Columbia University
Learning and Empirical Inference – Spring 2007
Term Project

Nonlinear Dimensionality Reduction Applied to climate Modeling

Carlos Henrique Ribeiro Lima

New York – April/2007

Outline

1. Goals
2. Motivation
3. Methodology
4. Results
5. Next Steps

1. Goals

1. Use of kernel PCA techniques (SDE and MVE) to reduce the dimensionality of climate data sets;

2. Draw inferences about the original space based on the behavior of the feature space;

3. Feature space as predictor for other climate variables;

2. Motivation

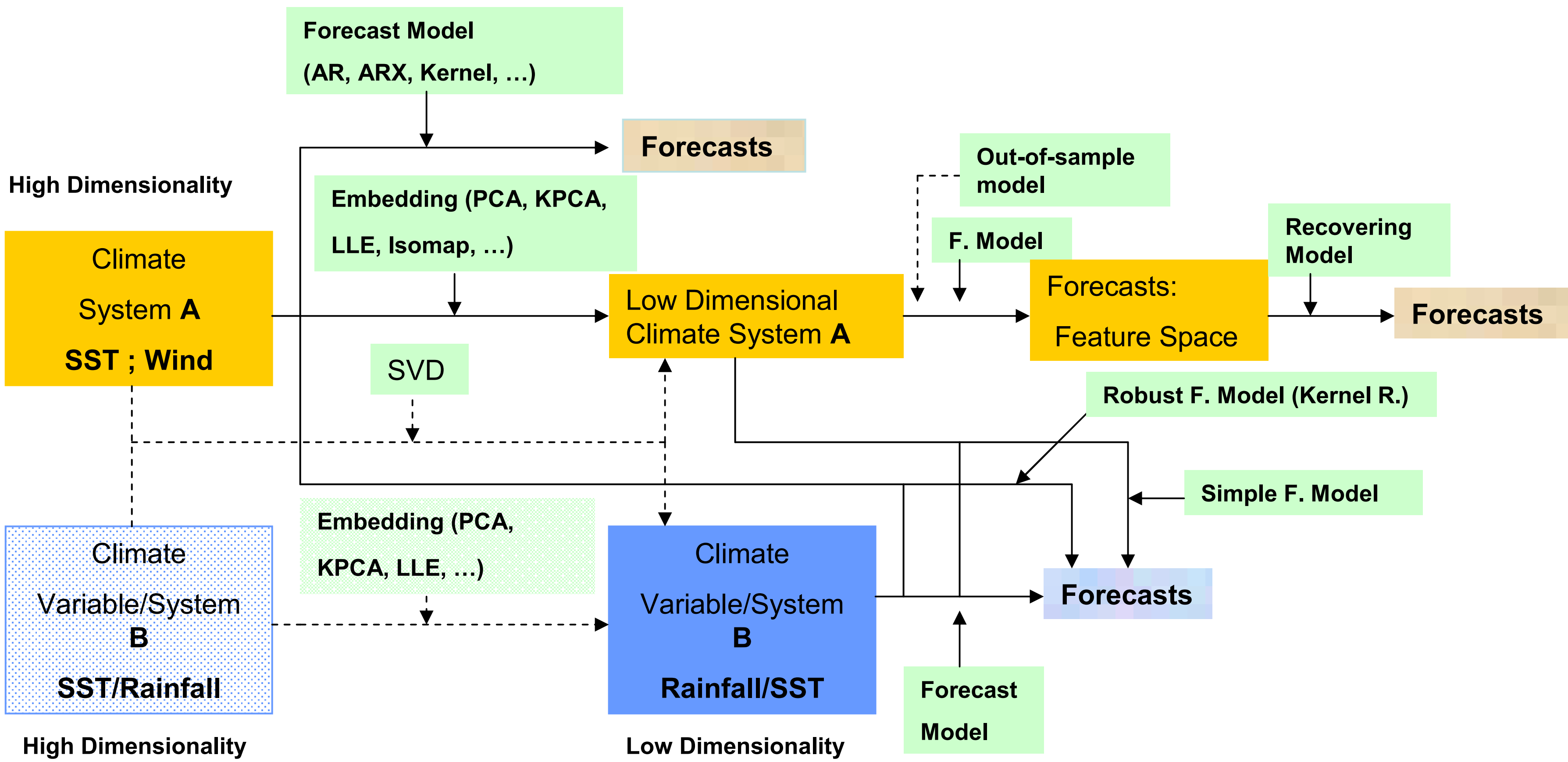
1. Visualization of complex (High dimensional) systems;

2. Needs to represent a multivariate system using just two or three variables → better understanding of the system complexities;

3. Importance of forecasts of key climate variables and phenomena (e.g. El Nino events) for the whole society.

3. Methodology

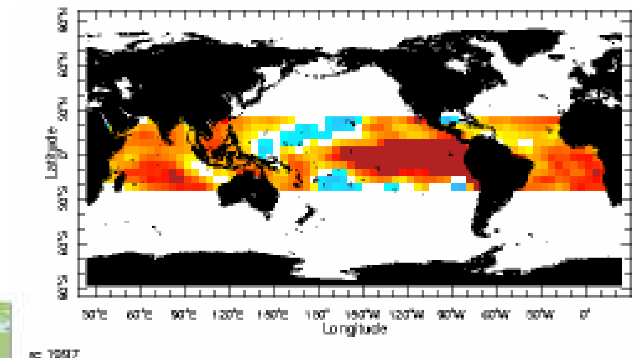
Climate Modeling



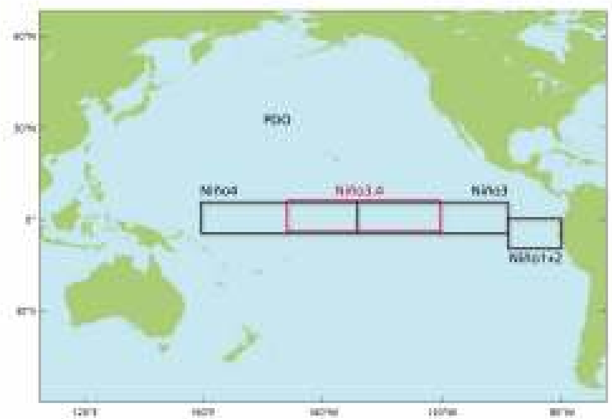
3. Methodology

Climate Variables & Concepts

1. Sea Surface Temperature (SST) →



2. NINO3 index →



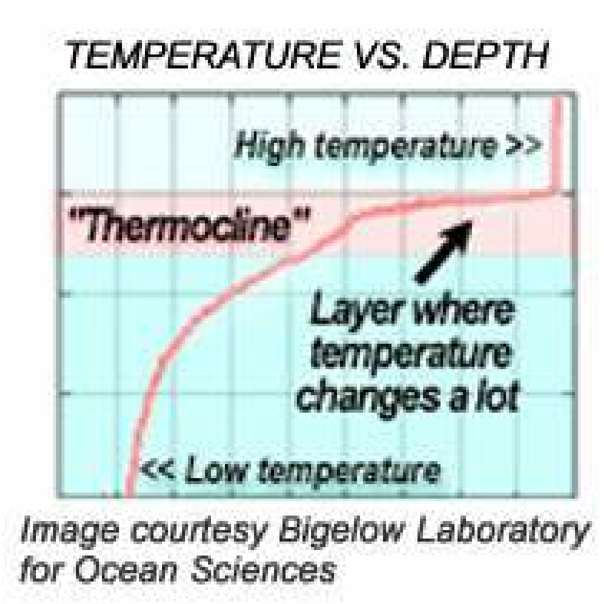
http://ioc3.unesco.org/oopc/state_of_the_ocean/sur/pac/

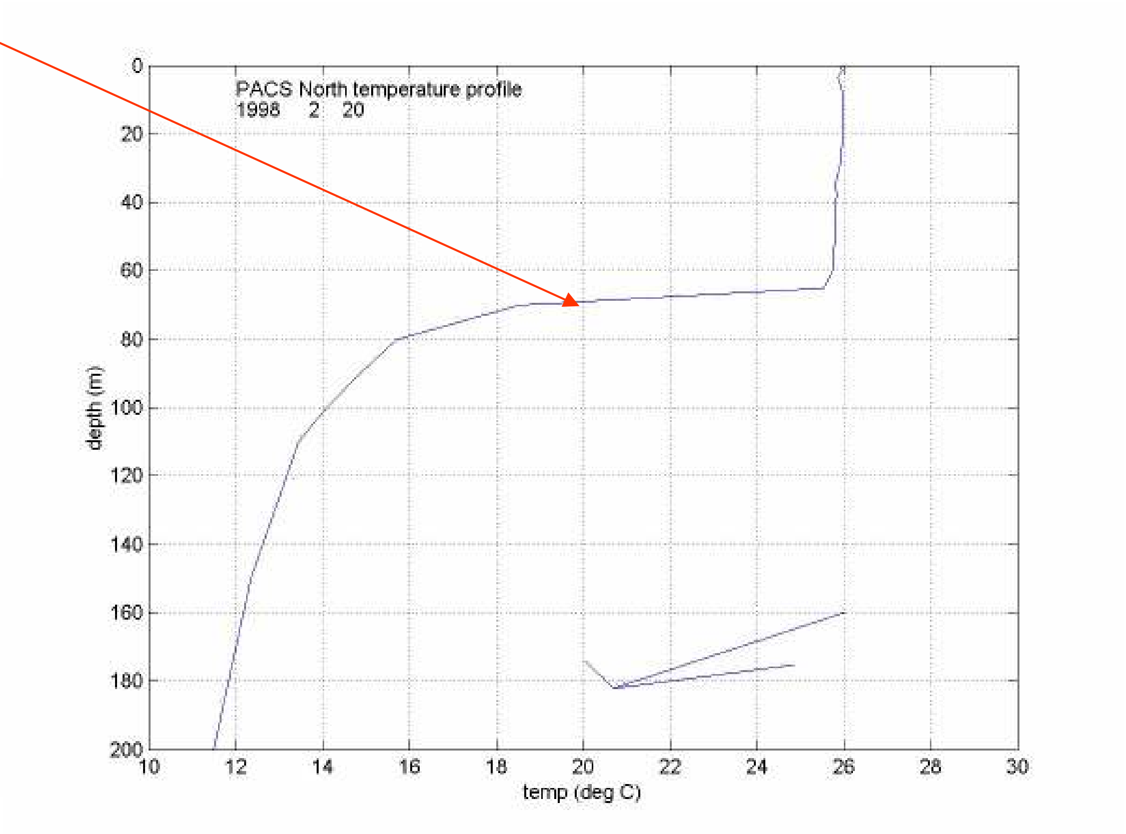
3. El Nino & La Nina Events

3. Methodology

Climate Variables & Concepts

3. Thermocline depth & D20



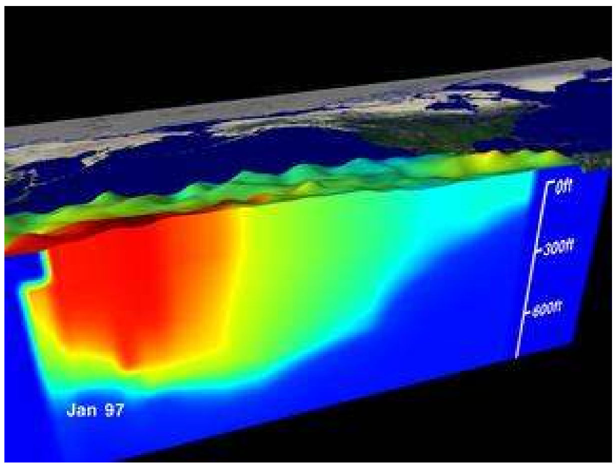


http://web.mit.edu/tomf/www/thcl.htm

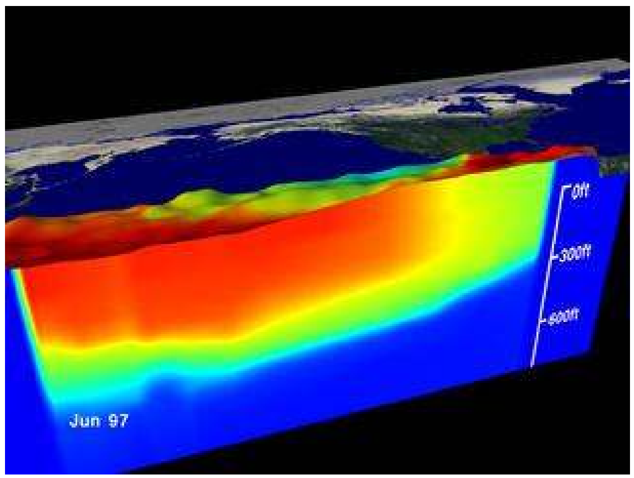
3. Methodology

Climate Variables & Concepts

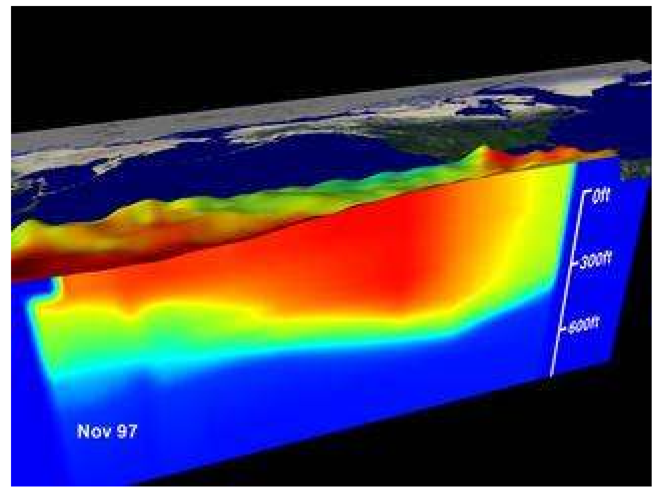
3. Thermocline depth & D20 → Importance for El Nino Events



Jan/1997



Jun/1997



Nov/1997

http://svs.gsfc.nasa.gov/vis/a000000/a000200/a000280/index.html

3. Methodology

1) Semidefinite Embedding (K. Q. Weinberger)

Maximize $\mathrm{Tr}(\mathbf{K})$ s.t.:

$$\mathbf{K} \geq 0. \quad (1)$$

→ **Semipositive definiteness**

$$\sum_{ij} K_{ij} = 0. \quad (2)$$

→ **Inner product centered on the origin**

$$K_{ij} + K_{ij} - K_{ij} - K_{ji} = G_{ii} + G_{jj} - G_{ij} - G_{ji},$$
$$\forall i, j \rightarrow \eta_{ij} = 1 \text{ or } [\eta^T \eta]_{ij} > 0. \quad (3)$$

→ **Isometry - local distances of the input space are preserved on the feature space**

where $G_{ij} = x_i \cdot x_j$ is the Gram matrix of the inputs and $K_{ij} = \Phi(x_i) \cdot \Phi(x_j)$ represents the Gram matrix of the features.

3. Methodology

How to compare the performances of dimensionality reduction methods (e.g. PCA and SDE)?

1) Variance Explained (Eigenvalues) → Quantitative;

2) Forecasts → Quantitative;

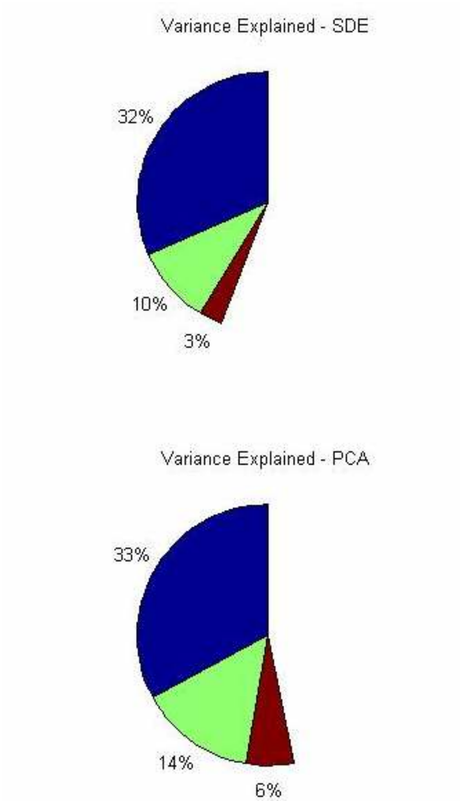
3) Representation of the main physical mechanisms of the climate system → Qualitative;

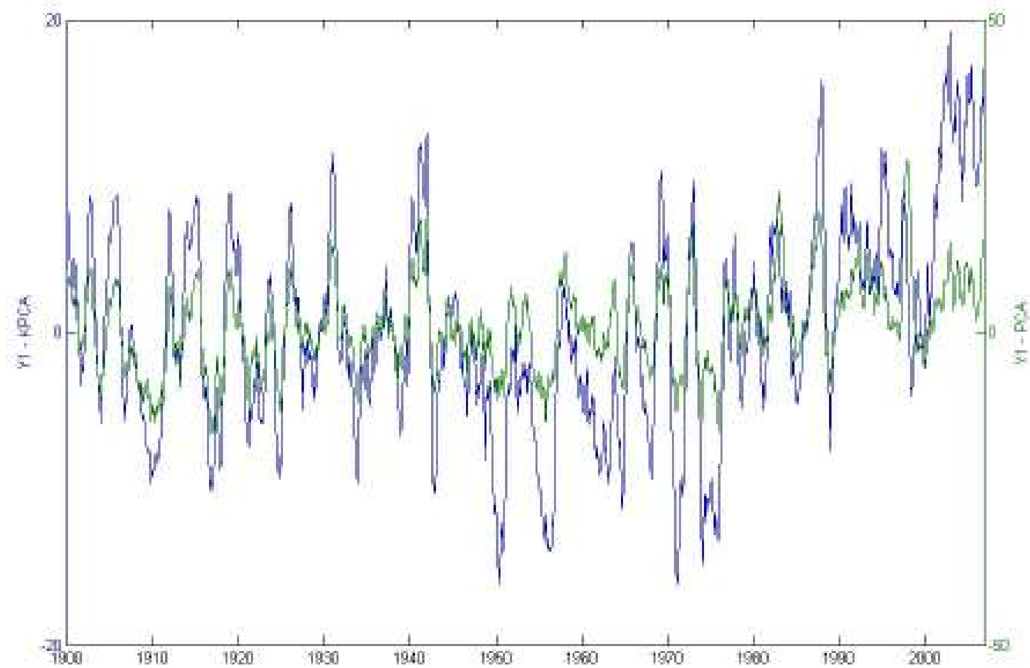
4) Good predictors of other climate variables (e.g. Thermocline system as predictor of the NINO3 index) → Quantitative/Qualitative;

4. Preliminary Results

Problem # 1

SDE applied to SST equatorial field in order to make forecasts for this field (T=1284, d = 599)





4. Preliminary Results

Problem # 1

SDE applied to SST equatorial field in order to make forecasts for this field

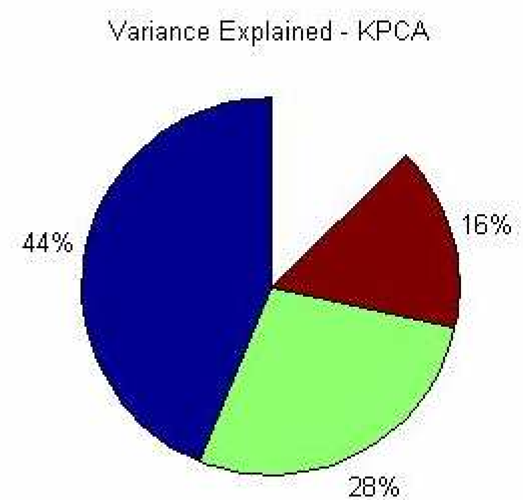
Some conclusions

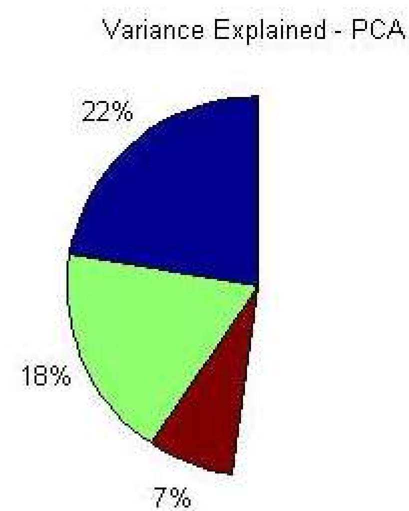
1. Y1 is high correlated with nino3 index for both PCA and SDE;
2. Almost same amount of variance captured by PCA and SDE;
3. High correlations among Y's from PCA and SDE;
4. Class forecasts (KNN) of nino3 give similar results for PCA and SDE;
5. System might behavior like a linear one (many authors agree with that);
6. Quantitative forecasts of the SST field have not been performed yet → Is there any advantage in using SDE (↑non-linearity ↓out-of-sample + recovering models) instead of PCA (↑original space ↓linear) ?

4. Preliminary Results

Problem # 2

SDE applied to the Pacific Thermocline Depth (T=326, d=4561) → Resulting feature space used as predictor for the nino3 index

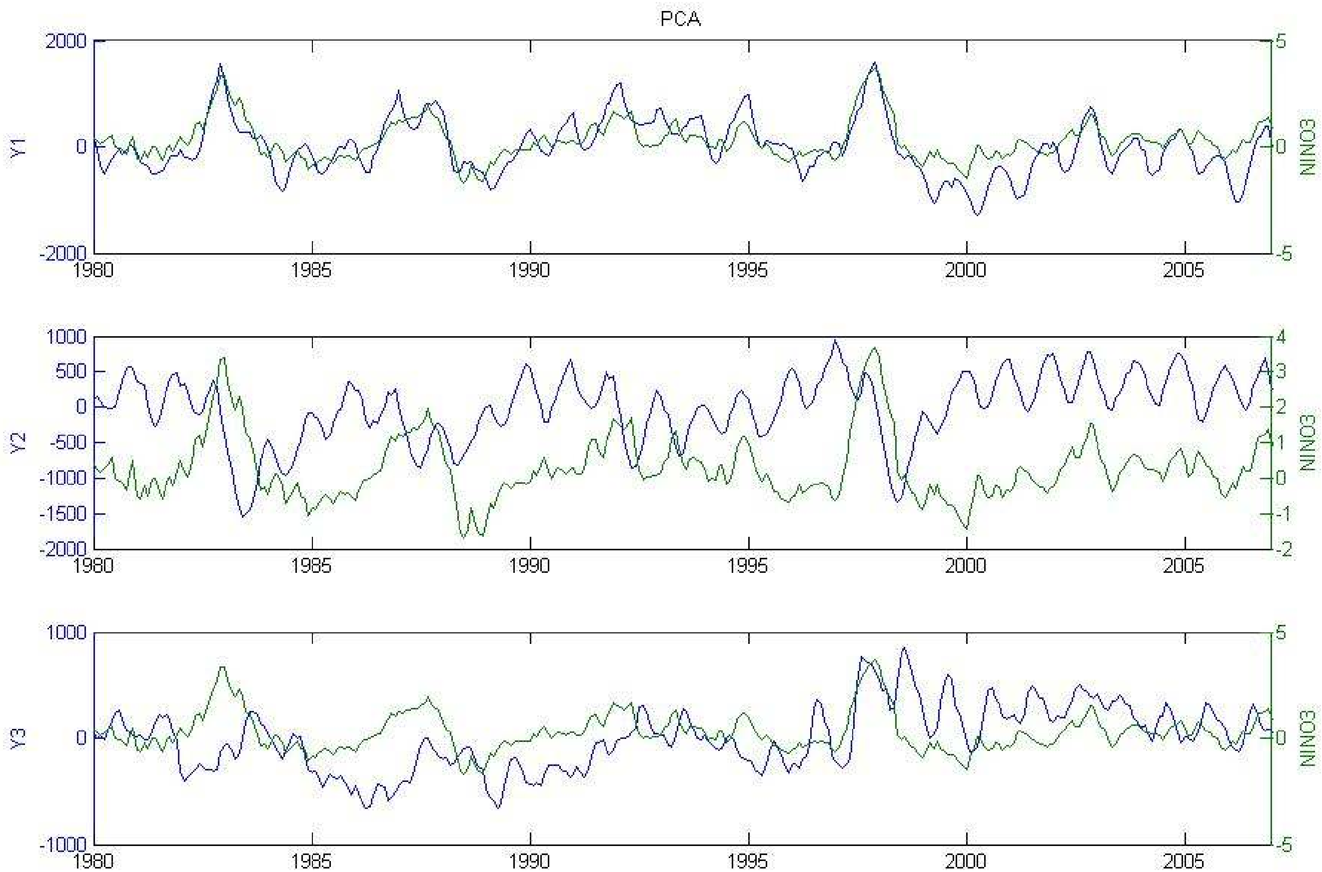




4. Preliminary Results

Problem # 2

PCA - Y's versus nino3



→ High correlated

Drosdowsky (2006) + many others

→ Some Lagged correlation ~ 9 months

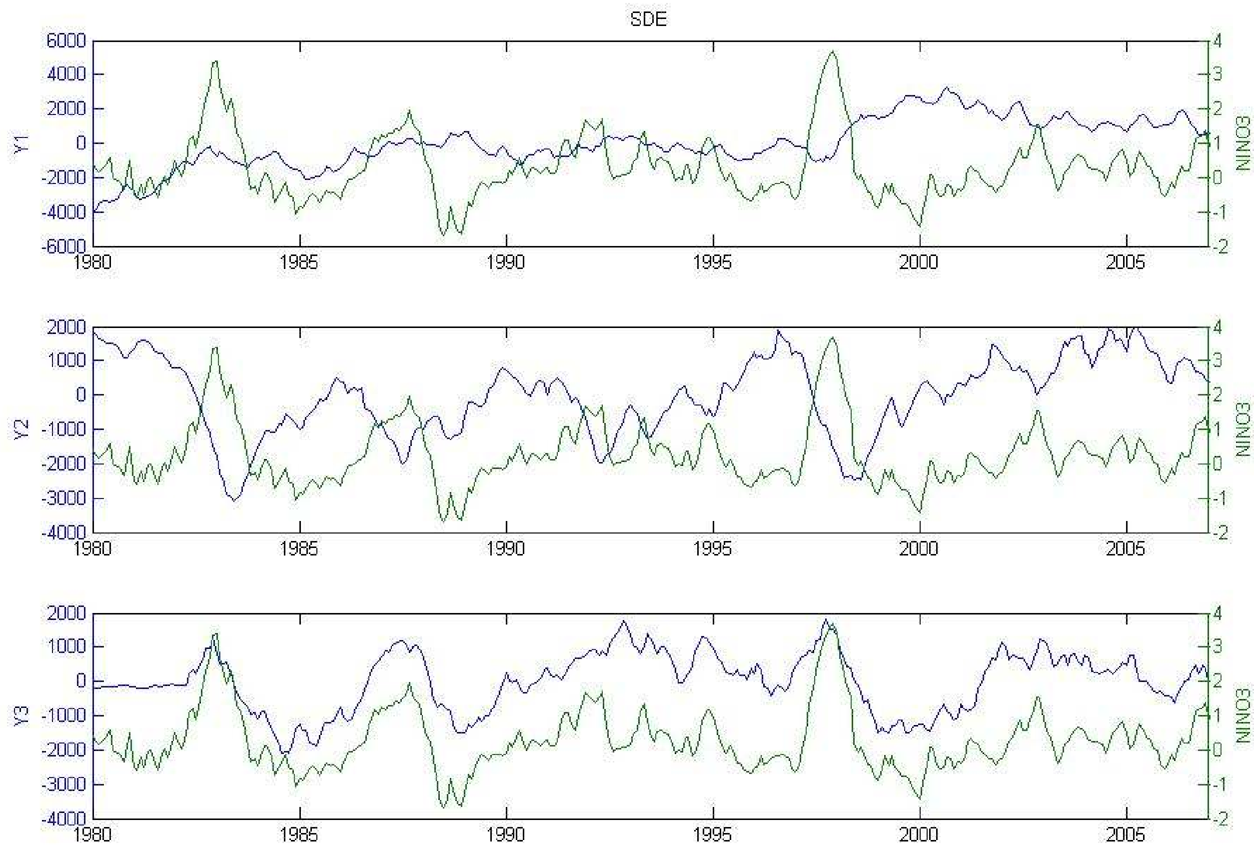
Drosdowsky (2006) + many others

→ Nothing interesting

4. Preliminary Results

Problem # 2

SDE - Y's versus nino3



→ Nothing interesting ?

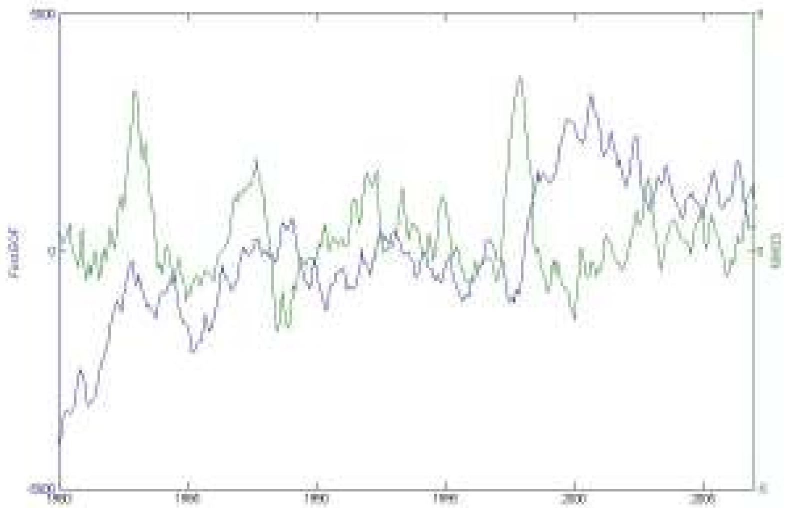
→ Some Lagged correlation ~ 18 months

→ High correlated

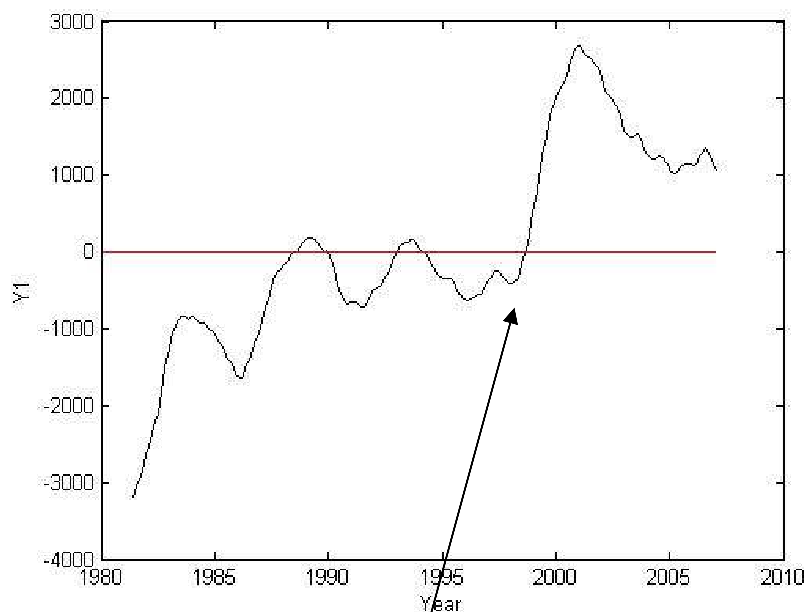
4. Preliminary Results

Problem # 2

Nothing interesting in SDE-Y1?



18-month low-pass filter



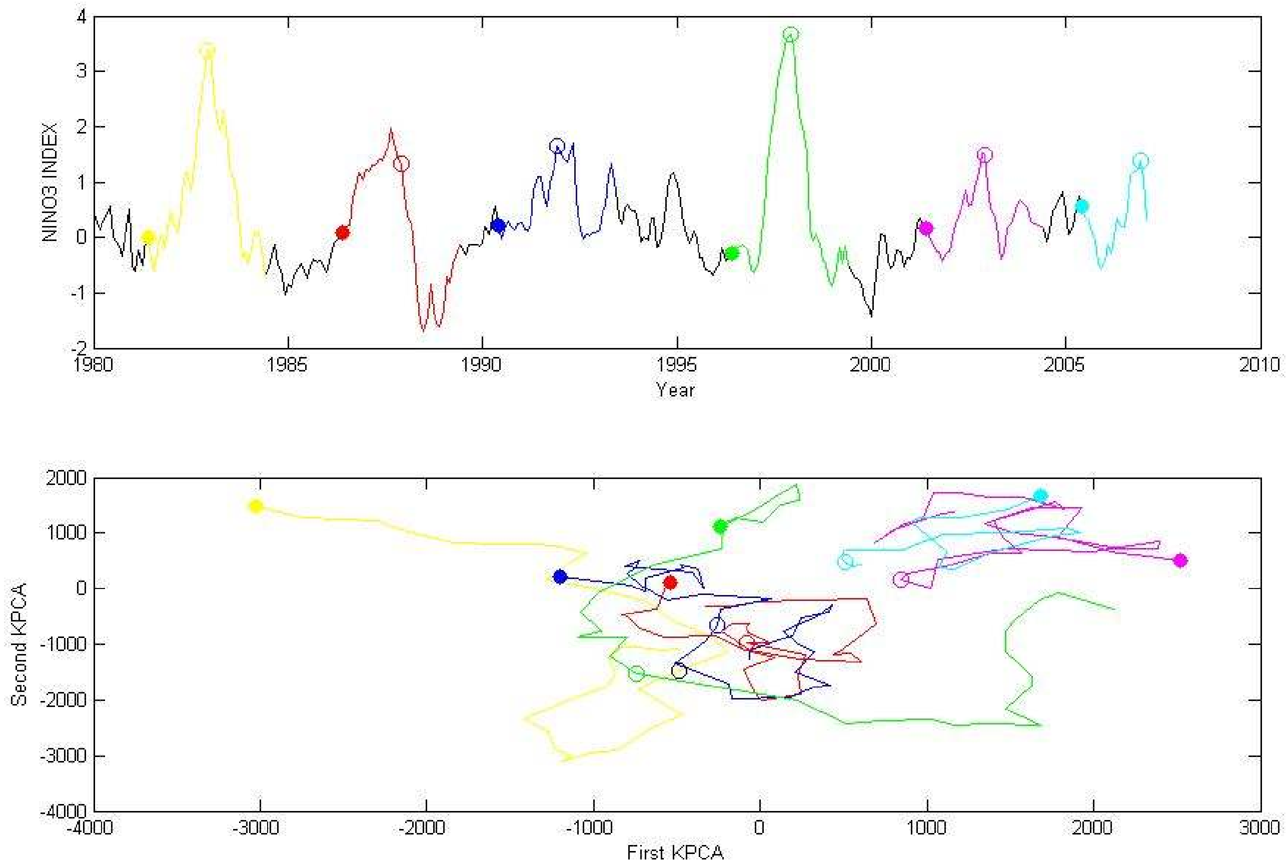
Shift in 1998-1999

Speculate by many authors→ there was a shift in the climate regime around this period (e.g. Chavez et al 2003)

4. Preliminary Results

Problem # 2

Nino3 and SDE – Y1 versus Y2



4. Preliminary Results

Problem # 2

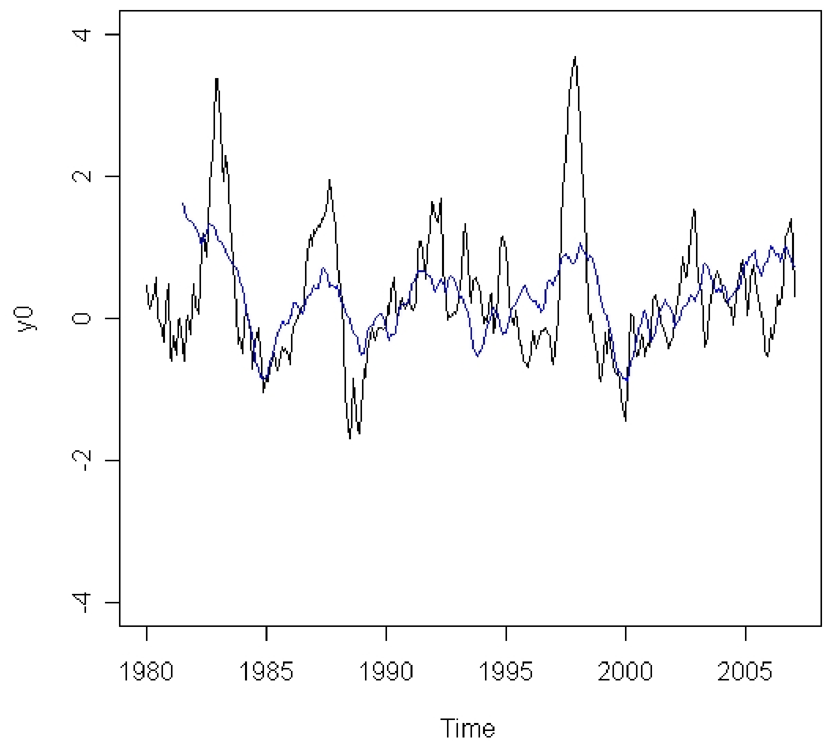
Predictive Model for nino3 index

Simple Linear Model: nino3 = f(Y1,Y2)

18 months lead time

Leave-one-out cross validation

r = 0.53



4. Preliminary Results

Problem # 2

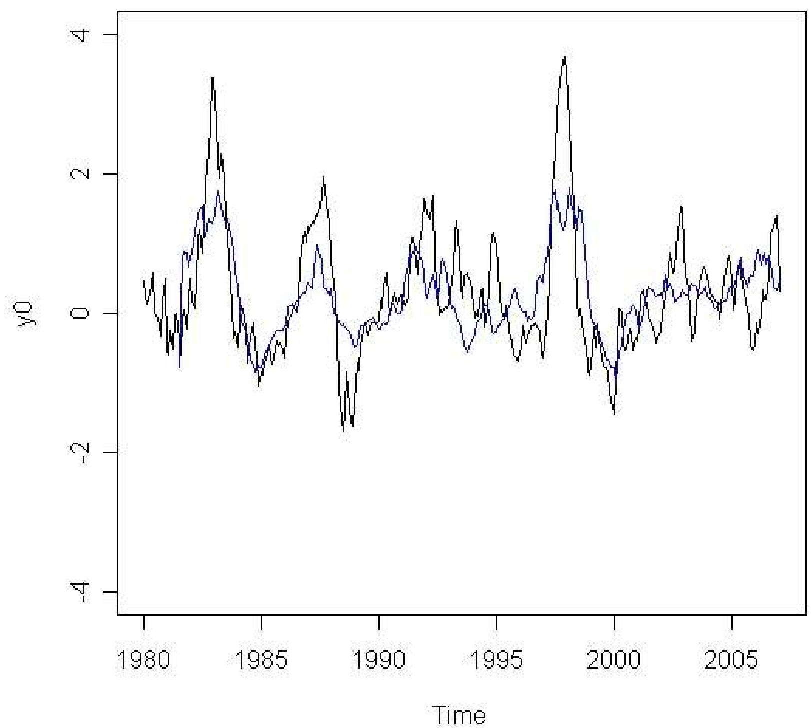
Predictive Model for nino3 index

Loess Model: nino3 = f(Y1,Y2)

18 months lead time

Leave-one-out cross validation

r = 0.66



4. Preliminary Results

Problem # 2

Some conclusions

1. Significant differences between SDE and PCA results;

2. Y1 from SDE shows a change in the end of 990’s → coherent with many other results (e.g. Chavez et al 2003); Not seen in PCA results;

3. Hypothesis1 → Both Y1 and Y2 influence nino3 index;

4. Hypothesys2 → Y1 modulates the intensity of nino3 → Reason why the period 1998-2007 didn’t show big El nino events, although Y2 presented very high values in this period;

5. Predictive model for nino3 shows very good results → very motivated!

5. Next Steps

1. Compare results with MVE;

2. Analyze other climate variables (long record) and compare with Y1;

3. Improve nino3 predictive model (SVM, Kernel Regression, …);

4. Finish the paper.