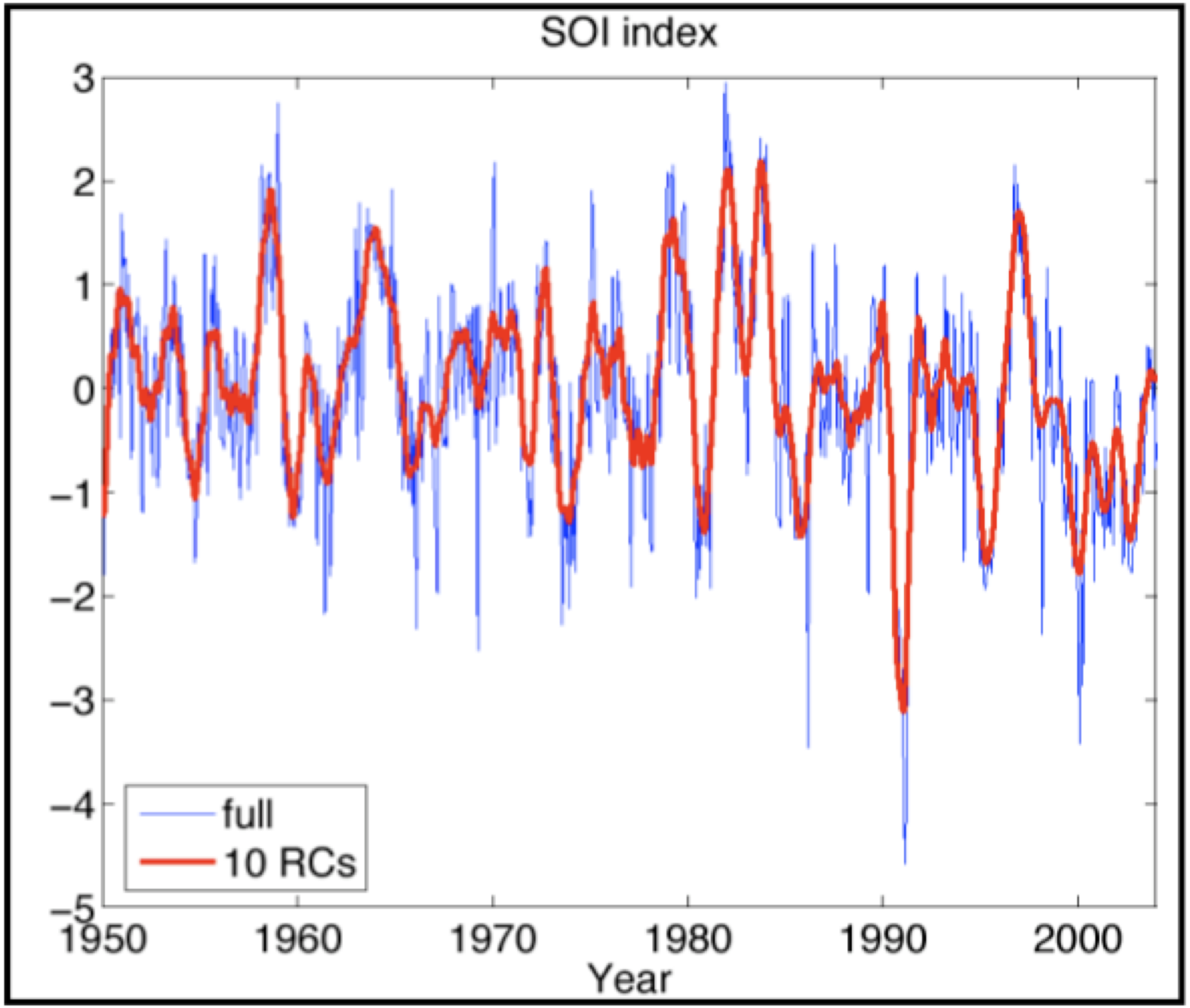


T-EOFs



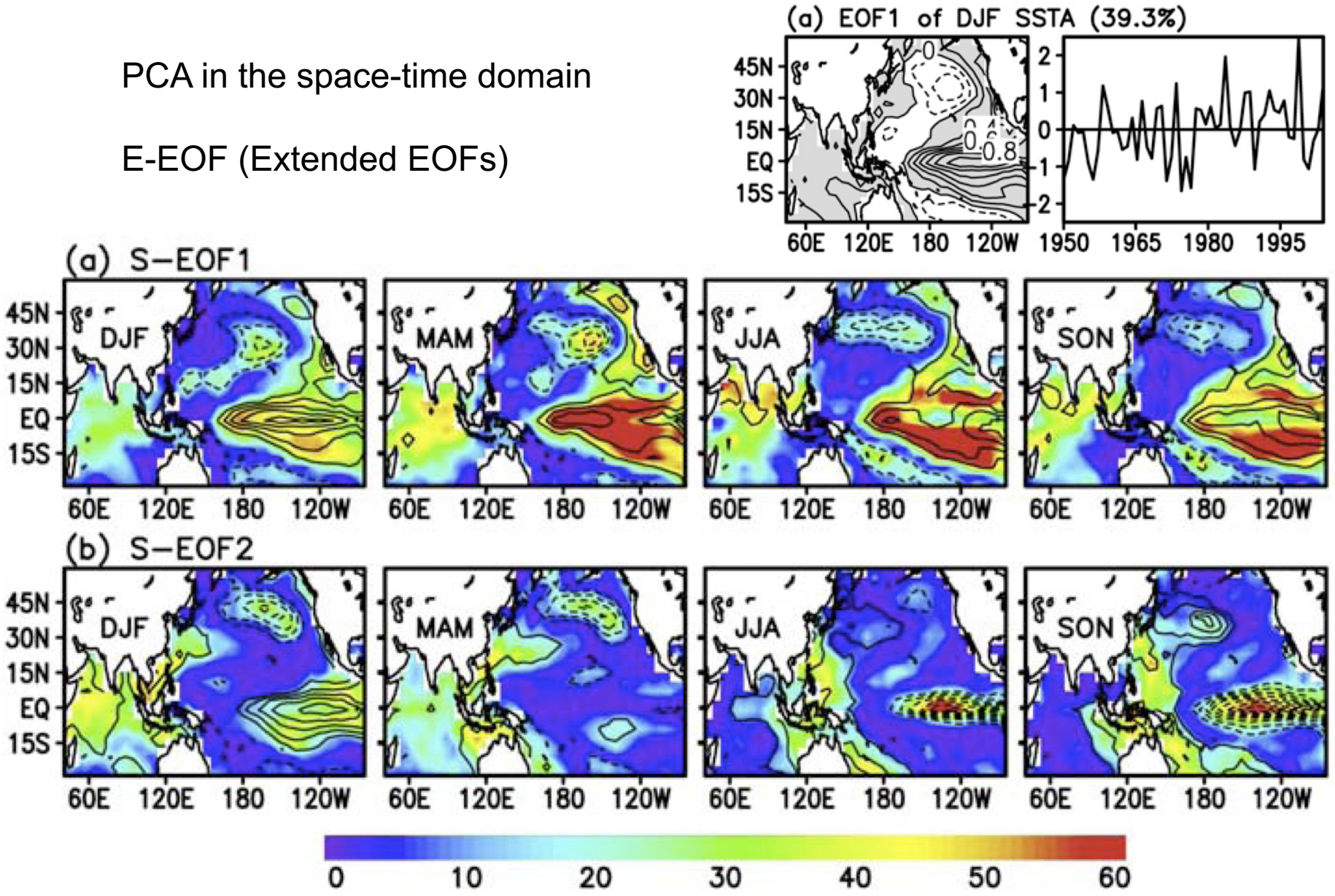
RCs




SOI index
3
2
1
0
−1
−2
−3
−4
−5
1950
1960
1970
1980
1990
2000
Year
full
10 RCs

PCA in the space-time domain

E-EOF (Extended EOFs)



Seasonal EOF modes of SST (4 fields per year)
Credit: Wang and An 2005 GRL.