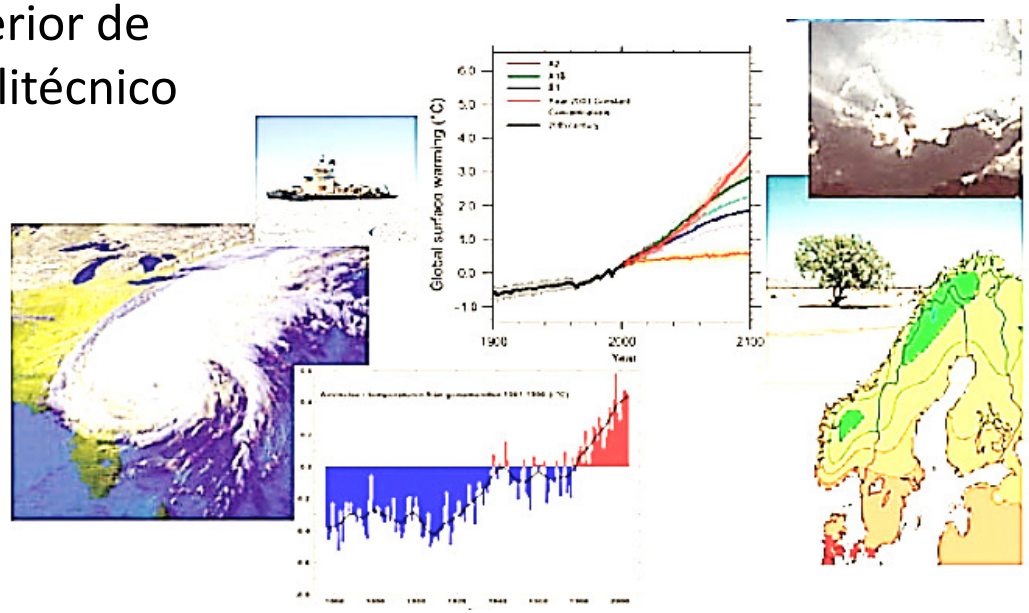
From PCA to ICA – Multivariate Techniques to Study Climate Data

Irene Oliveira, Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro, CITAB, ioliveir@utad.pt

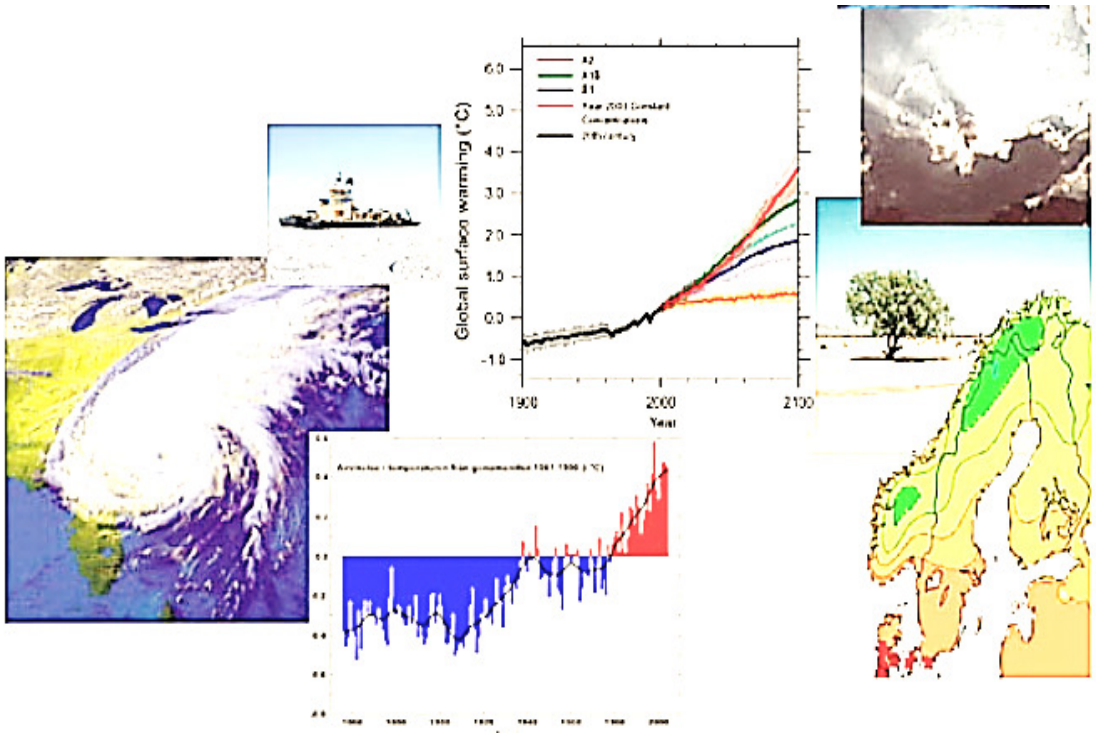
Fernando Sebastião, Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, fsebast@ipleiria.pt



II XORNADA DE USUARIOS DE R EN GALICIA
#Rgal15

Summary

- EOF (PCA) and Extented EOF for Climate Data
- Independent Component Analysis;
- Using ICA for Extended Time series
- Case study



II XORNADA DE USUARIOS DE R EN GALICIA
#Rgal15

(PCA) Empirical Orthogonal Function Analysis

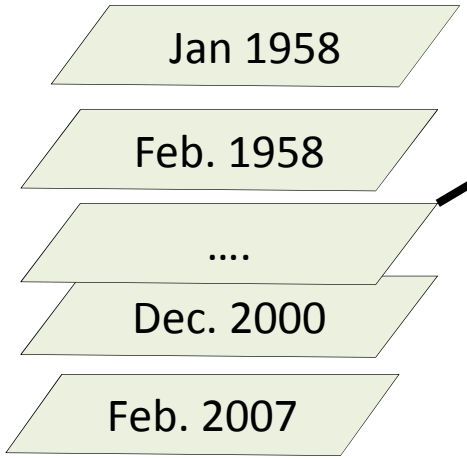
EOF/PC Analysis attempts to find a relatively **small number of linear transformations of initial variables** which convey as much of the original information as possible.

The method is being used **in climatology** for more than 60 years to extract the most valuable information from **larger spatiotemporal datasets** by **reducing the information to a few dominant space-time** patterns.

It is the classic **eigenvalue/eigenvector (or SVD) decomposition** of the correlation (or covariance) matrix of data.

EOF for climate data

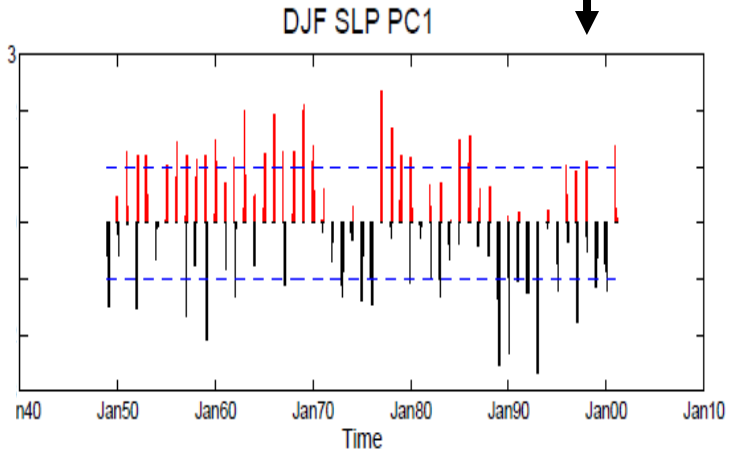
Usually data are time series and associated with **grid points data** (maps).



$$X=\begin{bmatrix} x_{11} & x_{12}\cdots & x_{1p} \\ x_{21} & x_{22}\cdots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2}\cdots & x_{np} \end{bmatrix} \xrightarrow{Cov} S=\begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

EOF Analysis

PCs scores or **PC time series** that are uncorrelated

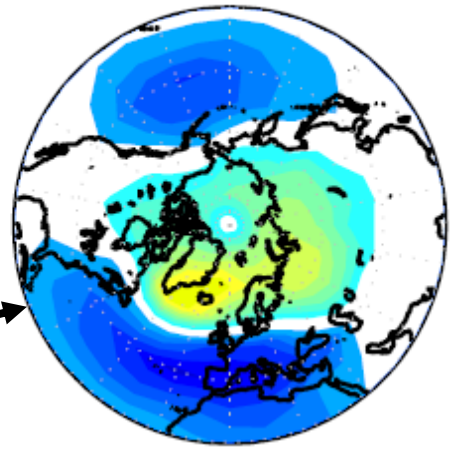


EOFs maps or PC loadings, **eigenvectors**, are the orthogonal set of spatial (maps) patterns (EOFs).

The first few EOFs account for as much as possible of the variation in the original data.

Eigenvalue

DJFSLP EOF1 (21%)



Winter monthly SLP over the Northern H. from 1948-2000 (NCAR/NCEP). **Ref: *Hannachi* A., 2004. A *primer* for *EOF* analysis of climate data**

Extending EOF in climate data. Why?

Since the climate data contain all sorts of features, e.g. stationary and propagating features, and **EOF only uses spatial correlation** of the field and we have **orthogonality in space and time**;

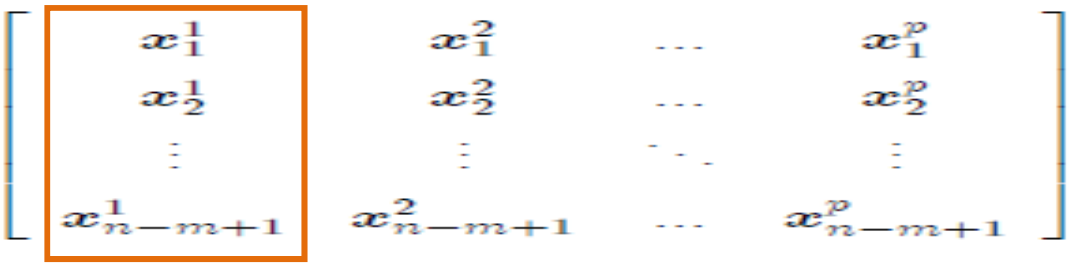
- Discussion about how to understand the ***physical meaning of EOFs.***

Modifications (extensions) of the classical EOF have been introduced to **use both spatial and time information,** in order to identify such propagating features:

e.g. Rotated EOFs (REOFs), **Extended EOFs**, Hilbert EOF, Principal Oscillation Patterns (POPs) ,…., **ICA…**

Extended EOF in climate data

Extended EOFs (MSSA) are simply **multivariate EOFs** in which the additional variables are lagged versions of the same process.
The p "elements" of the data matrix are (m-temporal) vectors.



The *original* matrix **X** is used to obtain a (*n*-m+1)×(*mp*) matrix **X'**, "**augmented matrix of lagged data**", which is significantly larger than the original matrix of data.

Pre-processing data: Previously to EEOF analysis, the data may be subjected to an EOF analysis to reduce its dimension.

Extended EOF problems

We extract information of Covariance matrix S' , where S_{ij} is the lagged covariance matrix up to lag *m*-1, between i^{th} and j^{th} gridpoints.

$$\begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \dots & \mathbf{S}_{1p} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \dots & \mathbf{S}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{p1} & \mathbf{S}_{p2} & \dots & \mathbf{S}_{pp} \end{bmatrix}$$

Interpretations of mp dimension EOFs and PCs time series

If data are describing points in a map then **EEOFs can be considered as m maps**, of dimension p, where propagating behaviour can be studied.

Caution must be used

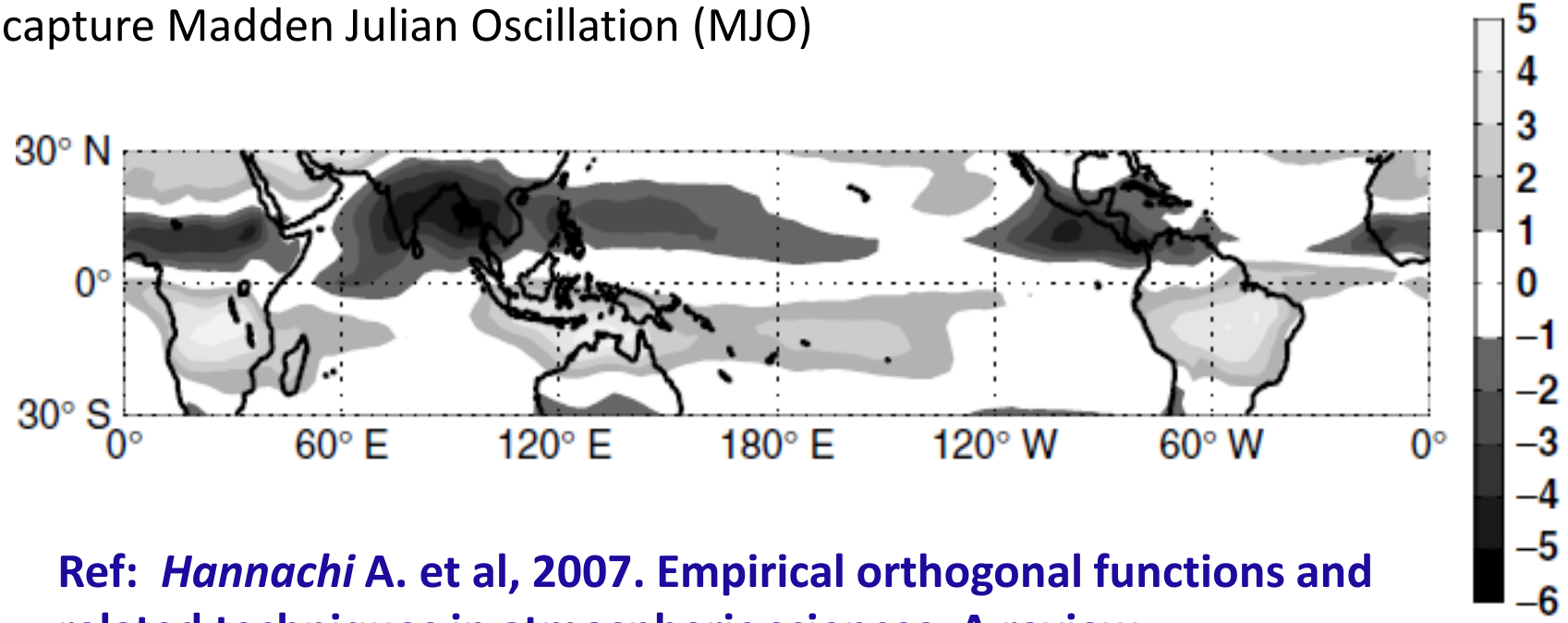
- **when we want to interpret the EEOFs** since **correlation among internal structures are not taken in account**

- **In deciding the m value**, the "best" size of the lagged vector.

EEOF example 1

The Outgoing Longwave Radiation (OLR) anomalies data (NCEP/NCAR) reanalyses over the tropical region from 30 °S to 30°N. Daily data from 1.jan.1996-31.dec. 2002

The leading 10 EOFs/PCs of the anomaly field was used as preprocessing the data to reduce the dimensionality of the data, and EEOF with M=80 days to try to capture Madden Julian Oscillation (MJO)

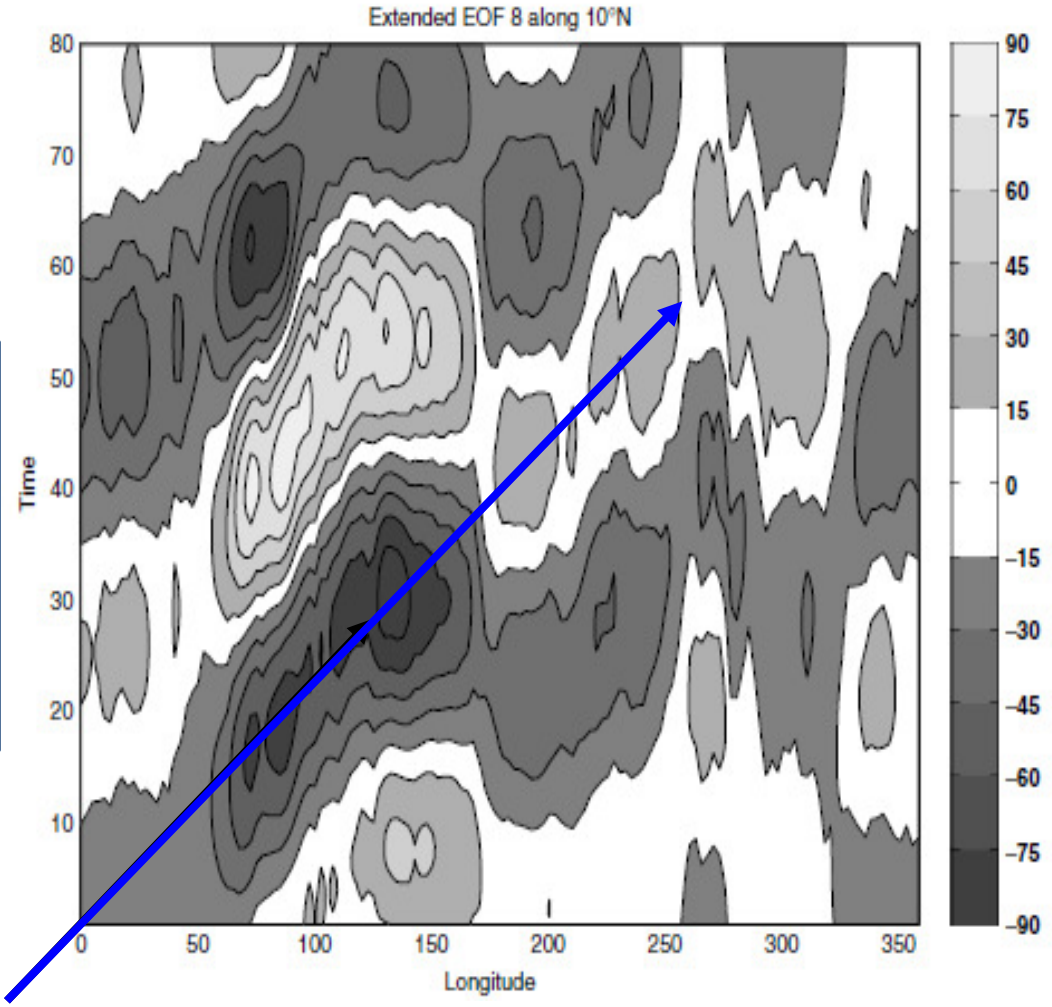


Ref: *Hannachi* A. et al, 2007. Empirical orthogonal functions and related techniques in atmospheric sciences. A review .

EEOF example1- Hovmoller diagram

We may use the EOFs that are linear combination of grid-points to rewrite EEOFs for grid-points.

EEOF 8 in 10 °N as a function of time lag.

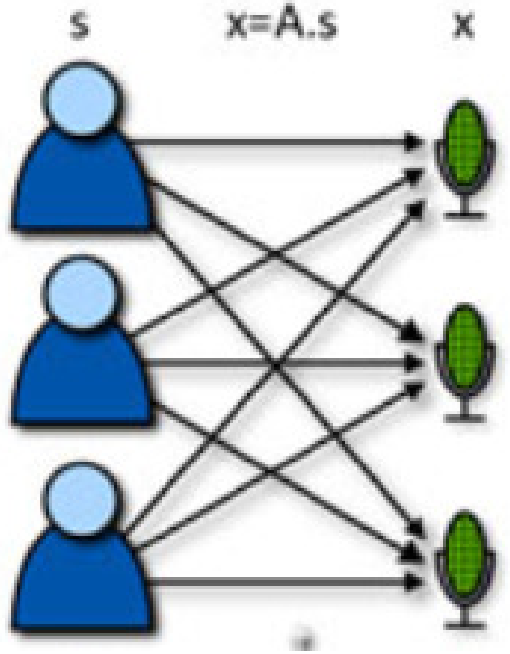


shows the **Madden–Julian oscillation** (MJO): eastward propagating structure with an average phase speed around 100°/23 day.

Independent Component Analysis

The main objective of ICA is **to find hidden components** or factors that relate sets of random variables, signals, time series.

In the model, we assume that those sources are **statistically mutually independent**, which can not be observed directly and are designated independent components.



In a matrix form by **X** = **SA'**:

$\mathbf{X}_{n \times p}$ observed data matrix

$\mathbf{S}_{n \times k}$ *matrix of* ***k*** **independent components**;

$\mathbf{A}_{p \times k}$ *is the* matrix of unknown parameters (columns linearly independent).

We must assume that components s_i are **statistically independent** for $i = 1, \ldots, k$ and at least $k - 1$ components of s_i have **nongaussian distributions.**

ICA ESTIMATION

Since in most applications it is impossible to derive exactly independent sources, ICA methods define **approximate measures of independence/ non-normality** as objective and then search for projections of the observations that optimize those measures:

- **Maximizing nongaussianity**:

Kurtosis, Negentropy and approximations of negentropy are used as measures of optimization of nongaussianity to estimate

- **(or) minimization of entropy-based mutual information**

the entropy-based mutual information is a common measure of independence

. Maximum Likelihood Estimation

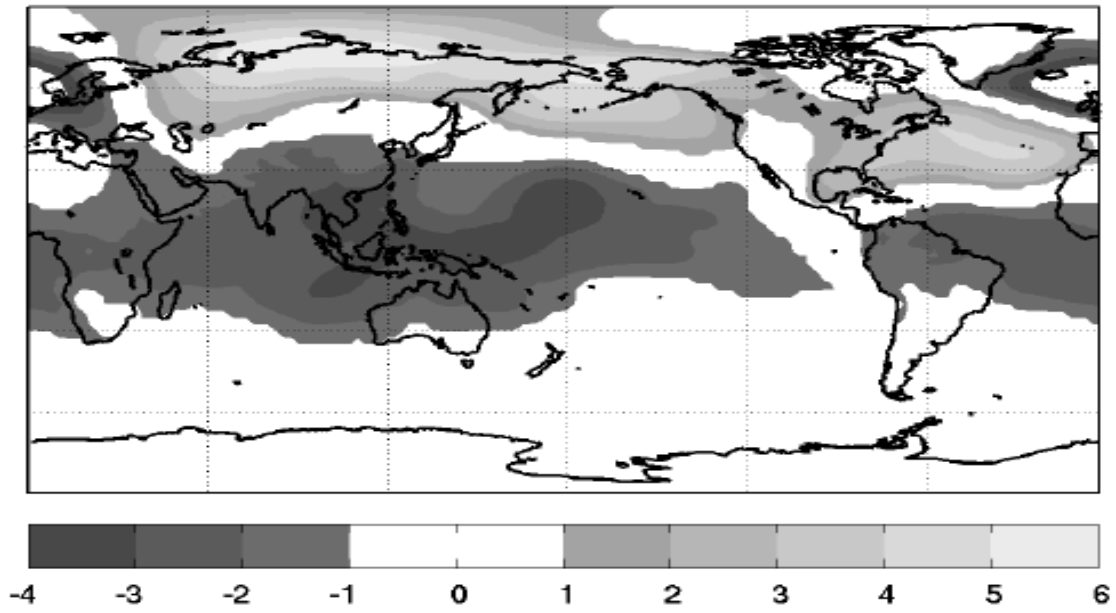
Several algorithms that allow the extraction of ICs (Hyvärinen *et al*., 2001)

- **FastICA** (efficient algorithm , fast convergence)
- AMUSE (when independent components have some temporal dependence)

ICA for Climate data

Independent components (ICs) obtained through rotation of the leading five PCs of monthly means SLP anomalies.

Weighted Monthly SLP anomalies data over the Northern H. Jan 48-Dec 06. Original data from (NCEP/NCAR) reanalysis.



significant correlations, at 1% level, multiplied by 10

Correlation map between the IC4 and the global monthly mean SLP field.

Ref. Hannachi, A., et al, 2009. *ICA of climate data: A new look at EOF rotation.*

➢ Monthly Mean Geopotencial Height

- DATA: 50-years set of the monthly mean geopotential height at 500 hPa, Jan. 58 to Dec. 07. (50*12=600 time values)

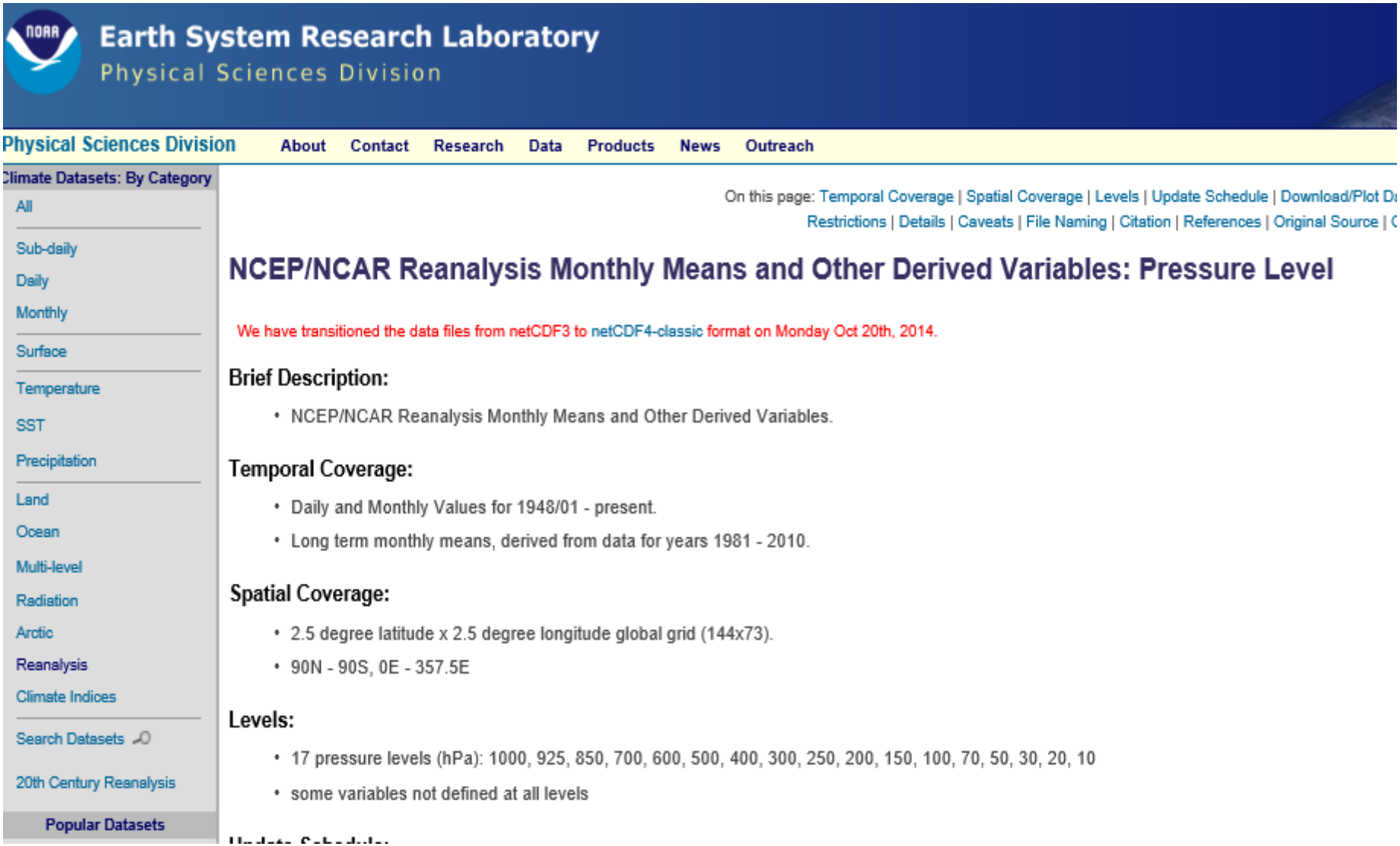
- SECTOR: covering a sector of the Northern Hemisphere for the domain (20^{o}N - 80^{o}N) and (0^{o}E - 357.5^{o}E)

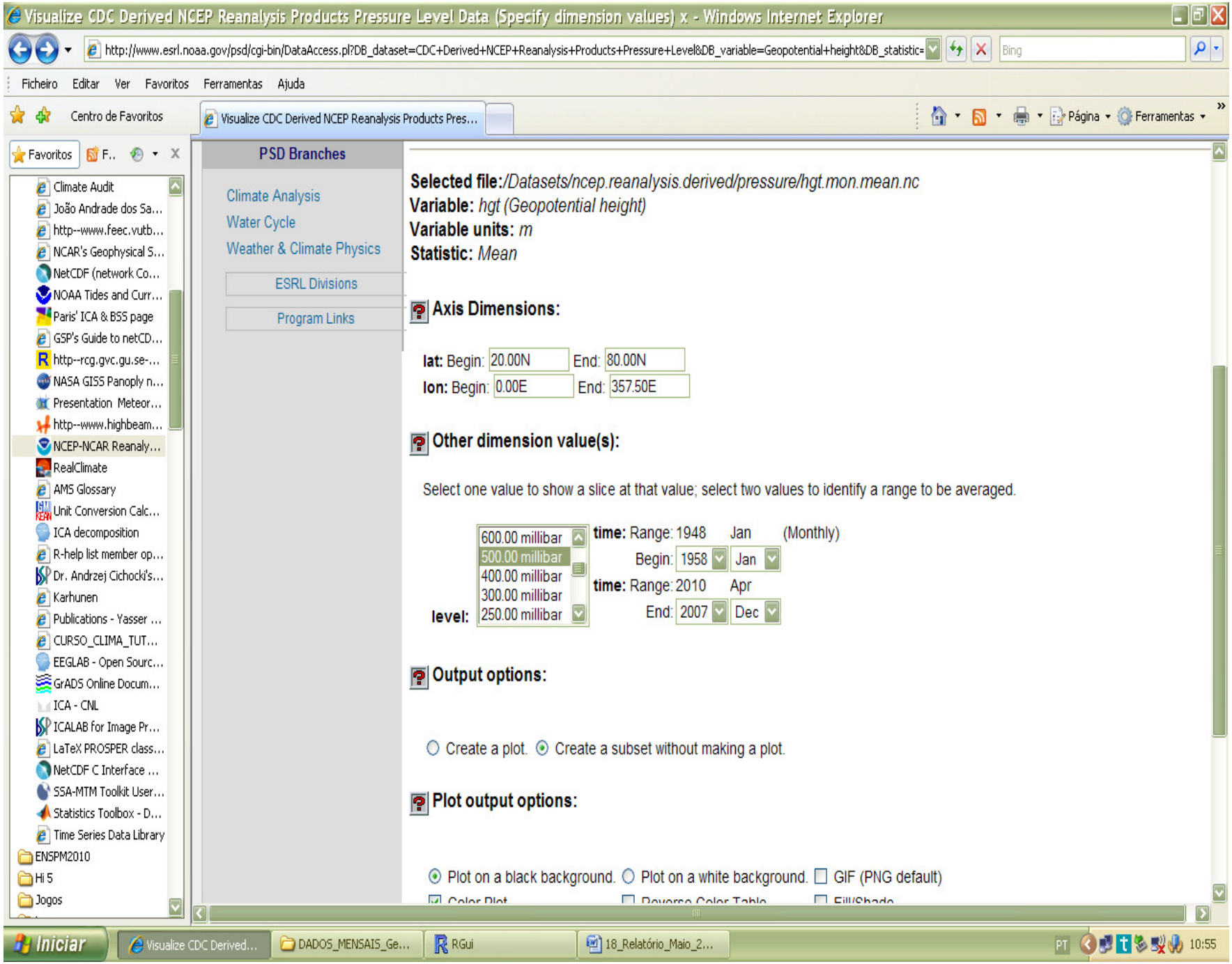
- SPATIAL GRID: an uniform spatial grid of 2.5^{o} in latitude and longitude (25 by 144 =3600 grid points)

- SOURCE: obtained from the NCEP/NCAR reanalysis archives (NOAA). *Provided by Physical Sciences Division, Earth System Research Laboratory, NOAA, Boulder, Colorado, from their Web site at* *http://www.esrl.noaa.gov/psd/*

http://www.esrl.noaa.gov/psd/data/gridded/reanalysis/

http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.derived.pressure.html




Visualize CDC Derived NCEP Reanalysis Products Pressure Level Data (Specify dimension values) x - Windows Internet Explorer
http://www.esrl.noaa.gov/psd/cgi-bin/DataAccess.pl?DB_dataset=CDC+Derived+NCEP+Reanalysis+Products+Pressure+Level&DB_variable=Geopotential+height&DB_statistic=
Bing
Ficheiro Editar Ver Favoritos Ferramentas Ajuda
Centro de Favoritos
Visualize CDC Derived NCEP Reanalysis Products Pres...
Página
Ferramentas
Favoritos
Climate Audit
João Andrade dos Sa...
http--www.feec.vutb...
NCAR's Geophysical S...
NetCDF (network Co...
NOAA Tides and Curr...
Paris' ICA & BSS page
GSP's Guide to netCD...
http--rcg.gvc.gu.se-...
NASA GISS Panoply n...
Presentation Meteor...
http--www.highbeam...
NCEP-NCAR Reanaly...
RealClimate
AMS Glossary
Unit Conversion Calc...
ICA decomposition
R-help list member op...
Dr. Andrzej Cichocki's...
Karhunen
Publications - Yasser ...
CURSO_CLIMA_TUT...
EEGLAB - Open Sourc...
GrADS Online Docum...
ICA - CNL
ICALAB for Image Pr...
LaTeX PROSPER class...
NetCDF C Interface ...
SSA-MTM Toolkit User...
Statistics Toolbox - D...
Time Series Data Library
ENSPM2010
Hi 5
Jogos
PSD Branches
Climate Analysis
Water Cycle
Weather & Climate Physics
ESRL Divisions
Program Links
Selected file: /Datasets/ncep.reanalysis.derived/pressure/hgt.mon.mean.nc
Variable: hgt (Geopotential height)
Variable units: m
Statistic: Mean
Axis Dimensions:
lat: Begin: 20.00N End: 80.00N
lon: Begin: 0.00E End: 357.50E
Other dimension value(s):
Select one value to show a slice at that value; select two values to identify a range to be averaged.
level:
600.00 millibar
500.00 millibar
400.00 millibar
300.00 millibar
250.00 millibar
time: Range: 1948 Jan (Monthly)
Begin: 1958 Jan
time: Range: 2010 Apr
End: 2007 Dec
Output options:
Create a plot. Create a subset without making a plot.
Plot output options:
Plot on a black background. Plot on a white background. GIF (PNG default)
Iniciar
Visualize CDC Derived...
DADOS_MENSAIS_Ge...
RGui
18_Relatório_Maio_2...
PT
10:55

Main packages

http://cran.r-project.org/web/packages/ncdf/index.html

This package provides a high-level R interface to Unidata's netCDF data files

David Pierce- Institution of Oceanography

https://cran.r-project.org/web/packages/RNetCDF/index.html

An interface to the NetCDF file format designed

https://cran.r-project.org/web/packages/RNCEP/index.html

This package contains functions to retrieve, organize, and visualize weather data from the NCEP/NCAR Reanalysis

Michael U. Kemp. **https://sites.google.com/site/michaelukemp/rncep**

https://cran.r-project.org/src/contrib/Archive/clim.pact/

For making climate analysis and downscaling of monthly mean and daily mean global climate scenarios. **Rasmus E. Benestad- http://www.realclimate.org/ and** Norwegian Met. Inst.

https://github.com/metno/esd

https://cran.r-project.org/web/packages/maps/index.html

(Draw Geographical Maps) Richard A. Becker and Allan R. Wilks

https://cran.r-project.org/web/packages/fields/index.html

(**Tools for Spatial Data**)

http://cran.r-project.org/web/packages/climatol/index.html

Jose A. Guijarro (jaguijarro@inm.es)

http://cran.r-project.org/web/packages/fastICA/index.html

FastICA is considered one of the most efficient algorithms, has a fast convergence and uses the classic method of approximating negentropy as a measure of nongaussianity to estimate the sample components).

http://research.ics.aalto.fi/ica/fastica/

http://cran.r-project.org/web/packages/ts/index.html

```
> length(AltGeop.500.MM.1958.2007.Lat20N80N$lon)
[1] 144
> length(AltGeop.500.MM.1958.2007.Lat20N80N$lat)
[1] 25
> length(AltGeop.500.MM.1958.2007.Lat20N80N$tim)
[1] 600 time=50 (yy)*12
```

DATA DIMENSION:(time, level, latitude, longitude)

```
> dim(AltGeop.500.MM.1958.2007.Lat20N80N$dat)
[1]  600    1   25  144
```

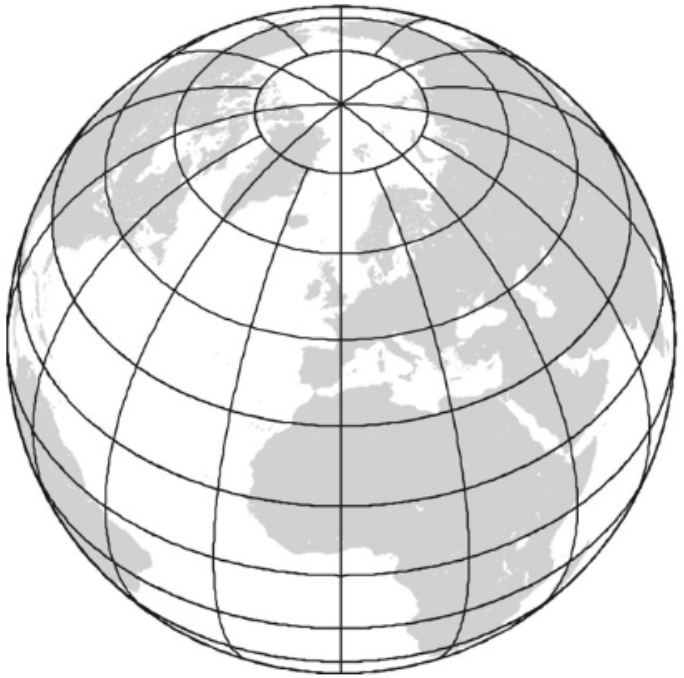
```
> AltGeop.500.MM.1958.2007.Lat20N80N$lon
 [1] -177.5 -175.0 -162.5  165.0  167.5  170.0 ..... [141]  172.5  175.0  177.5  180.0
> AltGeop.500.MM.1958.2007.Lat20N80N$lat
[1] 20.0 22.5 25.0 27.5 30.0 32.5 ..... [18]  62.5 65.0 67.5 70.0 72.5 75.0 77.5 80.0
> AltGeop.500.MM.1958.2007.Lat20N80N$lev
[1] 500          attr(,"unit")     [1] "millibar"
```

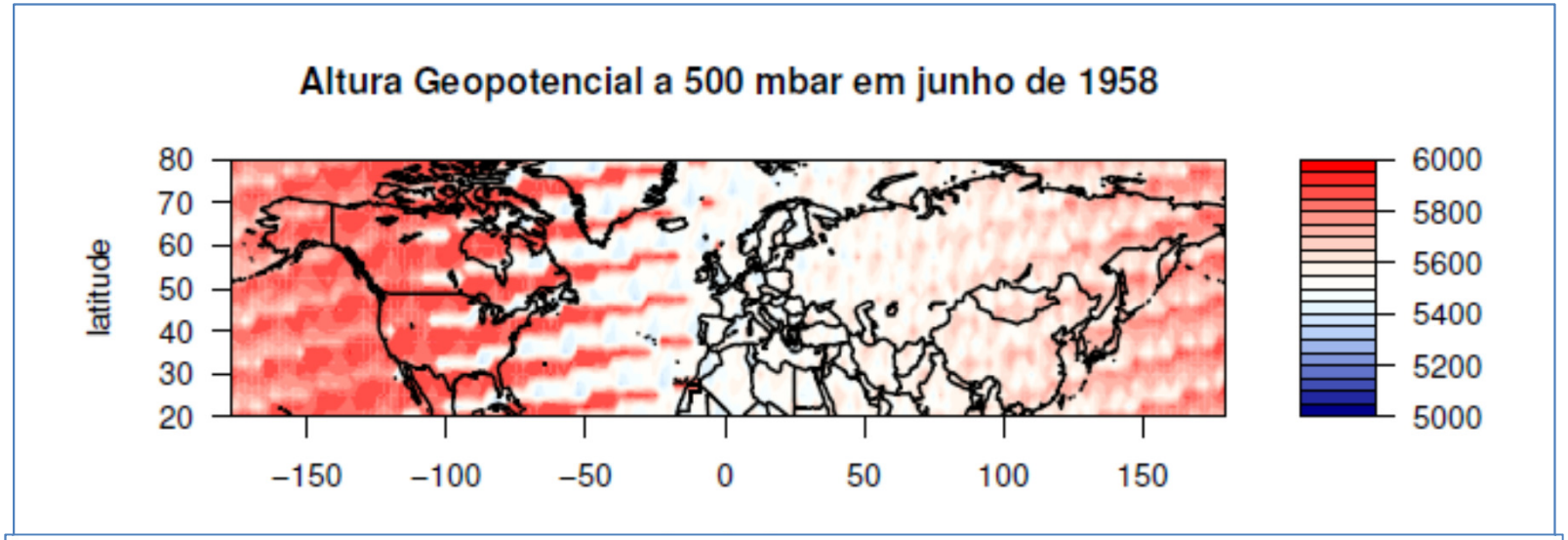
```
MAT2D.AltGeop<-function(X) {
# Function to obtain a matrix 2-D for data #
```

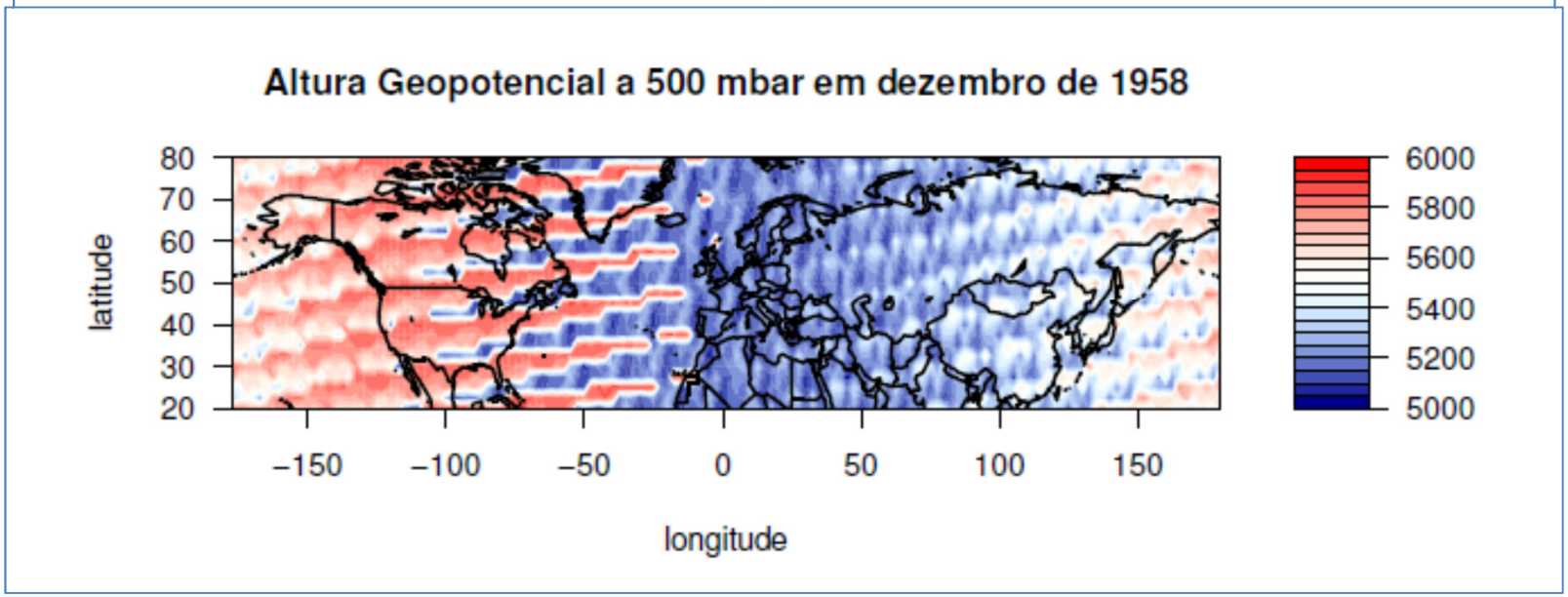
Weighted geopotential anomalies

- First we computed the geopotential **anomalies as departures from the mean annual cycle**.

- Then with the aim to **reduce the effect of high latitude data** that correspond to smaller grid sizes, **an area weighting was applied by multiplying the geopotential anomalies by the square root of the cosine of the corresponding latitude.**





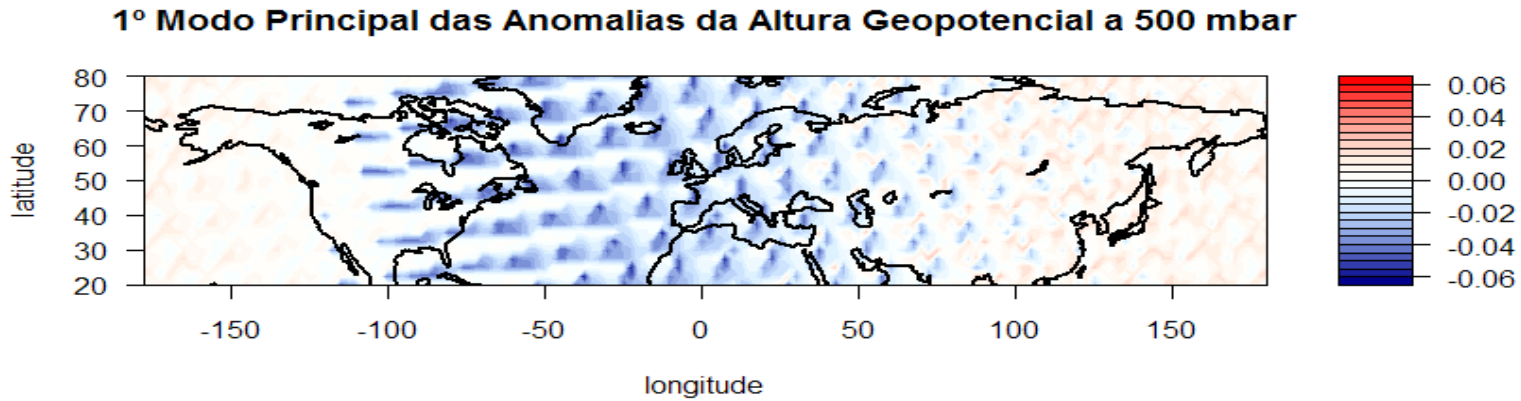


package graphics,
image.plot function and filled.contour,

addland()

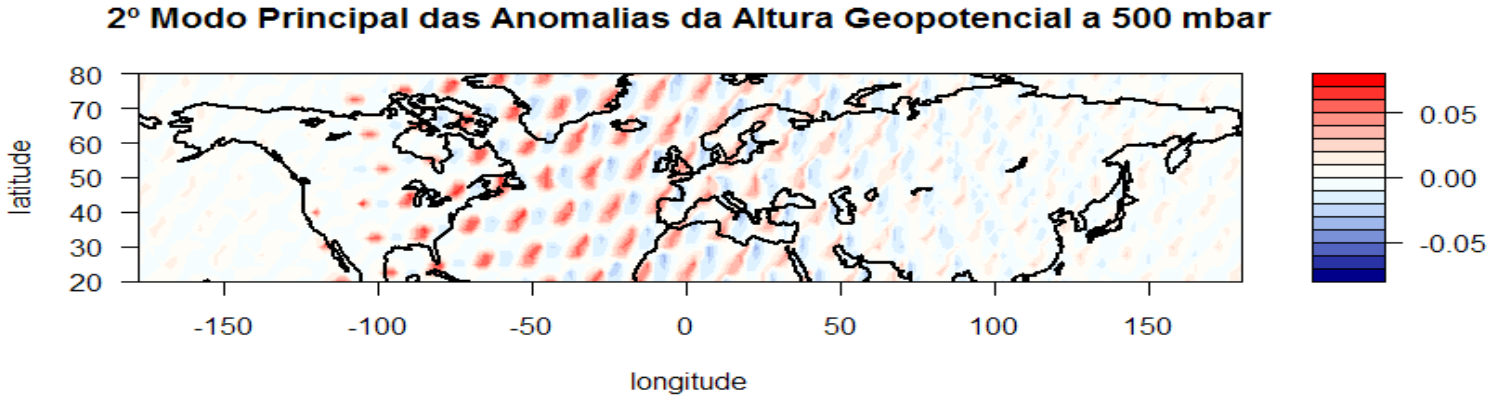
#heatcolors#
require(gplots)
require(RColorBrewer)

Variance explained - 15,5%



Correlation with *North Atlantic Oscillation* (*NAO*) index about 0.62.

Variance explained – 11,2%



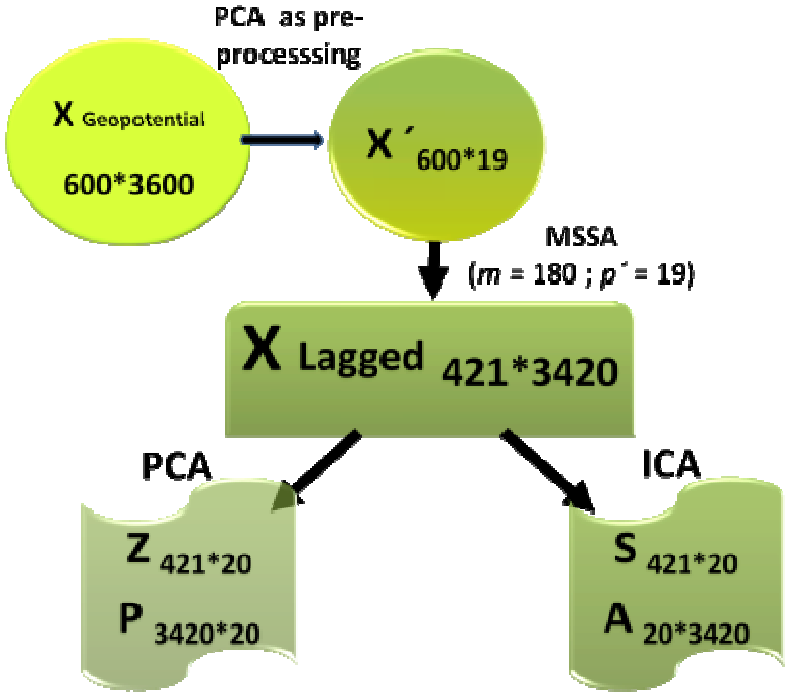
#define our palette#
coltab2<-two.colors(n=256, start="darkblue", end="red", middle="white", alpha=1.0)
coltab3<-two.colors(n=64, start=……)
imageFile2<-image(LON,LAT, **MATRIXPCi**), col=coltab2,
zlim=c(-max(abs(**MATRIXPCi**)*scaleF), max(abs(**MATRIX PCi**)*scaleF))
addland()

METHODOLOGY

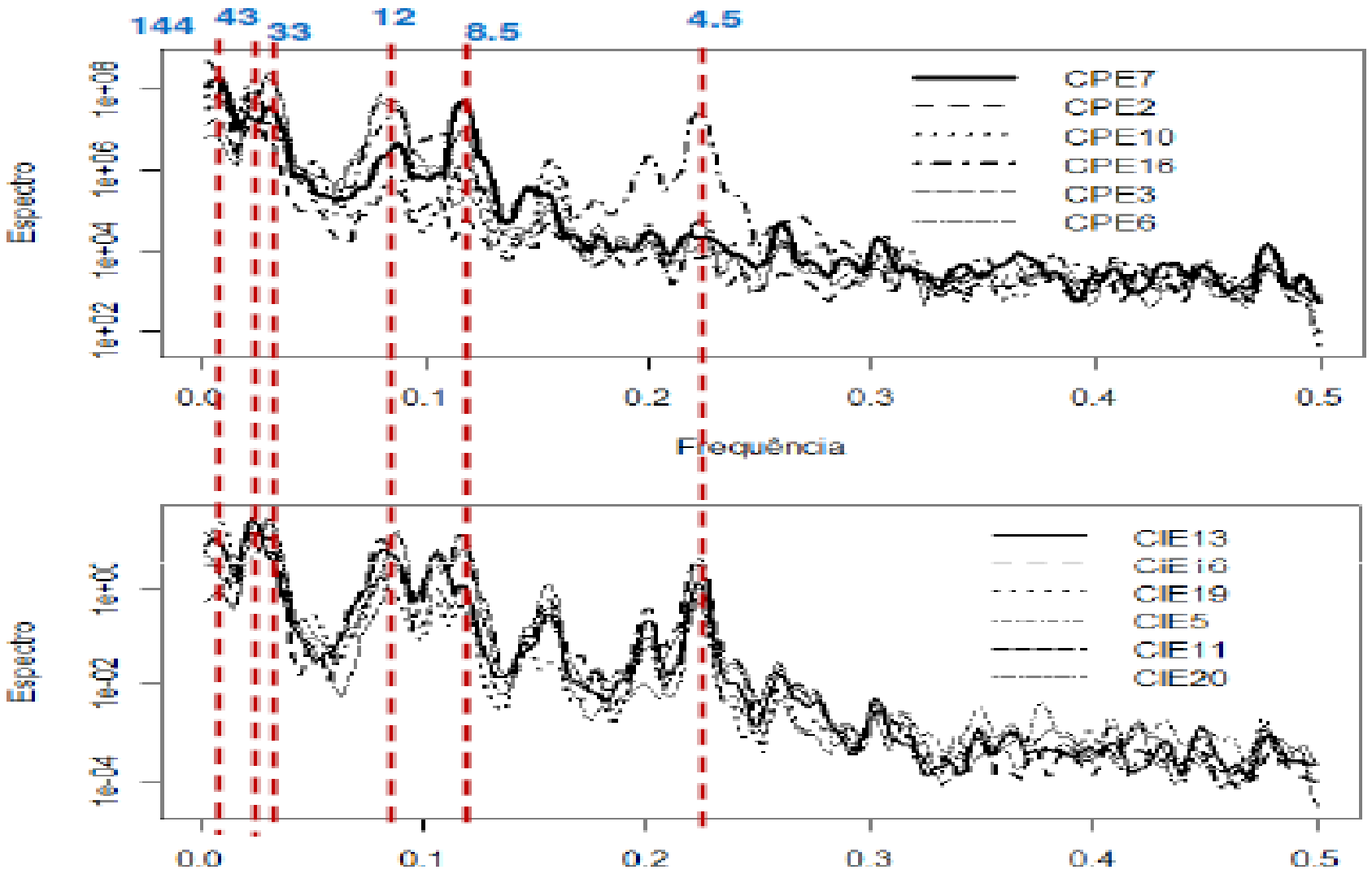
- PCA was used as a pre-processing method of retaining PCs. So the first 19 PCs (in a total of 600 PCs) were retained (which correspond to 87.17% of the explained variability in the data.

- Choice of a lag *m* = 180 months (15 years), to allow the distinction of oscillations with periods in the range (*m*/5, *m*) = (36, 180) (Plaut and Vautard, 1994).

- We decided to consider only final 20 PCs and 20 ICs.



Note that whereas PCs are ranked in descending order of the variance, usually ICs are not sorted out in a specific order.



Comparison between spectra with the PCs and ICs that correspond to the dominant periods of 144, 43 and 33 months after applying PCA and ICA in the matrix of lagged data with a lag m = 180 and 19 channels.

RESULTS

- The application of ICA reveals to have an interesting role as an alternative to the classical PCA (EEOF). Spectral analyses detects fundamentally the same dominant periods of oscillation along 50 years in the geopotential height.
- The main peaks of frequency showed to be coincident and the respective oscillations periods were 33, 43 and 144 months (12 years). **This period of 12 years of oscillation is very interesting, which is close to the 11-years solar cycle**.

Labitzke and Matthes (2005) refer that some solar cycles show high correlations between the 11-year solar cycle and the geopotential among others meteorological parameters, in the lower stratosphere and troposphere.

BIBLIOGRAPHY

1. Chen , J.M. and Harr, P.A. (1993). Interpretation of extended empirical orthogonal function (EEOF). Monthly Weather Review, 121, 2631-2636.
2. Golyandina, N. E., Nekruktin, V. V. and Zhigljavsky, A. A. (2001). Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall, Boca Raton.
3. Hannachi, A., Jolliffe, I. T. and Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: a review. International Journal of Climatology, 27, 1119-1152.
4. Hannachi, A., Unkel, S., Trendafilov, N. T. e Jolliffe, I. T. (2009). Independent componente analysis of climate data: a new look at EOF rotation. Journal of Climate, 22, 2797-2812.
5. Hyvärinen, A., Karhunen, J. and Oja, E. (2001). Independent Component Analysis. J. Wiley & Sons.
6. Jolliffe, I.T. (2002) 2nd ed. Principal Components Analysis. Springer-Verlag, New York.
7. Labitzke, K. e Matthes, K. (2005). Eleven-year solar cycle variations in the atmosphere: observations, mechanisms and models. The Holocene, 13(3), 311-317.
8. **Oliveira, I. (2003). Correlated Data in Multivariate Analysis**. Ph.D. Thesis .Univ. of Aberdeen, U.K.
9. **Sebastião, F. , Oliveira, I. (2013). Independent component analysis for extended TSeries in climate data.** Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications. Studies in Theoretical and Applied Statistics,427-436.
10. von Storch, H. and Zwiers, F.W. (1999). Statistical Analysis in Climate Research. Cambridge University Press, Cambridge.