Methods for finding coupled patterns in two data sets

Martin Widmann

VALUE training school, ICTP Trieste, 4. November 2014

Content

- patterns and time expansion coefficients in Principal Component Analysis

- Maximum Covariance Analysis (MCA) or Singular Value Decomposition (SVD)

- Canonical Correlation Analysis (CCA)

Courtesy for some slides

Jin-Yi Yu
Associate Professor, Earth System Science
School of Physical Sciences
University of California, Irvine

References

Books

Peixoto and Oort: Physics of Climate, appendix on EOFs.

Wilks: Statistical methods in the atmospheric sciences: an introduction

von Storch and Zwiers: Statistical Analysis in Climate Research

See also Dennis Hartmann‘s lecture notes (ATMS552, Objective Analysis)
http://www.atmos.washington.edu/~dennis/

Papers

Bretherton et al., 1992: An intercomparison of methods for finding coupled patterns in climate data. J. Climate, 5, 541-560.

DelSole and Yang, 2011: Field significance of regression patterns. J. Climate, 24, 5094-5107.

Hannachi et al. 2007: Empirical orthogonal functions and related techniques in atmosperic science: A review. Int. J. Climatol., 27, 1119-1152.

Tippett et al., 2008: Regression-based methods for finding coupled patterns. J. Climate, 21, 4384-4398.

Widmann 2005: One-dimensional CCA and SVD, and their relation to regression maps. J. Climate, 18, 2785-2792.

Principal Component Analysis (PCA)

or

Empirical Orthogonal Function (EOF) analysis

Nomenclature

Principal Component Analysis is also known as EOF analysis. Some authors use both names to distinguish whether the patterns have length 1 or length of square root of eigenvalue, but this is not generally followed.

The EOFs are sometimes called ‘Principal component loadings’.

The PCs are sometimes called ‘Principal Component scores’.

What does Principal Component Analysis do?

Reduction of datasets: attempts to find a relatively small number of variables that include as much as possible information of the original dataset.

Objective analysis of the structure of a dataset with respect to relationships between different variables.

What Does EOF Analysis do?

❑ In brief, EOF analysis uses a set of orthogonal functions (EOFs) to represent a time series in the following way:

$$Z(x,y,t) = \sum_{i=1}^{n} PC_i(t) \cdot EOF_i(x,y)$$

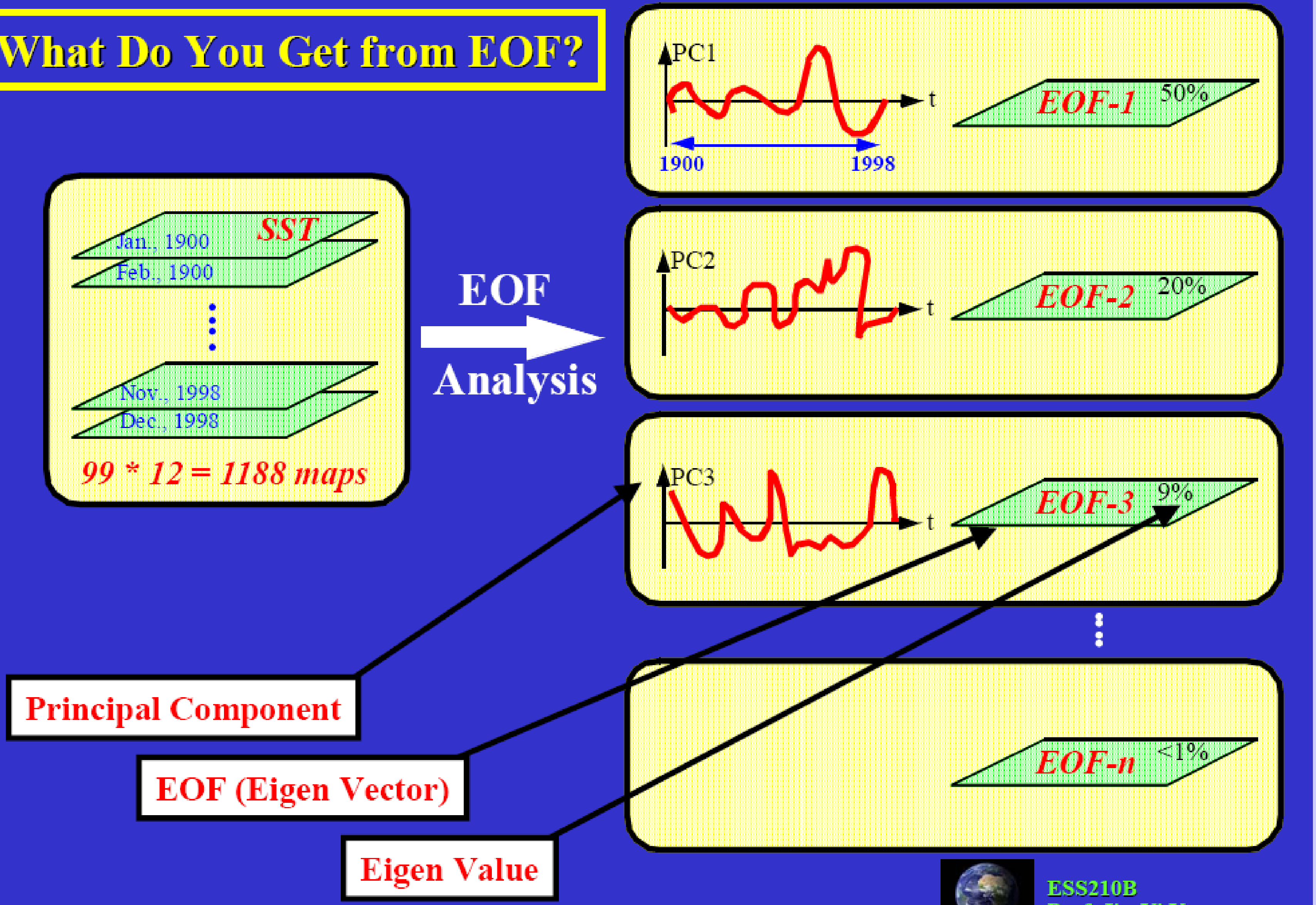
This is S-mode PCA

❑ Z(x,y,t) is the original time series as a function of time (t) and space (x, y).

EOF(x, y) show the spatial structures (x, y) of the major factors that can account for the temporal variations of Z.

PC(t) is the principal component that tells you how the amplitude of each EOF varies with time.

What Do You Get from EOF?



*99 * 12 = 1188 maps*

EOF Analysis

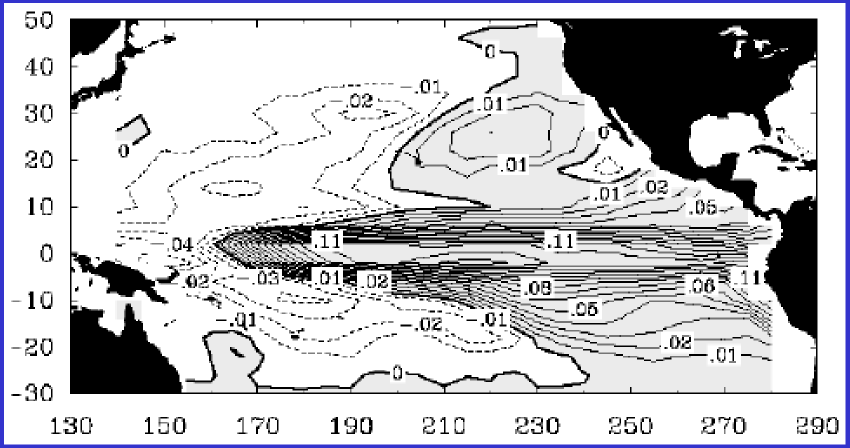
Principal Component

EOF (Eigen Vector)

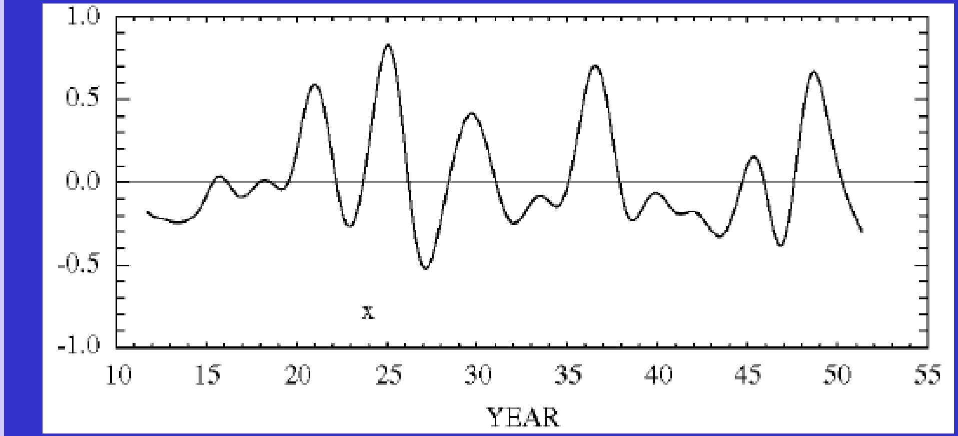
Eigen Value

An Example

Leading EOF Mode



Principal Component

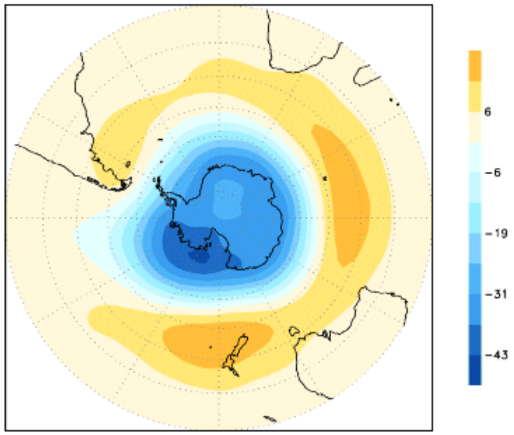


❑ We apply EOF analysis to a 50-year long time series of Pacific SST variation from a model simulation.

❑ The leading EOF mode shows a ENSO SST pattern. The EOF analysis tells us that ENSO is the dominant process that produce SST variations in this 50-year long model simulation.

❑The principal component tells us which year has a El Nino or La Nina, and how strong they are.

Southern Annular Mode Index (aka Antarctic Oscillation Index)

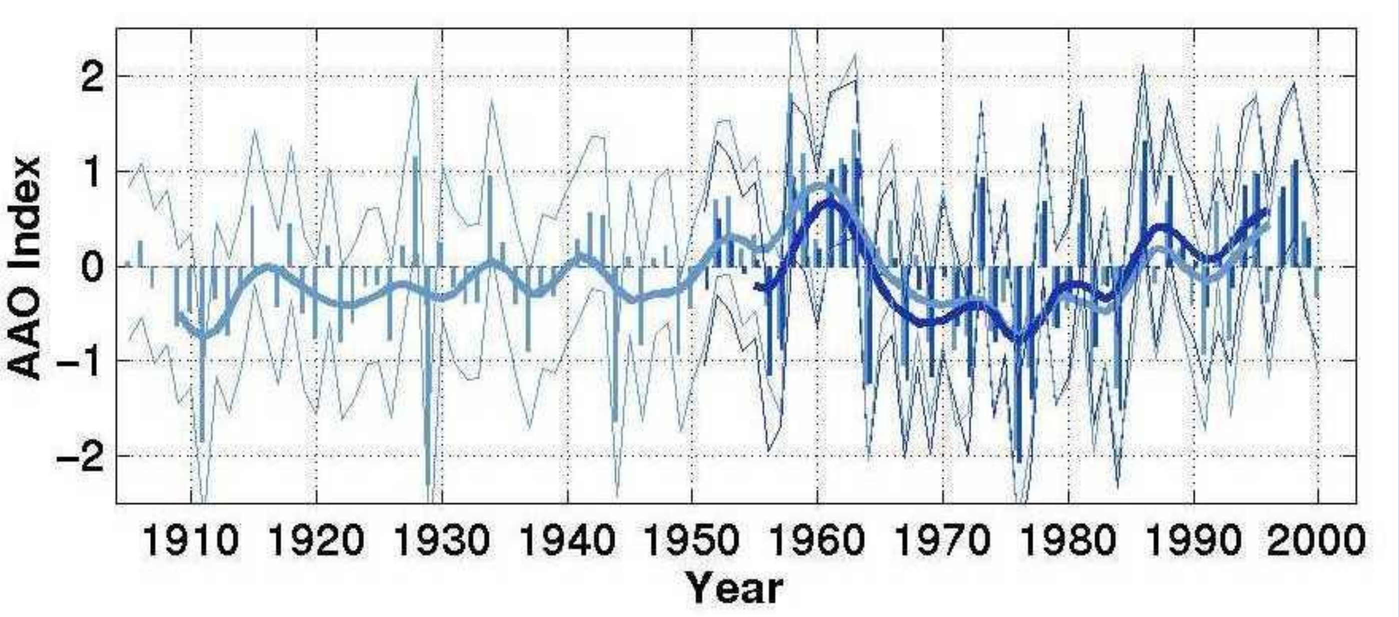


The Southern Hemisphere annular mode

The surface signature of the Southern Hemisphere annular mode. The SAM is defined here as the leading EOF of SH monthly-mean 850-hPa height anomalies. Units are m/std of the principal component time series.

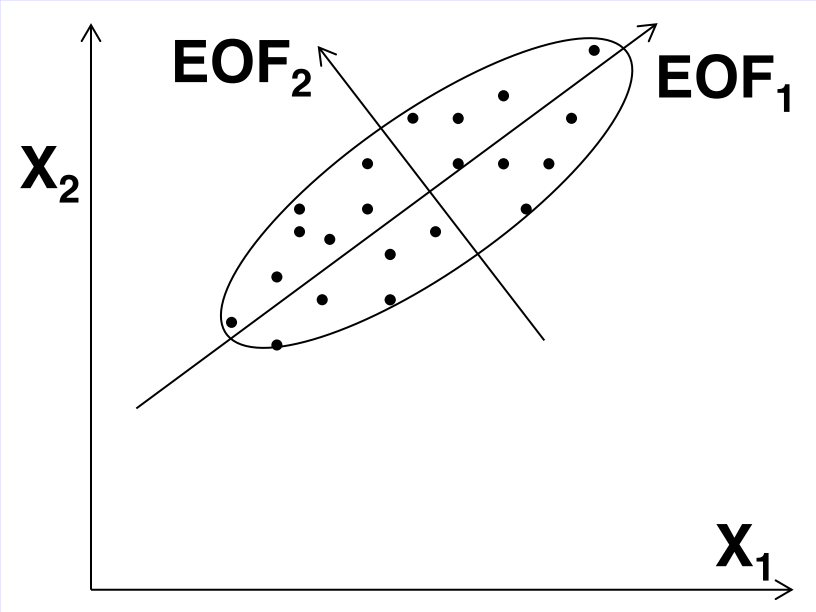
January/February mean SAM (AAO) Index

Reconstructions from two different sets of long pressure measurements



(from Jones and Widmann, *Nature*, 2004)

Principal Component Analysis, geometrical interpretation



- EOFs show the direction of axes of a fitted ellipsoid

- EOF indices are ordered such that the variability of the data along the corresponding axis decreases

- the EOFs are (unit) vectors, and thus can be expressed by their projections onto the original axes (the EOF loadings)

- the PCs are the projections of the data onto the EOFs

How to find PCs and EOFs?

The fitting outlined on previous slide is equivalent to

- choose EOF1 such that PC1 has maximum variance

- choose EOF2 orthogonal to EOF1 and such that PC2 has maximum variance

with PCs defined as the projection of the data onto the EOFs.

For higher dimensions the variances of the higher PCs are also maximised subject to the condition that the EOFs are mutually orthogonal.

This implies that an approximate expansion of the data using only n leading PCs and EOFs is the best approximation to the data
(it maximises the variance and minimises the error).

It can be shown that the EOFs are the eigenvectors of the covariance matrix.

It follows that the PCs are mutually uncorrelated.

The calculations have the simplest from (see later) when the EOFs have length one.

Eigenvectors of a Symmetric Matrix

❑ Any symmetric matrix **R** can be decomposed in the following way through a diagonalization, or eigenanalysis:

$$\mathbf{R}\mathbf{e}_i = \lambda_i \mathbf{e}_i$$

$$\mathbf{R}\mathbf{E} = \mathbf{E}\mathbf{L}$$

$$\mathbf{E}^T\mathbf{R}\mathbf{E} = \mathbf{L}$$

eigenvectors of symmetric matrices are orthogonal

❑ Where **E** is the matrix with the eigenvectors e_i as its columns, and **L** is the matrix with the eigenvalues λ_i, along its diagonal and zeros elsewhere.

❑ The set of eigenvectors, $\boldsymbol{e}_i$, and associated eigenvalues, λ_i, represent a coordinate transformation into a coordinate space where the matrix **R** becomes diagonal.

Note: the eigenvalues are sometimes denoted λ^2, because this avoids using roots in some equations (e.g. Hannachi et al. 2007).

Covariance matrix

The components are the covariances between the ith and the jth variable.

$$C_{xx} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ c_{n1} & \cdots & \cdots & c_{nn} \end{pmatrix}$$

with

$$c_{ij} = \frac{1}{T-1} \sum_{k=1}^{T} \left(x_i(t_k) - \bar{x}_i\right)\left(x_j(t_k) - \bar{x}_j\right)$$

Example: If there are 200 SST grid cells and 30 years of monthly data
n = 200 and T = 360

PCs as projections

If the k^th^ EOF is given by a vector with length one

$$EOF_k = \begin{pmatrix} eof_{1k} \\ eof_{2k} \\ \vdots \\ eof_{nk} \end{pmatrix}$$

$$\left\| EOF_k \right\|^2 = \sum_{i=1}^{n} eof_{ik}^2 = EOF_k^T \, EOF_k = 1$$

we get the PC time series through the projection

$$PC_k(t_j) = \sum_{i=1}^{n} x_i(t_j) \, eof_{ik}$$

For brevity we have used here the assumption that x are anomalies; this assumption will be used in all the following slides.

PCs as projections

If we arrange the data in a matrix containing n variables and T time steps

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{T1} & \cdots & \cdots & x_{Tn} \end{pmatrix}$$

the PCs can be expressed through a matrix multiplication

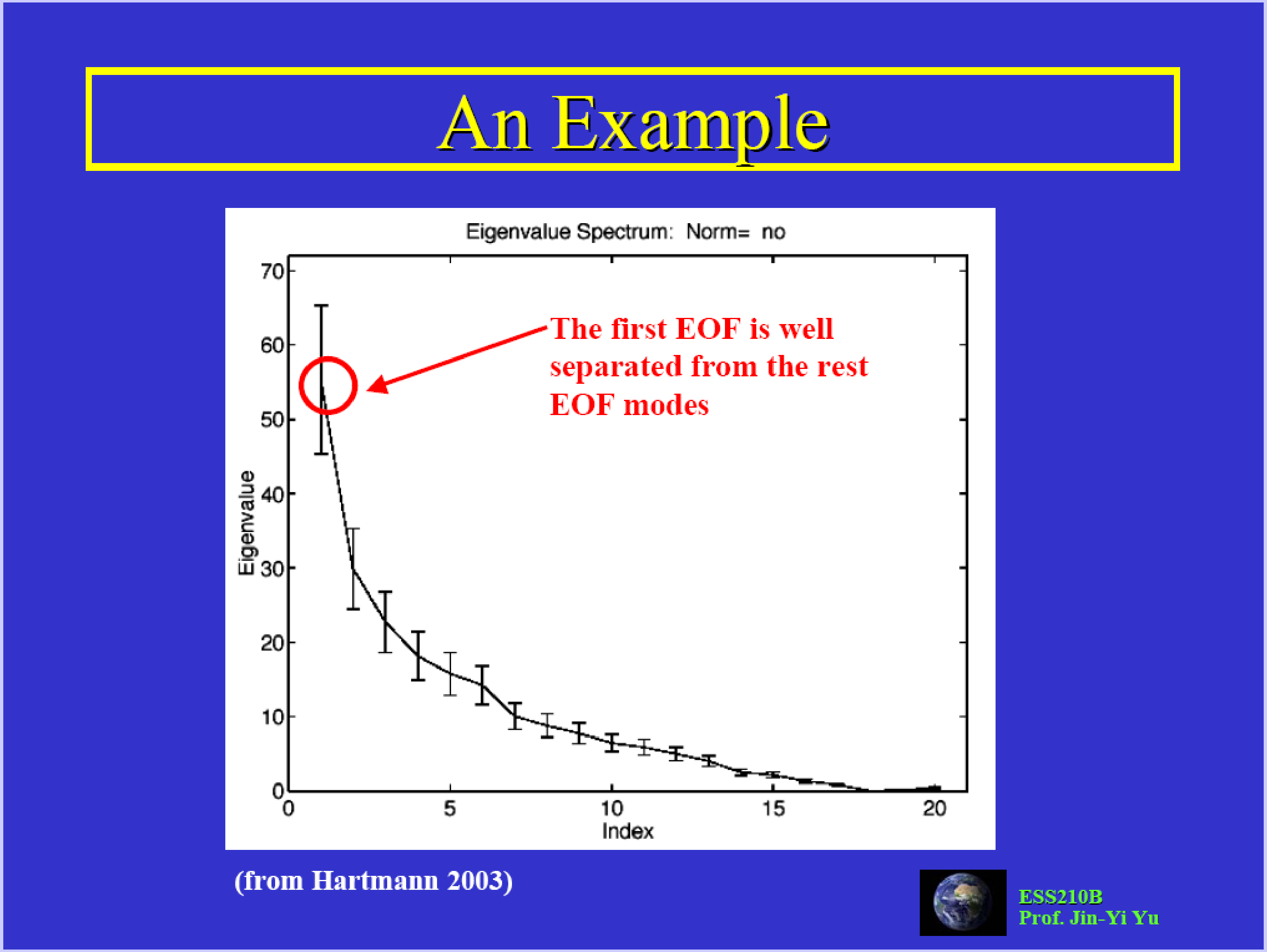
$$PC_k = X\ EOF_k$$

with

$$PC_k(t_j) = PC_{jk} = \sum_{i=1}^{n} x_{ji}\ eof_{ik}$$

Typical eigenvalue spectrum

The eigenvalues are the square roots of the variances of the PCs



An Example

The first EOF is well separated from the rest EOF modes

(from Hartmann 2003)

ESS210B
Prof. Jin-Yi Yu

Maximum Covariance Analysis (MCA)

and

Singular Value Decomposition (SVD)

Nomenclature

The statistical method should be called Maximum Covariance Analysis, and Singular Value Decomposition should be reserved for the algebraic operation. However, many older papers use SVD as a name for the statistical method.

What does Maximum Covariance Analysis do?

Objective analysis of the relationships between two sets of variables.

Finds patterns such that time expansion coefficients (which are given by projection onto the patterns) have maximum covariance and the patterns are orthogonal to each other.

These coupled patterns are often used to estimate one dataset from the other.

Patterns and time expansion coefficients in MCA

For data sets X (n variables) and Y (m variables) the patterns are denoted by

$$u_k = \begin{pmatrix} u_{1k} \\ u_{2k} \\ \vdots \\ u_{nk} \end{pmatrix}$$

and

$$v_k = \begin{pmatrix} v_{1k} \\ v_{2k} \\ \vdots \\ v_{mk} \end{pmatrix}$$

The time expansion coefficients (TECs) are given through projections

$$a_k(t_j) = \sum_{i=1}^{n} x_i(t_j)\, u_{ik}$$

$$b_k(t_j) = \sum_{i=1}^{m} y_i(t_j)\, v_{ik}$$

The first pair of patterns u$_1$, v$_1$ are chosen such that cov(a$_1$,b$_1$) is maximised (with the constraint that the patterns have length 1, which is u^T u = 1, v^T v = 1) .

The subsequent pairs of patterns are chosen such that they maximise the covariance of the time expansion coefficients subject to the constraint that they are orthogonal to the previous patterns. Note: TECs within the fields are correlated, TECs between fields for different modes are uncorrelated.

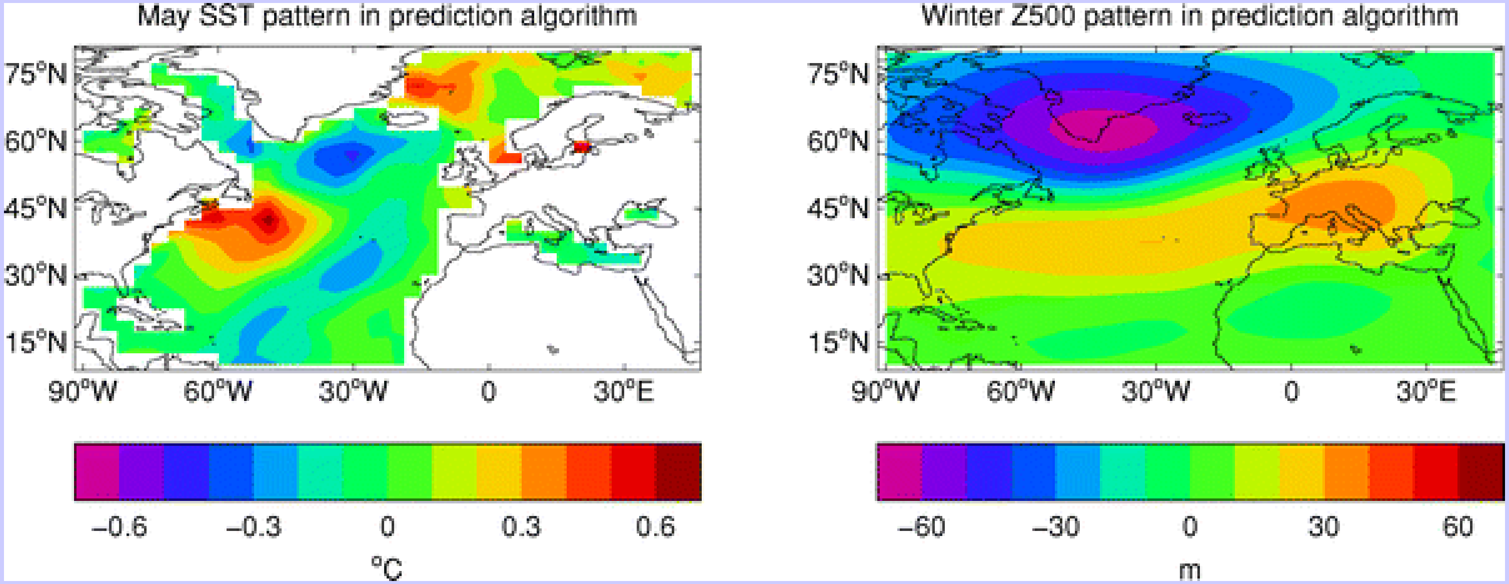
Approximate expansions

The approximate expansions of X and Y using the leading patterns and time expansion coefficients are given by

$$x_i(t_j) \approx \sum_{k=1}^{\tilde{n}} a_k(t_j)\, u_{ik}$$

$$y_i(t_j) \approx \sum_{k=1}^{\tilde{m}} b_k(t_j)\, v_{ik}$$

Coupled patterns of sea surface temperature and mid-tropospheric circulation used in the Met-Office statistical winter NAO forecast

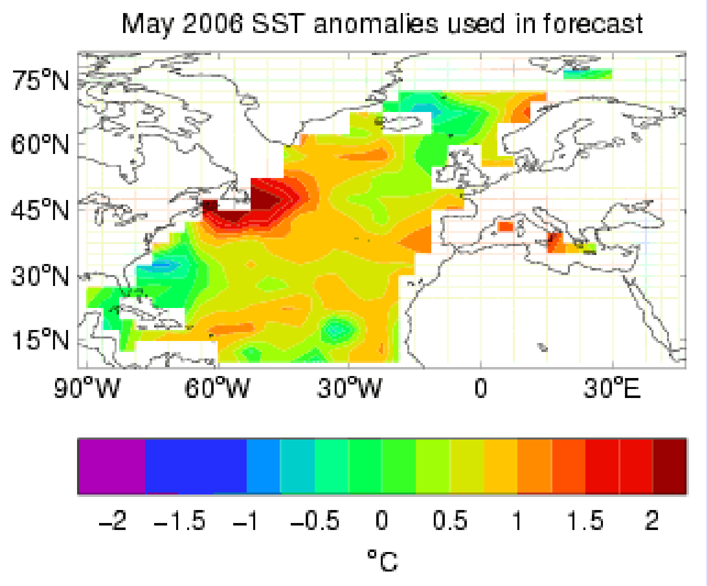


May SST pattern in prediction algorithm

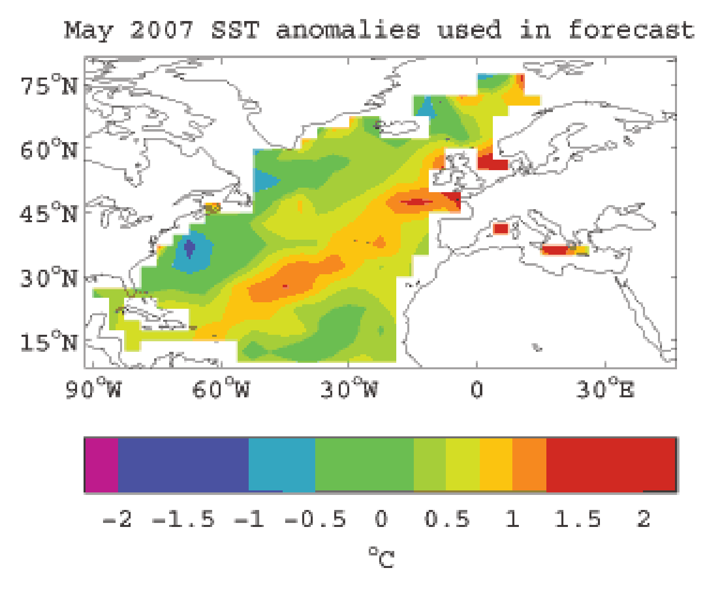
Winter Z500 pattern in prediction algorithm

coupled patterns

(MCA)



May 2006 SST anomalies used in forecast

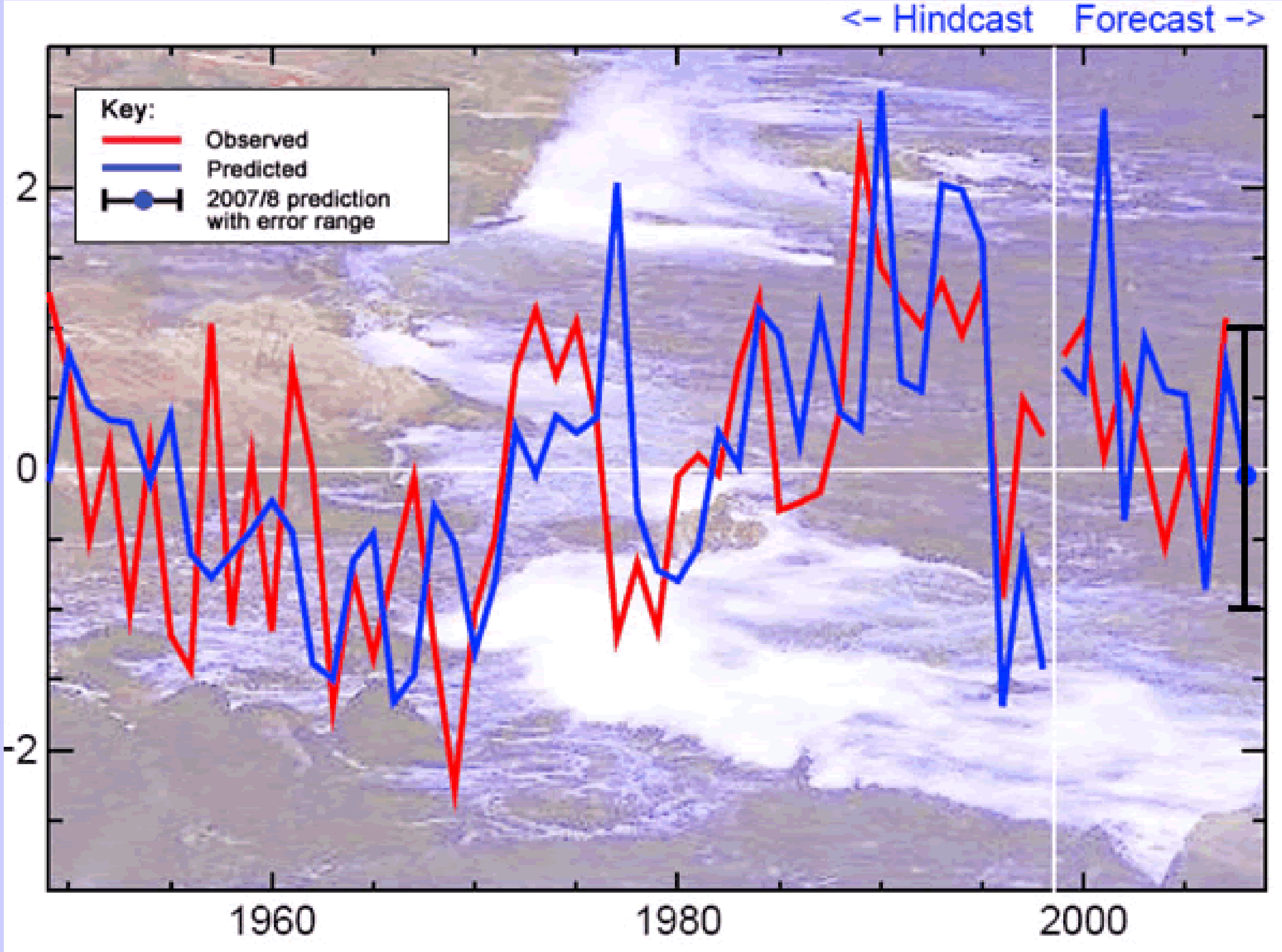


May 2007 SST anomalies used in forecast

sea surface temperature anomalies in May 2006 and May 2007

(http://www.met-office.gov.uk/research/seasonal/regional/nao/index.html)

NAO Index: Met-Office statistical prediction and observations



Skill

Correlation = 0.45

Correct sign 66%

(http://www.met-office.gov.uk/research/seasonal/regional/nao/index.html)

Details of method: Rodwell and Folland, 2002: Quarterly J. Royal Met. Soc., 128, 1413-1443.

Link SST and NAO: Rodwell et al., *Nature*, 1999, 398, 320-323.

Perfect Prog downscaling - estimating precip from pressure

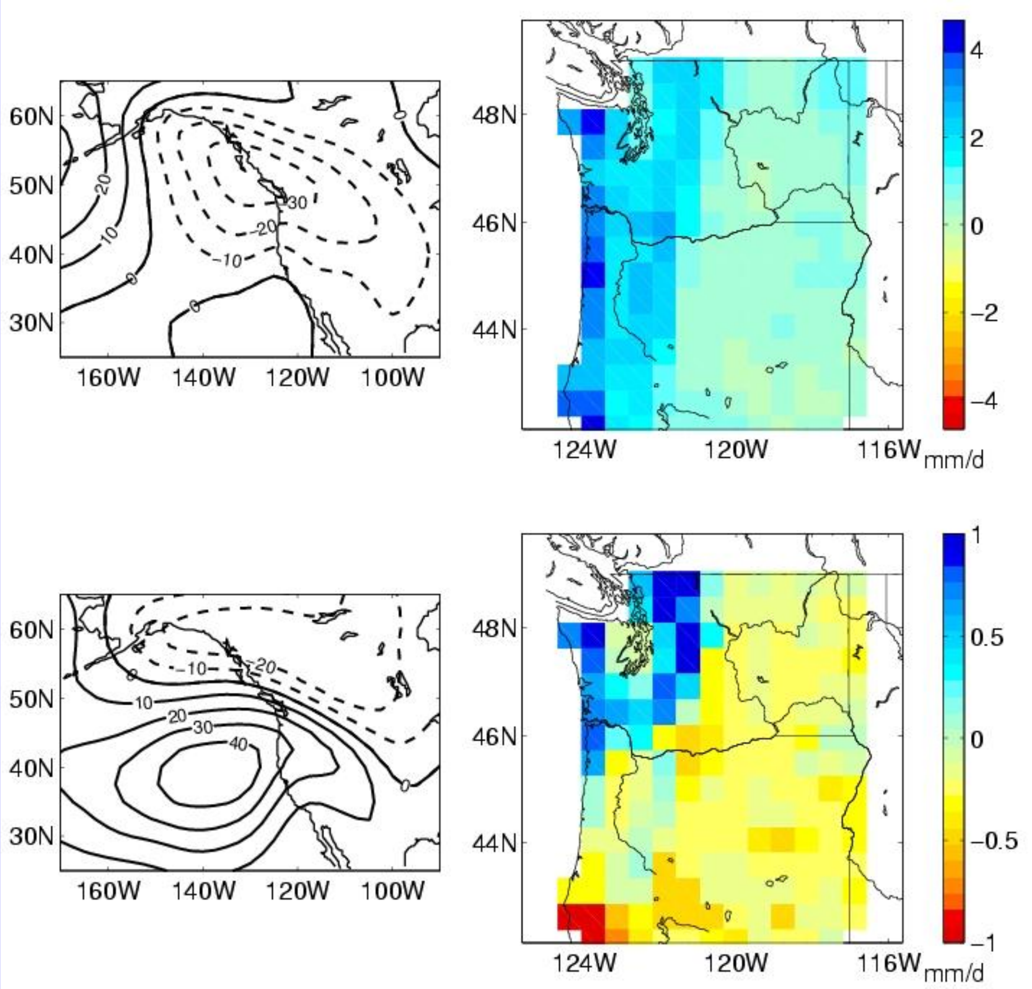
Coupled anomaly patterns (MCA) between DJF 1000 hPa geopotential height (NCEP) and daily preciptation

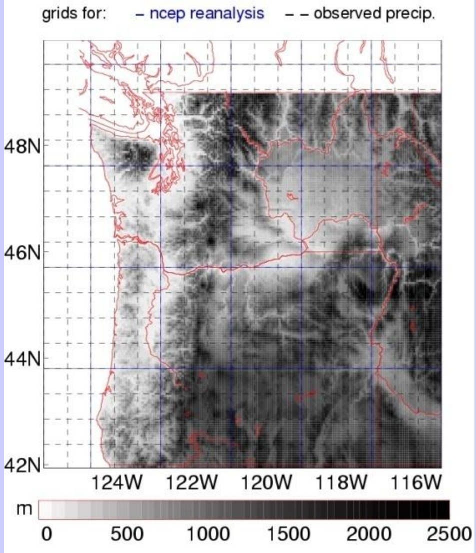
geopot. height (Z1000) precipitation

topography

pair 1

pair 2



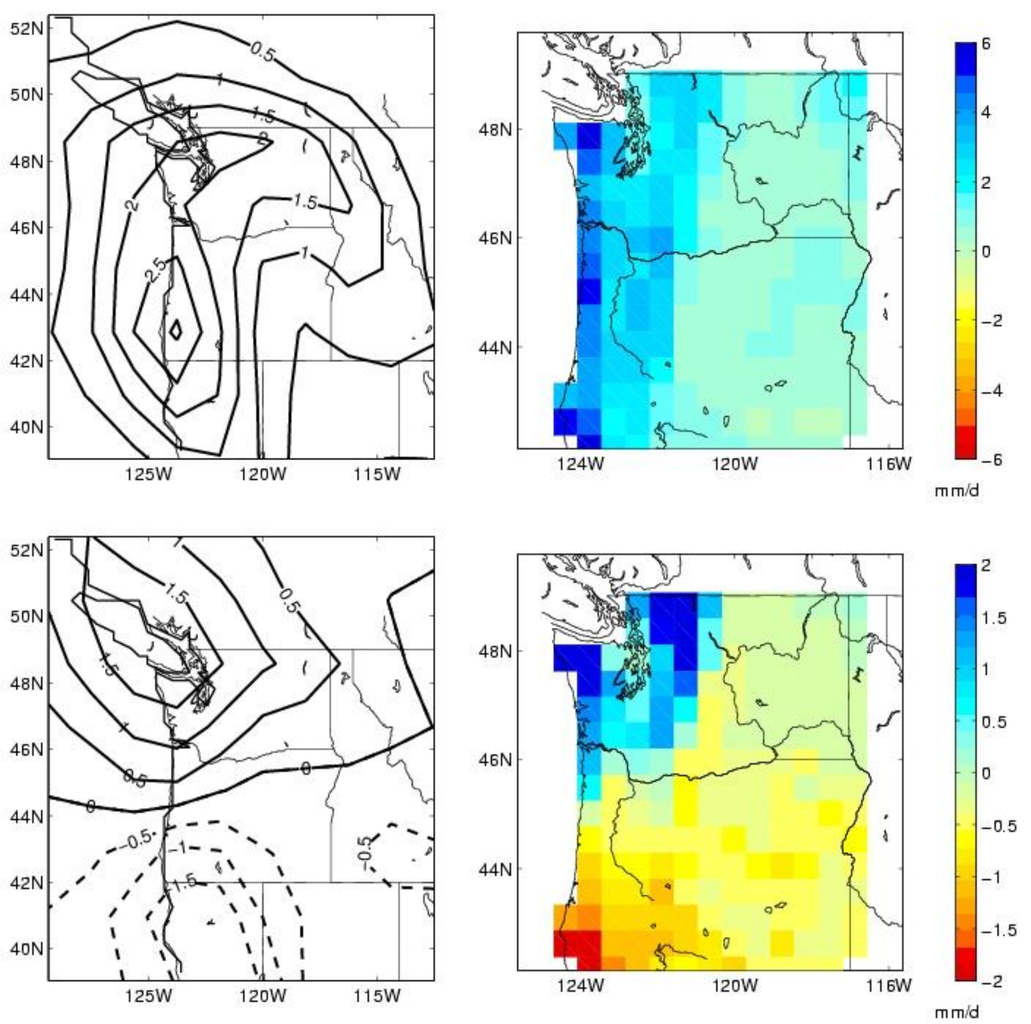


(Widmann and Bretherton, J. Climate 2000; Widmann et al., J. Climate, 2003)

Model Output Statistics - estimating true precipitation from simulated precipitation

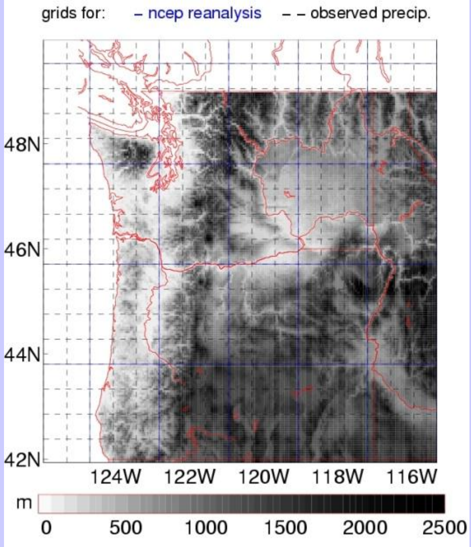
simulated precipitation (NCEP reanalysis)

observations



Coupled anomaly patterns (MCA) between DJF daily simulated (NCEP) and observed preciptation

topography



Singular Value Decomposition

The singular value decomposition of a matrix A is a generalisation of the eigenvalue problem to non-quadratic matrices and is given by

$A = U\,S\,V^T$ with U and V orthogonal matrices. If $n < m$ this is in components

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & \ddots & & & \vdots \\ \vdots & & & \ddots & \vdots \\ a_{n1} & & & \cdots & a_{nm} \end{pmatrix}$$

left singular vectors (columns of matrix)

singular values

right singular vectors (rows of matrix)

$$= \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ u_{n1} & \cdots & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} s_{11} & 0 & \cdots & & 0 & \cdots & 0 \\ 0 & s_{22} & & & & & \vdots \\ \vdots & & & & & & \vdots \\ 0 & \cdots & & s_{nn} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} & \cdots & & v_{m1} \\ v_{12} & \ddots & & & \\ \vdots & & & & \\ \vdots & & & \ddots & \vdots \\ v_{1m} & \cdots & & \cdots & v_{mm} \end{pmatrix}$$

(analogously for $n > m$, with zeros attached as rows)

Cross-covariance matrix and MCA

The components are the covariances between the i^th variable in the dataset X and the j^th variable in the dataset Y.

$$C_{xy} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ c_{n1} & \cdots & \cdots & c_{nm} \end{pmatrix}$$

Note this is in general a non-quadratic matrix

with

$$c_{ij} = \frac{1}{T-1} \sum_{k=1}^{T} \left(x_i(t_k) - \bar{x}_i\right)\left(y_j(t_k) - \bar{y}_j\right)$$

It can be shown that the MCA patterns are the left and right singular vectors of a SVD of the cross-covariance matrix.

Canonical Correlation Analysis (CCA)

What does Canonical Correlation Analysis do?

Same purpose as MCA: objective analysis of the structure of the relationships between two sets of variables.

But the selection criterion is different:

Finds projection vectors such that time expansion coefficients are uncorrelated within one dataset and have maximum correlation with the time expansion coefficient of the same index (mode) in the other dataset. TECs between the two fields for different indices are uncorrelated.

The patterns are obtained by minimising the error in an approximate expansion and are not orthogonal and not identical to the projection vectors.

The coupled patterns are often used to estimate one dataset from the other.

Distinction between projection vectors and patterns

Because the projection vectors used for calculating the time expansion coefficients from the data and the patterns used in the expansion are not identical (in contrast to PCA and MCA), we need to distinguish between them. We use u, v for the projection vectors, and p, q for the patterns. Note that the projection vectors are called weights in some papers, because they are the weights used to calculate the time expansion coefficients from the data. They are also sometimes called adjoint patterns.

	dataset X	dataset Y
data expansions using patterns p_k, q_k	$x_i(t_j) \approx \sum_{k=1}^{\tilde{n}} a_k(t_j)\; p_{ik}$	$y_i(t_j) \approx \sum_{k=1}^{\tilde{m}} b_k(t_j)\; q_{ik}$
time expansion coeffs. using projection (or weight) vectors u_k, v_k	$a_k(t_j) = \sum_{i=1}^{n} x_i(t_j)\; u_{ik}$	$b_k(t_j) = \sum_{i=1}^{m} y_i(t_j)\; v_{ik}$

Solution to the CCA problem

It can be shown that the projection vectors for the X dataset are given by the following eigenvector problem

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{xy}^{T} u_k = \lambda_k u_k$$

The projection vectors for Y are then given by

$$v_k = b\, C_{yy}^{-1} C_{xy}^{T} u_k$$

and the patterns by

$$p_k = C_{xx} u_k$$

$$q_k = C_{yy} v_k$$

CCA usually needs PCA prefiltering otherwise the inversion of the matrices becomes unstable:
Too many predictors lead to overfitting.

Example for CCA patterns between SLP and SST (Zorita et al. J. Climate 1992)

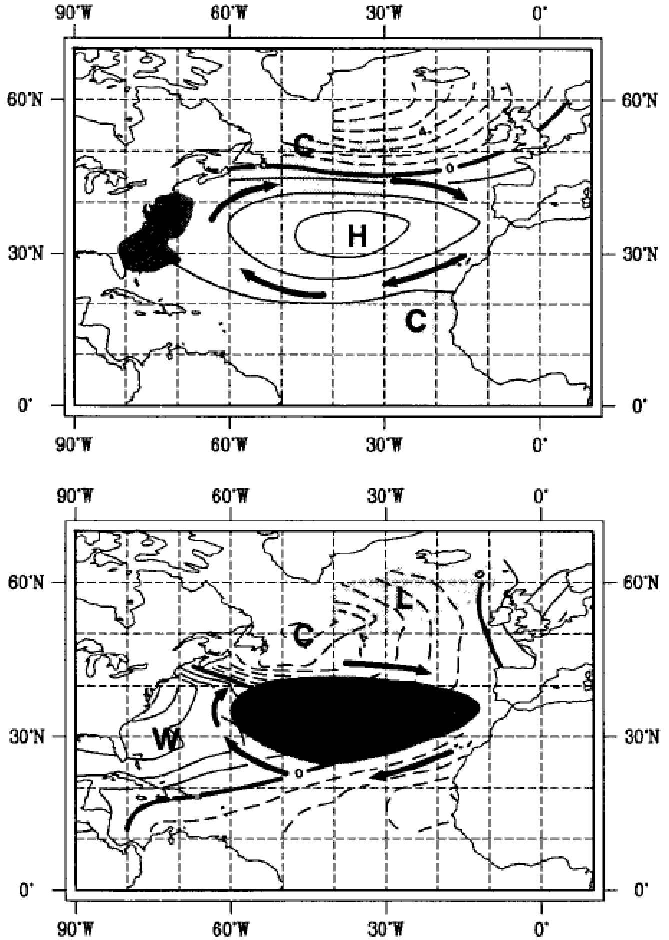


FIG. 5. The patterns of the first canonical pair of SLP (mb; countour interval 1 mb) and SST (K; contour interval 0.1 K) in the North Atlantic area. The correlation between the corresponding time components is 0.56. They explain 21% and 19% of the total variance. In each figure, hatched areas correspond to maxima or minima in the other figure. Continuous lines mark positive values, and dashed lines negative values. The zero line is in bold.

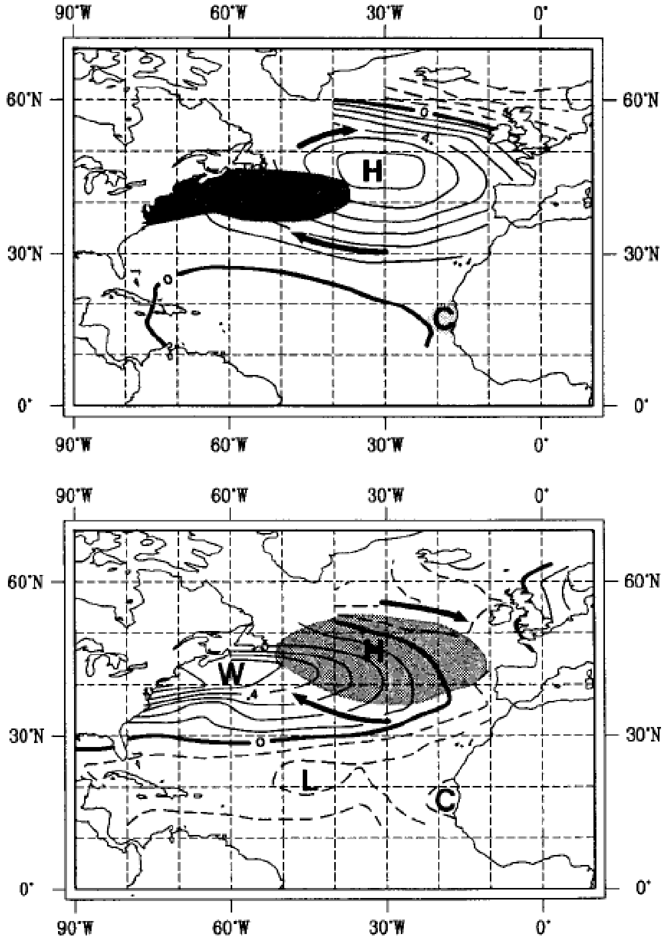


FIG. 6. The patterns of the second canonical pair of SLP (mb; contour interval 1 mb) and SST (K; contour interval 0.1 K) in the North Atlantic area. The correlation between the corresponding time components is 0.47. They explain 31% and 15% of the total variance. In each figure, hatched areas correspond to maxima or minima in the other figure. Continuous lines mark positive values, and dashed lines negative values. The zero line is in bold.

Example for CCA patterns between SST and precipitation (Zorita et al. J. Climate 1992)

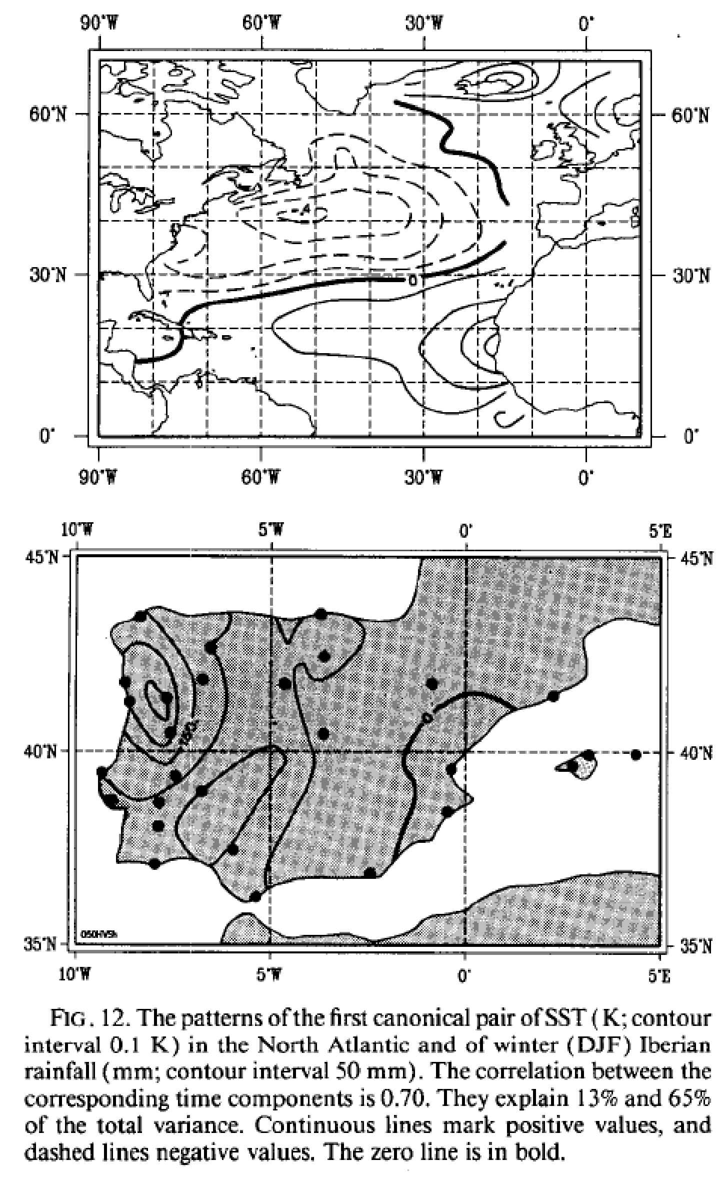
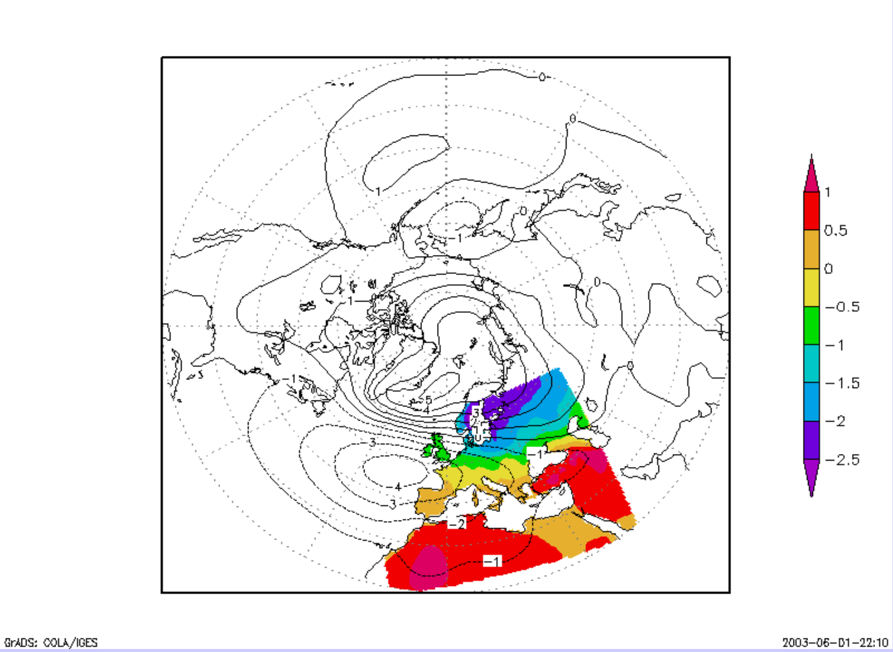


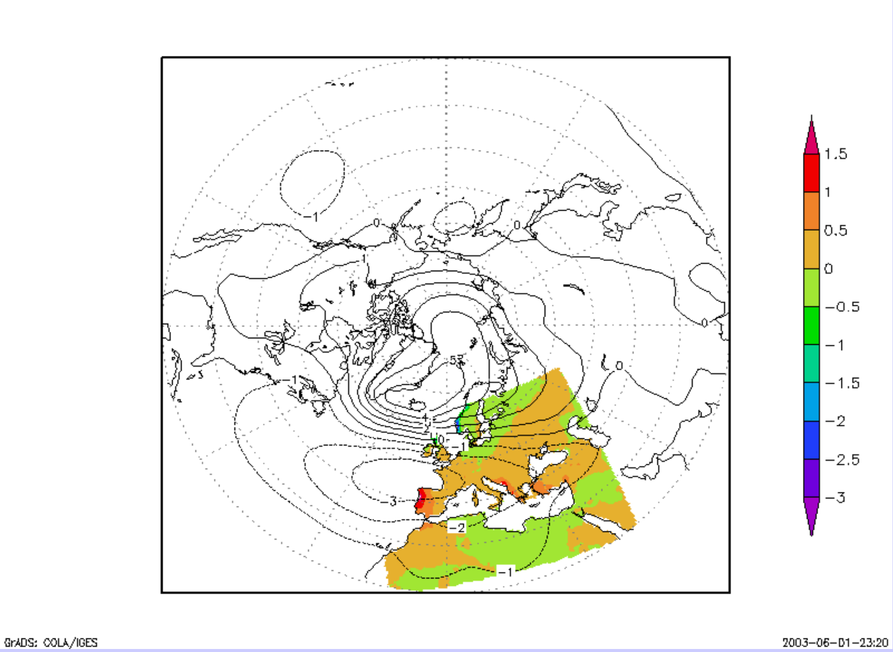
FIG. 12. The patterns of the first canonical pair of SST (K; contour interval 0.1 K) in the North Atlantic and of winter (DJF) Iberian rainfall (mm; contour interval 50 mm). The correlation between the corresponding time components is 0.70. They explain 13% and 65% of the total variance. Continuous lines mark positive values, and dashed lines negative values. The zero line is in bold.

First CCA patterns between SLP and temperature or precipitation from CRU data
(courtesy Roxana Bojariu and Lilana Vilea)

Air surface temperature (˚C) and SLP (hPa) anomalies



Precipitation (mm/day) and SLP (hPa) anomalies



Estimating one dataset from the other

Estimation of one dataset from the other one

The approximate expansion of Y using the leading patterns and time expansion coefficients is given by

$$y_i(t_j) \approx \sum_{k=1}^{\tilde{m}} b_k(t_j)\, q_{ik}$$

If we want to estimate Y from X, we use estimates for the TECs that are obtained through multiple linear regression from the TECs of X

$$\hat{y}_i(t_j) \approx \sum_{k=1}^{\tilde{m}} \hat{b}_k(t_j)\, q_{ik}$$

If the entire set of coupled patterns are used, the estimates obtained from MCA and from CCA are identical to the estimates based on Multiple Linear Regression.

If only a few leading modes are used the MCA, CCA, and MLR estimates are usually different (Tippet et al., J. Climate 2008).