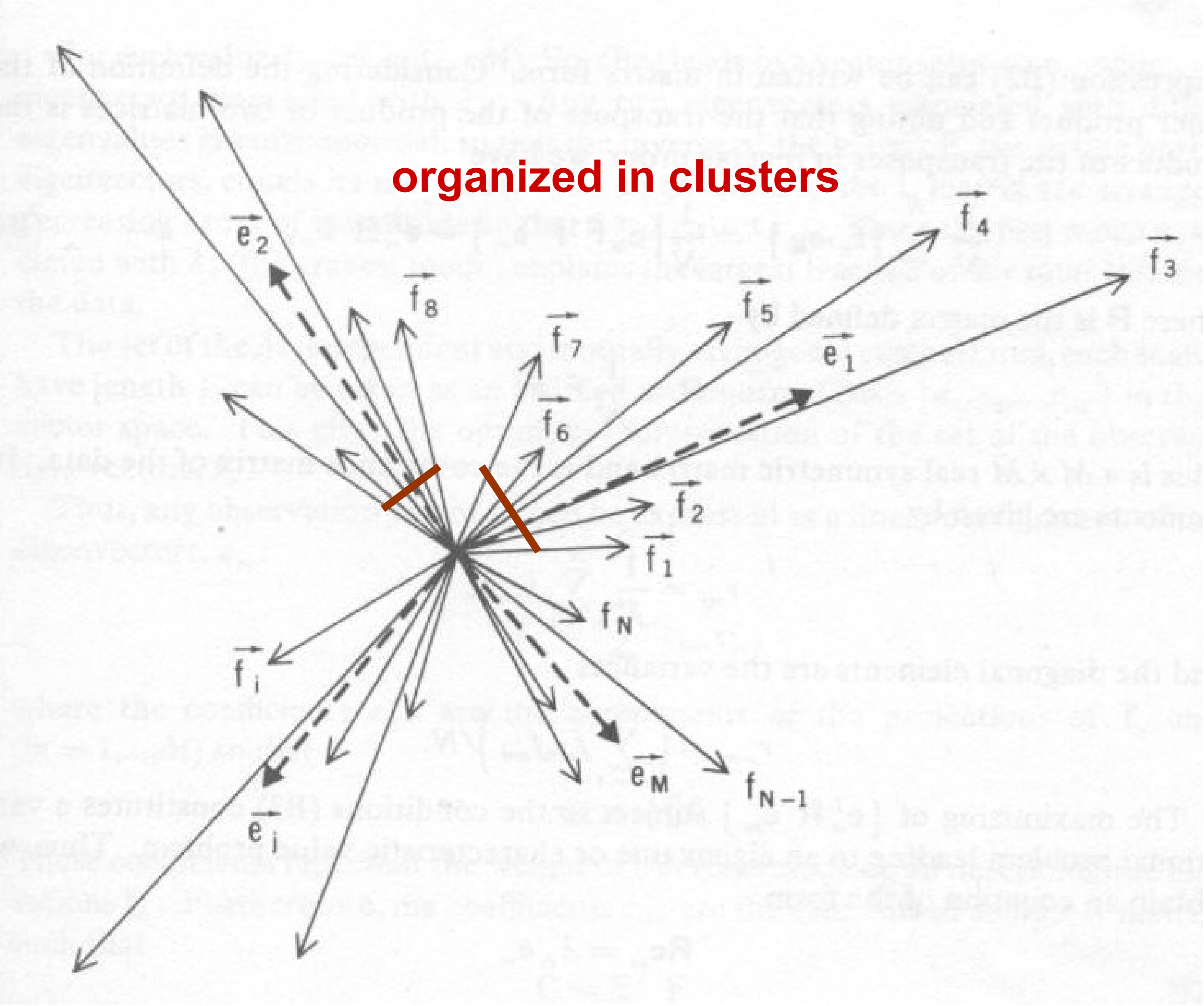
# PRINCIPAL COMPONENT ANALYSIS

*Willem A. Landman*

**AMS Meteorology Glossary:** Regression analysis to determine from a set of independent variables those contributing most to the explained variance.

**Lecture 2**




An example of a possible configuration of the station data vectors $\mathbf{f}_n$

(the subscript n = 1,…,N denotes a particular instant of time)

and the empirical orthogonal vectors $\mathbf{e}_m$ (m = 1,…M with, in general, M<<N)

# PCA

p x p symmetric, nonsingular matrix, i.e., covariance matrix **S**

▼

Pre- and post-multiply with orthonormal matrix **U**

$$\mathbf{U}^T\mathbf{S}\mathbf{U} = \mathbf{L}$$

orthonormal = orthogonal with unit length:
$u^T_1 u_1 = 1$; $u^T_2 u_2 = 1$; $u^T_1 u_2 = 0$

# Characteristic Equation: $|\mathbf{S} - l\mathbf{I}| = \mathbf{0}$

- Columns of **U**: $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ are the *eigenvectors* of **S**

- Diagonal elements of **L**: $l_1, l_2, \ldots, l_p$ are the *eigenvalues* of **S**
  - The ratio of each eigenvalue to the total will indicate the proportion of the total variability accounted for by each eigenvector

# Transformed variables…

- **x** = observations, p x 1 variables

- Transform p correlated variables $x_1$, $x_2$, … into p new UNcorrelated variables:

$$\mathbf{z} = \mathbf{U}^{\mathbf{T}}[\Delta\mathbf{x}],$$

the PRINCIPAL COMPONENTS

$$\Delta\mathbf{x} = \mathbf{U}\mathbf{z}$$

# Getting ready for eigenanalysis

- Covariance $S = X^TX$ or $S^* = XX^T$

- Covariance matrix requires:

$$X = \Delta\mathbf{x} \text{ (mean = 0)}$$

- Correlation matrix requires:

$$X = \Delta\mathbf{x} / \sigma \text{ (mean = 0; std =1)}$$

# In practice…

- Geophysical field, i.e., global SSTs of JFM

**X** =

| | 1950 | 1951 | … | n |
|---|---|---|---|---|
| grid point 1 | | | | |
| : | | | | |
| grid point p | | | | |

# Example (1)

- Matrix=
- 11.0000 22.0000 33.0000
- 3.3166 -4.6904 5.7446

- Normalised matrix=
- -1.0000 0 1.0000
- 0.3406 -1.1258 0.7852

- Standard deviations=
- 1
- 1

- Means=
- 0
- 0

# Example (2)

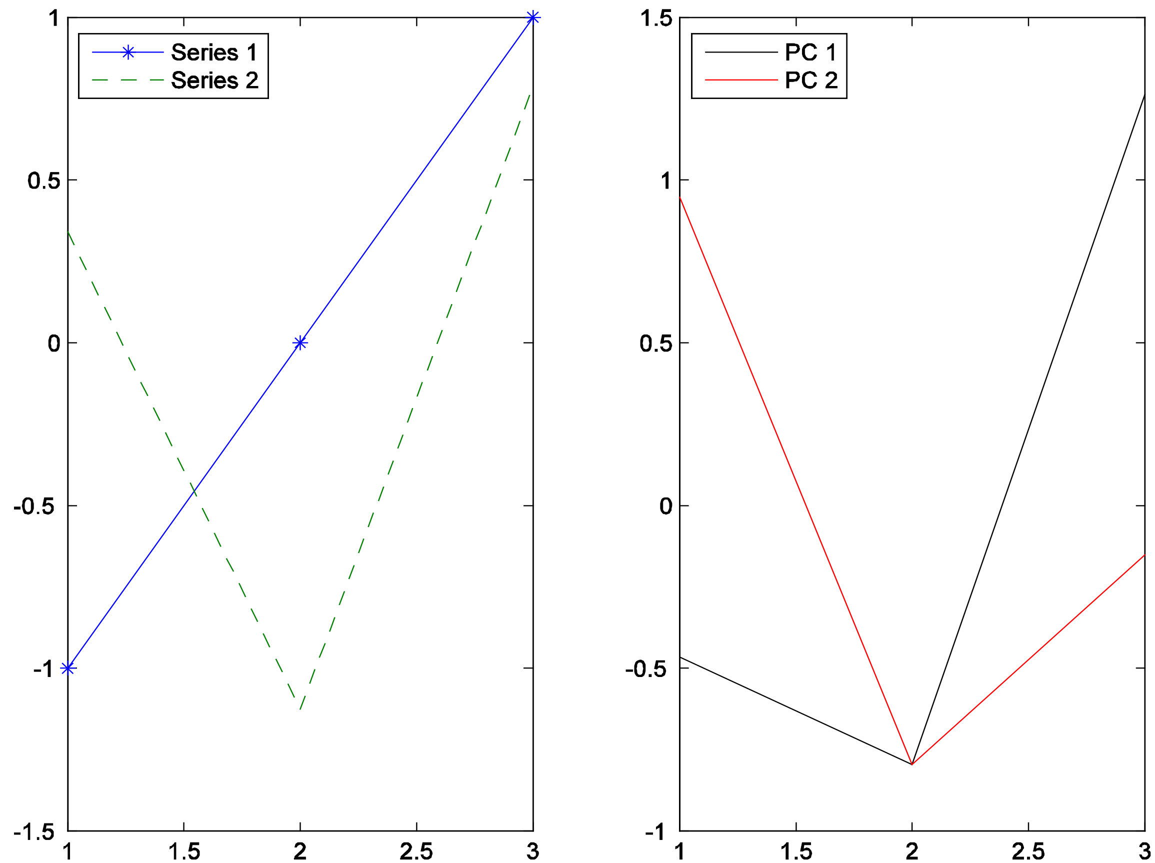
- S=1/(n-1)A*A'
- Correlation matrix=
- 1.0000 0.2223
- 0.2223 1.0000

- Unsorted Eigenvalues=
- 0.7777 0
- 0 1.2223

- Unsorted Eigenvectors
- *-0.7071* 0.7071
- *0.7071* 0.7071

- SORTED Explained variance=
- 61.1161 38.8839

- SORTED Eigenvectors=
- 0.7071 *-0.7071*
- 0.7071 *0.7071*




Principal Components= $\mathbf{U}^T[\Delta \mathbf{x}]$

| | | | | | |
|---|---|---|---|---|---|
| 0.7071 | 0.7071 | X | -1.0000 | 0 | 1.0000 |
| -0.7071 | 0.7071 | | 0.34066 | -1.1258 | 0.7852 |

# Modes of Variation: Eigenanalysis on $\Delta\mathbf{x}^T\Delta\mathbf{x}$

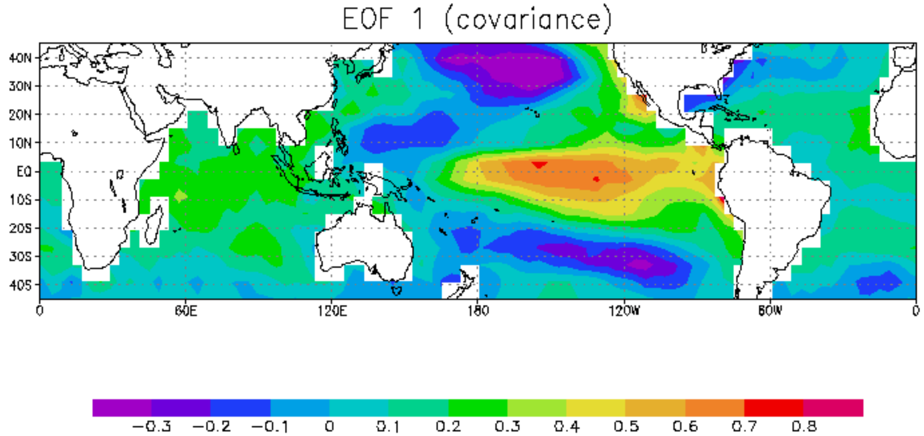
- S – mode:

  $[\Delta\mathbf{x}]=[n \text{ x } p]$; $[\Delta\mathbf{x}^T\Delta\mathbf{x}]=[p \text{ x } n][n \text{ x } p]=[p \text{ x } p]$

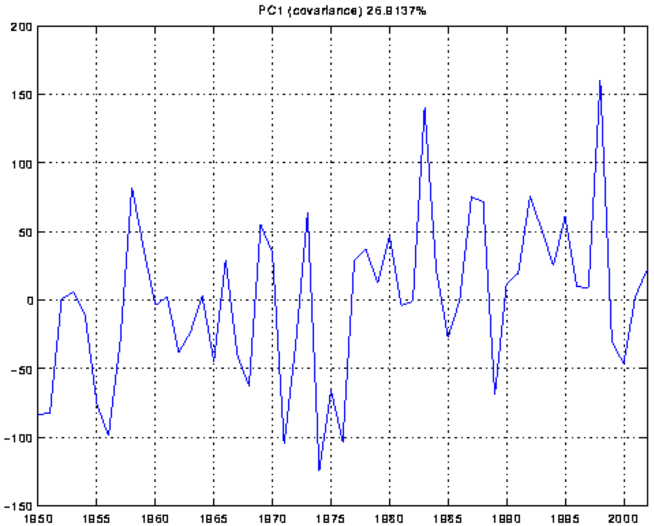
- T – mode:

  $[\Delta\mathbf{x}]=[p \text{ x } n]$; $[\Delta\mathbf{x}^T\Delta\mathbf{x}]=[n \text{ x } p][p \text{ x } n]=[n \text{ x } n]$

# Covariance

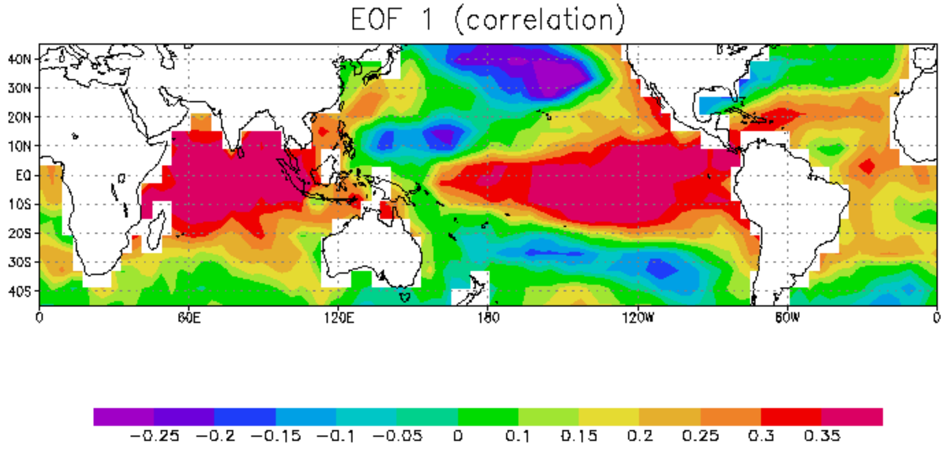


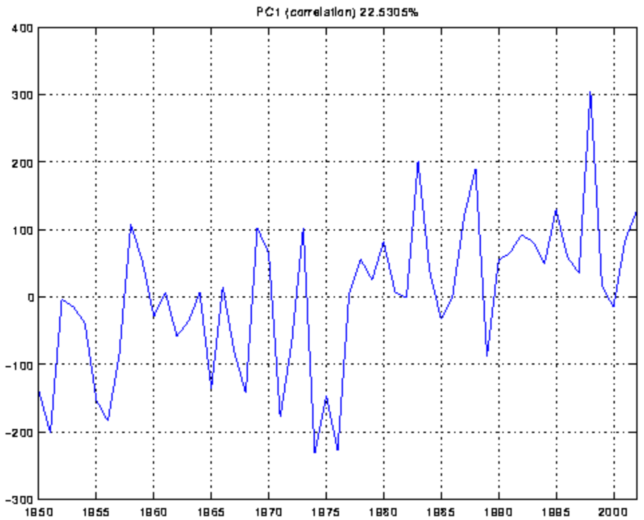
EOF 1 (covariance)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0
60E
120E
180
120W
60W
0
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

PC1 (covariance) 26.8137%
200
150
100
50
0
−50
−100
−150
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
2000

# Correlation



# When the covariance matrix CANNOT be used…

- When variances differ widely:

  Eq. IO SST var. $\ll$ Eq. PO SST var.

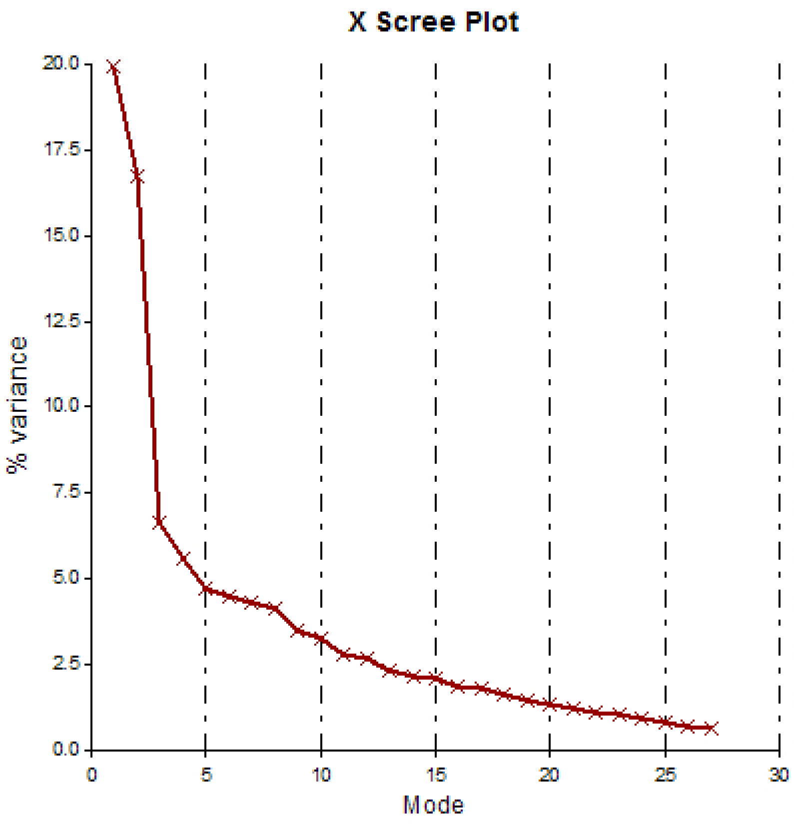
- When units are different:

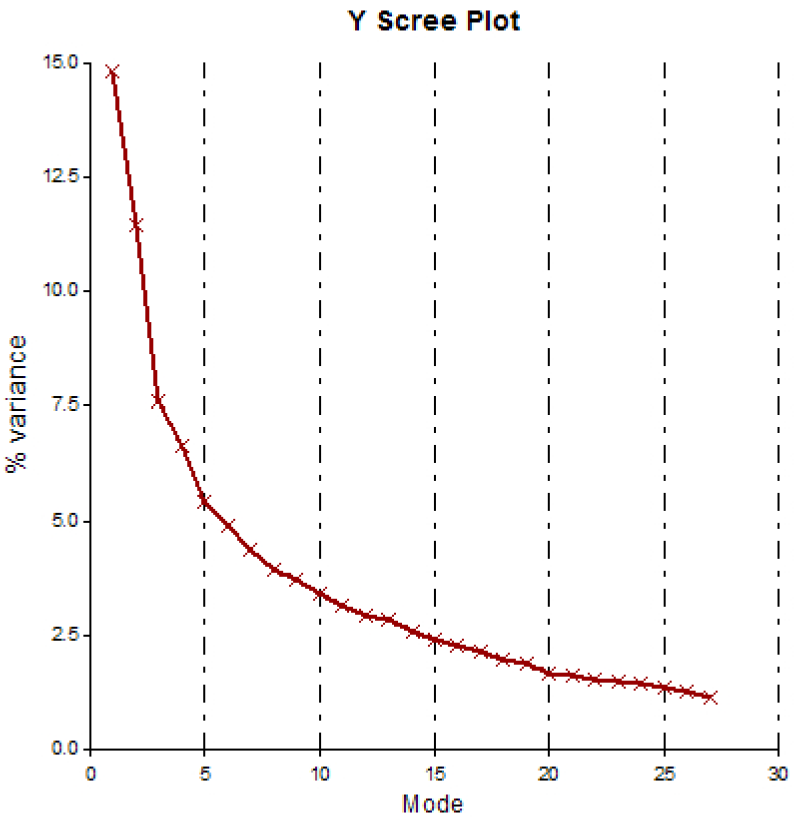
SSTs combined with 200 hPa gpm heights

# How many modes?

- Proportion of variance
- SCREE test
- Average root (Guttman – Kaiser)
- Sensitivity tests (used in statistical modelling)

# The SCREE test



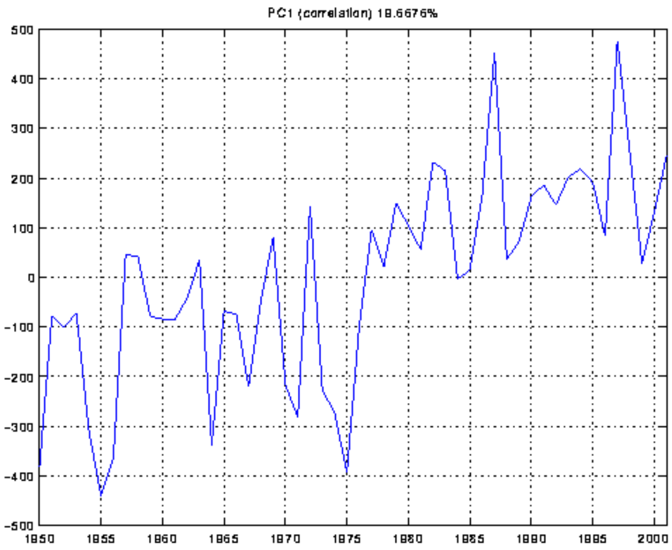
# Extended EOFs (EEOFs)

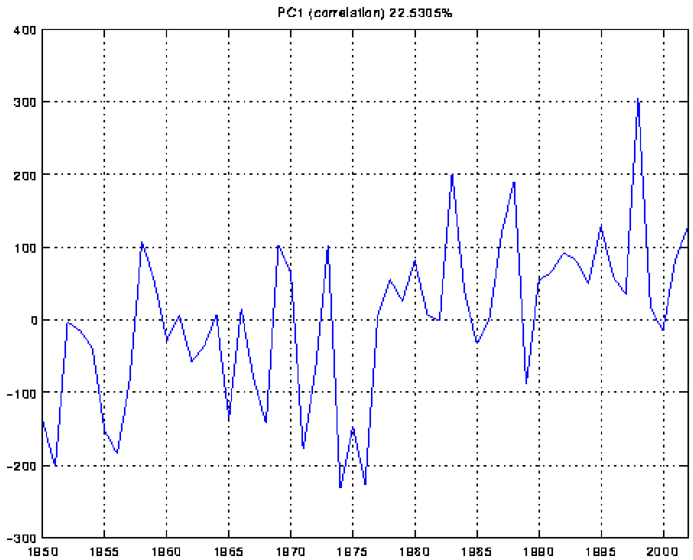
Example of time evolution:

AMJ–JAS–OND–JFM SSTs combined into one field **x**

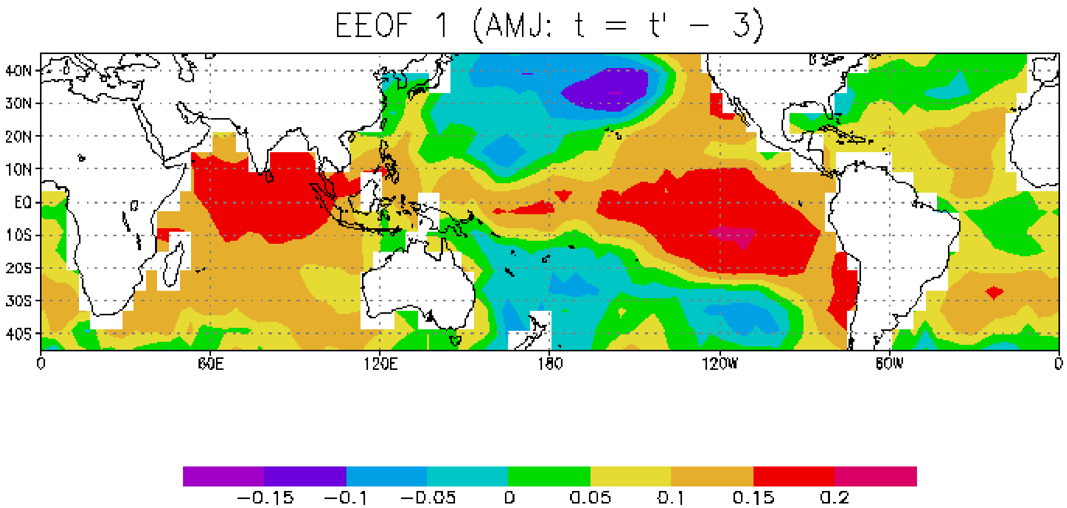
Resulting PCA will result in evolutionary or steady-state features of SSTs

# EEOF1

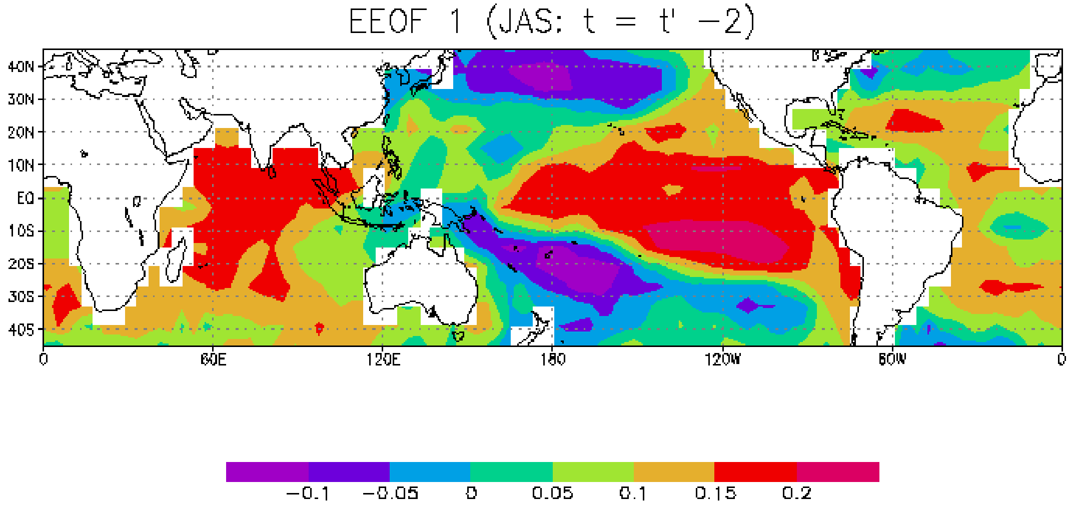

PC1 (correlation) 18.6676%
500
400
300
200
100
0
-100
-200
-300
-400
-500
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000


PC1 (correlation) 22.5305%
400
300
200
100
0
-100
-200
-300
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000

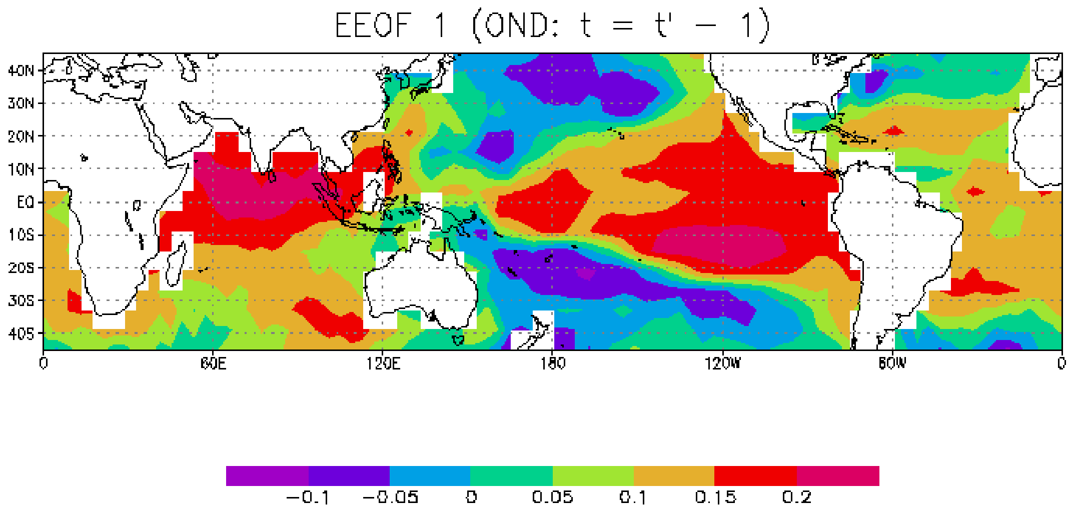
# EEOF1


EEOF 1 (AMJ: t = t' − 3)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0
60E
120E
180
120W
60W
0
−0.15
−0.1
−0.05
0
0.05
0.1
0.15
0.2

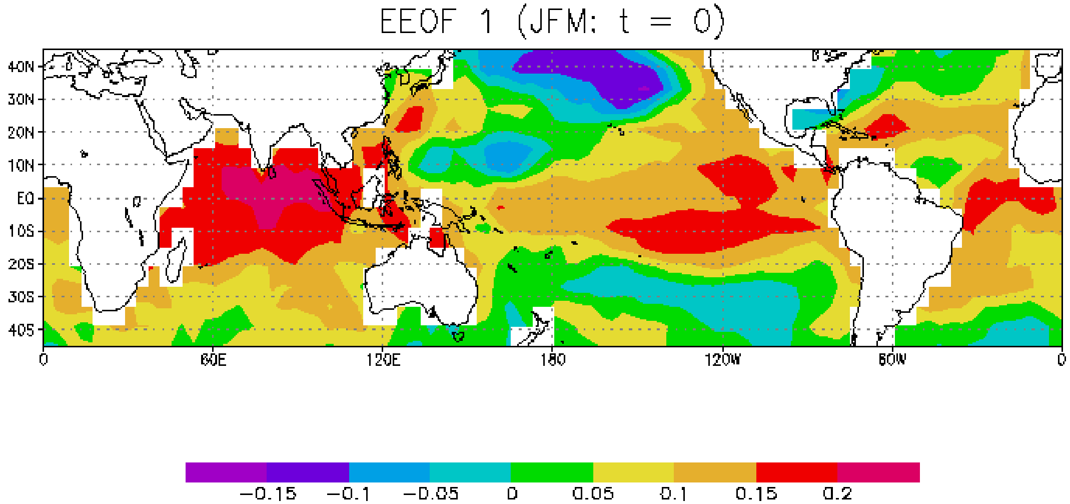
# EEOF1


EEOF 1 (JAS: t = t' −2)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0
60E
120E
180
120W
60W
0
−0.1
−0.05
0
0.05
0.1
0.15
0.2

# EEOF1


EEOF 1 (OND: t = t' − 1)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0
60E
120E
180
120W
60W
0
−0.1
−0.05
0
0.05
0.1
0.15
0.2

# EEOF1


EEOF 1 (JFM: t = 0)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0
60E
120E
180
120W
60W
0
−0.15
−0.1
−0.05
0
0.05
0.1
0.15
0.2

# ROTATION

- Rotation required when Buell patterns exist (Richman 1986)
- Same amount of variation explained after rotation
- Two types of rotation
    - Orthogonal – varimax
    - Oblique
- Rotation might be useful for a group of PCs whose eigenvalues are similar

## Interpreting the Principal Components

Principal components are notoriously difficult to interpret physically.

The weights are defined to maximize the variance, not maximize the interpretability!

With spatial data (including climate data) the interpretation becomes even more difficult because there are geometric controls on the correlations between the data points.

**Simon Mason**

# Buell patterns

Imagine a rectangular domain in which all the points are strongly correlated with their neighbours.

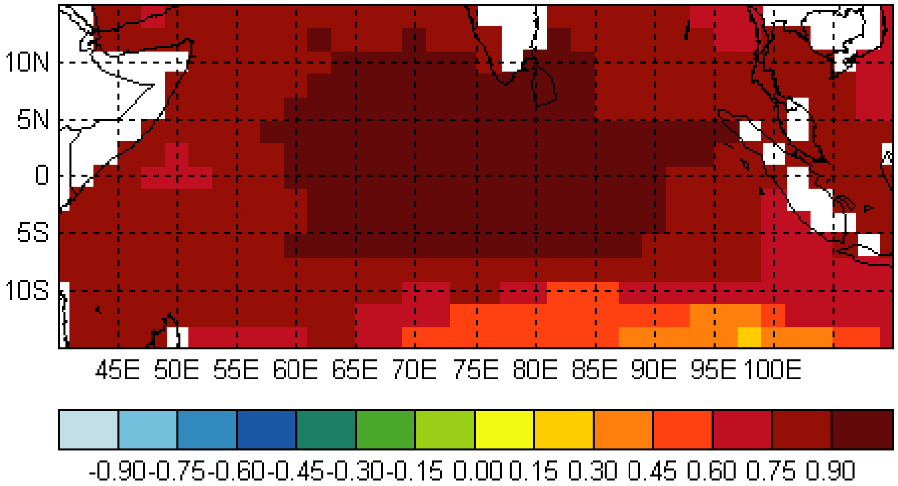
The points in the middle of the domain will have the strongest average correlations with all other points. This is often represented by PC 1.

The points furthest apart will have the strongest negative correlations. This is often represented by PC 2.

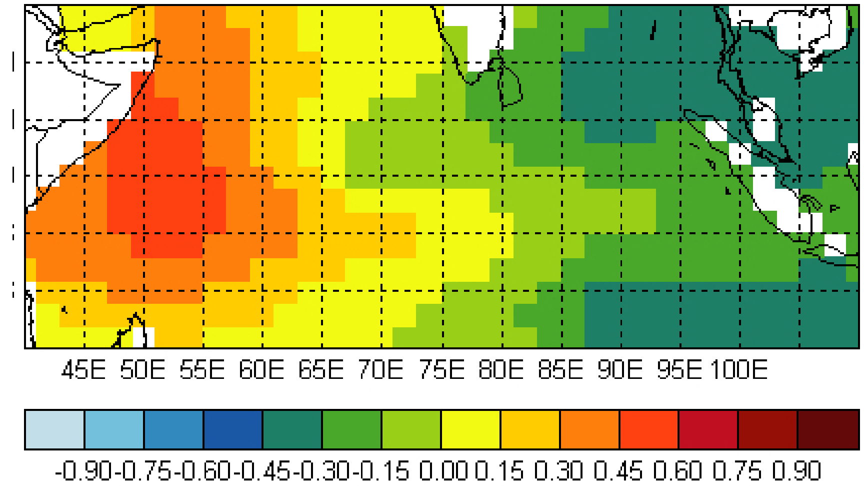
# Buell patterns?

Are these real, or are they a function of the domain shape?

X Spatial Loadings (EOF1)




X Spatial Loadings (EOF2)




**Simon Mason**

INTERNATIONAL RESEARCH INSTITUTE

# Buell patterns

Because of domain shape dependencies:

1. the first PC frequently indicates positive loadings with strongest values in the centre of the domain;
2. the second PC frequently indicates negative loadings on one side and positive loadings on the other side in the direction of the longest dimension of the domain.

Similar kinds of problems arise when using:

1. gridded data with converging longitudes, or simply with longitude spacing less than latitude spacing;
2. station data.

# Rotation

The weights are defined to maximize the variance, not maximize the interpretability!

The weights could be redefined to meet alternative criteria. Rotation is sometimes performed to maximize the weights of as many metrics as possible, and to minimize the weights of the others.

The variances of the first few principal components are reduced after rotation.

**Simon Mason**

# Rotation

Rotation does NOT solve Buell pattern problems, nor station and uneven gridded data problems, it only reduces them.

These problems are only of concern for <u>interpretation</u>.

# Canonical Correlation Analysis (CCA)

# CCA…

- Identifies new variables that maximize the interrelationship between **two** data sets
- This is in contrast to the patterns describing the internal variability within a **single** data set identified in PCA
- It is in this sense that CCA is referred to as "double-barreled" PCA

# CCA…

- In multiple regression, the predictand is a scalar
- CCA can also be viewed as an extension of multiple regression to the case of a vector-valued predictand variable
- The predictor: vector of SSTs, SLP, etc.
- The predictand: vector of rainfall stations, etc.
- Widely applied to geophysical data in the form of fields

# The Mathematics of CCA

Consider two vector variables **X** (predictor) and **Y** (predictand):

$\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, and
$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_q)^T$

where $q \leq p$

We will consider the case where both **X** and **Y** are centered data sets, i.e., the mean has been removed

# With PCA…

$$\mathbf{U} = \mathbf{E}^{\mathsf{T}}\mathbf{X}$$

**E**: eigenvectors

**X**: centered data

# With CCA…

$$\mathbf{V} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$$

$$\mathbf{W} = \mathbf{B}^{\mathrm{T}}\mathbf{Y}$$

each is a linear combination of elements of the respective data vectors **X** and **Y**

**A**: corresponding vectors of weights of **X**,

**B**: corresponding vectors of weights of **Y**, (called canonical vectors)

**X** and **Y**: centered data

# Properties of CCA…

- $\mathrm{corr}[V_1,W_1] \geq \mathrm{corr}[V_2,W_2] \geq \ldots \geq \mathrm{corr}[V_M,W_M]$

- $\mathrm{corr}[V_k,W_m] = r_c, \quad k = m,$
  $\qquad\qquad\qquad = 0, \quad k \neq m$

- $\mathrm{var}[V_m] = \mathrm{var}[W_m] = 1,\ m=1,\ldots,M$

  Each of the M successive pairs of canonical variates exhibits a weaker correlation than the previous pair

  Canonical correlations, $r_c$, are correlations between the pairs of canonical variates

# Computing CCA…

- A joint variance-covariance matrix of the variables **X** and **Y** is constructed
- These two data vectors are combined into one single vector $\mathbf{C}^T = [\mathbf{X}^T,\mathbf{Y}^T]$,

- The variance-covariance matrix of **C**:

$$[\mathbf{S}_C] = [\mathbf{C}]^T[\mathbf{C}]/(n-1)$$

# Correlation Matrix

$$[\mathbf{R}] = [\mathbf{Z}]^T[\mathbf{Z}]/(n\text{-}1)$$

dim $[\mathbf{Z}]$ = n x K

n: “occasions”

K: number of variables

dim $[\mathbf{R}]$ = K x K

# Structure of the Correlation Matrix for PCA

(of fields with observations of more than one variable at each location)

$$
\mathbf{[R]} = \begin{bmatrix} [\mathbf{R}_{1,1}] & [\mathbf{R}_{1,2}] & \ldots & [\mathbf{R}_{1,L}] \\ [\mathbf{R}_{2,1}] & [\mathbf{R}_{2,2}] & \ldots & [\mathbf{R}_{2,L}] \\ \ldots & \ldots & \ldots & \ldots \\ [\mathbf{R}_{L,1}] & [\mathbf{R}_{L,2}] & \ldots & [\mathbf{R}_{L,L}] \end{bmatrix}
$$

**The sub-matrices located on the diagonal contain ordinary correlation matrices for each of the L variables**

L: variables of which multiple observations exist

K: locations

KL: dimensionality of the data vector: the first K elements are observations of the first variable, the second K elements are observations of the second variable, etc…

dim[R] = MxM, where M = KL

# Analogous to PCA with more than one variable…

$$[\mathbf{R}] = \begin{bmatrix} [\mathbf{R}_{1,1}] & [\mathbf{R}_{1,2}] \\ [\mathbf{R}_{2,1}] & [\mathbf{R}_{2,2}] \end{bmatrix}$$

$$[\mathbf{S}_C] = \begin{bmatrix} \underset{(I\mathrm{x}I)}{[\mathbf{S}_{XX}]} & \underset{(I\mathrm{x}J)}{[\mathbf{S}_{XY}]} \\ \underset{(J\mathrm{x}I)}{[\mathbf{S}_{YX}]} & \underset{(J\mathrm{x}J)}{[\mathbf{S}_{YY}]} \end{bmatrix}$$

# Algebraic problem to solve for A and B…

$\psi = \mathbf{A}^T\mathbf{S}_{XY}\mathbf{B} - ½\, \lambda(\mathbf{A}^T\mathbf{S}_{XX}\mathbf{A} - 1) - ½\, \mu(\mathbf{B}^T\mathbf{S}_{YY}\mathbf{B} - 1)$,

(λ and μ are Langrangian multipliers)

$$\partial\psi/\partial\mathbf{A} = \mathbf{S}_{XY}\mathbf{B} - \lambda\mathbf{S}_{XX}\mathbf{A} = \mathbf{0}$$

$$\partial\lambda/\partial\mathbf{B} = \mathbf{S}^T{}_{XY}\mathbf{A} - \mu\mathbf{S}_{YY}\mathbf{B} = \mathbf{0}$$

**…and after a lot more incredible algebra…**

# The Mathematics of CCA

The CCA eigenvalue problem:

$$(\mathbf{S}_{\mathbf{xx}}{}^{-1}\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yy}}{}^{-1}\mathbf{S}_{\mathbf{yx}} - \boldsymbol{\lambda}^2\mathbf{I})\mathbf{A} = \mathbf{0}$$

$$(\mathbf{S}_{\mathbf{yy}}{}^{-1}\mathbf{S}_{\mathbf{yx}}\mathbf{S}_{\mathbf{xx}}{}^{-1}\mathbf{S}_{\mathbf{xy}} - \boldsymbol{\lambda}^2\mathbf{I})\mathbf{B} = \mathbf{0}$$

The largest eigenvalue $\lambda^2{}_1$ is associated with the first eigenvector $\mathbf{A}_1$ or $\mathbf{B}_1$

$(\boldsymbol{\lambda}^2 = \mathbf{r_c})$

# Analogies between PCA and CCA

- Define a new set of variables (PCs) that optimally describes variance in a single data set
- The PCs are based on the eigenvalue problem of the covariance matrix
- The eigenvalues represent the relative variance with each EOF

- Define a new set of variables that optimally describes the cross-correlation between two different data sets
- The set of variables is based on the coupled eigenvalue problem of the cross-covariance and auto-covariance matrices between two data sets
- The eigenvalues represent the level of correlation between patterns of predictor variables and patterns of predictand variables

# PCA version of CCA

- In practice: sometimes useful to “prefilter” the two fields (predictor and predictand) of raw data
- The two analyses may be truncated at different numbers of principal components
- BEWARE: important information could be lost when truncating the PCA
- PCA necessary when there is strong spatial correlation within fields
- With small sample size (number of years), PCA pre-filtering tends to improve stability – necessary for forecasting independent data

# Origins and Levels of Monthly and Seasonal Forecast Skill for United States Surface Air Temperatures Determined by Canonical Correlation Analysis

T. P. BARNETT

*Climate Research Group, Scripps Institution of Oceanography, La Jolla, CA 92093*

R. PREISENDORFER

*Pacific Marine Environment Laboratory, National Oceanographic and Atmospheric Administration, Seattle, WA 98115*

(Manuscript received 2 July 1986, in final form 4 February 1987)

## ABSTRACT

Statistical techniques have been used to study the ability of SLP, SST and a form of persistence to forecast cold/warm season air temperatures over the United States and to determine the space–time evolution of these fields that give rise to forecast skill.

It was found that virtually all forecast skill was due to three climatological features: a decadal scale change in Northern Hemisphere temperature, ENSO-related phenomena, and the occurrence of two distinct short-lived, but large-scale, coherent structures in the atmospheric field of the Northern Hemisphere. The physical mechanisms responsible for the first two signals are currently unknown. One of the large-scale, coherent features seems largely independent of the ENSO phenomena, while the second is at least partially related to ENSO and may be part of a recently discovered global mode of SLP variation. Both features resemble various combinations of known teleconnection patterns. These large-scale coherent structures are essentially stationary patterns of SLP variation that grow in place over two to three months. The structures decay more rapidly, typically in 1 month, leading to a highly asymmetric temporal life cycle.

The average forecast skills found in this study are generally low, except in January and February, and are always much lower than expected from studies of potential predictability. Increase in the average skills will require new information uncorrelated with any of the data used in this study and/or prediction schemes that are highly nonlinear. However, the concept of an average skill may be misleading. A forecast quality index is developed and it is shown that one can say in advance that some years will be highly predictable and others not. Use of the classical definition of "winter" in forecast work may not be advisable since each of the months that make up winter are largely uncorrelated and predicted by different atmospheric features.

Monthly Weather Review, 115.

# CCA as analysis tool

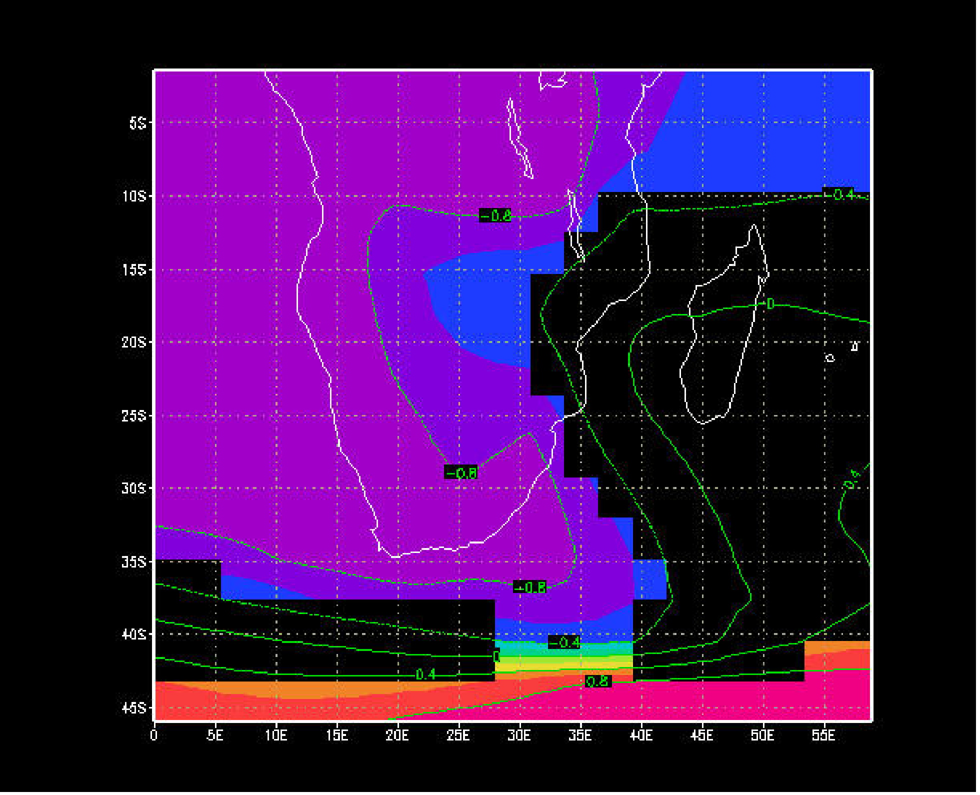
$$\boldsymbol{X}(x,t) = \Sigma_j \mathbf{r}_j(t)\mathbf{g}_j(x),\ j = 1, 2, \dots , p$$

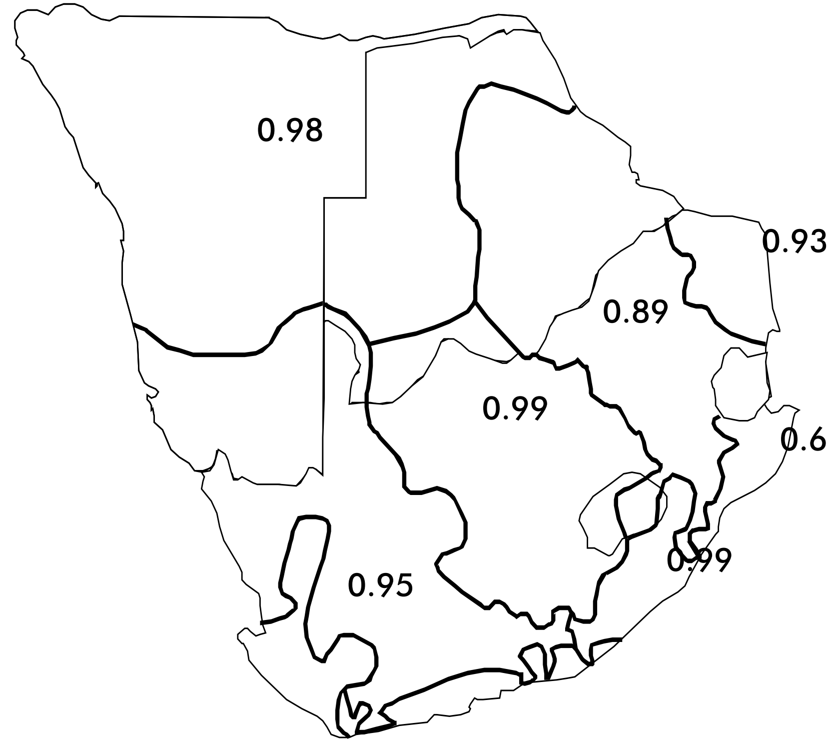
$$\boldsymbol{Y}(x',t) = \Sigma_k \mathbf{s}_k(t)\mathbf{h}_k(x'),\ k = 1, 2, \dots , q$$

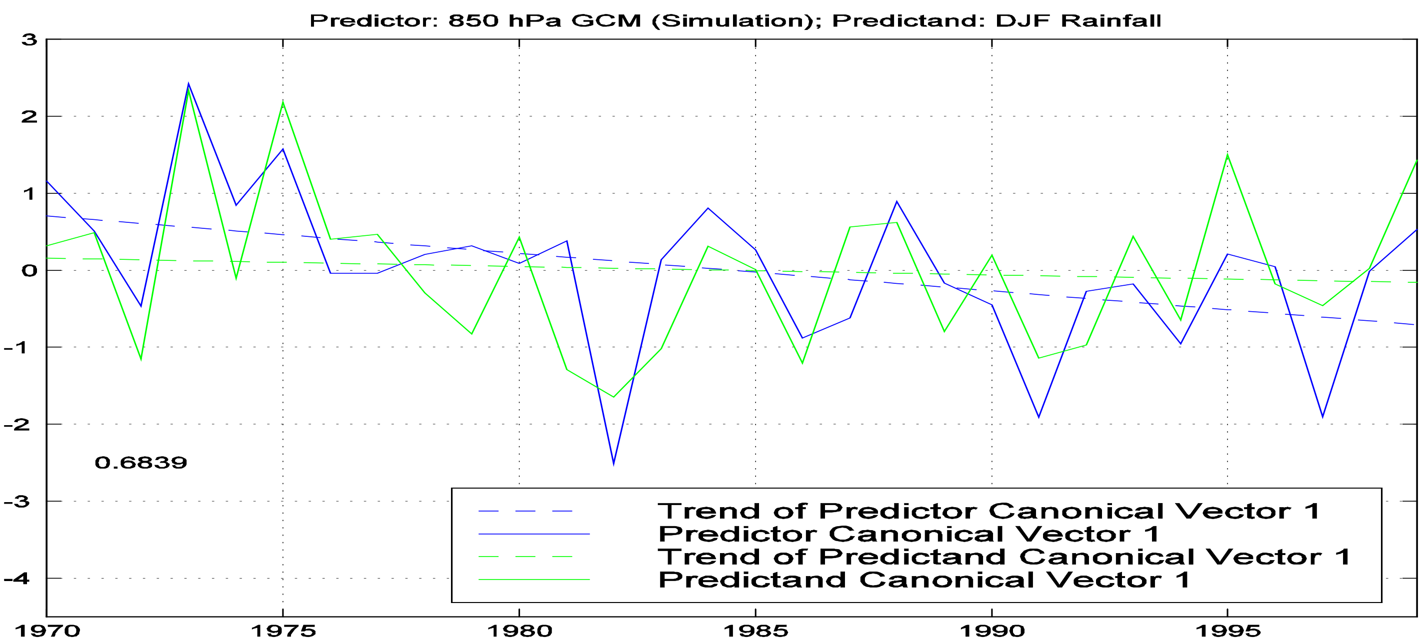
$\mathbf{g}_j$ and $\mathbf{h}_k$ are vectors whose components show the correlation at a specific location between the predictor or the predictand and their respective canonical component time series ($\mathbf{r}_j$ and $\mathbf{s}_k$)

# Example

- Predictor: DJF 850 hPa geopotential heights forecast by the ECHAM3.6 GCM
- Predictand: Southern African DJF observed rainfall indices
- g-map: correlations between the 850 hPa heights and the predictor canonical component time series
- h-map: correlations between the rainfall indices and the predictand canonical component time series

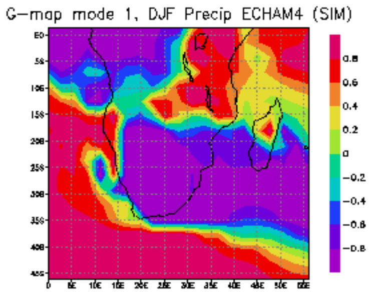

5S
10S
15S
20S
25S
30S
35S
40S
45S
0
5E
10E
15E
20E
25E
30E
35E
40E
45E
50E
55E
-0.8
-0.4
0
0.4
0.8

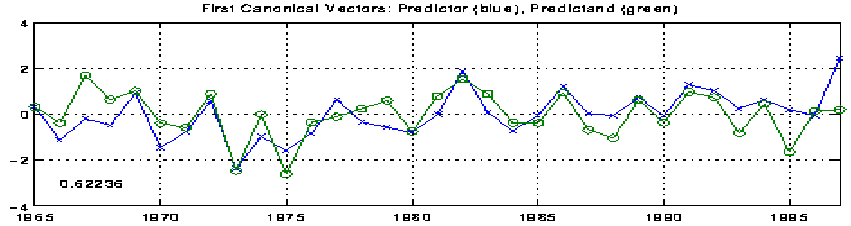

0.98
0.93
0.89
0.99
0.61
0.99
0.95

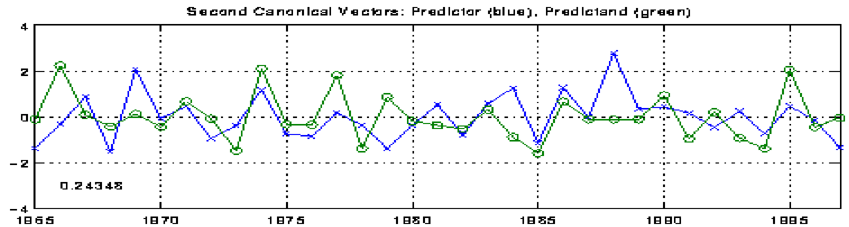

Predictor: 850 hPa GCM (Simulation); Predictand: DJF Rainfall
3
2
1
0
-1
-2
-3
-4
1970
1975
1980
1985
1990
1995
0.6839
Trend of Predictor Canonical Vector 1
Predictor Canonical Vector 1
Trend of Predictand Canonical Vector 1
Predictand Canonical Vector 1

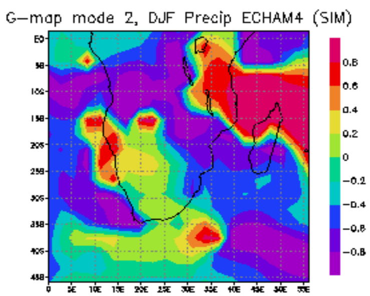
# Example

- Predictor: DJF rainfall simulated by the ECHAM4.5 GCM
- Predictand: Southern African DJF observed station rainfall
- g-map: correlations between the GCM-rainfall and the predictor canonical component time series
- h-map: correlations between the station rainfall and the predictand canonical component time series
- hn-map: correlation between the station rainfall and the predictor canonical component time series

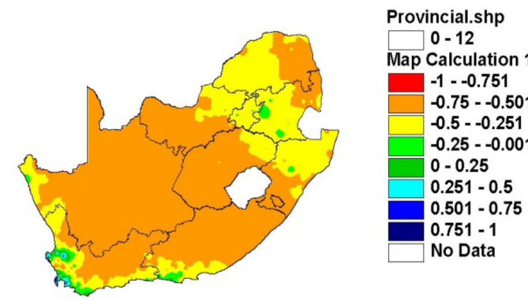


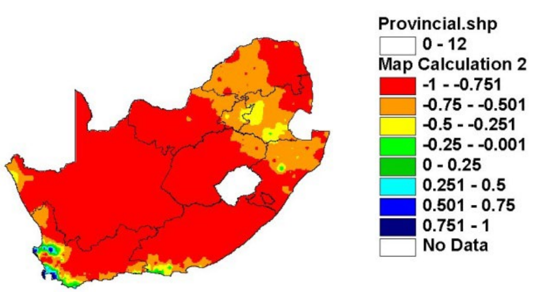



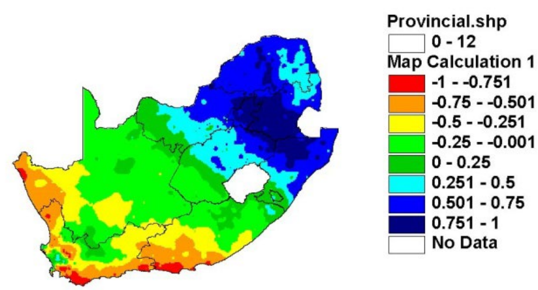

## Hn map CCA Mode 1



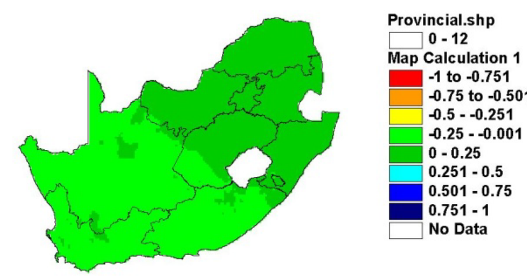

## H map CCA Mode 1




## H map CCA Mode 2



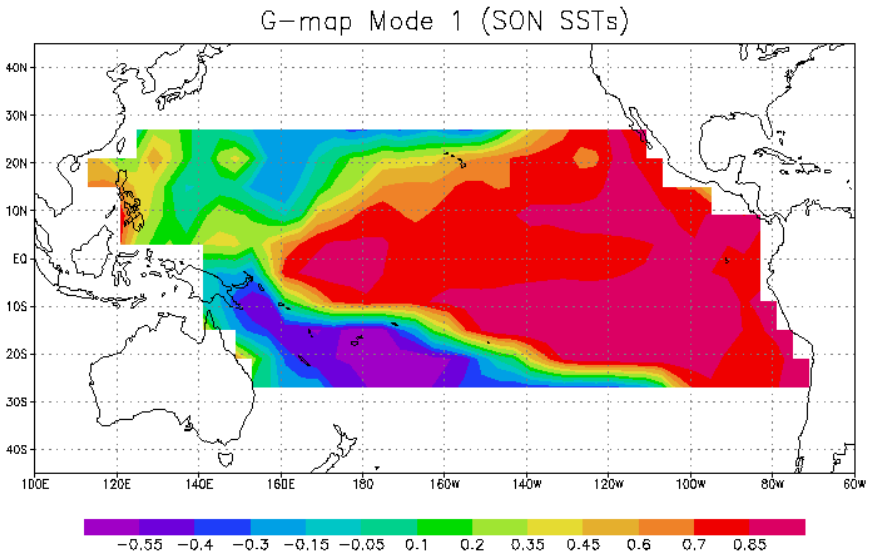

## Hn map CCA Mode 2

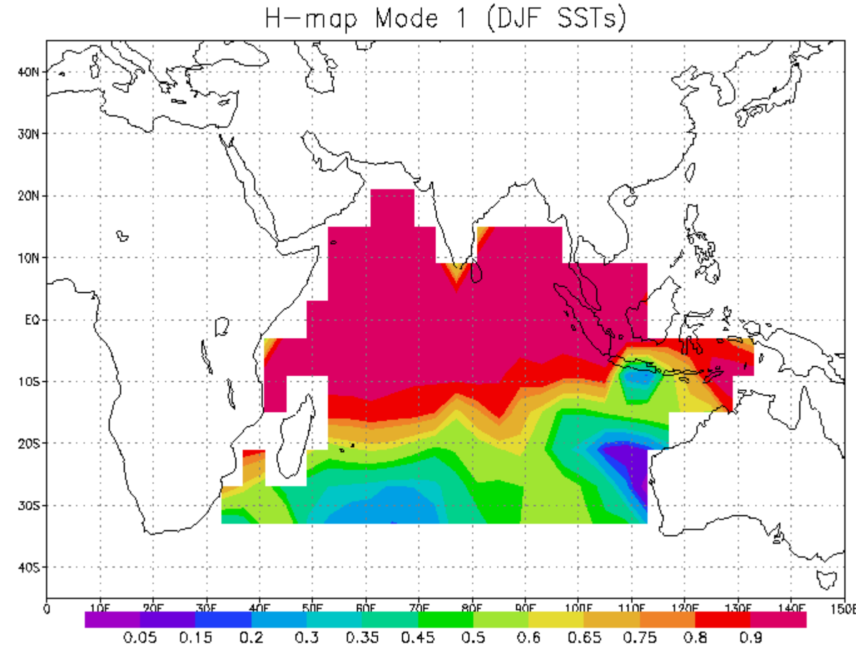


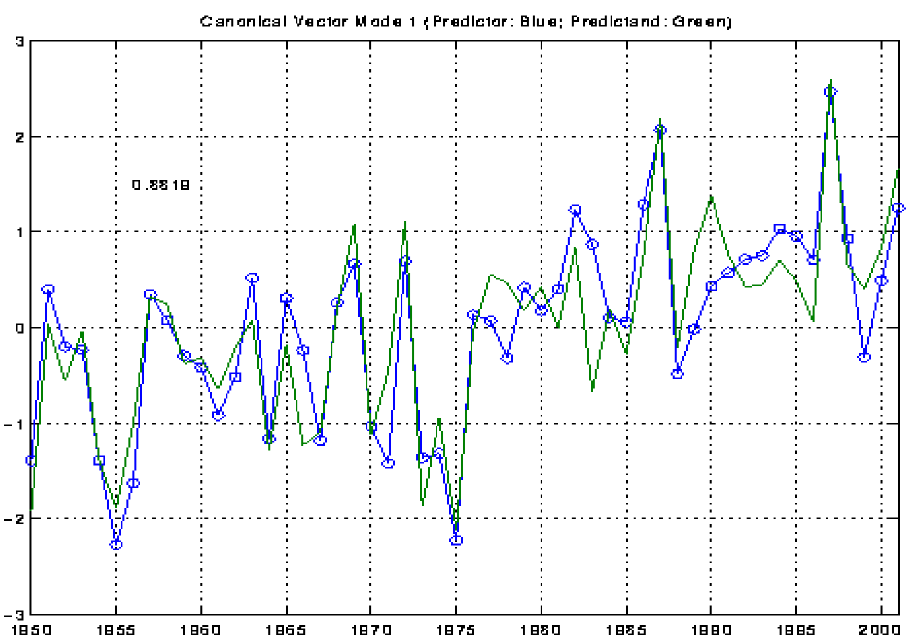
# SST model of the Indian Ocean

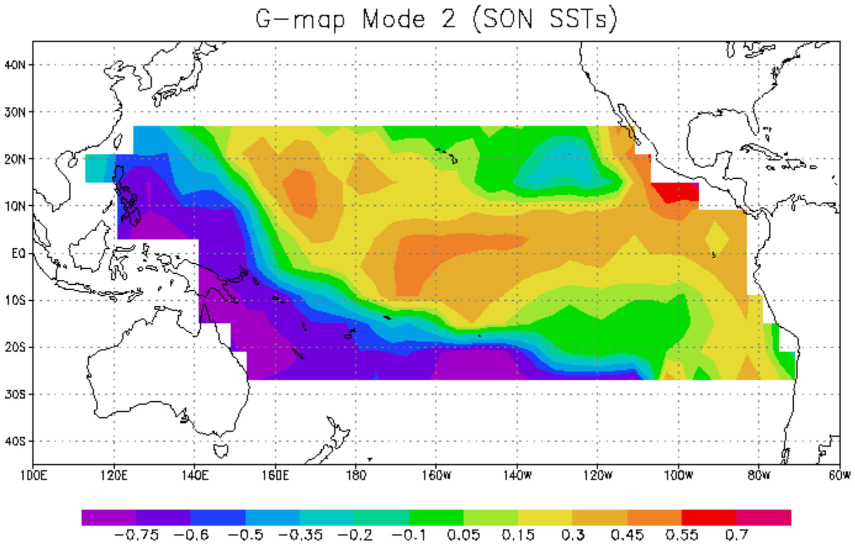
- Demonstration of a very simple model predicting DJF SSTs of the Indian Ocean using SON SSTs of the Pacific Ocean

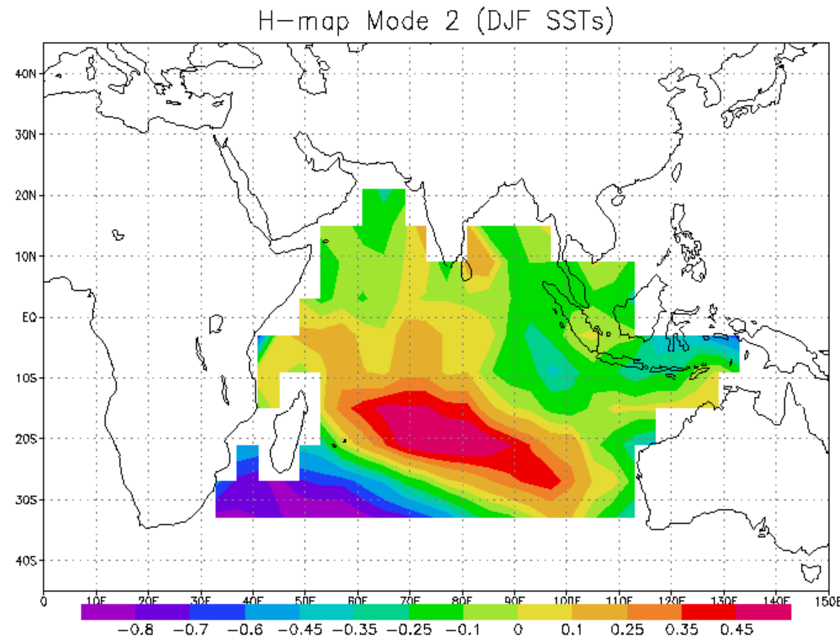
- Only the spatial patterns will be shown…

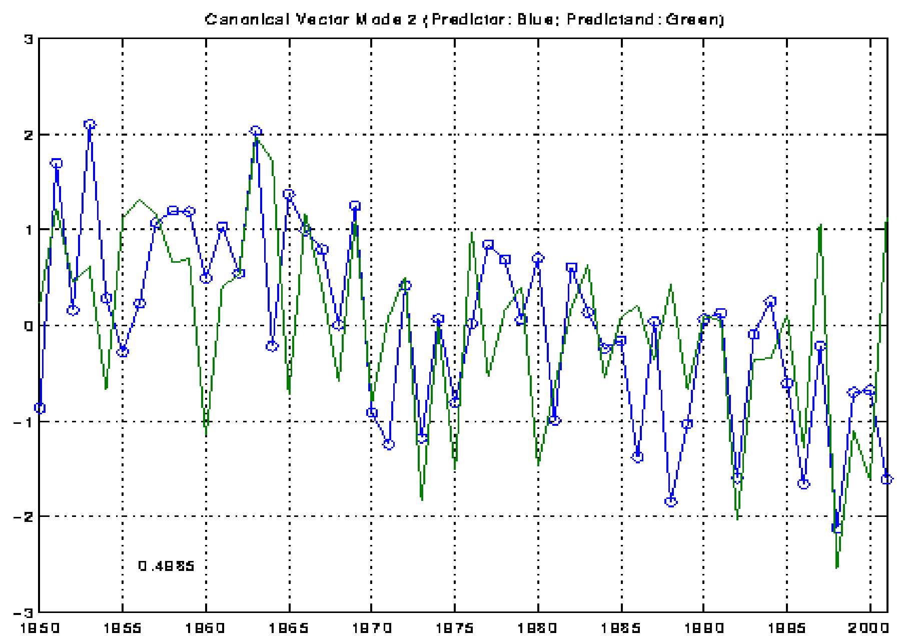

G-map Mode 1 (SON SSTs)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
100E 120E 140E 160E 180 160W 140W 120W 100W 80W 60W
-0.55 -0.4 -0.3 -0.15 -0.05 0.1 0.2 0.35 0.45 0.6 0.7 0.85


H-map Mode 1 (DJF SSTs)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0 10E 20E 30E 40E 50E 60E 70E 80E 90E 100E 110E 120E 130E 140E 150E
0.05 0.15 0.2 0.3 0.35 0.45 0.5 0.6 0.65 0.75 0.8 0.9


Canonical Vector Mode 1 (Predictor: Blue; Predictand: Green)
0.8818
3
2
1
0
-1
-2
-3
1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000


G-map Mode 2 (SON SSTs)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
100E
120E
140E
160E
180
160W
140W
120W
100W
80W
60W
−0.75 −0.6 −0.5 −0.35 −0.2 −0.1 0.05 0.15 0.3 0.45 0.55 0.7


H-map Mode 2 (DJF SSTs)
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0
10E
20E
30E
40E
50E
60E
70E
80E
90E
100E
110E
120E
130E
140E
150E
−0.8 −0.7 −0.6 −0.45 −0.35 −0.25 −0.1 0 0.1 0.25 0.35 0.45


Canonical Vector Mode 2 (Predictor: Blue; Predictand: Green)
3
2
1
0
−1
−2
−3
0.4985
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000

# HOMEWORK

Consider the maps and time series of the previous Indo-Pacific Ocean example. Using CCA diagnostics, explain how the SON SSTs of the equatorial Pacific represented by the first CCA mode are linearly related to the DJF SSTs of the equatorial Indian Ocean.