

**EOF, SVD and SSA Analysis**

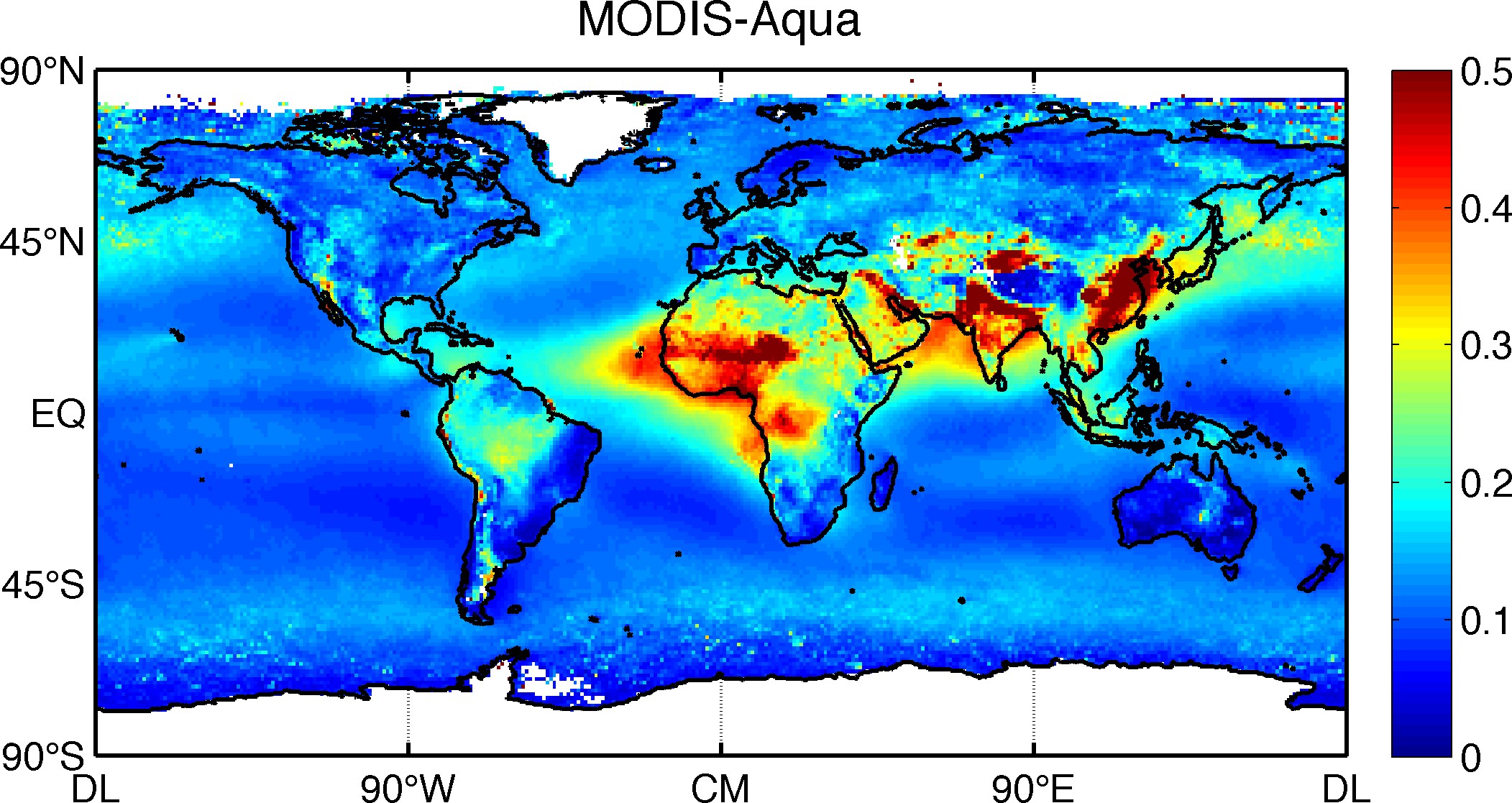
Jing Li PKU Fall 2018

* The atmospheric and oceanic system vary in both **space**

and **time**

* For example, AOD data:

North Africa



0.4

550 nm AOD

0.2

0

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

0.1

AOD Anomaly

0

−0.1

03 04 05 06 07 08 09 10 11 12

Years

0.3

550 nm AOD

0.2

0.1

South America

1

550 nm AOD

0.5

India

0

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

0

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

0.2

AOD Anomaly

AOD Anomaly

0.2

0 0

−0.2

03 04 05 06 07 08 09 10 11 12

Years

−0.2

03 04 05 06 07 08 09 10 11 12

Years

#### How to extract the major variability of the dataset?

* Fourier uses fixed sines and cosines basis
* We need something related to the actual behavior of the data
* Empirical Orthogonal Functions (EOF), a.k.a. Principal Component Analysis (PCA) decompose the data according to its variance
* Organize the data as follows:

⎥

⎡ *x*11

⎢

*x*12

! *x*1*m* ⎤

##### ⎢

**X**  ⎢

##### ⎢

⎢

⎣

*x*21 *x*22 ! *x*2 *m* ⎥

##### ! ! " ! ⎥

⎥

*xn*1 *xn*2 ! *xnm* ⎥

⎦

* *n* is the number of spatial locations and *m* is the number of measurements at each location (usually the length of the time series)
* The mean (sometimes seasonal cycle) is first removed from each row
* Problem is to find the eigenvalues and

eigenvectors of the data covariance matrix:

**R**  1

*n*  1

**XX***T*

* We then find the eigenvalue and eigenvectors of

**R** using eigenvalue decomposition:

## RC = C

*  is a diagonal matrix containing the eigenvalues *i*

of **R**

* The *i*th column vector of **C** is the eigenvector

corresponding to *i*

* Each Ci can be reshaped back to a map
* They are called the EOFs
* Order the EOFs according to the value of the

value of *s*

* The variance explained by each EOF is :

*n*

*Fi*  *i* /

*i*

*i*1

* Each EOF is associated with a time series (Principal Component, PC), found by:

**P = CTX**

* Each row of P is the PC associated with each EOF
* So we can reconstruct the original dataset as:

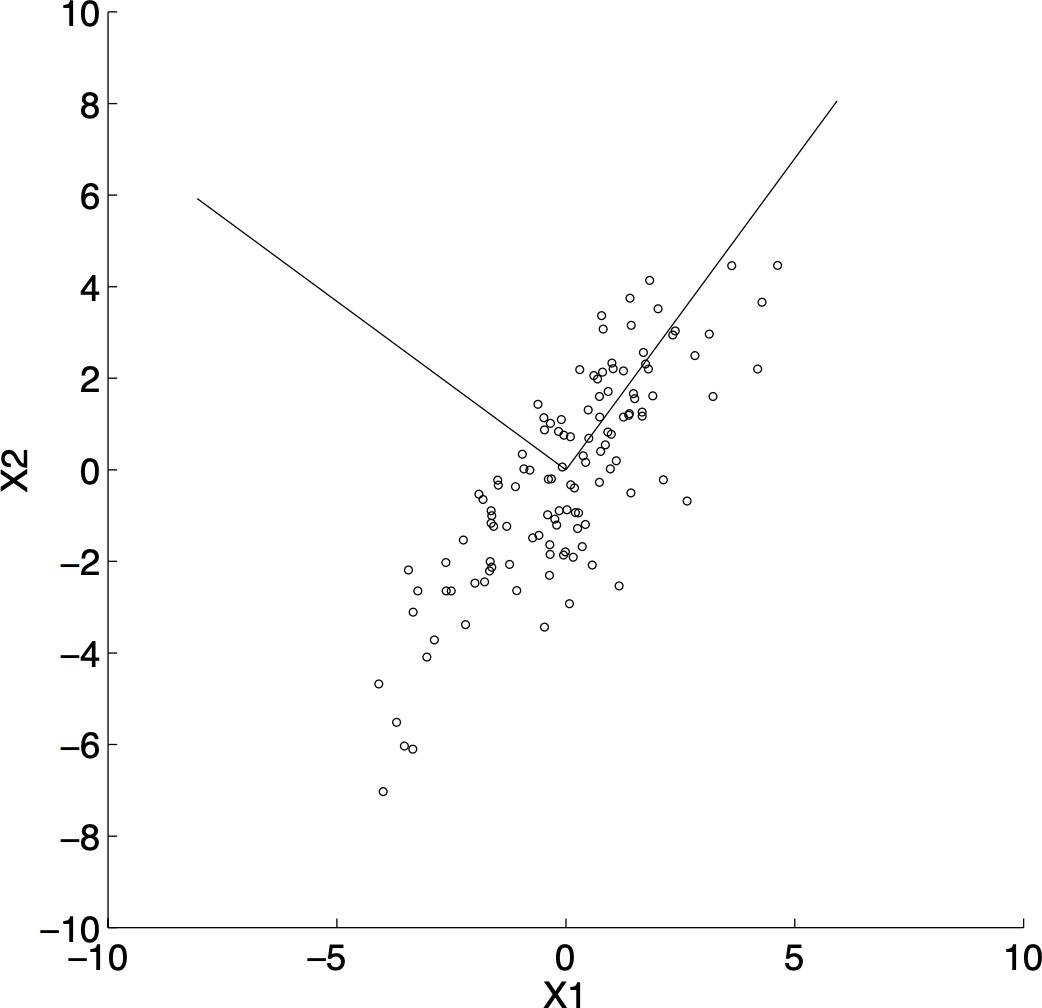
*i i*

**X**  

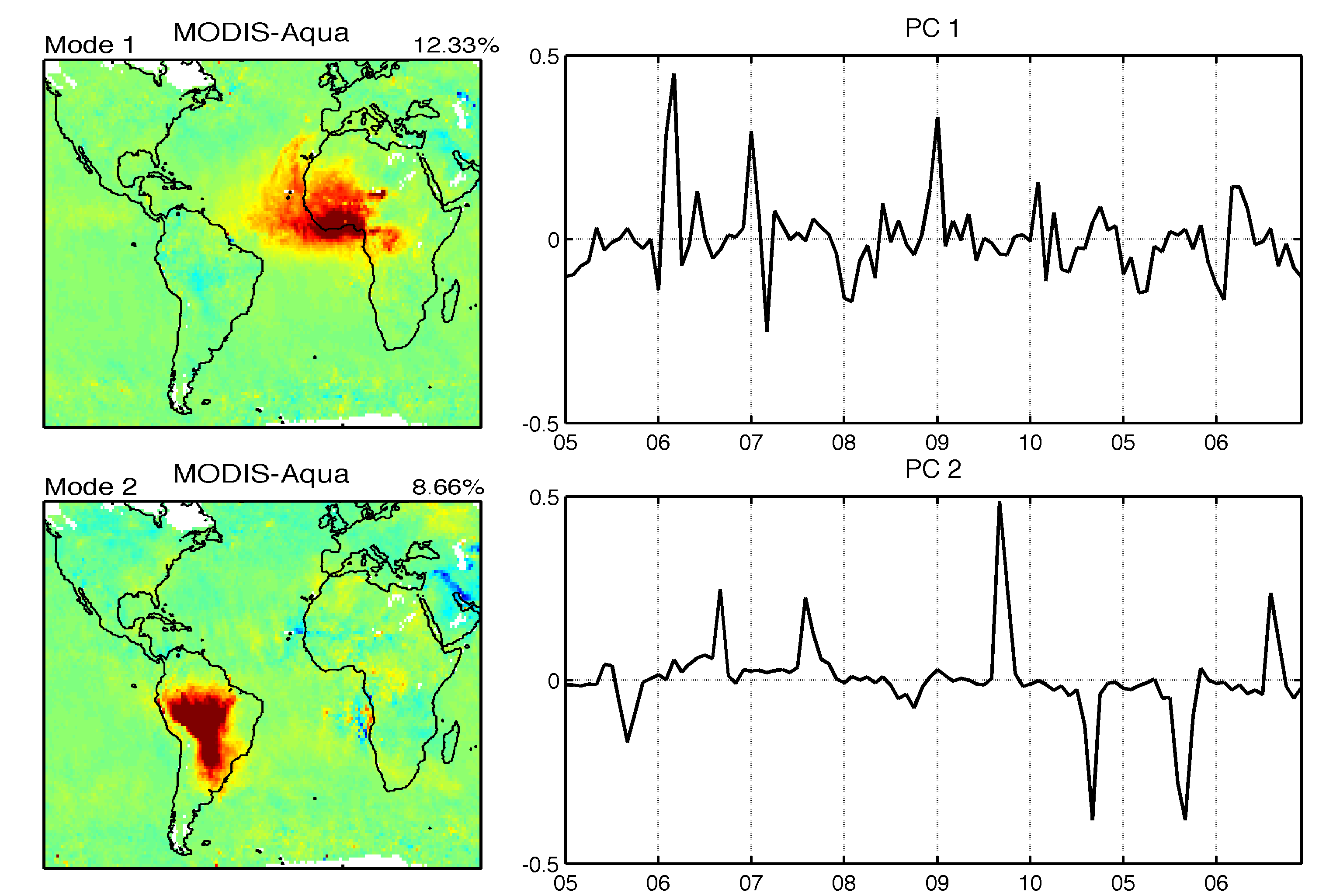
*n*

*i*1

*EOF* \**p*!

* EOF is a dimension reduction technique because the first few modes can usually represent the bulk of the variance
* For better presentation, sometimes it’s helpful to normalize the PCs to unit variance and project the data matrix onto each PC to get the EOFs
* The first eigenvector points down the direction of the most variability
* The second is orthogonal to the first

### EOF Modes of the AOD Dataset



Li et al., 2013

## EOF using the SVD method

* The EOFs and PCs can also be calculated using singular value decomposition (SVD)
* Any *n* by *m* matrix **X** can be decomposed as:

**X = U****VT**

* U contains eigenvectors of **XXT** and V contains eigenvectors of **XTX**
* The columns of U are the EOFs and those of V are

the PCs

* The singular values on the diagonal of  are the

square roots of nonzero eigenvalues of both **XXT**

and **XTX**

### Equivalence of the Two Methods

* Eigenvalue decomposition of covariance matrix **R**:

**R**  1

*n*  1

**XX***T*

**RC = C**

**R = C****CT**

* Substitute **X** with

**X = U****VT**

and calculate **R** as:

**R**  1

*n*  1

**XX***T*  1

*n*  1

**U****VT** (**U****VT** )**T**

= 1

*n*  1

**U****VTV****TUT**  1

*n*  1

**U****TUT**

* + Usually the space dimension is much larger than time dimension, i.e., n>>m
  + The n by n covariance matrix may be too large to handle
  + Decompose the spatial covariance matrix instead:

**L**  1

*m*  1

**XTX**

**LD = D**

* + The eigenvalues of L are also those of R
  + The eigenvectors of L are the PCs
  + The EOFs are found by

**C = XD**

* + Pay attention to the data grids. The data needs to be regularly spaced. If not, interpolate to regular grid
  + Usually needs to weigh each grid by the sqrt(area)
  + Missing data
    - If not many, interpolate to fill the gaps
    - If difficult to interpolate, calculate the covariance specifically for each grid (each element of R)

⎥

⎡ *r*11

⎢

*r*12

! *r*1*n* ⎤

⎢

*R*  ⎢

⎢

⎢

⎣

*r*21 *r*22 ! *r*2 *n* ⎥

" " # " ⎥

⎥

*rn*1 *rn*2 ! *rnn* ⎥

⎦

* + - If too many missing data in a row, remove that row
  + Using the major principal components as regressors
  + Address the multicollinearity problem
  + Procedure:
  + For the linear regression model:

**Y = X****  **

* + **Y** and **X** need to be centered by removing the mean
  + Organize all **X**s into the n by m data matrix and perform PCA
  + Obtain a new regressor basis using the first few major PCs
  + Regression **Y** against these PCs

## How Many Modes?

* + **Visual inspection** of the variance explained curve:

Global Full Dataset

35

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

30



25

% Variance Explained

20

15

10

5

0

2 4 6 8 10 12 14 16 18 20

Mode Number

## How Many Modes?

* + **Rule of thumb by North** (1982): Δ**  **

2

*N* \*

* + If the difference between adjacent eigenvalues is

smaller than Δ** , this mode and higher order

ones are not significant

* + *N\** is the degree of freedom, calculated as:
  + First calculate the lag-1 autocorrelation ****n**

each location

**   1

 

log *n*  max(*n* ) / 2

*N* \*  *n*

2**

for

* + **Monte Carlo:**
    - Generate random data matrices with the same size and variance as the original dataset
    - Calculate the eigenvalues of each random

matrix

* + - For each eigenvalue, only when * j* exceeds

the 95% of the ** values of the random

matrices it is considered significant

#### PCA maximize the variance explained by the first few factors

* + It tends to produce global features that may be difficult to explain
  + For more localized pattern it would help to rotate the factors
  + Rotation usually relax the orthogonal constraint in space but keeps it in time
  + So the modes are still uncorrelated in time
  + Better approximates the simplicity of the structure
  + Simplicity is defined as the square of its loadings (PCs)

2

*V* \*  1

∑*m* *b*2 2

 1 ⎧⎪∑*m b*2 ⎫⎪

*p*  1,2,!, *m*

*jp*

*m*

*j*1

2 ⎨

*m* ⎩ *j*1

⎪

⎭

*jp*⎬⎪

* + *b* are the elements of the rotated PC matrix
  + When *V\** is at maximum, the *b*s tend to be ones or zeros, corresponding to simple structure
  + The normalized Variamax criterion:

*b*

2

*m*∑*M* *b*2 / *h*2 2

*m*

 ⎡ *M* 2

/ *h*2 ⎤

*V*  

*p*1

*jp*

*j*1

*j* ⎢ *jp j* ⎥

⎣ *j*1 ⎦

*m*2

* + *hj* is the amount of the *j*th variables variance

## Combined PCA (CPCA)

* + Sometimes we need to examine the co- variability between two fields
  + One method is to combine them and perform PCA on the combined data matrix:

⎡**X1** ⎤

⎢ ⎥

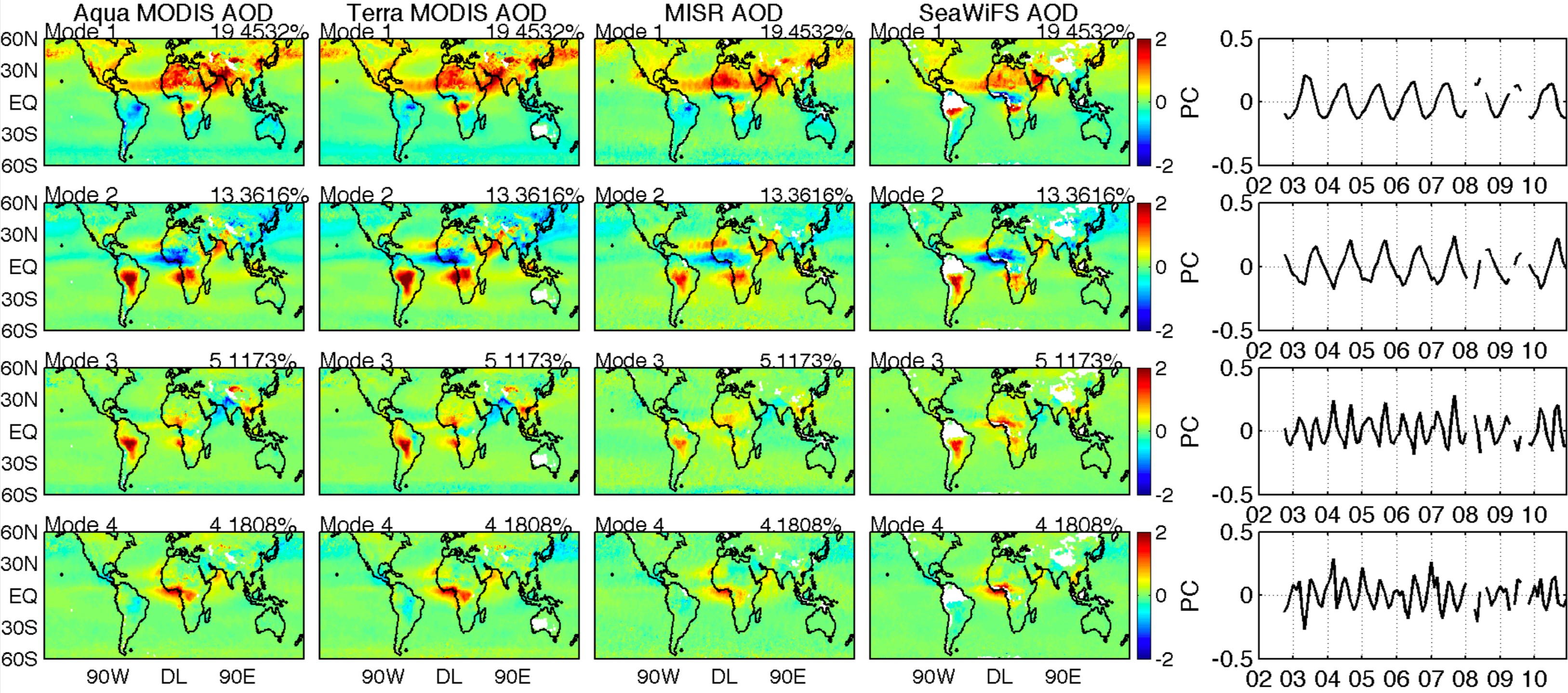
**X =** ⎢! ⎥

⎢**X** ⎥

⎣ **n** ⎦

* + Different fields should have the same time series resolution and length
  + Spatial locations should be similar
  + For different variables usually need to normalize

## CPCA Example



Li et al., 2014

## Presentation of the Results

#### We can direct present the EOFs in their normalized form, but sometimes we also want to show the amplitude

* + Normalize the PCs to unit variance (divide by stand deviation) and regress the original dataset on them
  + The new spatial patterns have the same structure as the EOF but the amplitude show the real amplitude
  + For two fields of different parameters, what is their comment modes of variability?
  + A.k.a. Maximum Covariance Analysis (MCA)
  + Algorithm: the same as PCA except the matrix to decompose is the covariance matrix between the two fields
  + For two datasets X of size n by m and Y of size p by m, the cross covariance matrix is:

**CXY**

 1

*m*  1

**XYT**

#### **Cxy** is of n by p dimension

* + Singular value decomposition of **Cxy**:

**C**  **U****VT**

**XY**

* + Columns of U are the EOFs of X
  + Columns of V are the EOFs of Y
  + The PCs are found by:

**P = UTX P = VTY**

**Y**

**X**

#### Fraction of squared covariance explained is:

*n*

*i*

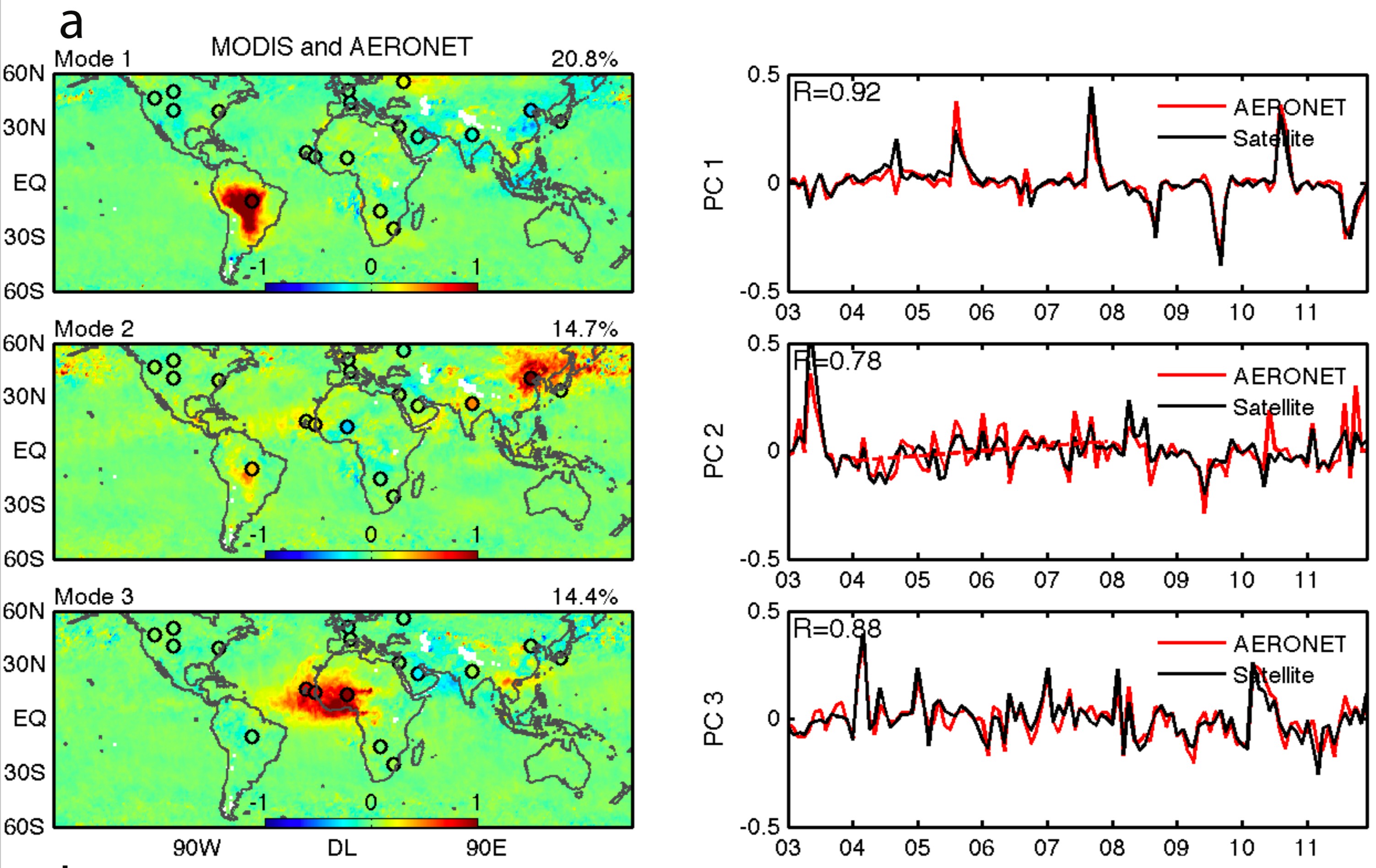
*SCFi*

 *l*2 /

*l*2

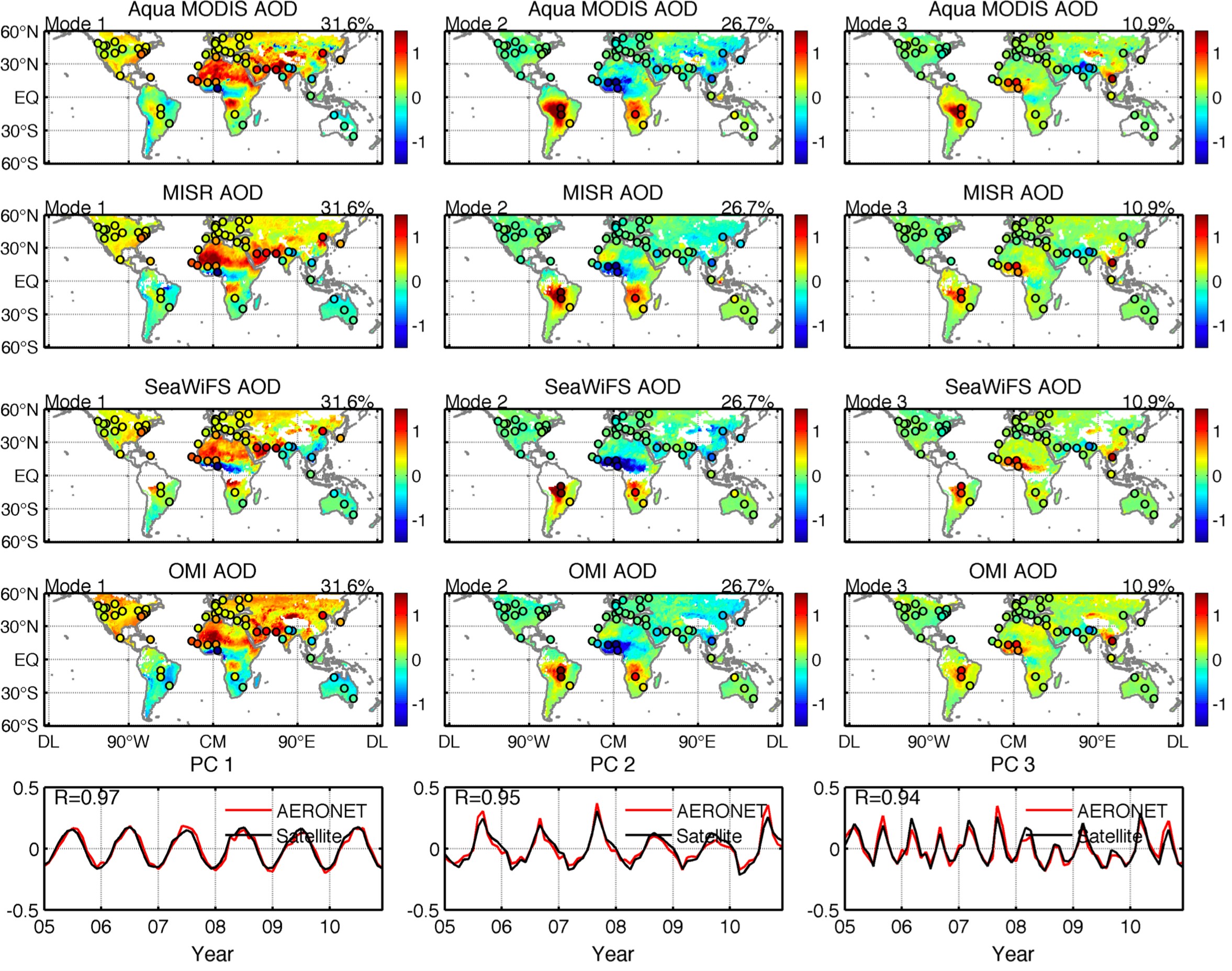
*j*1

*j*

* + SVD between satellite and ground based AOD measurements:

Li et al., 2014

* + Combine CPCA and MCA
  + Do SVD between one combined field (e.g., satellite datasets) and one independent field (e.g., ground based):

Li et al., 2014

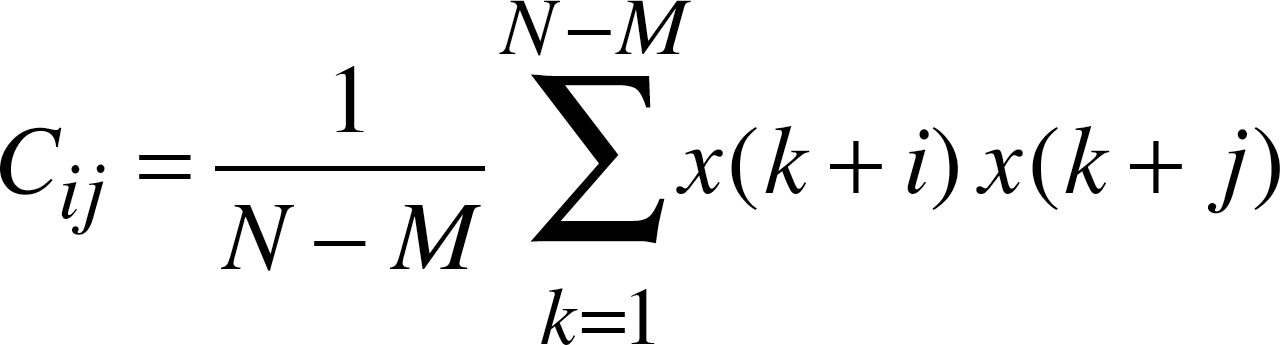
**Canonical Correlation Analysis**

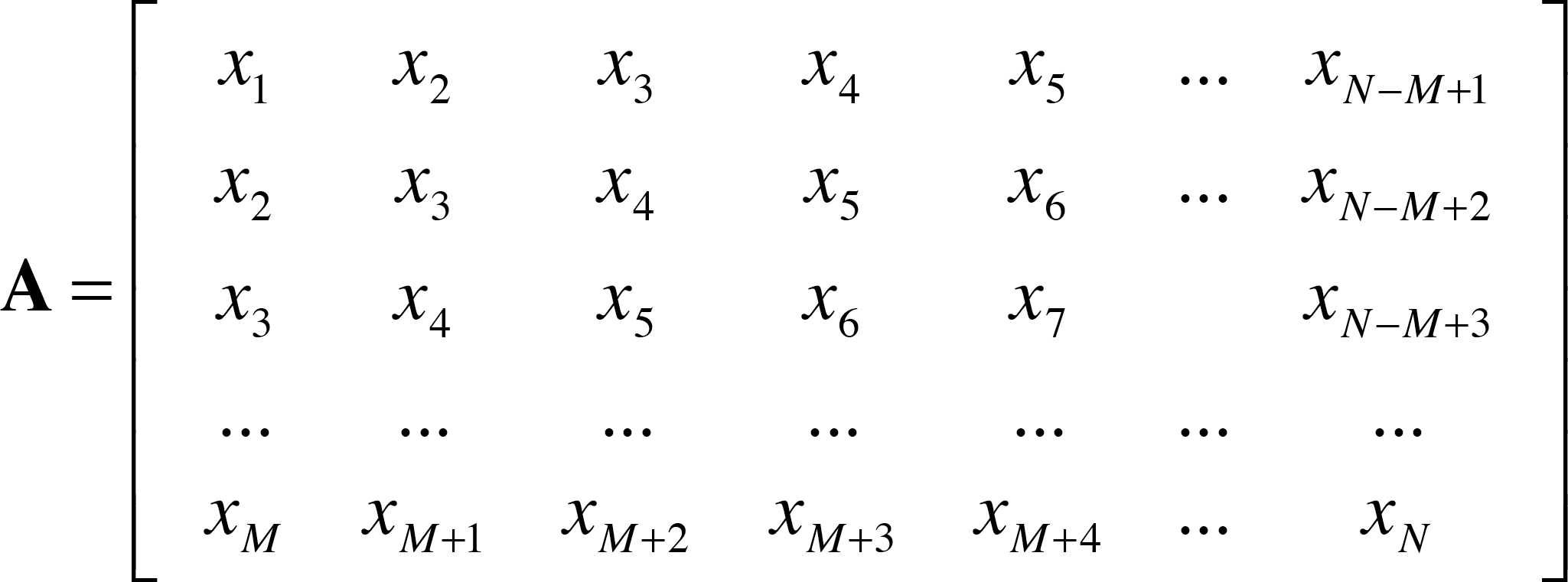
* + MCA or SVD can be influenced by modes with large variance but low correlation
  + In CCA, EOF is first performed, then SVD is performed on the normalized PC matrices
  + Steps:

1. Perform EOF of each of the two datasets
2. Select the major modes to reduce degrees of freedom
3. Normalize the PCs
4. Perform SVD on the two truncated, normalized PC matrices
   * The singular values are called “canonical correlations”

#### For one time series, use empirical function as the basis for the decomposition, instead of predefined functions such as sines and cosines

* + Achieved by doing an eigenanalysis of the lagged covariance matrix



* + Lagged data matrix:

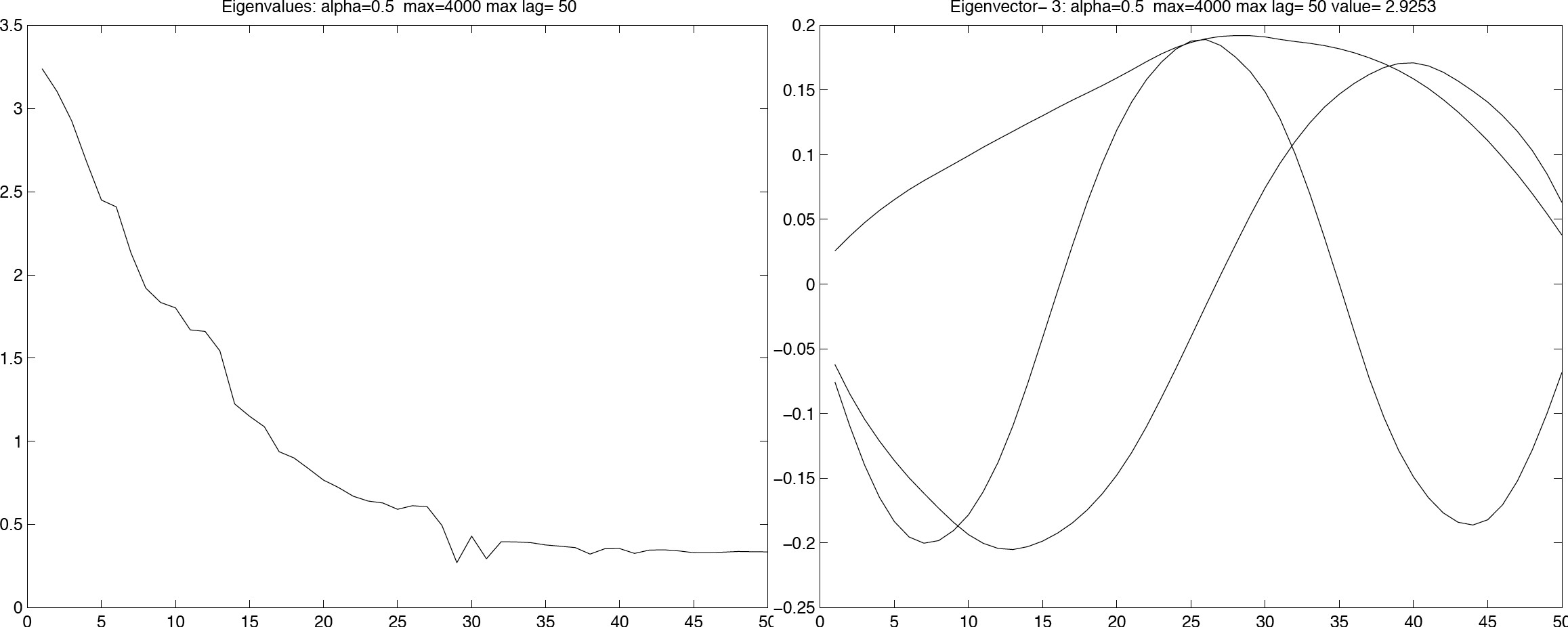
# 







#### SSA example of red noise:



**Matlab Commands**

* + Eigenvalue decomposition: [v,d] = eig(x);
  + Largest eigenvalue and eigenvectors: eigs
  + Singular value decomposition: [u,s,v] = svd(x)
  + Largest singular values: svds
  + EOF rotation: B = rotatefactors(A)