Aus dem Institut für Physik der Universität Potsdam

Application of nonlinear dimensionality reduction to climate data for prediction

Dissertation

Zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)
in der Wissenschaftsdisziplin Theoretische Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Antonio Juan Gámez López

Potsdam, im Juli 2006

Contents

Chapter 1

Introduction

The purpose of this work is to apply recently developed nonlinear dimensionality reduction methods to real climate data, in order to build low dimensional models for climate prediction. Reducing the number of dimensions could also offer us a clearer view of how climate systems evolve, which could lead us to a better interpretation of the physical processes.

Many physical systems, although multidimensional, exhibit a low dimensional dynamics that makes modelisation more affordable. An extreme example could be a vibrating string which is softly perturbed from its rest position. Although the string is, ideally, infinite dimensional, only a finite number of points is needed to characterise the full system and even predict its subsequent behaviour. In more complicated systems, as climate ones, we can observe similar properties. For example, El Niño, which is an extraordinary phenomenon of heating in the Eastern tropical Pacific Ocean, is usually represented with a one dimensional variable called the Southern Oscillation Index (SOI), therefore receiving the name of El Niño/Southern Oscillation, or ENSO, for the whole process. The underlying physical idea can be better understood with a thought experiment: let us consider a very long water pipe of circular section. When the water flows slowly, the velocity pattern along the pipe is quite simple due to symmetries and the low speed approximation. Hence, only a few variables can represent the whole flow. However, when the water flows more rapidly, the flow becomes turbulent, and more and more dimensions are needed to represent the complexity of the flow. In other words, if we represent the system in a properly chosen coordinate space, the point that represents it will move from a low dimensional manifold to a high dimensional manifold. Consequently, coordinates that were not needed in our previous low speed representation will become more and more important and,

consequently, not negligible.

Another example, which is closely linked with this work, refers to the dynamics of the sea surface temperature (SST) in the tropical Pacific Ocean. It is reasonable to suppose that the behaviour of the SST in each macroscopic point of the surface of the sea is closely tied to the points around. This means that a useful representation of the system can be obtained by giving a discrete grid of finite size. Moreover, the dynamics and thermodynamics of the fluid will link the behaviour of the set, therefore relating points with their neighbours and, of course, reducing the information amount needed to represent the whole system.

These examples take us from the physical to the mathematical problem: how can we reduce the dimensionality of a system, capturing the essential features and neglecting the irrelevant ones? The answer is not simple, of course, as there are many ways of performing the operation. In other words, the transformation of coordinates from the original physical variables to the transformed mathematical (and therefore, not necessarily physical) ones is not unique. For our aims, we could divide the bunch of methods that are commonly used in two types: linear and nonlinear methods of dimensionality reduction, in the sense that the original and new space of representation can be transformed using a linear or a nonlinear function. In climate research, for instance, a widely used linear method is the principal component analysis (PCA) (Joliffe, 1986). In PCA, the system under study is approximated by a linear combination of steady spatial patterns with time dependent coefficients. The relevant number of dimensions is determined from the cumulative explained variance of the PCA modes and their physical interpretation. This variance is closely related to the eigenvalues of the spectral decomposition. Unfortunately, for many physical systems, these eigenvalues show a slow convergence that hampers the selection of a minimum number of dimensions.

In this work, we shall apply a nonlinear method of dimensionality reduction to the observed sea surface temperature (SST) data in the tropical Pacific Ocean. In the equatorial Pacific, the SST evolution is characterised by a nonlinear superposition of two different oscillatory phenomena, ENSO and the Annual Cycle. Attempts to reconstruct ENSO's attractor have been made using different nonlinear methods (Monahan (2001), Grieger and Latif (1994)). The statistical analysis of ENSO is mostly based on SST anomalies which are obtained by subtracting a mean Annual Cycle from the monthly averaged SST data. Extracting a time-varying Annual Cycle and an ENSO mode in a multivariate way from SST data is not simple and the results may depend strongly on the assumptions used by different methodologies. In

particular, linear methods may fail to disentangle both modes since ENSO and the Annual Cycle exhibit in some sense a joint synchronised behaviour: ENSO amplitude is strong during the boreal winter season. This behaviour is reminiscent of an interactive coupling between the two modes (Pikovsky et al., 2001). For this reason, the study of the interaction is of great importance for understanding the variability of ENSO. In the last years, several articles discussing how ENSO and the Annual Cycle interact in the tropical Pacific Ocean have been published (*e.g.* Chang et al. (1994), Xie (1995), Jin et al. (1996), Tziperman et al. (1998), An and Wang (2001)). This type of interaction, which could be nonlinear, may lead to erroneous conclusions when subtracting a constant Annual Cycle from SST data under consideration, as it is usually done in the analysis of ENSO dynamics. Therefore, the data space cannot be decomposed into a sum of linear subspaces each of them containing an independent variable because of the existing interaction. So that, the separation of the SST into physically independent modes is not possible. Our aim will be, first, to extract a low dimensional manifold where the whole physical system could be embedded, and then, to make predictions of future dynamical states of the physical system.

In Chapter 2, the basis of dimensionality reduction will be developed. Linear and nonlinear techniques will be explained in detail. In particular, the multidimensional scaling (MDS) point of view will serve to unify most of the methods that will be used throughout this work.

Chapter 3 will be devoted to climate systems. More specifically, the state-of-the-art of the El Niño/Southern Oscillation (ENSO) and of the Annual Cycle in the tropical Pacific Ocean will cover the majority of the chapter. On one side, ENSO is the most dominant statistical and physical mode of climate variability on interannual timescales (Philander, 1990). Climate models of different complexity have been used to explore the origin of its oscillatory character, its period and skewness (Zebiak and Cane (1987), Tziperman et al. (1994), Jin (1997)). On the other side, the Annual Cycle in the tropical Pacific area originates from a complex interplay between semi-annual solar forcing and coupled air-sea instabilities (*e.g.* Li and Philander (1996), Xie (1994)). As the strength of these instabilities varies slowly in time, one may expect that the amplitude of the physical Annual Cycle is not stationary but time-dependent. Both oscillations will be explained in detail.

Chapter 4 will verse about the application of the mathematical methods developed in Chapter 2 to the physical systems explained in Chapter 3. This will help to compare the application of linear and nonlinear methods of dimensionality reduction to real data. We shall see how nonlinear methods

are more efficient in explaining the variability and the physical behaviour of the data. The advantages and disadvantages of both approaches will be commented.

The goal of the previous chapters will be accomplished in Chapter 5. There, prediction of real states of the coupled model will be computed by using two different methods. First, a model built by adjusting the data of the reduced system found in Chapter 4 using a backfitting algorithm; and, on the other, an ensemble of linear and nonlinear models. Both approaches will be used to predict future Annual Cycle and El Niño events as a joint system. We shall see how the models, despite their simplicity, predict future states with a lead time of several months. The limits of the model will be discussed as well.

Finally, this work will conclude in Chapter 6, where a broad view of the whole Thesis will serve as first step to point out future directions of work.

Chapter 2

An overview of dimensionality reduction

In this Chapter, a general theory of dimensionality reduction is presented. Although a general framework of the problem, the Multidimensional Scaling (MDS), can be used to explain most of the methods used in this work, we shall start with a classical explanation of them. Beginning from the most used techniques of linear analysis, we shall take a step forward into a more general framework.

2.1 Linear methods of dimensionality reduction

In this section, we shall use the following notation. Let us define a matrix $\mathbf{X}_{\mathrm{n}\times\mathrm{m}}$ of data, in such a way that $\mathbf{X}_{\mathrm{ij}}$ is an observation of a certain physical variable $\mathbf{X}_{\mathrm{j}}$ at time t_{i}. Usually, the set of variables $\{\mathbf{X}_{\mathrm{j}}\}_{\mathrm{j}=1}^{\mathrm{m}}$ shares the same dimensionality. If not, an interesting problem arises, as we should adimensionalise the variables either using statistical quantities or physical parameters, although there is no unique way of doing it in general. We shall suppose that the variables $\{\mathbf{X}_{\mathrm{j}}\}_{\mathrm{j}=1}^{\mathrm{m}}$ form a field in a certain physical space, meaning that the variable takes different values in different spatial points $\mathbf{r}_{\mathrm{j}}$. We could then rename $\mathbf{X}_{\mathrm{ij}}$ as $\mathbf{X}(\mathbf{t}_{\mathrm{i}}, \mathbf{r}_{\mathrm{j}})$. Let $\mathbf{Y} = \mathbf{X}^{\mathbf{T}}$. Then, $\{\mathbf{X}_{\mathrm{j}}\}$ is a vector of m unidimensional variables measured at the same time, and $\{\mathbf{Y}_{\mathrm{i}}\}$, which is a column of the matrix $\mathbf{X}$, will be a time series of n measures of the variable ordered at position i. Summarising, the set of observations at different times $t_1, ..., t_n$ builds up the matrix $\mathbf{X}_{\mathrm{n}\times\mathrm{m}}$, so each row will correspond to an obser-

vation of the set of variables at a certain time, $\{\mathbf{X}_\mathrm{j}\}$, and each column will represent the time series of a particular unidimensional variable $\{\mathbf{Y}_\mathrm{i}\}$. Unless otherwise stated, the temporal means are always set to zero in this section, so $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{X}_{\mathrm{ij}} = 0$. Moreover, we shall set $\sum_{\mathrm{j}=1}^{\mathrm{m}} \mathbf{X}_{\mathrm{ij}} = 0$ in section 2.1.1.

2.1.1 Principal Component Analysis

The Principal Component Analysis (PCA) is one of the most common used methods of dimensionality reduction. PCA has been widely applied in many fields of science. For this reason, a handful of names is used to refer to it: empirical orthogonal function (EOF) analysis in the geoscience context (von Storch and Zwiers, 1999), proper orthogonal decomposition (POD) in fluid dynamics (Holmes et al., 1997) or Karhunen-Loève (KL) decomposition if the data are in continuous form (Karhunen (1946) and Loève (1945)). In this work, we shall use the name PCA, which is the most used in the bibliography when analysing time series of data.

The mathematical idea of PCA is to span the original data space into a sum of orthogonal subspaces with a particular optimal property. In this case, the basis of the subspaces are the directions of maximal statistical variability of the data. Let us define $\{\mathbf{P}_\mathrm{k}\}_{\mathrm{k}=1}^{\mathrm{m}}$ as a orthogonal basis of the data space. Any point $\{\mathbf{X}_\mathrm{i}\}$ can be represented as

$$\mathbf{X}_\mathrm{j} = \sum_{k=1}^{m} a_k(t_j)\mathbf{P}_\mathrm{k} \tag{2.1}$$

where $a_k(t_j)$ are called *principal components* and hold the time variability, while $\mathbf{P}_\mathrm{k}(\mathbf{r})$ are called *patterns* or *empirical orthogonal functions*. In general, the patterns are associated with the weight of the different variables. In this case, they give information about the variation of the variables in space $\mathbf{r}$. Therefore the name *pattern*, a picture that represent a sort of spatial structure which is independent of time.

Although there are algorithms that extract the whole basis of patterns at the same time, we shall follow a step-by-step procedure. The first pattern P_1, will be the vector that minimises

$$S_1 = < \mathbf{X}^\mathrm{T}\mathbf{X} > - < (\mathbf{X}\cdot\mathbf{P}_1)^\mathrm{T}(\mathbf{X}\cdot\mathbf{P}_1) > \tag{2.2}$$

where $< \mathbf{X}^\mathrm{T}\mathbf{X} >$ is the statistical variance of $\mathbf{X}$ and we shall be denoted as $\mathbf{V}$. It is noteworthy that $\mathbf{V}$ is positive definite. Eq. (2.2) means that

the projection of the data onto the vector $\mathbf{P}_1$ takes the maximum possible variance. The second step of the orthogonalisation minimises a function S_2 that considers the part of the data that was not projected onto the $\mathbf{P}_1$ and so on. We can summarise this in the formula

$$S_{\mathrm{j}} =< \mathbf{X}^{\mathrm{T}}\mathbf{X} > -\sum_{\mathrm{k=1}}^{\mathrm{j}} < (\mathbf{X}\cdot\mathbf{P}_{\mathrm{k}})^{\mathrm{T}}(\mathbf{X}\cdot\mathbf{P}_{\mathrm{k}}) > \tag{2.3}$$

The minimisation of Eq. (2.3) can be achieved using a Lagrange multiplier λ, with a constraint $\mathbf{P_j}^{\mathrm{T}}\mathbf{P}_{\mathrm{j}}-1=0$. So that

$$\frac{\partial}{\partial \mathbf{P}_{\mathrm{kj}}}[(\mathbf{X}\cdot\mathbf{P}_{\mathrm{j}})^{\mathrm{T}}(\mathbf{X}\cdot\mathbf{P}_{\mathrm{j}}) + \lambda_{\mathrm{j}}(\mathbf{P_j}^{\mathrm{T}}\mathbf{P}-1)] = 0 \quad \text{for k=1,...,n} \Rightarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{P}_{\mathrm{j}} =$$

$$\mathbf{V}\mathbf{P}_{\mathrm{j}} = \lambda_{\mathrm{j}}\mathbf{P}_{\mathrm{j}} \tag{2.4}$$

This means that the directions of maximum variance are the eigenvectors of the correlation matrix of data. Then, projecting the data into the directions of the eigenvectors, we get the principal components $a_k(t_j)$, for $k = 1, ..., m$. As $\mathbf{V}$ is definite positive, the eigenvalues are always non-negative.

There are many interesting properties of the eigenvalues, patterns and principal components of the PCA decomposition. Here, we shall enumerate some:

1. $\mathbf{V} = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathbf{P_j}^{\mathrm{T}}\mathbf{P}_{\mathrm{j}}$
2. $\sum_{\mathrm{j=1}}^{\mathrm{m}} < \mathbf{P_j}^{\mathrm{T}}\mathbf{P}_{\mathrm{j}} >= \sum_{\mathrm{j=1}}^{\mathrm{m}} \lambda_{\mathrm{j}}$
3. $\lambda_{\mathrm{j}} = \mathbf{P_j}^{\mathrm{T}}\mathbf{V}\mathbf{P}_{\mathrm{j}}$
4. $a_j(t) = \mathbf{X}\mathbf{P}_{\mathrm{j}}$
5. $< a_i(t)a_j(t) >= 0$

Property 3. states that the eigenvalue associated with a certain subspace measures the variance of the projection of the data into that subspace. Therefore, as we can order all the eigenvalues, the bigger will correspond to the subspace of maximum variance. Let us order them in the following way: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{\mathrm{m}} \geq 0$. For this reason, we could carefully choose a certain number of subspaces and approximate,

$$\mathbf{X} \approx \sum_{\mathrm{j=1}}^{\mathrm{M}} a_j(t)\mathbf{P}_{\mathrm{j}} \tag{2.5}$$

There are many criteria for choosing a proper number of representative subspaces. This will be explained in detail in Chapter 4.

Rotation of EOFs

When the patterns found by principal component analysis are difficult to interpret in terms of physics, a useful, although controversial, mathematical procedure is the rotation of the eigenvectors. The common procedure is first to conduct a regular PCA, retain some of the eigenvectors and then reconstruct the data using the truncated basis. Starting from this truncated basis, a new basis is found by maximising some function. There are several of these functions in literature: varimax, promax, etc... (Preisendorfer, 1988). Rotations can be made keeping the orthogonality or not. In any of the cases, the main property of the PCA is lost, that is the principal components are no longer uncorrelated. It is important to note that rotation of eigenvectors adds a new component of decision in the data analyst. There is no unique way of doing this operation. As an example, the varimax rotation (Kaiser, 1958) tries to concentrate the explained variance in one of the new vectors, $\mathbf{R}_j$,$l = 1, ..., p < n$ by minimising the function

$$\mathbf{f} = \sum_{j=1}^{p}(\frac{1}{m}\sum_{k=1}^{m}\mathbf{R}_{kj}^4 - \frac{1}{m^2}(\sum_{k=1}^{m}\mathbf{R}_{kj}^2)^2) \tag{2.6}$$

which can be seen as a spatial variance of the squared vectors. But this is, again, based on statistical ideas, not in physical reasoning in general.

2.1.2 Multidimensional Scaling

Consider again the matrix of data $\mathbf{X}_{\mathrm{n\times m}}$. We can think that the matrix is built up of n points $\{\mathbf{X}_{\mathrm{i}}\}$ in a $\mathbb{R}^m$ space, and that these points belong to a trajectory as they are ordered in time.

In this space, for every two points $\{\mathbf{X}_{\mathrm{i}}, \mathbf{X}_{\mathrm{j}}\}$ in $\mathbb{R}^m$, the metric or distance $d_{ij} = d(\mathbf{X}_{\mathrm{i}}, \mathbf{X}_{\mathrm{j}})$ can be defined as a function onto nonnegative real numbers that obeys the following rules:

1. $d_{ij} = 0$ if and only if $i = j$
2. $d_{ij} = d_{ji}$
3. $d_{ij} + d_{jk} \geq d_{ik}$ which is popularly known as the triangle inequality.

The key idea of the multidimensional scaling (MDS) (Borg and Groenen, 1997) is that the distance can be interpreted as a measure of similarity: if $d_{ij} < d_{ik}$ we shall say that the physical state represented by $\mathbf{X}_\mathrm{i}$ is closer to $\mathbf{X}_\mathrm{j}$ than $\mathbf{X}_\mathrm{k}$. Let us define

$$(\mathbf{D^{(2)}})_{\mathrm{ij}} = {d_{ij}}^2 \tag{2.7}$$

as a matrix of squared distances. This matrix is symmetric and positive definite. Now let us transform $\mathbf{D^{(2)}}$ via an operation called double-centring. More precisely, let us define a matrix $\mathbf{J}$ as

$$\mathbf{J}_{\mathrm{ij}} = \delta_{\mathrm{ij}} - 1/\mathrm{n} \tag{2.8}$$

where δ_{ij} is the Kronecker delta and a second matrix

$$\mathbf{Z^{(2)}} = \ -\tfrac{1}{2}\mathbf{J}\mathbf{D^{(2)}}\mathbf{J} \tag{2.9}$$

Then, $\mathbf{Z^{(2)}}$ can be decomposed into its eigenvalues and eigenvectors,

$$\mathbf{Z^{(2)}}\mathbf{q}_\mathrm{i} = \lambda_\mathrm{i}\mathbf{q}_\mathrm{i} \tag{2.10}$$

Defining $\Lambda_{\mathrm{ij}} = \lambda_\mathrm{i}\delta_{\mathrm{ij}}$ with $\lambda_i > \lambda_{i+1}$ and $Q_{ij} = q_{ij}$ the component j of vector $\mathbf{q}_\mathrm{i}$, then $\mathbf{Z^{(2)}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}$. If the set of eigenvalues have some leading components $\{\lambda_1, ..., \lambda_M\}$ with $M < m$, and the rest decays very fast, meaning $\lambda_M >> \lambda_{M+1}$, we could approximate $\mathbf{Z^{(2)}}$ to its projection into the subspace spanned by the eigenvectors $\{\mathbf{q}_1, ..., \mathbf{q}_\mathrm{M}\}$.

The MDS framework is a generalisation of the PCA. In particular, defining $\mathbf{D^{(2)}}$ using the definition of euclidean distance will take us to the same algebraic results as in PCA, as we shall show in section 2.1.3. Moreover, a step forward into a nonlinear generalisation can be taken with a suitable definition of distance in Eq. (2.7).

2.1.3 Principal Coordination Analysis, or PCA revisited

Let us define the scalar product in the data space as,

$$\mathbf{X}_\mathrm{i} \cdot \mathbf{X}_\mathrm{j} = \ \sum_{k=1}^{m} \mathbf{X}_{\mathrm{ik}}\mathbf{X}_{\mathrm{jk}} \tag{2.11}$$

A space where the scalar product is defined as in (2.11) is called euclidean. We also define the norm of a vector as

$$\|\mathbf{X}_\mathrm{i}\| = \sqrt{\mathbf{X}_\mathrm{i} \cdot \mathbf{X}_\mathrm{i}} = \sqrt{\sum_{k=1}^{m} \mathbf{X}_\mathrm{ik}\mathbf{X}_\mathrm{jk}} \tag{2.12}$$

and the distance between two points $\mathbf{X}_\mathrm{i}$ and $\mathbf{X}_\mathrm{j}$ as the norm of their difference

$$d_{ij} = \|\mathbf{X}_\mathrm{i} - \mathbf{X}_\mathrm{j}\| = \sqrt{\sum_{k=1}^{m} (\mathbf{X}_\mathrm{ik} - \mathbf{X}_\mathrm{jk})^2} \tag{2.13}$$

which is the well known expression of euclidean distance. Therefore, the matrix of squared distances $\mathbf{D^{(2)}}$ can be written in the form

$$\mathbf{D^{(2)}} = \mathbf{1c}^\mathrm{T} + \mathbf{c1}^\mathrm{T} - 2\mathbf{X}\mathbf{X}^\mathrm{T} \tag{2.14}$$

where the components of the vector $\mathbf{c}$ are $c_i = \sum_{k=1}^{m} X_{ik}^2$ and $\mathbf{1}$ is a vector formed by ones and dimension n. Then, applying the centring and eigendecomposition as explained in section (2.1.2), we obtain,

$$\mathbf{Z^{(2)}} = -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{J}\mathbf{D^{(2)}}\mathbf{J} = \mathbf{X}\mathbf{X}^\mathrm{T} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\mathrm{T} = \mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{Q}^\mathrm{T} \tag{2.15}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues and $\mathbf{Q}$ is the matrix of eigenvectors of $\mathbf{Z^{(2)}}$. This allows us to make a representation of $\mathbf{X}$ in terms of $\mathbf{Q}_\mathrm{i}\lambda_\mathrm{i}^{1/2}$ which is called principal coordinates analysis (PCO or PCoA) (Gower, 1966). If the data have zero mean and variance one, the algebraic correspondence with PCA arises. We remind that the correlation matrix $\mathbf{V}$ can be written as,

$$\mathbf{V} = \mathbf{X}^\mathrm{T}\mathbf{X} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\mathrm{T} = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{Q}^\mathrm{T}\mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{P}^\mathrm{T} \tag{2.16}$$

where $\mathbf{Q}^\mathrm{T}\mathbf{Q} = \mathbf{I}$. Finally, we can associate, by singular value decomposition, $\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}^{1/2}\mathbf{P}^\mathrm{T}$, where $\mathbf{Q}\mathbf{\Lambda}^{1/2}$ is the principal components matrix, $\mathbf{\Lambda}$ is the diagonal matrix formed by the eigenvalues of the correlation matrix and $\mathbf{P}$ is the matrix of patterns. Therefore, PCA and PCO get the same representation in terms of components $\mathbf{Q}\mathbf{\Lambda}^{1/2}$. This way we can reduce the dimensionality of the data by taking $\mathbf{Q}_\mathrm{i}\lambda_\mathrm{i}^{1/2}$ from $i = 1$ to some value $M < m$, which could be $M << m$.

In other words, PCA can be regarded as an euclidean MDS for normalised

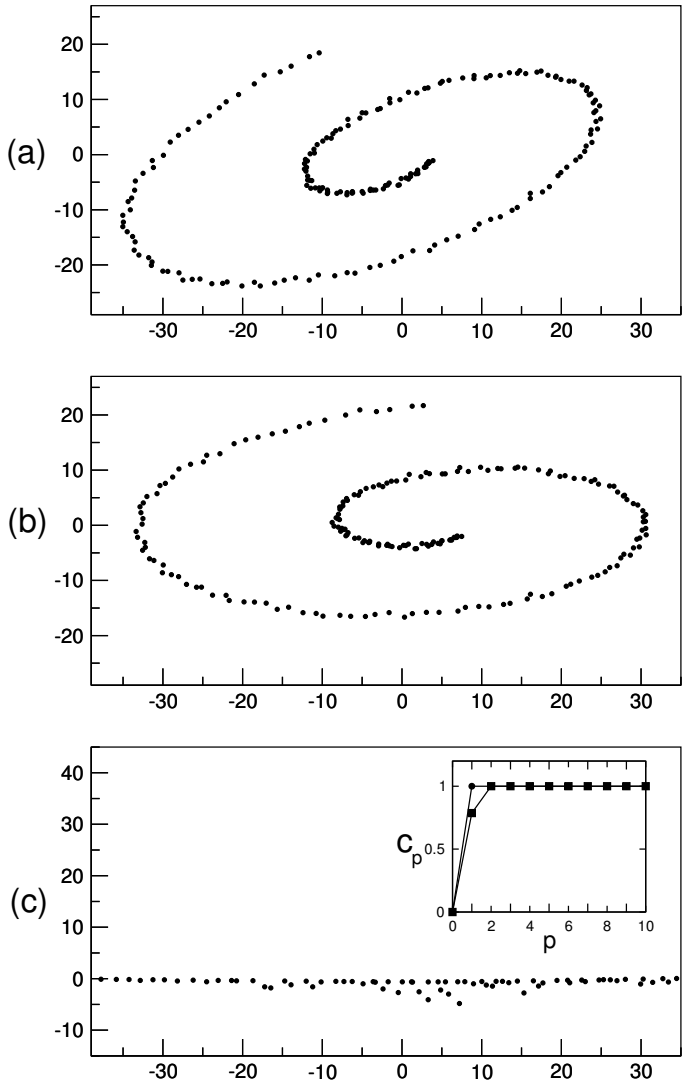

Figure 2.1: (a) Spiral with uniform noise (arbitrary units). (b) Result for the spiral after PCA, the method cannot guess the true dimensionality (arbitrary units). (c) Result for the spiral after Isomap, now the method unfolds the one dimensional trajectory (arbitrary units). In the inner box, the cumulative variances c_p against the number of eigenvalues considered for Isomap (•) and PCA (■).

data. We would like to stress that PCA is a linear method of decomposition, where the data are projected into orthonormal linear subspaces. However, if the data points belong to a nonlinear manifold, the orthonormal projection spreads contributions to the variance onto the different principal components. In that case, reducing the dimensionality of a physical system in which the dynamics is not governed by linear processes or where there are nonlinear relations between variables could lead to a wrong interpretation of the dimensionality. As a simple example, we can apply PCA to a simple nonlinear dataset. For instance, we construct a trajectory of a particle moving on a spiral with some added uniform noise (Fig. 2.1a). In this example we can observe (Fig. 2.1b) that the PCA do not reconstruct the one-dimensional trajectory. To reproduce the real dynamics we need, in this context, to use a definition of distance that captures the nonlinear structure of the manifold. The geodesic distance is the proper metric for measuring distances on nonlinear manifolds, as we shall discuss in section 2.2.

2.2 Nonlinear methods of dimensionality reduction

In this section, two methods of nonlinear dimensionality reduction are presented. These methods, Isomap (Tenenbaum et al., 2000) and Local Linear Embedding (Roweis and Saul, 2000) were originally developed in the context of cognitive sciences as a way of representing the reduction of dimensionality in visual perception (Seung and Lee, 2000). It is important to stress why the word *nonlinear* is used. Generally speaking, dimensionality reduction is a transformation of coordinates $\mathcal{T}$ from a data space $\mathbf{X}$ to a new one of lower dimensionality $\mathbf{\Theta}$, therefore $\mathbf{\Theta} = \mathcal{T}(\mathbf{X})$. If $\mathcal{T}$ is not a linear transformation connecting the original and new spaces, we shall say that the reduction is nonlinear.

2.2.1 Isometric Feature Mapping or Isomap

The natural metric for nonlinear manifolds is the geodesic distance (for a complete review on differential geometry, see Do Carmo (1976)). Let us have an euclidean space of data $\mathbf{X}$ of coordinates $\{\mathbf{X}_1, ..., \mathbf{X}_\mathrm{m}\}$. Suppose there is a manifold $\mathbf{\Theta} \subseteq \mathbf{X}$ represented by the coordinates $\{\Theta_1, ..., \Theta_p\}$. The metric of $\mathbf{\Theta}$ is the matrix $\mathbf{g}$ with elements defined by,

$$g_{ij} = \sum_{k=1}^{m} \frac{\partial \mathbf{X}_\mathrm{k}}{\partial \Theta_i} \frac{\partial \mathbf{X}_\mathrm{k}}{\partial \Theta_j} \tag{2.17}$$

For a general metric $\mathbf{g}$, the distance between two points θ_1 and θ_2 in $\mathbf{\Theta}$ is then

$$d(\theta_1, \theta_2) = \|\theta_1 - \theta_2\| = \int_{\theta_1}^{\theta_2} \sqrt{\sum_{i,j=1}^{m} g_{ij} d\Theta_i d\Theta_j} \tag{2.18}$$

In the case of a euclidean manifold, $g_{ij} = \delta_{ij}$ and we recover Eq. (2.13) for the discrete case.
For example, let us think of a particular manifold $\mathbf{\Theta}$, in our case a spiral similar to that one shown in Fig. 2.1a. The equations of the spiral are:

$$x = t \sin t$$
$$y = t \cos t$$

where t belongs to a real interval. The euclidean distance d_E between two points $(x_1, y_1) = (x(t_1), y(t_1))$ and $(x_2, y_2) = (x(t_2), y(t_2))$ is $d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = \sqrt{t_1^2 + t_2^2 - 2t_1 t_2 \cos(t_1 - t_2)}$. The path from one point to another one is a straight line that does not belong to the spiral. But, if we restrict ourselves to a path inside the spiral, the geodesic distance can be calculated as $d_S = \int_{t_1}^{t_2} \sqrt{1 + t^2} dt = \frac{1}{2}\{t_2\sqrt{1 + t_2^2} - t_1\sqrt{1 + t_1^2} + \text{arcsinh}(t_2) - \text{arcsinh}(t_1)\}$. It is interesting to note that small euclidean distances may correspond to large geodesic distances. For this reason, measuring similarities based on inadequate distances could lead to misleading results. Consequently, the idea is to substitute the euclidean distance in the MDS method (hence, the scalar product defined by the statistical correlation in the case of PCA) by the geodesic distance between each pair of points. Tenenbaum et al. (2000) proposed Isomap, a method for computing geodesic distances through graph distances. The method could be divided in several phases:

1. In a first step, Isomap approximates the geodesic distance using a graph constructed by connecting nearest neighbours in the euclidean $\mathbb{R}^m$ space. More specifically, we shall say that a point $\mathbf{X}_\text{i}$ is one of the nearest neighbours of $\mathbf{X}_\text{j}$ if it belongs to a ball centred on $\mathbf{X}_\text{j}$ with radius ϵ. Alternatively, we could also define $\mathbf{X}_\text{i}$ as a nearest neighbour of $\mathbf{X}_\text{j}$ if it is one of the K closest points (measured by the euclidean distance) to $\mathbf{X}_\text{j}$ in the set.

2. After the nearest neighbours are defined, they are connected via weighted edges where the weight is the euclidean distance between connected points.

3. Then, the minimum graph distance between each pair of points is computed.

This distance is used as a fair approximation to the true geodesic distance (for discussion and proofs, see Tenenbaum et al. (2000) and references therein). The crucial point lies on finding an interval of ϵ, or a number of K, where the solution is robust. Low values of ϵ or K will not connect all the points, while too many will overestimate the dimension of the manifold. After the new matrix of squared distances is computed, the MDS procedure is applied, starting at Eq. (2.9). The dimensionality of the manifold (the optimum number of dimensions needed to capture the variability of the data) can be measured via the eigenvalues of the MDS procedure. These eigenvalues are a measure of the error made when we project the whole dataset onto the directions defined by the corresponding eigenvectors. The cumulative variance c_p, of dimension p is defined using $c_p = \frac{\sum_{i=1}^{p} \Lambda_{ii}}{Tr(\Lambda)}$, and takes the value $c_p = 0$ if no statistical variance is explained, and $c_p = 1$ if all the variability is taken into account.
The Isomap algorithm has two computational bottlenecks (Silva and Tenenbaum, 2003). The first is the calculation of the n^2 shortest-paths distance matrix. The second is the eigenvalue calculation after the double-centring operation in Eq. (2.9) of the n^2 rank matrix $\mathbf{Z^{(2)}}$. These two inefficiencies can be avoided by designating $n' < n$ landmark points. Instead of computing the whole set of distances, only the $n \times n'$ matrix of distances from each data point *to* the landmarks are calculated. Of course, if $n' << n$, a lot of computing time is saved. The fact that the use of landmarks is feasible can be justified from the assumption that the data are embedded in a low-dimensional manifold (Silva and Tenenbaum, 2003).
Using the same example as in the preceeding section, we apply Isomap ($k = 5$) to the spiral set with noise represented in Fig. 2.1a. For the spiral, the dimensionality is one (Fig. 2.1c, inner box), in contrast with the results offered by PCA. We can observe how the spiral is unfolded into an approximately one dimensional set in the Fig. 2.1c. Comparison of the eigenvalues calculated by PCA and Isomap confirms that the dimensionality of the spiral is correctly guessed by Isomap, and that PCA overestimates the true dimensionality (Fig. 2.1c, inner box). A second academic and, nevertheless, clarifying example is the so-called swiss roll (Fig. 2.2). In Fig. 2.3 (left), we observe how Isomap can unfold the original three dimensional data. It is important to note that this method is rotationally invariant, as distance is an invariant measure for these transformations. Therefore, rotating the resulting data does not change

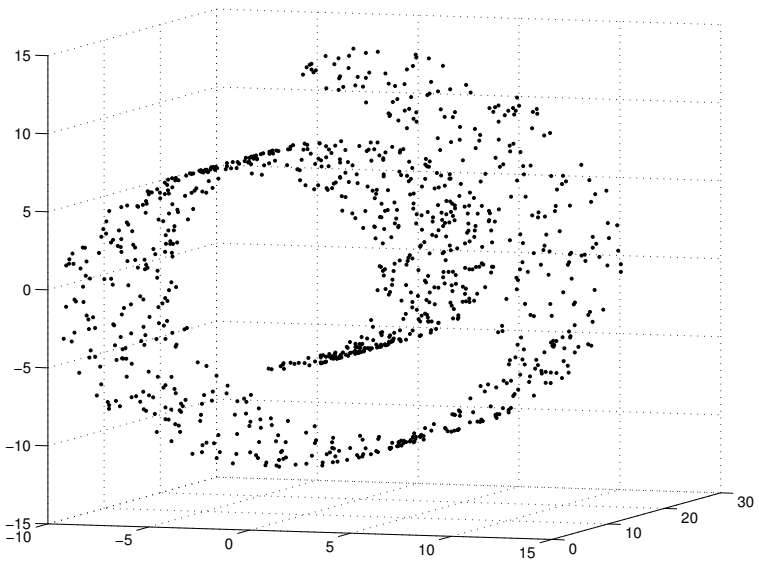

Figure 2.2: Three dimensional plot of the swiss roll data.

the results, only the physical interpretation of them.

2.2.2 Local Linear Embedding

The objective of the Local Linear Embedding (LLE) is to reconstruct a lower dimensional manifold $\mathbf{\Theta}$ from a high dimensional one $\mathbf{X}$ by imposing a linear metric in the neighbourhood of each point of the manifold $\mathbf{X}$. The steps of LLE are the following:

1. First, we assign neighbours to each data point $\{\mathbf{X}_\mathrm{i}\}$ using, for example, an euclidean metric.

2. Then, we introduce the assumption that neighbouring points can be embedded in a linear manifold if the data are well sampled. This is done by computing the weights W_{ij} that minimises the error

$$\epsilon(\mathbf{W}) = \sum_{\mathbf{i}} |\mathbf{X}_\mathrm{i} - \sum_{\mathrm{j}} \mathrm{W}_{\mathrm{ij}}\mathbf{X}_\mathrm{j}|^2 \tag{2.19}$$

   subject to two constraints: first, that each data point $\{\mathbf{X}_\mathrm{i}\}$ is reconstructed only from its neighbours, so $W_{ij} = 0$ if $\{\mathbf{X}_\mathrm{j}\}$ is not a neighbour

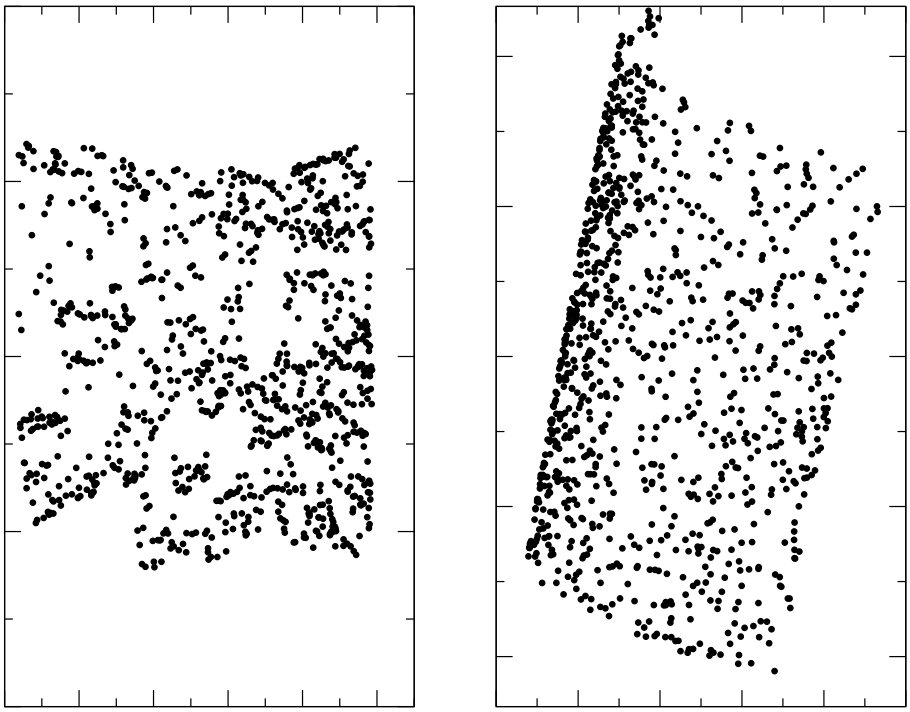

Figure 2.3: Isomap (left) and LLE (right) two dimensional embeddings for the swiss roll data. Actual values of the coordinates are not shown.

of $\{\mathbf{X}_\mathrm{i}\}$; second, that $\sum_j W_{ij} = 1$. The minimisation of Eq. 2.19 is a least-squared problem.

3. Finally, compute the new points $\{\mathbf{\Theta}_\mathrm{i}\}$ by minimising a cost function

$$\Phi(\mathbf{\Theta}) = \sum_{\mathbf{i}} |\mathbf{\Theta}_\mathrm{i} - \sum_{\mathrm{j}} \mathrm{W_{ij}} \mathbf{\Theta}_\mathrm{j}|^2 \tag{2.20}$$

As with the previous method, Isomap, there is a rotational, translational and scaling invariance. A fundamental difference is that the LLE algorithm needs the embedding number of dimensions as input for Eq. 2.20. The example of the swiss roll is successfully disentangled by LLE, as observed in Fig.2.3(right), but, on the other hand, the number of dimensions had to be used as input to reach this result.

2.2.3 Kernel PCA is a kernel view

The nonlinear algorithms explained in this section share a common characteristic: they induce a local neighbourhood structure to construct a global map from a high dimensional space to a lower one. The main difference is the type of algorithm used to find the global embedding. There is an interesting way of interpreting the different classes of algorithms as kernel methods (Ham et al., 2004). Summarising, considering the data space $\mathbf{X}$, a positive definite kernel k is a real-valued function on $\mathbf{X} \times \mathbf{X}$ with the property that there exists a map Φ into a dot product space, $\mathbf{P}$, $\Phi : \mathbf{X} \rightarrow \mathbf{P}$ such for all $x, x' \in \mathbf{X}$, there exists $\langle x, x' \rangle = k(x, x')$. With this definition, the PCA problem is transformed into a problem that can be solved in terms of a kernel k in the following way. In PCA, as explained in section 2.1.1, the key operation is the diagonalisation of the covariance matrix $\mathbf{V}$. Therefore, the patterns can be written as $\mathbf{P_j} = \frac{1}{\lambda}\sum(\mathbf{X_i} \cdot \mathbf{P_j})\mathbf{X_i}$, so the patterns are spanned into a sum of the data vectors. Let us consider a covariance matrix $\mathbf{V} = \sum_{j=1}^{m} \Phi(\mathbf{X_j})^{T}\Phi(\mathbf{X_j})$ as a result of the map Φ. This matrix can be diagonalised, $\lambda\mathbf{P} = \mathbf{VP}$. Now, $\mathbf{VP} = \lambda\mathbf{P} = \lambda\sum\alpha_\mathbf{j}\mathbf{\Phi}(\mathbf{X_j})$ and substituting in the eigenvalue equation, $\mathbf{VP} = \sum\sum\alpha_\mathbf{j}\mathbf{\Phi}(\mathbf{X_i})\Phi(\mathbf{X_j})^{T}\Phi(\mathbf{X_j}) = \sum\sum\alpha_j K(\mathbf{X_i}, \mathbf{X_j})\Phi(\mathbf{X_j})$ and we can rewrite a new eigenvalue problem $K\alpha = \lambda\alpha$. These calculations were done supposing zero mean for the variables. In a more general case, the new kernel can be written as $K' = (\mathbf{I} - \mathbf{11}^{T})K(\mathbf{I} - \mathbf{11}^{T})$, where $\mathbf{I}$ is the identity matrix of dimension n.

In the case of Isomap, the kernel can be defined as $K_{Isomap} = -\frac{1}{2}(\mathbf{I} - \mathbf{11}^{T})\mathbf{D^{(2)}}(\mathbf{I}-\mathbf{11}^{T})$, meaning that we centre the squared-distance matrix, while in the case of LLE, the weight matrix $\mathbf{W}$ is the one that defines the kernel $K_{LLE} = \lambda_{max}\mathbf{I}-(\mathbf{I}-\mathbf{W})(\mathbf{I}-\mathbf{W}^{T})$ where λ_{max} is the maximum eigenvalue. In conclusion, Isomap and LLE can be seen as kernel PCA with particular kernel matrices. This is the starting point of an increasing amount of literature about generalisation of these nonlinear methods. In particular, a continuum generalisation of Isomap (Zha and Zhang, 2003) could be interesting for finding reduced equations for classical infinite dimensional equations.

Chapter 3

A short introduction to geophysics in the tropical Pacific Ocean

In this Chapter, an overview about two physical phenomena occurring in the tropical Pacific Ocean will be given. First of all, we shall focus on the physics of El Niño phenomenon. After the main description is given, the annual oscillation in the tropical Pacific Ocean will be described. There is no current mathematical model for it, although many physical features are being better understood over the last years. In the last part of the Chapter, three climate models of different complexity will be explained in detail: the Zebiak-Cane (ZC) model (Zebiak and Cane, 1987), the delayed oscillator (DO) model (Tziperman et al., 1994) and the recharge oscillator (RC) model (Jin, 1997). These two phenomena join together in a complex oscillation that embeds most of the variability of the physical variables in that region of the Earth. The interaction between them is still not well known. One of the problems that difficults the modeling of the interaction is the separation of the two oscillations. In other words, how can we distinguish which contribution from the total value of the variable corresponds to the effect of the solar forcing, the Annual Cycle, and which to the interannual oscillation, El Niño.

3.1 A general physical framework for the tropical Pacific Ocean

The atmosphere and the ocean are essentially composed by fluids. Therefore, the Navier-Stokes equations are fundamental in the study of physical phenomena concerning both systems. Let us first start with a quite general Navier-Stokes equation on a rotating frame at constant angular velocity Ω,

$$\frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{V} + \mathbf{2\Omega}\times\mathbf{V} = -\frac{1}{\rho}\nabla \mathrm{p} + \nu\nabla^2\mathbf{V} + \mathbf{g}' \tag{3.1}$$

where $\mathbf{V}$ is the velocity field of the fluid, ρ is the density of the fluid, p is the pressure, g' is the gravity minus the centripetal force (which can usually be neglected) and $\nu = \eta/\rho$ is the dynamical viscosity, being η the shear viscosity. Let us apply this equations to the particular case of a thin layer of fluid on the Earth, i. e. the scale of horizontal motions are much larger than the scale of vertical motions. Thus we assume that the fluid is confined to a layer which is much thinner than it is wide. A second assumption is that vertical motions only change the layer depth and that the pressure gradient is near hydrostatic. In the case, the equations are usually called the Laplace tidal equations and, in spherical coordinates λ (longitude) and ϕ (latitude) are (Gill, 1982),

$$\frac{\partial u}{\partial t} + (\mathbf{V}\cdot\nabla)u - \left(2\Omega + \frac{u}{r\cos\phi}\right)v\sin\phi = -\frac{g}{r\cos\phi}\frac{\partial \eta}{\partial \lambda} \tag{3.2}$$

$$\frac{\partial v}{\partial t} + (\mathbf{V}\cdot\nabla)v + \left(2\Omega + \frac{u}{r\cos\phi}\right)u\sin\phi = -\frac{g}{r}\frac{\partial \eta}{\partial \phi} \tag{3.3}$$

$$\frac{\partial \eta}{\partial t} + \frac{1}{r\cos\phi}\left\{\frac{\partial}{\partial\lambda}[(H+\eta)u] + \frac{\partial}{\partial\phi}[(H+\eta)v]\right\} = 0 \tag{3.4}$$

where u is the horizontal velocity in the longitudinal direction, taking the eastward flow as positive, v is the meridional flow, taking the northward flow as positive and r is the radius of the Earth. Vertical motion is neglected except when affecting η, the thickness of the water layer which has an equilibrium depth of H in absence of motion. Thus, the ∇ operator is defined as $\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$. It is convenient to define a new variable $h = g\eta$ that will help to simplify further equations. Although the system we are dealing with is spherical, if we restrict ourselves to the tropical region, we can approximate the spherical coordinates to a local cartesian frame (x, y, z) fixed to the Earth surface. Near

the Equator, ϕ is small so $\sin\phi \approx \phi$ and $\cos\phi \approx 1$. We shall take the variable $x = r\lambda$ as positive in the East direction, the variable $y = r\phi$ is positive in the North direction, and the z following the cross product rule. The Coriolis parameter f is defined as $f = 2\Omega\sin\phi$. We define the β parameter as,

$$\beta = \frac{1}{r}\frac{df}{d\phi} = \frac{2\Omega\cos\phi}{r} \tag{3.5}$$

so we approximate $f = \beta y$ near the Equator. The parameter $\beta = 2.3 \times 10^{-11} m^{-1}s^{-1}$ is exact at the Equator and approximately constant in the rest of the tropics. This approach is called the *Equatorial Beta plane* approximation. Substituting in Eqs. (3.2)-(3.4) and assuming that the velocities are small enough to consider that the squared velocities are much smaller than the velocities themselves we get the following equations:

$$\frac{\partial u}{\partial t} - \beta y v = -\frac{\partial h}{\partial x} \tag{3.6}$$

$$\frac{\partial v}{\partial t} + \beta y u = -\frac{\partial h}{\partial y} \tag{3.7}$$

$$\frac{\partial h}{\partial t} + gH\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = 0 \tag{3.8}$$

These equations are known as the linear shallow water equations. After some complex manipulation (see Gill (1982)), they can be reduced to a single equation for v,

$$\frac{\partial^3 v}{\partial x^2 \partial t} + \frac{\partial^3 v}{\partial y^2 \partial t} + \beta\frac{\partial v}{\partial x} - \frac{1}{gH}\frac{\partial^3 v}{\partial t^3} - \frac{\beta^2 y^2}{gH}\frac{\partial v}{\partial t} = 0 \tag{3.9}$$

A general solution of this equation is of the form $v = V(y)\exp(ikx - i\omega t)$. Substituting this in Eq.(3.9) gets the equation,

$$\frac{\partial^2 V}{\partial y^2} + \left(\frac{\omega^2}{gH} - k^2 - \frac{\beta k}{\omega} - \frac{\beta^2 y^2}{gH}\right)V = 0 \tag{3.10}$$

This is the well-known Schrödinger equation for the harmonic oscillator. There is a solution for $V = 0$ that gives the dispersion relation $\omega^2 = (gH)k^2$. It is clear now that the constant $c = \sqrt{gH}$ is a natural velocity of the wave equation. Only the positive solution is physical because the negative one gives unbounded values for u. For $V \neq 0$, the discrete eigenvalues are,

$$\frac{\omega^2}{gH} - k^2 - \frac{\beta k}{\omega} = \frac{\beta}{gH}(2n+1) \quad \text{for} \quad n \in \mathbb{Z}^+ \tag{3.11}$$

The corresponding eigenfunctions are the well-known parabolic cylinder functions $D_n = \exp(-\beta/(2c^2 y))H_n(\sqrt{\beta/c}y)$ where H_n are Hermite polynomials. We shall not go further on them, but it is interesting to say they are orthogonal, and symmetric around the Equator for n even and antisymmetric for n odd. The natural scaling distance $R = \sqrt{c/\beta}$ is called the equatorial radius of deformation. Of course, having v we can compute h and u simply by substituting v into Eqs. (3.6-3.8). h and u are also related to the Hermite polynomials and are symmetric when v is antisymmetric and viceversa.

Now we shall focus on some particular types of waves. The dispersion relation $\omega = ck$ is associated to a wave called *Kelvin wave*. Its main property is that it has no meridional velocity, as $v = 0$. It is easy to see that the propagation of this wave is toward the east with the natural velocity c, as the dispersion relation is $\omega = ck$, with k positive. The low frequency solutions are called *Rossby waves*. Their existence is due to the fact that the Coriolis parameter f is not constant. An approximate dispersion relation can be obtained from Eq.(3.11) by neglecting the squared term in ω,

$$\omega = \frac{-\beta k}{k^2 + \frac{\beta(2n+1)}{c}} \tag{3.12}$$

The limit of small wavenumber is very interesting. The dispersion relation can be approximated to,

$$\omega = \frac{-ck}{(2n+1)} \tag{3.13}$$

so these waves are nondispersive and propagate towards the west. For $n = 1$, the speed is one third of that of the Kelvin wave. It is noteworthy that the Rossby waves have a maximum frequency. In the case $n = 1$, the corresponding period is of around fifteen days. Thus, a forcing of a greater frequency will not excite Rossby waves, but Kelvin waves instead. Finally, for long frequencies, the dispersion relation is, approximately,

$$\omega^2 = c^2 \left(k^2 + \frac{\beta}{c}(2n+1) \right) \tag{3.14}$$

These waves are called *gravity waves*.

3.1.1 Forced shallow water equations

There are two main forces that could affect the ocean dynamics: diabatic heating, which is affecting the density and pressure variables, and the wind

stress, which affects the momentum balance. A useful approximation that considers both effects modify the shallow water equations in this simple way,

$$\begin{aligned} \frac{\partial u}{\partial t} - \beta y v &= -\frac{\partial h}{\partial x} + \frac{\tau_x}{H} \\ \frac{\partial v}{\partial t} + \beta y u &= -\frac{\partial h}{\partial y} + \frac{\tau_y}{H} \\ \frac{\partial h}{\partial t} + gH\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) &= -sQ \end{aligned} \tag{3.15}$$

where τ_x and τ_y are the wind stress components and sQ is a scaled diabatic heating component. The equations can now be solved by expanded parabolic cylinder equations.

3.1.2 Boundary reflections

In ocean dynamics (and in general wave dynamics), reflection of waves is an important process for the dynamic adjustment of the tropical basin. For the region we are considering in this work, coasts parallel to the Equator will not be considered. In the low frequency regime, there are three types of waves: the eastward Kelvin wave, which is trapped in the equator; the long wave westward Rossby waves, and the short wave eastward Rossby waves. When a Kelvin wave reaches a eastern coast, it is reflected as a coastal Kelvin wave (Philander, 1990). It propagates poleward at a speed c and it is confined to a coastal zone with width of the order of the radius of deformation R. The short wave Rossby waves are also trapped near the coast and they travel poleward as well. At low frequencies, the westward Rossby waves are reflected as an equatorial Kelvin wave with a smaller amplitude. The rest of the energy is reflected as shortwave Rossby waves that are dissipated before they move back far from the boundary. The islands do not affect the dynamics in any of the cases, because their latitudinal scale is small relative to the radius of deformation, as is the case of the Galápagos and Gilbert islands in the tropical Pacific Ocean.

3.2 The Annual Cycle in the tropical Pacific Ocean

The Annual Cycle in the tropical Pacific area originates from a complex interplay between semi-annual solar forcing and coupled air-sea instabilities (*e.g.*

Li and Philander (1996), Xie (1994)). As the strength of these instabilities varies slowly in time, one may expect that the amplitude of the physical Annual Cycle is not stationary but time-dependent. In this section, several general features about the behaviour of the seasonal cycle will be explained. The most relevant processes in the atmosphere of the tropical Pacific region are known as the Hadley and Walker circulation. The Hadley circulation creates a cell of air that rises in the equatorial region and descends at higher latitudes. The cell is driven by the latent heat that occurs in the Equator. Due to the conservation of angular momentum, the mean winds are easterly (i.e. going from east to west) at the sea surface near the Equator. These winds are named trade winds. The Walker circulation operates on the Equatorial plane. There is air rising over the West Pacific area and air sinking over the East Pacific that complete the Walker cell with the westward winds we mentioned before. The second relevant characteristic of the tropical Pacific Ocean atmosphere is the existence of a band with persistent cloudiness and heavy precipitation, the Intertropical Convergence Zone (ITCZ). This band is located north of the Equator and is characterised by high moisture and convection. During the northern summer, the ITCZ is strong all across the Pacific. The transition to winter shows a weaker convection in the western Pacific with a minimum in January. Moreover, the ITCZ is located around 12^oN in August and September, while it moves south to the vicinity of the Equator in March and April. Location of the ITCZ may vary depending on variations in solar forcing (Xie (1996), Xie (1996)). In the convective zones, the surface pressure is low and rainfall is high. The winds generally blow from the high pressure zones towards the convective regions. Therefore, the seasonal cycle also implies changes in the intensity and even in the direction of the winds. On average, there is a northward surface flow across the Equator due to the asymmetry of the ITCZ. The trade winds attain their maximum speeds around 15^oN and 15^oS. In the North, the maximum is attained in March and April, when the ITCZ is close to the Equator. In the South, the maximum happens in September. Concerning the SST, the moist air rises where the SST is highest while dry air subsides where waters are cold. For this reason, the warmest waters are in the western tropical Pacific and in a band of latitudes just north of the Equator, where the ITCZ is located. Seasonally, SST in the eastern tropical Pacific is at its maximum during the northern spring when the southeast trades are relaxed and are at a minimum during the northern summer and autumn when these winds are intense. The description in terms of the thermocline (which is the isotherm with a sea temperature of 20^oC) is that it is deeper in the west than in the east, and it

moves up and down depending on the season.

3.3 El Niño/Southern Oscillation

El Niño is the most dominant statistical and physical mode of climate variability on interannual human timescales (Philander, 1990). Their effects reach worldwide extension and can be reflected even in human health (Kovats et al., 2003). A phenomenological description needs the consideration of the main elements of the mean state of the equatorial Pacific ocean and atmosphere that play a role in its dynamics. Then, the El Niño/Southern Oscillation (ENSO) could be seen as a perturbation of the mean state described in section 3.2. Concerning the atmosphere, El Niño is characterised by the Southern Oscillation Index (SOI). The SOI is defined as the difference of pressure at the sea level between Darwin (Australia) and Tahiti. In normal conditions, the pressure in Tahiti is higher than in Darwin (winds go from east to west). An accepted theory is that El Niño is created by a perturbation located in the central Pacific Ocean that weakens the trade winds, eventually reverting them. For this reason, when El Niño is not present, the SOI is negative, while it is positive if El Niño is on stage. The wind stress perturbation creates a Kelvin wave that moves to the east, deepening the thermocline. This causes SST to increase in the South American coast and a local increase in moist convection. This involves a diabatic heating anomaly of the atmosphere. The Kelvin wave is, after some time, reflected from the eastern boundary as Rossby waves which will reinforce the instability growth in the east but will not travel back as the move poleward. At the same time in the west, westward Rossby waves are created and reflected in the western boundary as a Kelvin wave that moves fast to the east and helps to end the process. There are several models that describe the phenomenon mathematically. In the following subsections, we shall describe the most widely accepted schemes from different levels of complexity: from the coupled ocean-atmosphere model of Zebiak and Cane to the heuristic description of the oscillations given by the delayed and the recharge oscillators.

3.3.1 The Zebiak-Cane model

The Zebiak-Cane (ZC) model (Zebiak and Cane, 1987) is an ocean-atmosphere coupled model. The dynamics of the atmosphere follows the steady-state linear shallow-water on an equatorial beta plane approximation (Gill, 1980). A Rayleigh friction and a Newtonian cooling are used for linear dissipation. A

heating anomaly is used to force circulation. This anomaly depends on local heating and moisture convergence. The convergence feedback is introduced by iterating the set of equations using in each step n the convergence field in $n-1$. The feedback is nonlinear because the moisture-related heating is operative only when the total wind field is convergent, and this depends not only on the convergence anomaly, but also on the mean convergence (Eq. 3.18). The effect of the feedback is to focus the atmospheric response to SST anomalies into or near the regions of mean convergence, in particular, the Intertropical Convergence Zone (ICTZ) and the South Pacific Convergence Zone (SPCZ). The whole set of equations for the atmosphere are (Zebiak, 1986),

$$\epsilon u_a^n - \beta_0 y v_a^n = -(p/\rho_0)_x \tag{3.16}$$

$$\epsilon v_a^n + \beta_0 y u_a^n = -(p/\rho_0)_y \tag{3.17}$$

$$\epsilon(p/\rho_0) + c_a^2[(u_a^n)_x + (v_a^n)_y] = -\dot{Q}_s - \dot{Q}_1^{n-1} \tag{3.18}$$

$$\dot{Q}_s = (\alpha T) exp[(\bar{T} - T_a)/T_b] \tag{3.19}$$

$$\dot{Q}_1^n = \beta[M(\bar{c} + c^n) - M(\bar{c})], \tag{3.20}$$

where

$$M(x) = \begin{cases} 0, & x \leqslant 0, \\ x, & x > 0. \end{cases} \tag{3.21}$$

In this equations, $\bar{T}$ is the mean monthly SST, T is the anomaly in SST, $\bar{c}$ is the anomalous convergence. At iteration n, its value is,

$$\bar{c}^n = -(u_a^n)_x - (v_a^n)_y \tag{3.22}$$

The ocean basin is a rectangle at 124E to 80W and 29N to 29S. The dynamics are represented by a linear reduced gravity model, commonly used for simulating thermocline depth anomalies and surface pressure changes during El Niño events. Such models produce only depth-averaged baroclinic currents, but the surface current is usually dominated by a frictional component. Therefore, a shallow frictional layer of constant depth of 50 m is added to simulate the surface intensification of wind-driven currents in the real ocean. The dynamics of this layer are kept linear, and the Rayleigh friction is used to stand in for nonlinear influences at the equator (Eqs. 3.27 and 3.28). The influence of temperature changes occurring in the surface layer alone is neglected, so

only the thermocline depth variation is taken into account.

$$u_t - \beta_0 y v = -g' h_x + \tau^{(x)}/\rho H - ru \tag{3.23}$$

$$\beta_0 y u = -g' h_y + \tau^{(y)}/\rho H - rv \tag{3.24}$$

$$h_t + H(u_x + v_y) = -rh, \tag{3.25}$$

where

$$\mathbf{u} = \mathrm{H}^{-1}(\mathrm{H}_1\mathbf{u_1} + \mathrm{H}_2\mathbf{u_2}) \tag{3.26}$$

Indexes 1 and 2 refer to the surface layer and the underlying layer, respectively. The border between them is the thermocline. The equations for the interphase are,

$$r_s u_s - \beta_0 y v_s = \tau^{(x)}/(\rho H_1) \tag{3.27}$$

$$r_s v_s + \beta_0 y u_s = \tau^{(y)}/(\rho H_1) \tag{3.28}$$

$$\tag{3.29}$$

where

$$\mathbf{u_s} = \mathbf{u_1} - \mathbf{u_2} \tag{3.30}$$

The thermodynamics describe the evolution of temperature anomalies in the model surface layer. It includes advection by the mean and anomalous currents. The surface heat flux anomaly is proportional to the local SST anomaly, acting to adjust the temperature field toward its climatological mean state. The temperature equation for the surface layer is,

$$\frac{\partial T}{\partial t} = -\mathbf{u_1} \cdot \nabla(\bar{T} + T) - \bar{u_1} \cdot \nabla T - [M(\bar{w_s} + w_s) - M(\bar{w})]\,\bar{T_z} \tag{3.31}$$

$$-M(\bar{w_s} + w_s)\frac{T - T_e}{H_1} - \alpha_s T \tag{3.32}$$

where $\mathbf{u_1}$ and w_s are the mean horizontal surface currents and upwelling, respectively, $\bar{T}$ is the mean SST, which is prescribed, and $\bar{T_z}$ is the prescribed mean vertical temperature gradient. The function M accounts for the fact that surface temperature is affected by vertical advection only in the presence of upwelling. The anomalous vertical temperature gradient, T_z, is defined by

$$T_z = (T - T_e)/H_1 \tag{3.33}$$

where H_1 is the surface layer depth and T_e is the temperature anomaly entrained into the surface layer. This entrainment temperature anomaly, T_e, is defined by

$$T_w = \gamma T_{sub} + (1 - \gamma)T \tag{3.34}$$

The model parametrises subsurface temperature anomalies T_{sub} in terms of thermocline motions. More specifically, the parametrisation assumes a fixed vertical temperature profile for the thermocline structure and simply displacing this profile up and down with the thermocline depth determined by the shallow water dynamics (Eqs. 3.23-3.25). This temperature profile is fitted to a simple functional form based in observations,

$$\begin{aligned} T_{sub} = T_+\{\tanh[b_1(\bar{h}+h)] - \tanh(b_1\bar{h})\}, h > 0 \\ T_-\{\tanh[b_2(\bar{h}-h)] - \tanh(b_2\bar{h})\}, h < 0 \end{aligned} \tag{3.35}$$

where $\bar{h}$ is the prescribed mean upper layer depth. For the coupling, the ZC model considers that the ocean is forced by surface wind stress anomalies. This stress anomalies are generated by a standard bulk formula. The atmosphere model is steady-state and is run previously with specified monthly mean SST anomaly to simulate the wind anomalies. There are two approaches to simulate the changes. One is to calculate the steady response to the SST anomaly field at each time step, assuming that the atmosphere adjusts very rapidly (2-3 days) to the changes in the boundary forcing. This assumption cannot be justified. The second is to add the time dependence explicitly to the model. This is computationally costly, as inertial gravity waves requires a time resolution of two hours. The alternative proposed in the ZC model is to allow time dependence only in the moisture convergence component of the heating. Typical parameter values for this model are $\epsilon = 0.5\ day^{-1}$, $c_a = 60\ ms^{-1}$, $\alpha = 0.031\ m^2 s^{-3}/^o C$, $\beta = 1.6 \times 10^4\ m^2 s^{-2}$, $r = 0.4\ year^{-1}$, $c = 2.9\ ms^{-1}$, $H = 150\ m$, $H_1 = 50\ m$, $r_s = 0.5\ day^{-1}$, $\alpha_s = 0.008\ day^{-1}$, $\gamma = 0.75$, $T_a = 30\ ^oC$, $T_b = 16.7\ ^oC$, $T_1 = 28\ ^oC$, $T_2 = -40\ ^oC$, $b_1 = 1/80\ m^{-1}$ and $b_2 = 1/33\ m^{-1}$.

3.3.2 The delayed oscillator model

The delayed oscillator (DO) model (Tziperman et al., 1994) (a seminal DO model can be found in Suarez and Schopf (1988)), is a low-order chaotic system driven by a external force associated to the seasonal cycle. The basic idea is that the equatorial Pacific ocean-atmosphere oscillator can go into nonlinear resonance with the 12-month frequency oscillator and, with strong enough coupling between the ocean and the atmosphere, the system may become chaotic as a result of irregular jumping of the system among different resonances. The idea below the basic equation is the following: a positive SST perturbation along the eastern tropical Pacific weakens the easterly winds above the equator. The change in the winds excite a downwelling wave in the

thermocline that travels eastward as equatorial Kelvin waves and an upwelling signal that travels westward as equatorial Rossby waves. The Kelvin wave heatens the South American coast, what is considered as El Niño. The Rossby waves are reflected from the western boundary of the Pacific Ocean as an upwelling Kelvin wave, which collapses with the first downwelling Kelvin wave, ending the event. The heuristic equation of the thermocline depth anomalies h is,

$$\frac{dh(t)}{dt} = aA[h(t - \frac{L}{2C_K})]$$
$$-bA[h(t - (\frac{L}{C_K} + \frac{L}{2C_R}))] + c\ cos(\omega_a t) \tag{3.36}$$

where L is the basin width and ω_a is the annual frequency of the seasonal forcing. The first term on the right represents a wind-forced Kelvin mode that travels at a speed C_K, so it takes a time $L/(2C_K)$ to reach the eastern boundary from the middle of the basin. The second term is due to the westward-traveling Rossby wave of speed C_R excited by the wind a time $t - L/C_K - L/(2C_R)$ and reflected as a Kelvin wave. Considering that the speed of a Kelvin wave is three times faster than that of the Rossby wave, and defining the delay times by $\tau = \frac{L}{C}$, we can rewrite the model as,

$$\frac{dh(t)}{dt} = aA[h(t - \tau_1)] - bA[h(t - \tau_2)] + c\ cos(\omega_a t) \tag{3.37}$$

where $\tau_1 = 1.15$ months, $\tau_2 = 5\tau_1$, $a = 1/180$ days^{-1} and $b = 1/120$ days^{-1}. The function $A(h)$ relates wind stress to thermocline depth and SST. Its nonlinear form reflects the nonlinear relationships between the physical variables in the ocean and represents the gradient of the vertical temperature profile in it. The particular form of $A(h)$ is of little interest for this work. We shall only point that is an hyperbolic tangent modified by a straight-line segment near the origin (Tziperman et al., 1998). The nonlinearity is measured by a parameter κ. The slope of $A(h)$ at $h = 0$ is a measure of the strength of the coupling between ocean and atmosphere. In this model, as the nonlinearity measured by κ is increased, the system follows the quasi-periodicity route to chaos. For small κ, the solution is periodic with a frequency of ω_a, corresponding to the Annual Cycle. As κ increases, a second frequency ω_n associated to the natural frequency of the ocean-atmosphere system appears. The first and second frequencies are not commensurate, therefore an irregular, but not chaotic, oscillation arises. These irregular oscillations are not locked to the seasonal cycle. For even bigger κ, the system is again locked to a rational

multiple of the annual frequency, showing a nonlinear resonance between ω_n and ω_a. Finally, for sufficiently large nonlinearity, the system becomes chaotic and the phase-space reconstruction is that of a strange attractor. Nevertheless, the system stays partially locked to the frequency ω_a. Typical values of κ for the four cases described are 0.9, 1.2, 1.5 and 2.0 respectively.

3.3.3 The recharge oscillator model

The recharge oscillator (RO) model (Jin, 1997), is a low-order ENSO model derived from a simplified version of the ZC model using a two-box and a two-strip approximation (Jin (1997)). The upper ocean is a box model version of a shallow-water model for the equatorial ocean combined with a mixed layer of fixed depth. The atmosphere is approximated in terms of a linear relationship between surface winds and the SST gradient. A thermal relaxation towards a radiative-convective equilibrium temperature T_r is assumed. The heat budget can be expressed as:

$$\frac{dT_1}{dt} = -\alpha(T_1 - T_r) - \frac{u(T_2 - T_1)}{(L/2)} \tag{3.38}$$

$$\frac{dT_2}{dt} = -\alpha(T_2 - T_r) - \frac{w(T_2 - T_{sub})}{(H_m)} \tag{3.39}$$

where T_1 and T_2 represent the eastern and western equatorial temperature, respectively. $1/\alpha$ measures a typical thermal damping timescale, the term T_{sub} denotes the temperature being upwelled into the mixed layer, w is the upwelling velocity, u is the zonal advection velocity, which is taken as proportional to the wind stress. Finally, H_m and L denote the depth of the mixed layer and the basin width, respectively. As the equatorial upwelling is essentially due to Ekman flow divergence (the flow caused in the upper layer by wind stress) and its dominating meridional component is proportional to zonal wind stress τ we obtain a second pair of equations:

$$\begin{aligned} \frac{u}{(L/2)} &= \epsilon\beta\tau \\ \frac{w}{(H_m)} &= -\zeta\beta\tau \end{aligned} \tag{3.40}$$

where ϵ and ζ measure the strength of the zonal and vertical advection, respectively, and can be used as bifurcation parameters. Neglecting the role of the Hadley circulation and other external sources for wind stress anomalies,

the wind stress related to the Walker circulation can be expressed as,

$$\tau = \mu(T_1 - T_2)[\gamma cos(\omega_a t) + \sigma\xi - 1]/\beta \tag{3.41}$$

where $\sigma\xi$ represents a Gaussian white noise of variance σ^2 and γ is the strength of the Annual Cycle of frequency ω_a. A typical value for γ is 0.2. T_{sub} depends strongly on the thermocline depth, and can be parametrised as

$$T_{sub} = T_r - \frac{T_r - T_{r0}}{2}\left[1 - \frac{\tanh(H + h_2 - z_0)}{h^*}\right] \tag{3.42}$$

In this equation, h_2 is the depth departure of the eastern equatorial thermocline from its reference depth H, z_0 is the depth at which w takes its characteristic value and h^* measures the sharpness of the thermocline. The east-west contrast of thermocline depth is determined by the Sverdrup balance between the pressure gradient and equatorial wind stress. An instantaneous adjustment of the thermocline gradient to wind stress changes is assumed. This is a reasonable approximation because the traveling time of Kelvin waves in the equatorial Pacific is short compared to the dynamical adjustment time of the western equatorial Pacific thermocline. Its dynamics come from zonally integrated Sverdrup meridional mass transport resulting from wind-forced Rossby waves with timescale $1/r$. The dynamical equations for the thermocline are,

$$\begin{aligned} h_2 &= h_1 + bL\tau \\ \frac{dh_1}{dt} &= r\left(-h_1 - \frac{bL\tau}{2}\right) \end{aligned} \tag{3.43}$$

where h_1 and h_2 are the western and eastern thermocline depth anomalies respectively. The parameter b considers the efficiency of the windstress τ in driving the thermocline tilt. It is important to note that the wind stress equation is not (3.41) but $\tau = \mu(T_1 - T_2)/\beta$, where the Annual Cycle and the noise component are not considered because the wave propagation processes are filtered out. Hence, this model considers noise and seasonal forcing only in the temperature equations (3.38) and (3.39). Typical parameter values are, in this model, $T_{r0} = 16\ ^oC$, $T_r = 29.5\ ^oC$, $\alpha = 1/180\ day^{-1}$, $r = 1/400\ day^{-1}$, $H_m = 50\ m$, $H = 100\ m$, $z_0 = 75\ m$, $h^* = 62\ m$, $\mu = 0.0026\ K^{-1}\ day^{-1}$, $\mu bL/\beta = 22\ mK^{-1}$, $\xi = 1.3$ and $L = 15 \times 10^6\ m$. In this model, the zonal advection strength ϵ acts as bifurcation parameter. For small ϵ a periodic oscillation of period 2.5 years is observed. When ϵ reaches the value 0.082, a second oscillation appears, showing a period doubling situation. Increasing to

0.086 reveals irregular oscillations with decadal amplitude changes. A further increase to ϵ=0.12 gets back a periodic orbit of 4-5 years. The decadal changes are not forced externally, as previous models. This dynamical system can be understood as follows (Timmermann and Jin (2002),Timmermann (2003)): ENSO undergoes long-term amplitude modulations as well as period doubling bifurcations; the dynamical system (3.38) - (3.43) exhibits chaotic dynamics in a homoclinic/heteroclinic chaos scenario, where a homoclinic orbit associated with a saddle focus is characterised by an oscillation which grows in amplitude in an unstable manifold until a critical value is attained. Eventually, the orbit grows in a stable third direction, reducing the amplitude of the original oscillation around the tropical mean state.

3.4 Motivation for constructing a new model

As we have seen, the most commonly used ENSO models introduce the Annual Cycle *ad hoc*, usually as a combined effect of a sinusoidal wave plus atmospheric data. Our purpose is to find a combined model for both El Niño and the Annual Cycle without using external inputs, in order to achieve a better comprehension of the interaction between both oscillations. As it was previously stated, the separation of variables is a complicated task, as linear methods usually do not disentangle both modes because of the interactive coupling between the two oscillations. Our plan will be, first, to separate the different physical phenomena using new nonlinear methods of dimensionality reduction; second, to analyse the resulting data in order to find a suitable model for representation of the system. Once the model is found, we shall have a clearer view of how El Niño and the Annual Cycle interact. Finally, to predict future states and compare them with the real physical measured variables.

Chapter 4

Application of dimensionality reduction to climate data

This chapter is centred on the application of the nonlinear methods explained in chapter 2 to the climate phenomena explained in chapter 3. We shall show how linear methods fail in separating the oscillations corresponding to El Niño and the Annual Cycle. However, some nonlinear methods, in particular, Isomap, successfully disentangle both oscillations, therefore projecting the multivariate data space into a much lower dimensional space.

4.1 Separation of mathematical modes

As we previously explained, the statistical analysis of ENSO is mostly based on SST anomalies which are obtained by subtracting a mean Annual Cycle from the monthly averaged SST data. Extracting a time-varying Annual Cycle and an ENSO mode in a multivariate way from SST data is not simple and the results may strongly depend on the assumptions used by different methodologies. In particular, linear methods may fail to disentangle both modes since ENSO and the Annual Cycle exhibit in some sense a joint synchronised behaviour, as ENSO amplitude is strong during the boreal winter season. This behaviour is reminiscent of an interactive coupling between the two modes (Pikovsky et al., 2001). For this reason, the study of the interaction is of great importance for understanding the variability of ENSO. In the last years, several articles discussing how ENSO and the Annual Cycle interact in the tropical Pacific Ocean have been published (*e.g.* Xie (1995), Jin et al. (1996), Tziperman et al. (1998), An and Wang (2001)). There has been also some attempt to analyse the variability of the Annual Cycle

due to its interaction with El Niño by using linear methods of decomposition (Tozuka and Yamagata, 2003). This type of interaction, which could be non-linear, may lead to erronous conclusions when subtracting a constant Annual Cycle from SST data under consideration, as usually done in the analysis of ENSO dynamics. Therefore, the data space cannot be decomposed into a sum of linear subspaces each containing an independent variable because of the existing interaction. So that, the separation of the SST into physically independent modes is not possible. Our aim is to extract a low dimensional manifold where the whole physical system could be embedded. First of all, we shall apply the usual linear methods, as PCA. Later, we will consider nonlinear methods of dimensionality reduction.

4.1.1 Selection of datasets

We have taken Sea Surface Temperature (SST) data from public databases (*e.g.* `http://ingrid.ldeo.columbia.edu/`). More precisely, we have made use of two databases:

1. the Reynolds-Smith (RS) database (Reynolds and Smith, 1995) for the region limited by 89.5W to 149.5E and 20.5N to 20.5S, from November 1981 to October 2002, with a resolution of one month in time and one degree in space. This means that there are 5,124 spatial measurements for 252 months.

2. the Kaplan Extended (KE) database (Kaplan et al., 1998) for the region limited by 87.5W to 147.5E and 17.5N to 17.5S, from January 1856 to October 1981, with a resolution of one month in time and five degrees in space. In this case, the spatial resolution is 208 points with 1,762 time points.

RS is used to complement the data of KE in the November 1981 to October 2002 time range. The time evolution of the temperature on the Equator is shown in Fig. 4.1a. If we focus on a fixed position in the ocean, we can observe two principal oscillations, a rather regular one associated with the Annual Cycle, and an irregular one associated with ENSO (Fig. 4.1b).

There are several reasons to select these databases and these particular ranges of data. The reason to choose SST is, basically, because the phenomena

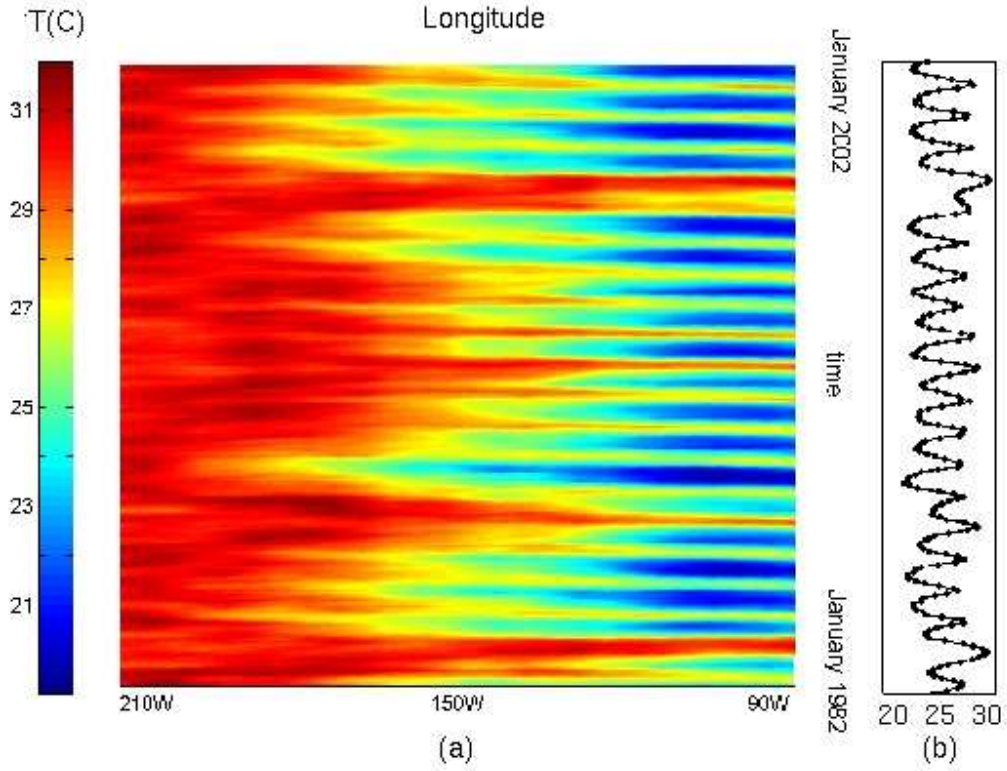

Figure 4.1: (a) Evolution of the temperature pattern along the Equator from January 1982 to January 2002 (resolution one month). We can observe the warm pool in the west and the cold tongue in the east punctuated by ENSO events (marked with arrows). (b) Time evolution of the temperature at the point (115E, 0N).

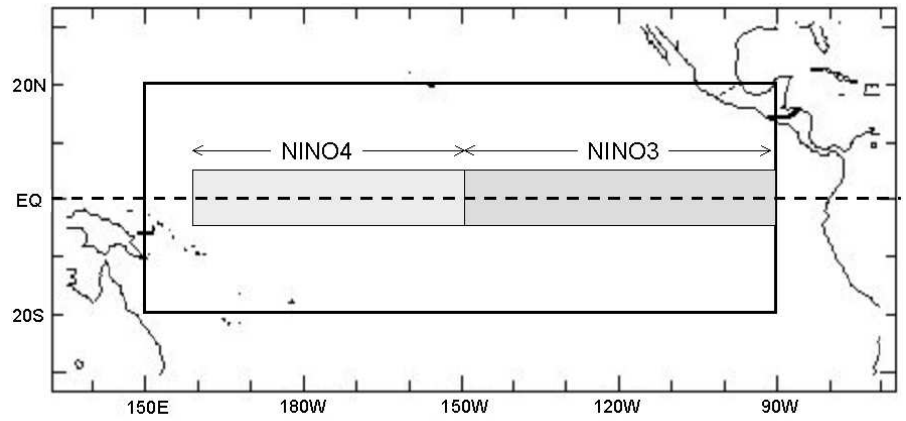

Figure 4.2: Map of the Pacific Ocean. The big rectangle shows the region under study in comparison to the regions where NINO3 and NINO4 are defined in the literature.

are very well reflected in the SST. Moreover, the ocean drives the atmosphere in the tropics (Patil et al., 2001). So we expect that the SST will be a good representation of the whole phenomenon. Focusing on the SST, we chose the reconstructed KE database because it is the longest time series available with a good resolution. Proxy data are not useful in our case because they lack of spatial regularity (Cobb et al., 2003). The region chosen in Fig. 4.2 is the maximum available if we want to explore the combined effect of ENSO and the Annual Cycle. If we extend the data further north and south, we shall overpass the tropics, where the physical processes are different. In particular, the Annual Cycle show a different behaviour as the ITCZ does not exist and ENSO is a second order oscillation.

The following results are obtained by using the normalised (unit variance and zero mean) KE database. This allows us to compare PCA and Isomap using the same normalised data. It is important to stress that the results obtained with both databases are similar. This means that the results are consistent independently of the length of the time series, supposing it is sufficiently large. A second point to consider is the planar approximation. We are assuming that the grid is cartesian, meaning that all the points are equally separated. Of course, due to the spherical shape of the Earth, this is not rigorously true, but the approximation is good to a 1% in the worst case of a point located in the tropical line.

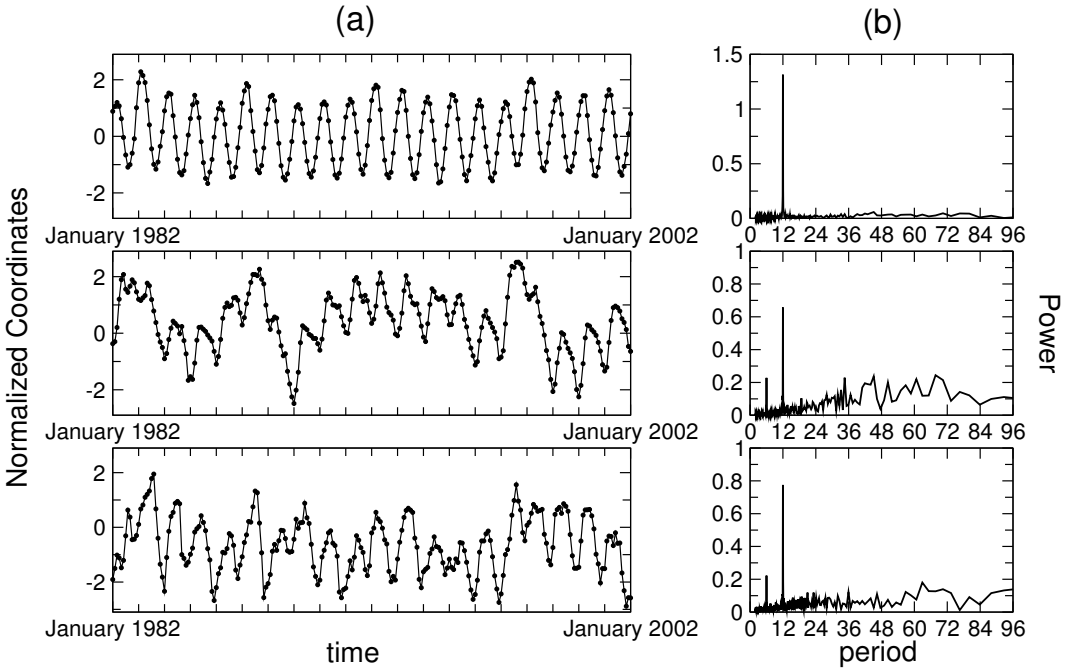

Figure 4.3: (a) Coordinates and (b) power spectra of, from top to bottom, the first, second and third principal components for the KE dataset from January 1982 to January 2002 (resolution one month) computed via PCA.

4.1.2 Linear methods

The analysis of SST with PCA shows an annual oscillation (Fig. 4.3a) which is present in all the principal components, as we check using the Fourier spectrum for each variable (Fig. 4.3b). The convergence of c_p with the number of dimensions is slow (Fig. 4.4). For this reason, it is difficult to select the number of dimensions that best describes the physical process. Depending of the criterion selected, we get different cut-offs in the number of relevant components. It is important to note that any available criterion is not based on mathematical or physical reasons, so there is an important lack of rigourousness when selecting a particular number of components.

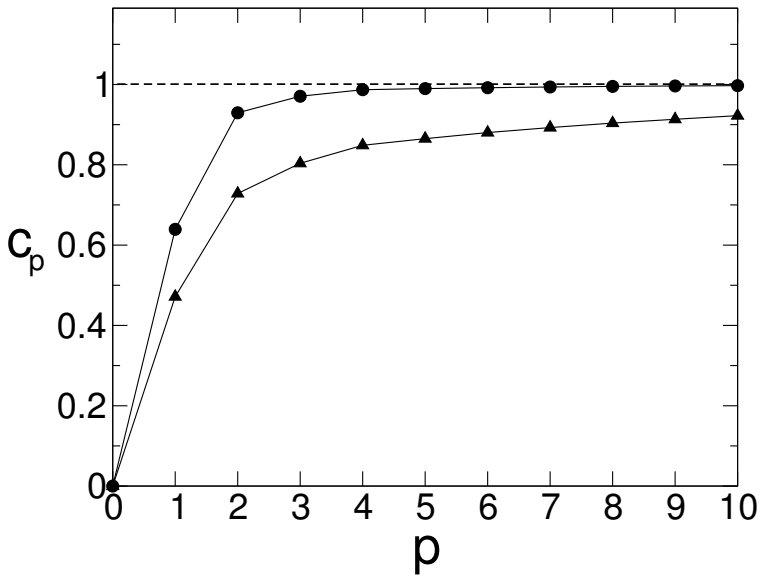

Figure 4.4: Cumulative variances c_p against the number of eigenvalues considered for Isomap (•) and PCA (▲).

4.1.3 Nonlinear methods. The mathematical modes made physical

We turn now our efforts to Isomap to compare its results in terms of dimensionality. For the KE database, the radius of the ball that defines the nearest neighbours was taken as $\epsilon = 9$. Similar results were found when ϵ was ranging from 9 to 30. Although the use of landmark points is suggested for $n > 1000$, we used the whole set of data points for calculating the eigenvalues of $\mathbf{Z^{(2)}}$. This was done so because, for the KE database, $n = 1762$ points, and the use of the Floyd's algorithm (Floyd, 1962) when computing distances saved enough time of computation without sacrificing accurateness in the calculation of the eigenvalues of $\mathbf{Z^{(2)}}$, as they are needed for finding the dimensionality of the data.

The cumulative variances calculated by PCA and Isomap are shown in Fig. 4.4. We observe that the dimensionality found by Isomap is three, while the convergence of PCA's cumulative variance is much slower. The first three

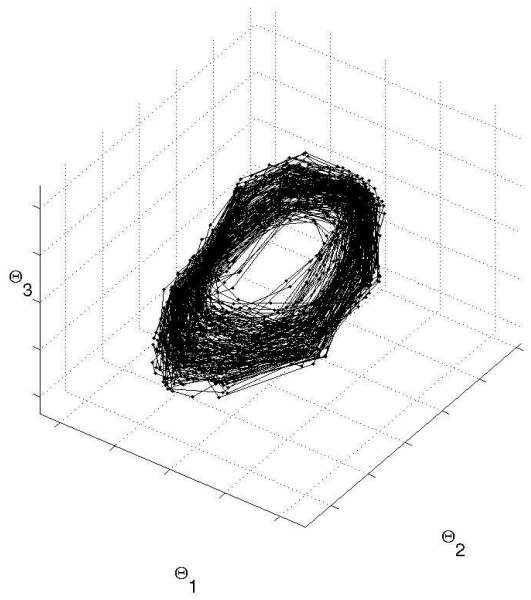

Figure 4.5: Embedding of the KE SST into the dimensions defined by the first three eigenvalues.

components found by Isomap are,

$$\begin{aligned} \Theta_1(\mathrm{t_j}) &= \lambda_1^{1/2}\mathbf{p}_{1\mathrm{j}} \\ \Theta_2(\mathrm{t_j}) &= \lambda_2^{1/2}\mathbf{p}_{2\mathrm{j}} \\ \Theta_3(\mathrm{t_j}) &= \lambda_3^{1/2}\mathbf{p}_{3\mathrm{j}} \end{aligned} \tag{4.1}$$

which we can plot in a three dimensional coordinate space (Fig. 4.5). We can observe a twelve-month oscillation in a plane with some deviation in a coordinate perpendicular to that plane. To isolate the twelve-month oscillation we can rotate the new three dimensional points because the representation in terms of distances is equivalent to any orthonormal transformations. The optimal plane can be found by computing the plane that best embeds a mean 12-month cycle over the whole trajectory. After applying this particular rotation,

$$(\Theta_1, \Theta_2, \Theta_3) \longrightarrow (x_1, x_2, x_3) \tag{4.2}$$

the three time series shown in Fig. 4.6a are extracted. We can see that there are major differences between the second and third components of Fig. 4.3a

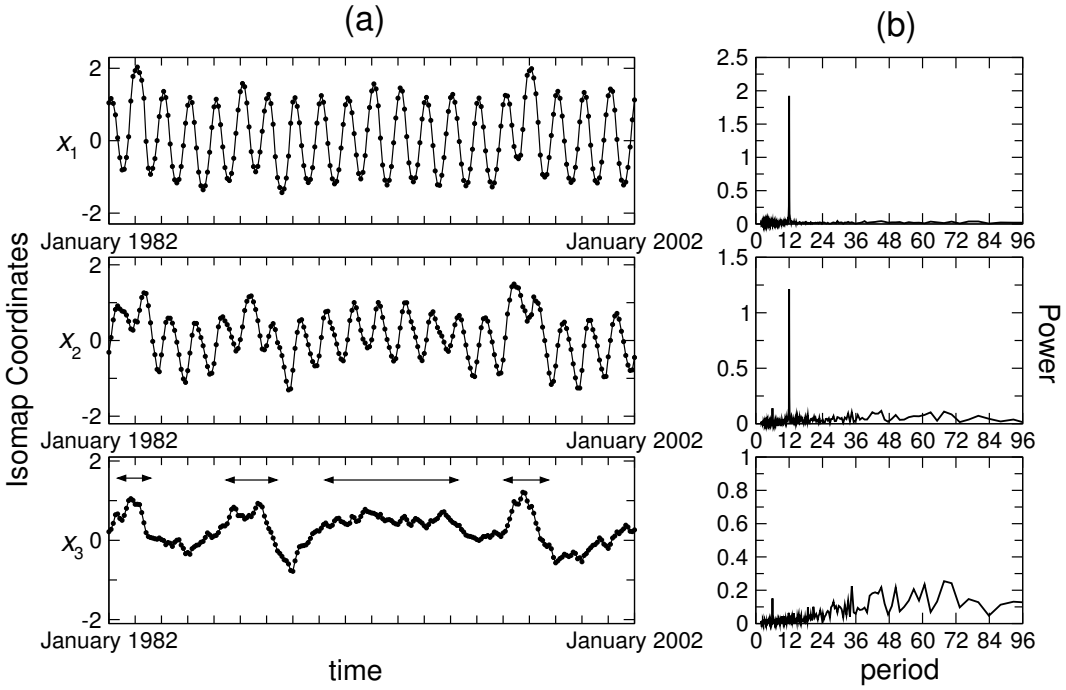

Figure 4.6: (a) Coordinates and (b) power spectra of, from top to bottom, the first, second and third principal components for the KE dataset from January 1982 to January 2002 (resolution one month) computed via Isomap. The double headed arrows indicate the approximate duration of El Niño events.

and Fig. 4.6a. Moreover, the third Isomap component faithfully represents ENSO, as all the well known events have their corresponding peak in the time series of the third component. Unfortunately, there is some correlation between the third variable and the amplitude of the oscillations of the first and, specially, the second variables. This suggests that the separation cannot be complete due to the interaction between oscillations, which is particularly represented in the second variable.

The results show that the complex dynamics due to the interaction between ENSO and the Annual Cycle can be well approximated by a three dimensional manifold. It is worth noting that the selection of the number of variables comes naturally from the method of dimensionality reduction used. No criterion was chosen to cut the series of components other than the stability of the variance explained. In fact, if we go back to the PCA results in section 4.1.2 and we make use of the information obtained with Isomap, we can obtain a similar three dimensional space where, with a carefully selected

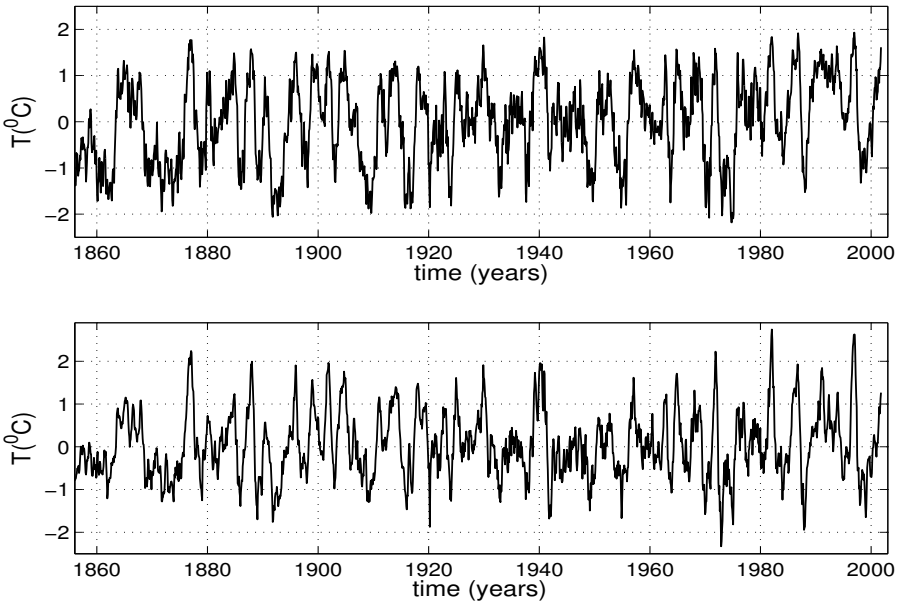

Figure 4.7: Third normalised Isomap coordinate (top) in comparison with the index NINO3.4 (bottom). The amplitude of the events changes slightly from one method to the other.

rotation, similar components to those in 4.2 can be found. This suggests that the interaction between the Annual Cycle and ENSO, if nonlinear, shows only a weak nonlinearity.

We would like to observe that PCA introduces a particular normalisation in the data because the matrix of the eigendecomposition is the correlation, which is naturally normalised to variance one. But we can apply MDS and Isomap to the raw data, without any other mathematical operation. The results are similar to the usual indexes that describe ENSO, as we can see in Fig. 4.7. This plot shows how this decomposition provides a useful way to characterise ENSO by using the third Isomap component. Moreover, we observe that the occurrence of the events is essentially preserved, although the amplitude and the probability distribution function found by the decomposition are slightly different. This is due to the fact that now the Annual Cycle is not approximated by a periodic function, as in the NINO3.4 or other indexes.

In summary, we have shown that the SST in the tropical Pacific Ocean can be described by a three dimensional system. Contrary to other nonlinear methods of dimensionality reduction (Hsieh, 2001), Isomap offers a physi-

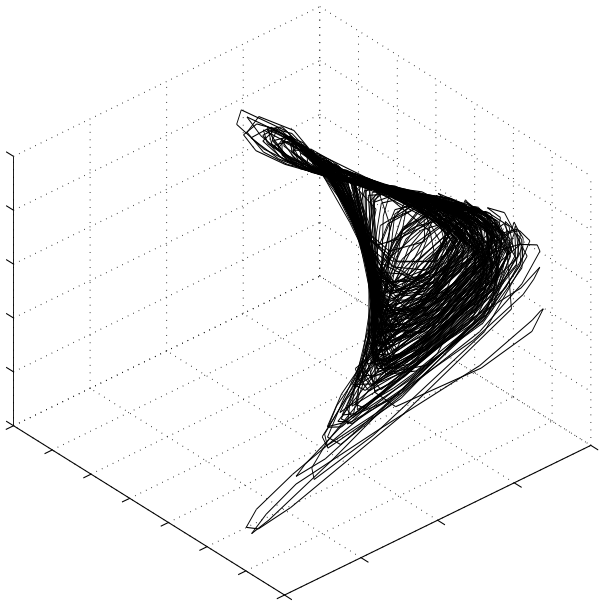

Figure 4.8: Embedding of the KE SST into the dimensions defined by the first three eigenvalues. Arbitrary units not shown.

cally appealing method to face the problem. One could ask if LLE is also suitable for accomplishing this task. The answer comes by inspection of Fig. 4.8. We can observe how the components found by LLE are somewhat twisted and, because of that, they cannot be decomposed in a physically meaningful way as Isomap components are. Therefore, our task, in the next chapter, will be to find a model that can be adjusted to our three new time series.

Chapter 5

Prediction of the coupled system El Niño-Annual Cycle

In this chapter, we use two different methods to predict states of the system composed by El Niño and the Annual Cycle (Gámez et al., 2006a). The first method tries to model a nonlinear dynamical system in order to represent the dynamics of the data. Once found, we shall try to reproduce the Isomap components found in chapter 4. The second method is based on ensemble prediction: the dataset will be used to construct an ensemble of models in order to predict states of the system. We shall point out the efficiency of the methods as compared to different methods and models found in the literature. The limits of the prediction will also be outlined.

5.1 Estimation of a model using nonparametric regression

Let us consider that each point $\boldsymbol{x}(t) = (x_1, x_2, x_3) \in \mathbb{R}^3$ of the output of Isomap is a realisation of the flow in phase space represented by the general equation,

$$\dot{\boldsymbol{x}} = \Phi(\boldsymbol{x}(t)) \tag{5.1}$$

where $\Phi : \mathbb{R}^3 \to \mathbb{R}^3$, one obtains as a best estimator of (5.1) in the least-square sense (Kantz and Schreiber, 1997),

$$\Phi_i = E\left[\dot{x}_i \,|\, x_1, \ldots, x_m\right] \ , \tag{5.2}$$

with $E[\cdot|\cdot]$ the conditional expectation value (CEV) operator. In practice the CEV is calculated by using the neighbours of $\boldsymbol{x}$, as in most nearest neighbour methods. Extracting analytical models from this formula is very hard, and visualization is obviously impossible for $m > 2$.

To encounter this drawback, we start building the model from the simplest version in terms of estimation and cast the r.h.s. of Eq. (5.1) into an additive model:

$$\Phi_i = \sum_{j=1}^{m} \phi_{ij}(x_j) \ . \tag{5.3}$$

After having estimated the components $\phi_{ij}(x_j)$, we can easily visualise the functions and try analytical formulae for the ϕ_{ij}. With (5.3), we have a less general model than (5.1), but a wider model class than in parametric methods, because we do not rely on a given set of basis functions. Rigorous results related to a nonparametric additive representation of a general function are given in a classical work of Kolmogorov and later on refined by Vitushkin (Vitushkin, 1977). If it turns out that the strictly additive model is not sufficient, successively more complicated dependencies will be taken into account.

The optimal estimate for the ϕ_{ij} is calculated by the backfitting algorithm (Haerdle and Hall, 1993), which is a class of the general Alternating Conditional Expectation value (ACE) algorithm (Breiman and Friedman, 1985). It works by alternately applying the CEV operator to projections of Φ_i on the coordinates: $\phi_{ij}(x_j) = E[\dot{x}_i - \sum_{k\neq j} \phi_{ik} \mid x_j]$, and is proven to converge to the global optimum. For the application to spatio-temporal data analysis, see Voss et al. (1998) and Abel (2004). The estimation of the CEV can be realised numerically by running mean, local linear fits, polynomials, etc... In Haerdle and Hall (1993) it is shown that smoothing splines are optimal for nonparametric regression; they are suitable for many applications because of their differentiability. It is important to note that the parameters used by splines, averages, etc... are method-inherent and not prescribed by a preselected model. In this sense, the model is parameter free.

As an overall quality measure, instead of the least-square error, we use the correlation coefficients C_{i0} between the r.h.s. and the l.h.s. in Eq. (5.3). The correlation C_{ij} between the jth term in the sum (5.3) with the sum of the remaining terms indicates its statistical weight for the model:

$$C_{ij} = C\left[\phi_{ij}\,;\,\Phi_i - \sum_{k\neq j}^{m} \phi_{ik}\right] \tag{5.4}$$

where C is the correlation function. A correlation close to 1 means a good model, close to 0 a bad one. It is worth noting the geometrical aspect of the approach: Eq. (5.1) defines a hypersurface in $\mathbb{R}^m$, approximated by the sum of the functions ϕ_{ij}, cf. (5.3). Naturally, the dynamical and topological properties of the original system must be mirrored in embedding space. Long-term predictions of the dynamics are thus possible on the basis of the obtained model.

In the particular case of the system composed by El Niño and the Annual Cycle, we shall consider the three time series as a minimal input for modeling a dynamical system that could represent the whole dataset. This system can be modeled as a dynamical system as follows,

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, x_3) \\ \dot{x}_2 &= f_2(x_1, x_2, x_3) \\ \dot{x}_3 &= f_3(x_1, x_2, x_3) \end{aligned} \tag{5.5}$$

This will be our first ansatz. In principle, f_1, f_2 and f_3 could be any non-linear functions of the variables x_1, x_2 and x_3, which are the three Isomap coordinates. This ansatz is essentially mathematical. But, to bring the model closer to physical reality, we could also admit the existence of a delayed term related to the second and third variables, $x_2(t-\tau)$ and $x_3(t-\tau)$. From now on, we rename $x_4(t) = x_3(t-\tau)$ and $x_5(t) = x_2(t-\tau)$ to simplify the notation. As the third variable is associated to El Niño process, and the second one is strongly correlated to it, the introduction of the delay terms is bond to the fact that the existence of Kelvin and Rossby waves are capital to the evolution of the process, as explained in the delayed oscillator scheme. Anyway, the method is able to neglect its effect if the correlation is low. So that, if the delay is not affecting the data evolution, the variables $\dot{x}_1$, $\dot{x}_2$ and $\dot{x}_3$ would not depend on x_4 or x_5. As we shall see later, the first variable is not directly affected by delayed variables, as x_1 is essentially capturing the effect of the Annual Cycle. The model then is transformed into

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, x_3) \\ \dot{x}_2 &= f_2(x_1, x_2, x_3, x_4, x_5) \\ \dot{x}_3 &= f_3(x_1, x_2, x_3, x_4, x_5) \end{aligned} \tag{5.6}$$

The model depicted by the Eqs. (5.6) represents an oscillatory process in the first two equations, with a main contribution by El Niño in the second. The third equation is, essentially, the evolution of El Niño. Here, two

contributions are considered, one is the delay and the second is the coupling to the Annual Cycle represented by the solutions of the first two equations. Even considering the new contributions, the model (5.6) is too complicated to be parametrised. Thus, more heuristic reasoning is required to simplify the equations. Obviously, a period of twelve months has to appear in the solution of variables x_1, x_2. We require therefore that the annual period is reproduced by the first two equations and assume that the contribution of x_1 to the first equation and of x_2 to the second one is linear. Furthermore, we shall restrict the model to be fully additive. Including these information we obtain the model

$$\begin{aligned}
\dot{x}_1(t) &= -c_1 x_2(t) + f_{12}(x_2(t)) + f_{13}(x_3(t)) \\
\dot{x}_2(t) &= c_2 x_1(t) + f_{21}(x_1(t)) + f_{23}(x_3(t)) + f_{24}(x_4(t)) + f_{25}(x_5(t)) \\
\dot{x}_3(t) &= f_{31}(x_1(t)) + f_{32}(x_2(t)) + f_{33}(x_3(t)) + f_{34}(x_4(t)) + f_{35}(x_5(t))
\end{aligned} \tag{5.7}$$

The frequency of the Annual Cycle can be analitically calculated as $\omega = \sqrt{c_1 c_2}$. The functions f_{ij} where $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$ are nonlinear corrections, not necessarily polynomials. This will be the input to the method we shall explain in the next chapter.

Considering that not all the variance of the data will be explained by Eqs.(5.7), a stochastic part μ_i of the measurement is added to represent the model uncertainty. The corresponding model reads,

$$\begin{aligned}
\dot{x}_1(t) &= -c_1 x_2(t) + f_{12}(x_2(t)) + f_{13}(x_3(t)) + \mu_1 \\
\dot{x}_2(t) &= c_2 x_1(t) + f_{21}(x_1(t)) + f_{23}(x_3(t)) + f_{24}(x_4(t)) + f_{25}(x_5(t)) + \mu_2 \\
\dot{x}_3(t) &= f_{31}(x_1(t)) + f_{32}(x_2(t)) + f_{33}(x_3(t)) + f_{34}(x_4(t)) + f_{35}(x_5(t)) + \mu_3
\end{aligned} \tag{5.8}$$

These equations can be numerically integrated using standard algorithms for stochastic equations. The Euler integration is sufficiently accurate for our purposes.

5.1.1 Results of the Regression

As a preprocessing, we have to calculate the derivative from the Isomap data series. We used a spectral estimator, which turned out to be numerically well behaved and fast. For details see Abel et al. (2005). For the delay, it turned out that a value of approximately $\tau = 6$ months yeld an optimal correlation

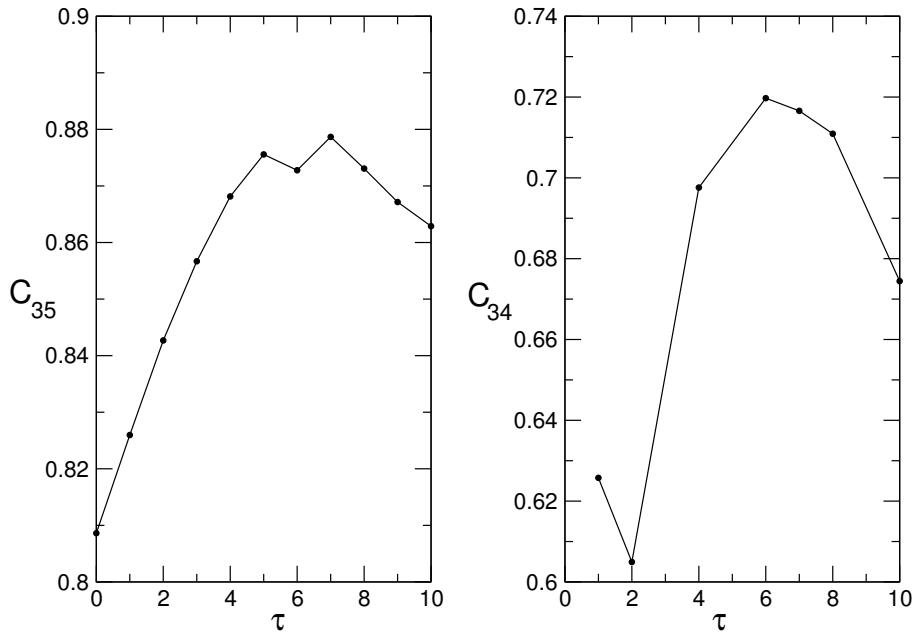

Figure 5.1: Variation of the correlation C in the model when introducing delay terms in the functional dependence. The left graph is for x_5 and right graph is for x_4

for the second and third variable, which is consistent with the time evolution of Kelvin waves in the Pacific ocean, as explained in Chapter 3. This is shown in Fig. 5.1.

Now, we shall take a closer look into the first equation. A scatter plot of $\dot{x}_1$ vs. x_2 yields Fig. 5.2a. A slope of value $c_1 = -1$ approximates quite well the optimal straight line through the data. A fully nonlinear analysis yields the thick line in Fig. 5.2a, which, up to the usual effects at the boundaries (Abel, 2004), follows quite nicely a curve with a strong linear behaviour. For c_2, although the linear analysis yields a value of 0.30, we fix it to be 0.27416 in order to get a main oscillation of annual period. The corresponding plot is in Fig. 5.2b. The reason to adjust these coefficients *ad hoc* is that, if we input the functions directly obtained by regression, we do not reproduce the annual oscillation because, due to the lack of data, the linear coefficients c_1 and c_2 are not accurate enough. This can be seen as a rough approximation to the data. Nevertheless, we see in Fig. 5.2 that a nonlinear polynomial fitting yields similar results up to the usual effects at the boundaries (Abel, 2004).

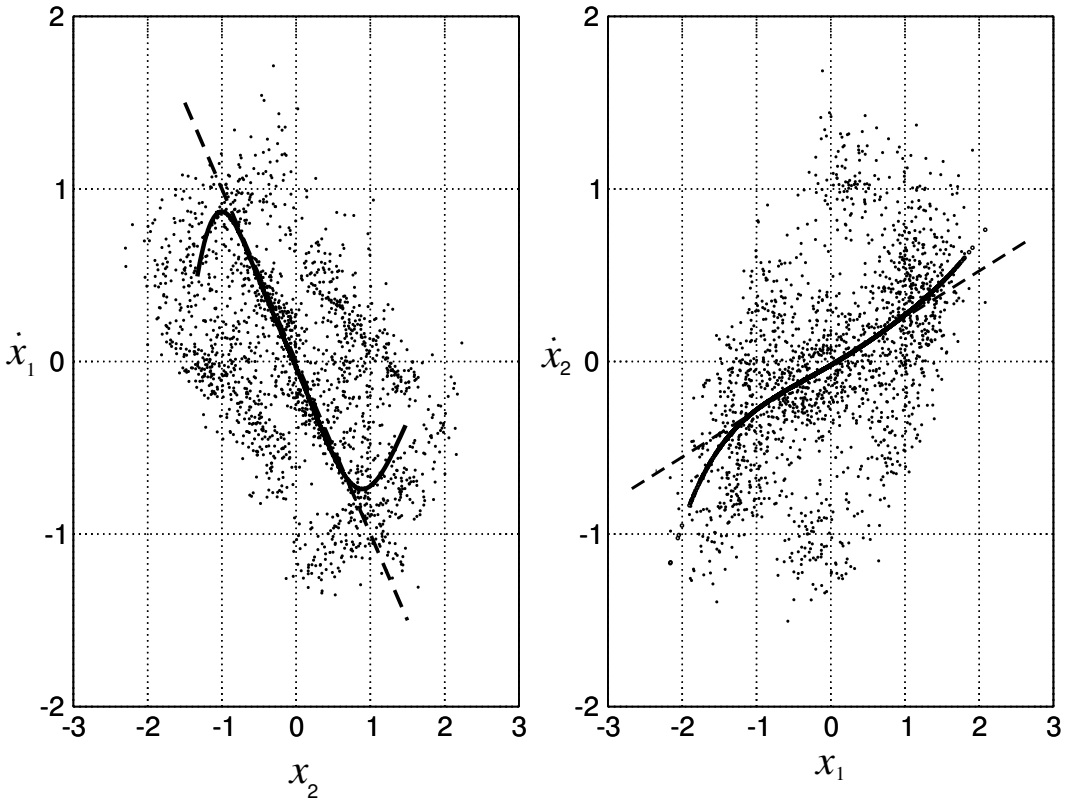

Figure 5.2: (a) Plot of $\dot{x}_1$ vs. x_2. A straight line of slope $c_1 = -1$ is a good approximation of the nonlinear function fit (thick line). (b) Plot of $\dot{x}_2$ vs. x_1. A straight line of slope $c_2 = 0.27416$ approximates the nonlinear function fit (thick line)

A nonlinear fit does not work particularly well due to insufficient number of data and the very inhomogeneous scatter. So that, for the reasons previously explained, we shall consider that the functions f_{12} and f_{21} are linear. The third component x_3 has a significant linear coupling in the first and second equation with functions shown in Fig. 5.3, while the rest of the variables are not relevant (not shown).

The most relevant equation in terms of the system and also in terms of the modeling difficulties is the third equation. It turns out that the x_1 and x_5 components are close to zero for central values. The deviation at the boundaries is mainly due to the inhomogeneous data distribution. The correlations, however are high. So that, there are two options for the modeling: the first could consider that the contribution of these components is not relevant, the

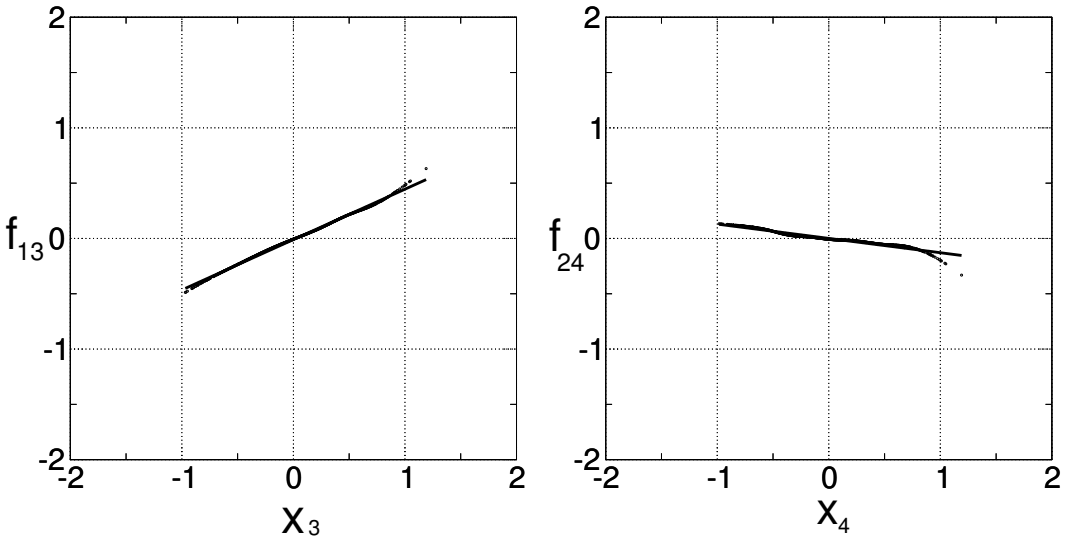

Figure 5.3: The functions representing the different functional dependencies between $\dot{x}_1$ and x_3(left) and between $\dot{x}_2$ and $x_4(t)$(right).

second would try to adjust some nonlinear functions to the data. When we run the model, we find that f_{31} is responsible for a rapid divergence of the solutions, while f_{32} is not affecting the stability of the system. Therefore, we shall assume that $f_{31} = 0$. For the x_3 and x_4 components, the contributions are relevant, in good agreement with the model of delayed oscillator (Tziperman et al., 1994). Summarising all these considerations, the regression of the functions represented in Fig. 5.4 gives the following results for the functions defined in the Eqs. (5.8):

$$\begin{aligned} c_1 &= -1 \\ c_2 &= 0.27416 \\ f_{12} &= 0 \\ f_{13} &= -0.01 + 0.45x_3 \\ f_{21} &= 0 \\ f_{24} &= -0.13x_4 \end{aligned}$$

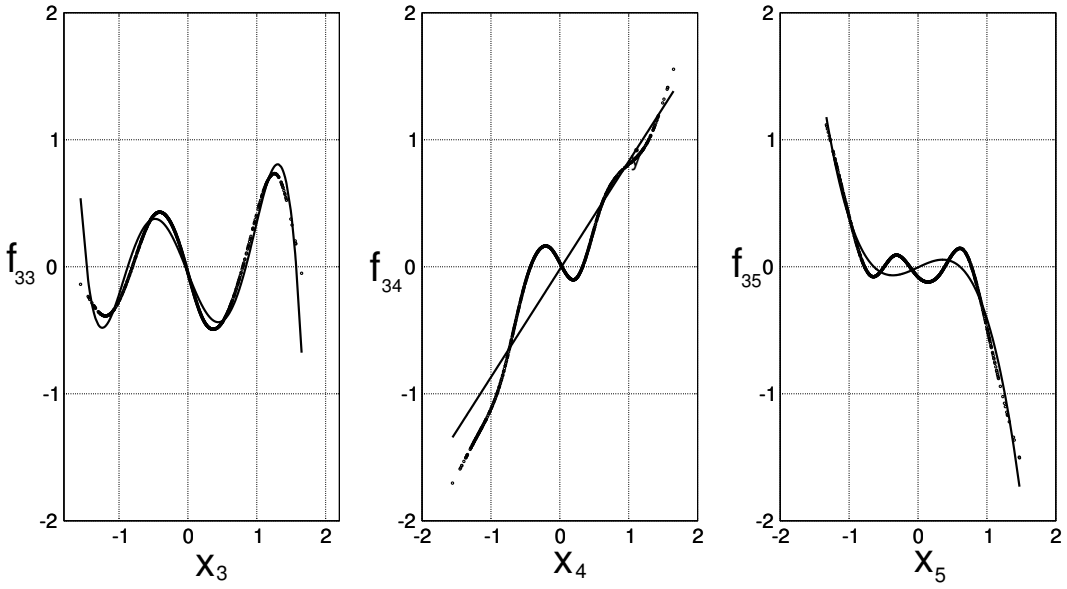

Figure 5.4: From left to right, the functions representing the three different functional dependencies in the $\dot{x}_3$ equation: x_3, x_4 and x_5.

$$\begin{aligned} f_{31} &= 0 \\ f_{33} &= 0.01 - 1.33x_3 + 2.35x_3^3 - 0.75x_3^5 \\ f_{34} &= -0.02 + 0.85x_4 \\ f_{35} &= 0.26x_5 - 0.01x_5^2 - 0.65x_5^3 \end{aligned}$$

The first equation is purely linear. It contains a term that develops a regular Annual Cycle and a linear contribution of El Niño that modulates it. The second equation contains the regular cycle term and a linear contribution from the delayed wave originated by El Niño. Finally, the third equation contains two terms that evoke the delayed oscillator paradigm with a nonlinear term and a complex nonlinear contribution from the interaction variable $x_2(t-\tau)$. This is the model that we are going to integrate in the next section.

5.1.2 Simulation of the Model

The method we are using for the integration is the classic Euler algorithm with a time step $\Delta t = 0.001$, meaning that a thousand integrations are needed to make the system evolve one month. This is sufficient for our purposes, as the Runge-Kutta algorithm is much more tricky due to the presence of noise

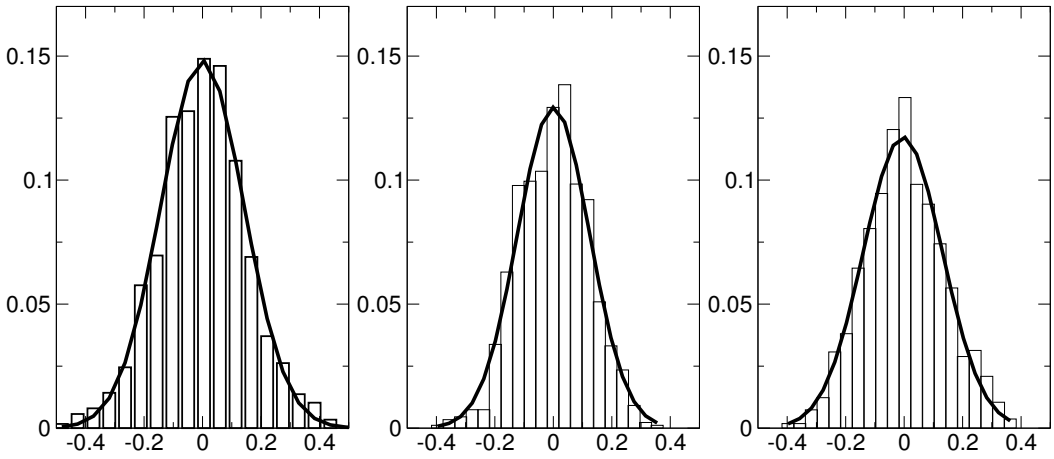

Figure 5.5: From left to right, normalised probability distribution functions corresponding to the $\dot{x}_1$, $\dot{x}_2$ and $\dot{x}_3$ equations. The lines correspond to the corresponding approximations to gaussian distributions.

and delay in the same equations. To model the stochastic part in Eqs. (5.8), μ_1, μ_2 and μ_3, we shall fit different functional forms of noise to the residuals of the backfitting algorithm. As we see in Fig. 5.5, a good approximation is the selection of gaussian white noise with variance σ_1, σ_2 and σ_3, that can be calculated directly from the fitting. The selection of functional forms that are different from the gaussian does not significantly change the results. We also consider different noise strengths in order to find the most adequate fitting to the results.

The strength of the noise is fundamental when analysing the stability of the model (5.8). The three components behave differently when noise is increased. For the first and the second, as representations of the Annual Cycle, increasing the noise implies an increase in the amplitude of the cycle, and the amplitude of the oscillations tends to infinity as time increases (not shown). For the third variable, which is a representation of the ENSO oscillation, the dynamical behaviour is richer. As we see in Fig. 5.6, in absence of noise, there is a relaxation to a fixed point with a transient time of approximately one year. As noise is increasing (variance 0.1, 0.5, 1.0 and 5.0), the system undergoes irregular oscillations at low ($\sigma_3 = 0.1$) variance first, then regular oscillations are created ($\sigma_3 = 0.5, 1.0$) that are smoothly destroyed when noise is increasing to reach irregular oscillations. We observe that a noise of $\sigma_3 = 5.0$ promotes irregular oscillations that are, in a mean, partially phase-locked to

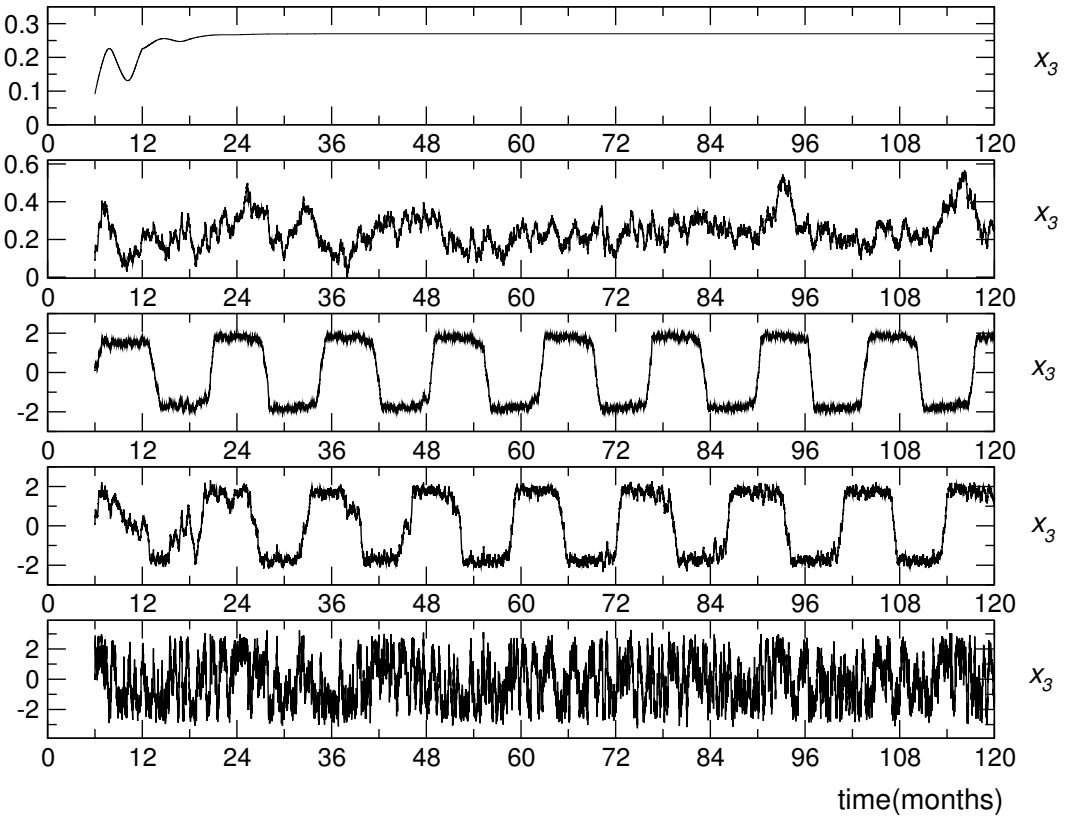

Figure 5.6: From top to bottom, sample simulation of the $\dot{x}_3$ equation for noise variances $\sigma_3 = 0, 0.1, 0.5, 1.0, 5.0$. The approximate value for σ_3 taken from the backfitting algorithm is equal to 0.136.

rational multiples of six months. Increasing the noise to values greater than 5.0 suppose non-physical solutions that we are not considering. In fact, the physical solutions are similar, in the long run, to the solutions found for noise around $\sigma_3 = 0.1$, which is, precisely, the value found by adjusting the variance unexplained by the backfitting algorithm (Fig. 5.5).

5.1.3 Prediction of El Niño events. Limits of the model

In this section, we are using the model depicted in section (5) to predict future states of the system. It is worth pointing out that the predicted points in the low-dimensional system can be converted into approximate spatio-temporal SST values by parametrising a function that transforms the spatio-temporal

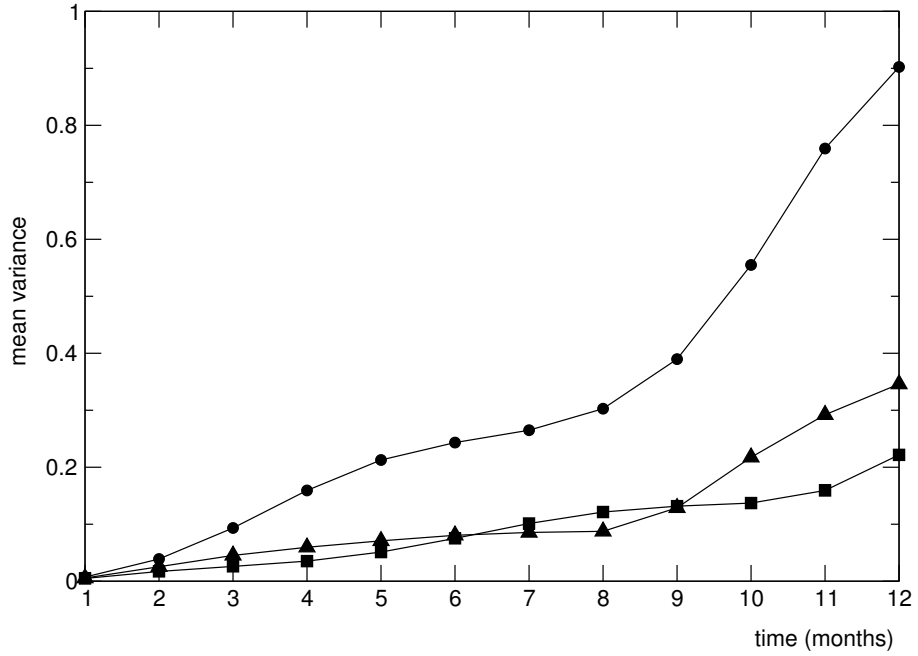

Figure 5.7: Mean variance of the first (− • −), second (−■−) and third (−▲−) variable for the fifty representations of the predicted data points for different time predictions.

data in the Isomap points and its inverse (see, for example, Poggio and Girosi (1990)).

We shall make calculations for prediction times ranging from one month to one year. The noise variance will be set to the adjusted values $\sigma_1 = 0.142$, $\sigma_2 = 0.123$ and $\sigma_3 = 0.136$. We run fifty realisations of noise in order to make a mean value for each predicted point. The mean variance of each predicted point can be calculated to get a measure of the probability distribution of the different realisations. This is done in Fig. 5.7. We observe how the highest variance for all prediction times stands for the first variable, while the second variable is the less spread in terms of noise dispersion. This is a reflection of the magnitude of the variables. The first variable is oscillating between -1.5 and 1.5, while the second and third oscillate around -1 and 1.

Two typical measures of the model prediction skills are the statistical correlation coefficients and the root mean square (r.m.s) error. In Fig. 5.8, we observe that, as expected, the prediction capabilities of the model worsens when the time for prediction is increased. This is reflected in the fact that

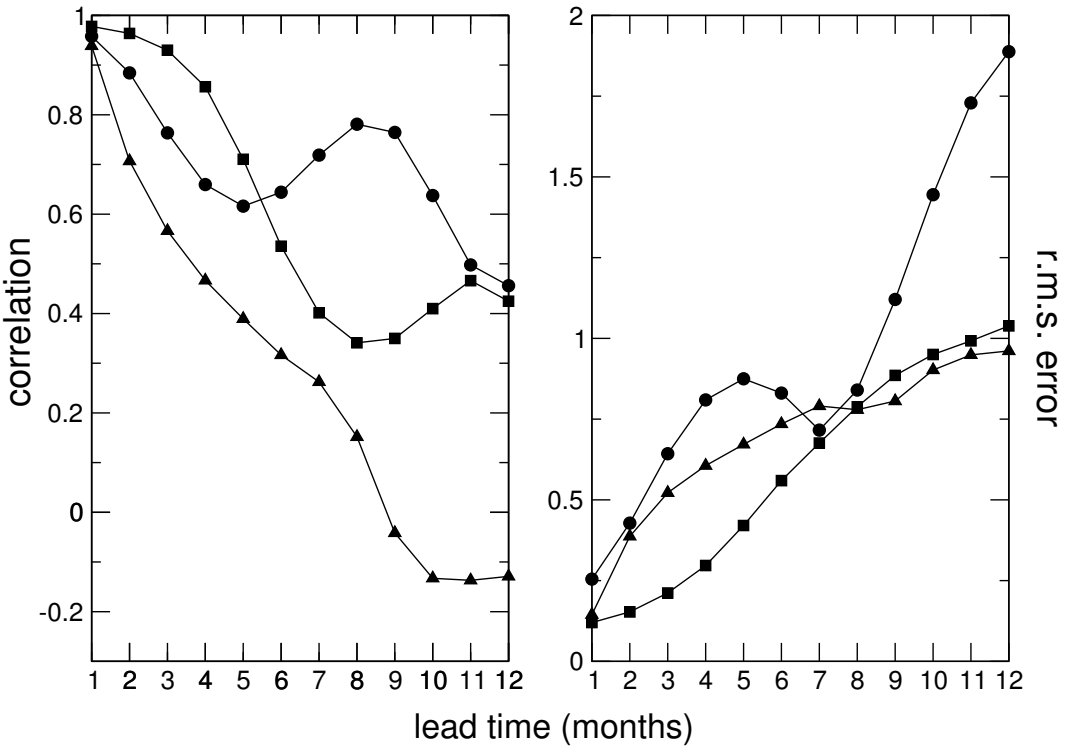

Figure 5.8: Correlation coefficients (left) and root mean square error (right) of the first (−•−), second (−■−) and third (−▲−) variable as a function of the lead time used for predicting the observed data.

the correlation coefficients are approaching to zero and the r.m.s errors are increasing. It is interesting to note that the correlation coefficients are independent of the magnitude of each variable (they are defined in the interval from -1 to $+1$). On the other hand, the r.m.s. error is magnitude dependent. Variables with higher values will tend to have higher r.m.s. errors, therefore the difference between the errors in the first variable as compared to the other two. We could also normalise the r.m.s. error by dividing by the variance, but this is not common in climate literature. Therefore, we shall keep the definitions as they are. It is noteworthy to say that the system works worse after certain conditions of high El Niño events. If we measure the correlation for different decades, we observe big differences in the predictability of the system. In particular, the 1900 and the 1980 decades were particularly bad for prediction, as we see in Fig. 5.9.

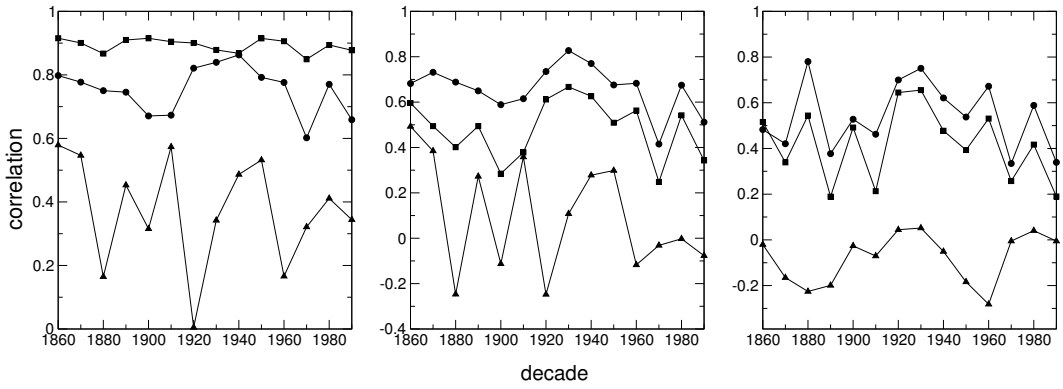

Figure 5.9: Correlation skills of the first (−•−), second (−■−) and third (−▲−) variable in different decades of the time series for different lead times: three months (left), six months (middle) and twelve months (right).

Summarising, the model faithfully represents the first and second variables for lead times under one year, while the third variable shows lower predictability in most of the decades for lead times bigger than six months. It is important to say that the dynamical system proposed here does not achieve the prediction capabilities for the third variable of some other ENSO models found in literature (see Latif et al. (1998), Latif et al. (1994) for reviews), although beats the delayed-oscillator model for particular decades (Kirtman and Schopf , 1998). Anyway, there are several drawbacks to consider in our case: first, there are few data points to input in the backfitting algorithm; second, the system is not trained; third, it is not artificially pumped with external data; fourth, it does not consider effects out of the boundaries of the tropics and is not sustained by the atmosphere. Therefore, we expect that the dynamical system could be improved by considering some atmospheric influence, more specifically the related with the creation of Kelvin waves by easterly winds in the western Pacific Ocean. It is interesting to note that the noise intensity is not improving the results. In Fig. 5.10 we observe how increasing the noise worsens the model, as expected. The somewhat surprising behaviour comes when analysing differences between the regimes of low noise and no noise. We observe almost no differences between the performance of the model. This suggest, as previously stated, that the unexplained variance using the backfitting algorithm cannot be fully explained by noise, but also by some external variable.

We can plot some predicted intervals in order to get a clearer view of what

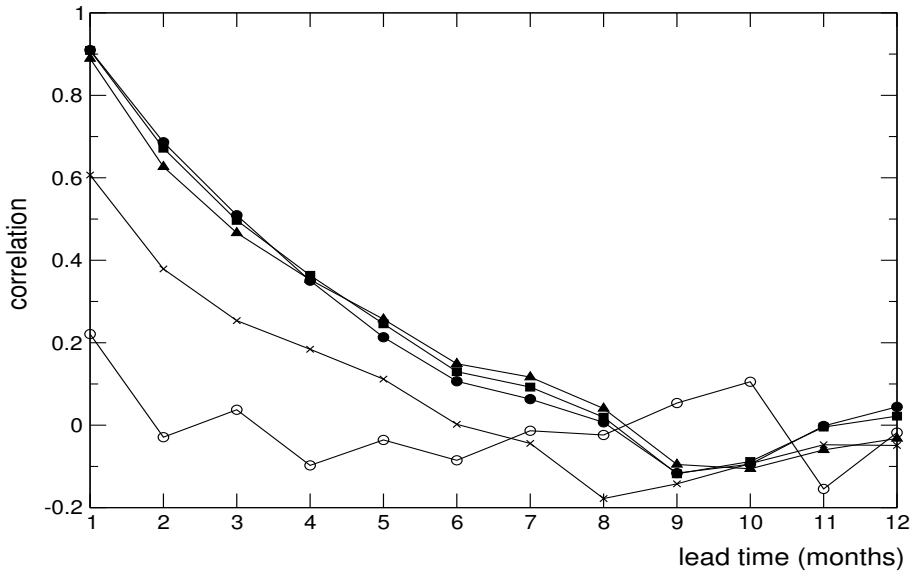

Figure 5.10: Correlation coefficients for the third variable for different one realisation noise intensities for the period 1976-2000: no noise ($-\bullet-$), 0.1 ($-\square-$), 0.2 ($-\triangle-$), 1.0 ($-\times-$) and 5.0 ($-\circ-$).

has been written till now. Fig. 5.11 shows the prediction series compared to the Isomap data for the decade of 1990. We observe years were prediction is good for the three variables even using a lead time of twelve months, while after an ENSO event like the one of 1997 or the short previous in the mid-nineties, the prediction skills for the three variables worsens dramatically even for three months forecasts. In particular, we also observe that the model does not reproduce the Annual Cycle immediately after a major ENSO event. We suggest that this is because the interaction between them during El Niño and Annual Cycle peaks (December-April) causes instabilities in the system. This could be related to the well-know spring barrier of prediction (Latif et al. (1998) and references therein). A final remark for this section is that the system prediction capabilities could be enhance by a careful tuning of the coefficients of the polynomials in the model (5.8). This path is far out of the philosophy of this work, although it could be interesting to explore it with greater detail.

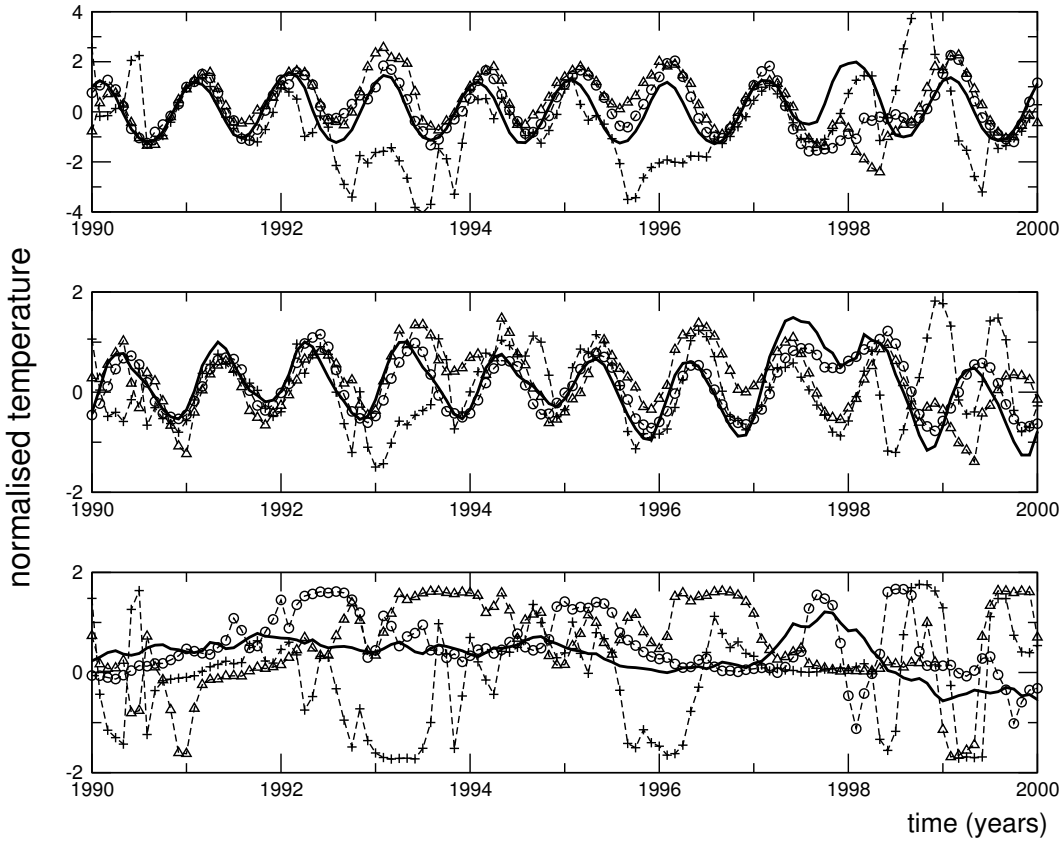

Figure 5.11: Prediction of the first (top), second (middle) and third (bottom) variables for lead times of three ($-\circ-$), six ($-\triangle-$) and twelve ($-+-$) months compared to the Isomap time series (thick line).

5.2 Prediction of El Niño events using ensemble model prediction

In this section, we are applying ensemble prediction to our three dimensional system. Ensemble building is a common way to improve the performance of a model for prediction tasks because an ensemble of individual predictors performs better than a single predictor in the average. This is based on the bias-variance decomposition of ensemble models (Krogh and Sollich, 1997). Usually an ensemble consists of models taken from one single class, as neural networks, although, in order to introduce more diversity, which is the central feature of the ensemble approach, several models from different model classes can be trained and combined to build the final ensemble (Wichard and Ogorzalek, 2004). The ensemble prediction can be mathematically described the following way:
Let us average the output $y(\boldsymbol{x})$ of several different models $f_i(\boldsymbol{x})$, and define the ensemble model as $\bar{f}(\boldsymbol{x}) = \sum_{i=1}^{K} \omega_i f_i(\boldsymbol{x})$. We assume that the model weights sum to one ($\sum_{i=1}^{K} \omega_i = 1$). If we define the ambiguity on $\boldsymbol{x}$ of a single member of the ensemble as $a_i(\mathbf{x}) = (f_i(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}))^2$, the ensemble ambiguity will be defined as,

$$\bar{a} = \sum_i \omega_i a_i(\boldsymbol{x}) = \sum_i \omega_i (f_i(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}))^2 \tag{5.9}$$

and quantifies the disagreement among the predictors on input $\boldsymbol{x}$. The quadratic errors of predictor i and ensemble are, respectively,

$$\epsilon_i(\boldsymbol{x}) = (y(\boldsymbol{x}) - f_i(\boldsymbol{x}))^2$$
$$\epsilon(\boldsymbol{x}) = (y(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}))^2$$

Adding and subtracting $y(\mathbf{x})$ in Eq. 5.9 yields,

$$\epsilon(\boldsymbol{x}) = \bar{\epsilon}(\boldsymbol{x}) - \bar{a}(\boldsymbol{x}) \tag{5.10}$$

where $\bar{\epsilon}(\boldsymbol{x}) = \sum_i \omega_i \epsilon_i(\boldsymbol{x})$ is the average error of the individual models. Essentially, Eq. 5.10 states that, the more the models in an ensemble differ, the less the error of the whole ensemble.

In our particular case, we shall use two types of ensembles in order to predict the three dimensional time series. First, we build an ensemble consisting in linear models

$$f_i(\boldsymbol{x}) = a_{0i} + \sum_{k=1}^{3} a_{ki} x_k \tag{5.11}$$

The coefficients are calculated with the standard method for ridge regression (Hastie et al., 2001). The optimal ridge parameter is evaluated by performing a cross-validation on the training data. Five cross-validation rounds were used in this case. The second type of model we are using is a nearest trajectory model, which is based on a strategy for time series prediction introduced by McNames (1998). It is based on the assumption that the time series stems from a dynamical system and the states can be reconstructed with a time delay embedding. The nearest trajectory model looks for the nearest trajectory segments in the reconstructed state space. The prediction is done with a local linear model of the closest trajectory points as described in McNames (1998). The number of neighbouring trajectories is chosen randomly at the start of the training algorithm.

The results for the Isomap data prediction are plotted in fig. 5.12. The first 300 points of the dataset were used for one step ahead prediction using 20-dimensional state space vectors. We observe that the linear ensemble model fits better than the nearest trajectories model ensemble in the majority of the cases. We suggest two possible explanations for this fact: first, we suspect that nonlinearities in the time series are not very strong, in accordance with model 5.8; second, the lack of a large number of data makes the nearest trajectories model less robust, as it is not always possible to find a nearest trajectory to predict from the existing dataset.

We can also observe how the dynamical system is able to achieve better predictions for short lead times. For longer times the linear model ensemble achieves results that are comparable to those found in the literature for more complicated models (Latif et al. (1998), Galanti et al. (2003), Chen et al. (2004)), although for longer lead times very complex atmosphere-ocean coupled models offer better results. Concerning the decadal variability (Fig. 5.13, we can appreciate that the differences are, as in the previous case, remarkable. As in the previous case, Figs. 5.14 and 5.15 show the prediction series compared to the Isomap data for the nineties. In these two cases, the system shows good prediction skills even for the short mid-nineties events, while it still worsens after the 1997 El Niño event. The shift in this peak is also common in several models found in the literature (Kirtman and Schopf (1998), (Tang and Hsieh, 2002)).

In summary, ensemble model prediction offers very good results even for long lead times for the first and second variables. The third variable shows lower predictability for lead times bigger than six months. In general, as we previously suggested, the linear ensemble offers better results due to its robustness. Compared to more complex ensemble model prediction schemes,

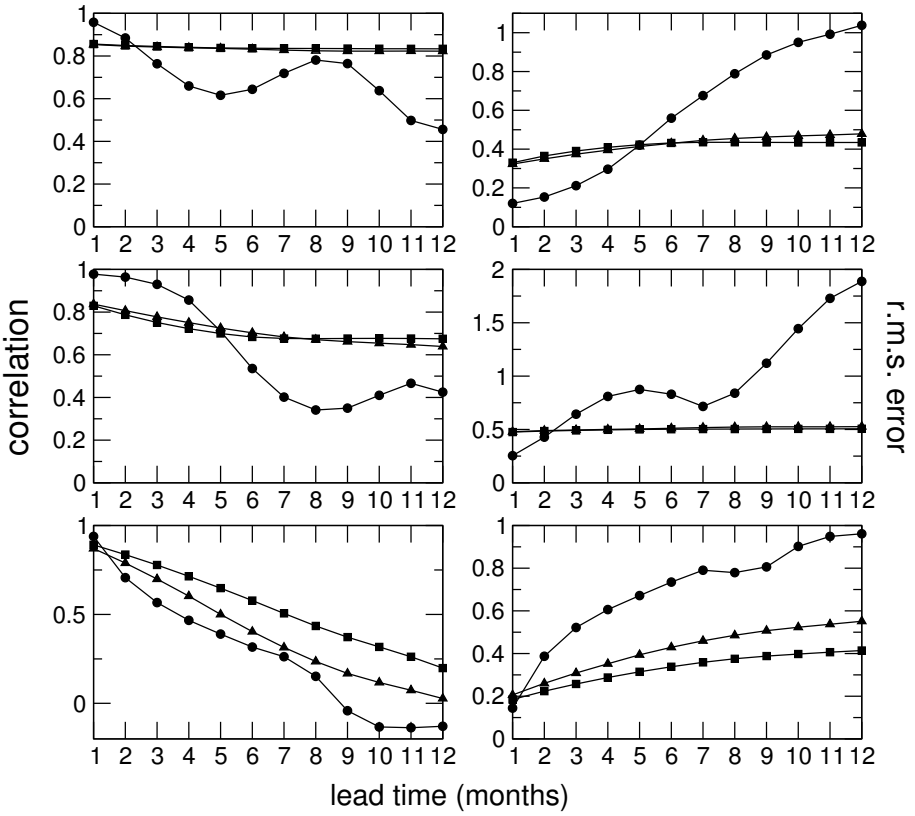

Figure 5.12: Correlation coefficients (left) and r.m.s. error (right) for the first (top), second (middle) and third (bottom) variables for the dynamical system model (− • −), linear model ensemble (−■−) and nearest trajectories model ensemble (−▲−).

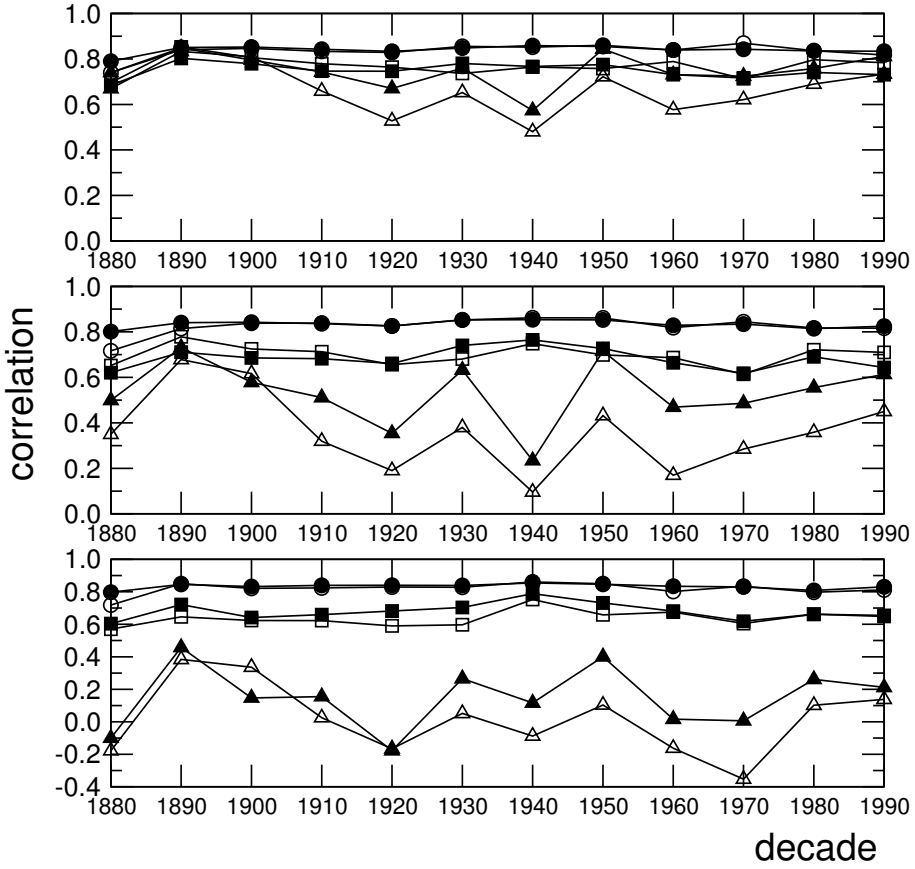

Figure 5.13: Correlation skills of the first (− • −linear), (− ○ −nonlinear), second (−■−linear), (−□−nonlinear), and third (−▲−linear), (−△−nonlinear) variable in different decades of the time series for different lead times: three months (top), six months (middle) and twelve months (bottom).

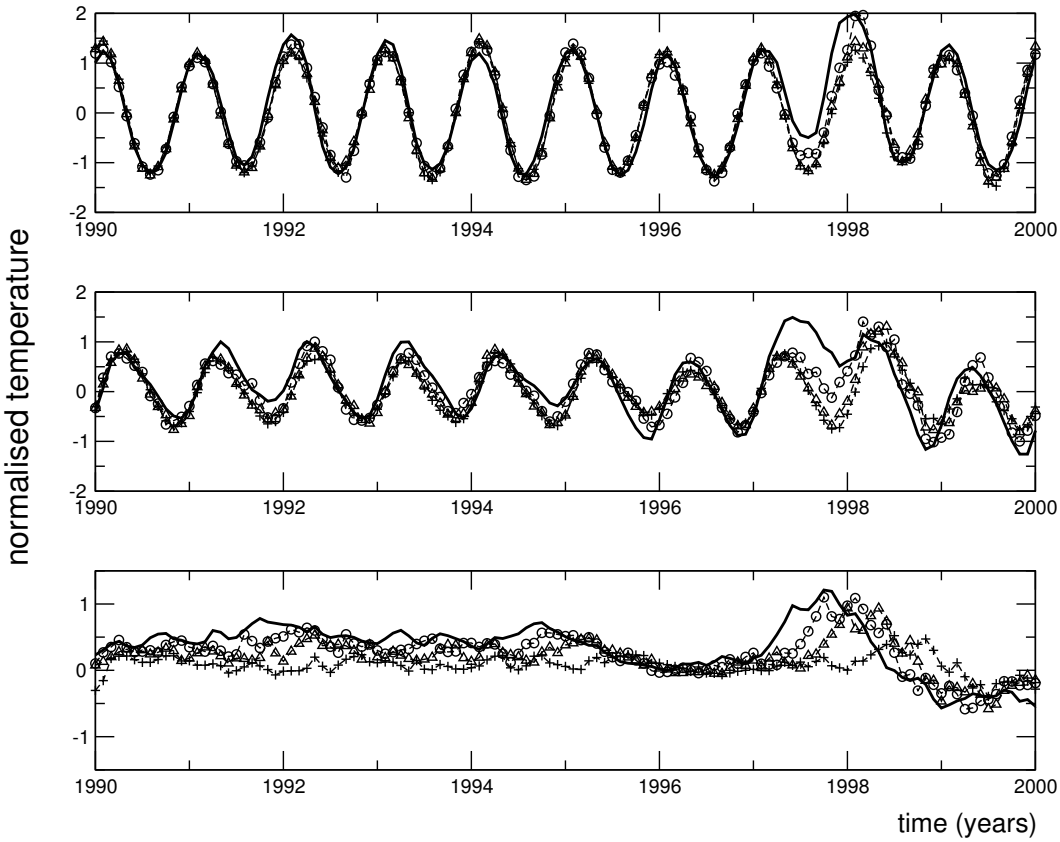

Figure 5.14: Prediction of the first (top), second (middle) and third (bottom) variables for lead times of three $(-\circ-)$, six $(-\triangle-)$ and twelve $(-+-)$ months compared to the Isomap time series (thick line).

as in Tang et al. (2005), the predictability is better. This could be a hint for what we previously suggested, that most of the physics of the system is related to the ocean rather than to the atmosphere. Although the dynamical system fails in predicting in a long lead time due to the little amount of data and the hidden variables, the ensembles chosen in this work are enough to predict with the same capability, at least, as that of more complex hybrid models (Tang et al., 2005).

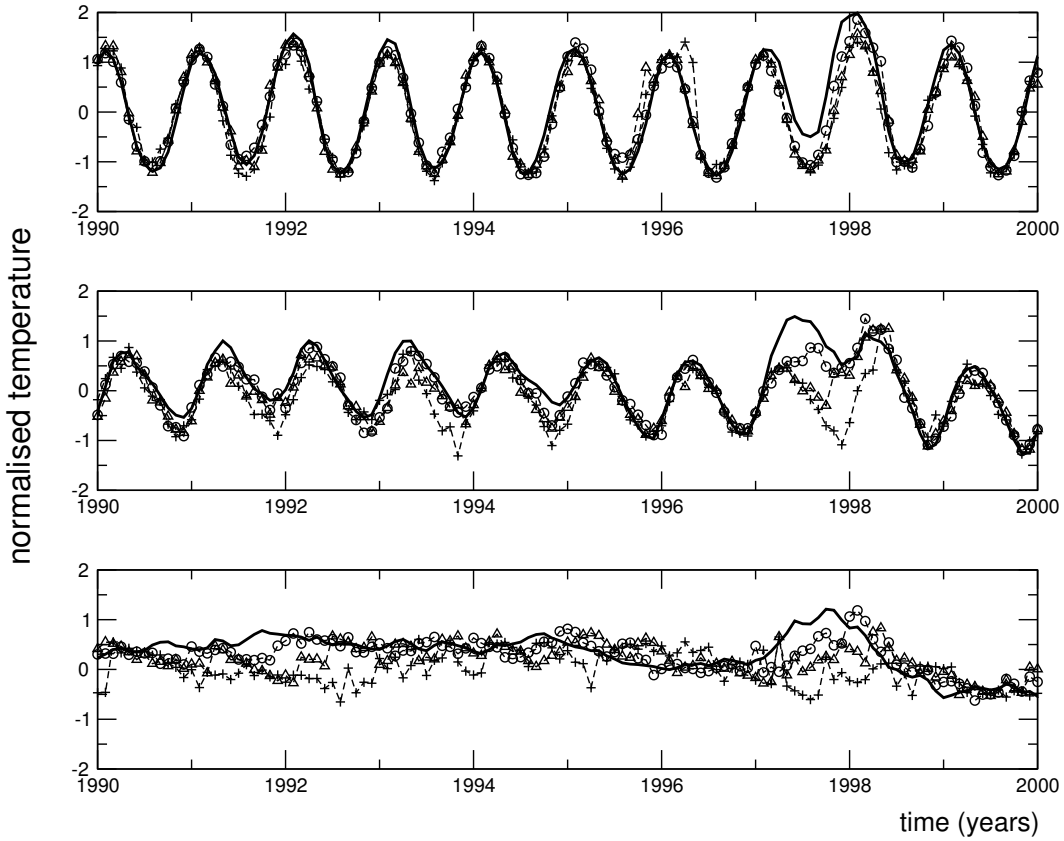

Figure 5.15: Prediction of the first (top), second (middle) and third (bottom) variables for lead times of three ($-\circ-$), six ($-\triangle-$) and twelve ($-+-$) months compared to the Isomap time series (thick line).

Chapter 6

Conclusions

This work was devoted to the study of the coupled system formed by El Niño/Southern Oscillation and the Annual Cycle. More precisely, the work was focused on two main problems:

• How to separate both oscillations into an affordable model for understanding the behaviour of the whole system.

• How to model the system in order to achieve a better understanding of the interaction, as well as predict states of the system.

The answers to these problems could be summarised in these two points:

• Linear methods are not suitable for characterising the dimensionality of the sea surface temperature in the tropical Pacific Ocean, therefore they do not help to separate the oscillations by themselves. Instead, nonlinear methods of dimensionality reduction are proven to be better in defining a lower limit for the dimensionality of the system as well as in explaining the statistical results in a more physical way (Gámez et al., 2004).

• A three dimensional dynamical system could be a starting point for describing the dynamics of the sea surface temperature in the tropical Pacific Ocean. Although not all the variability is explained, ensemble modeling prediction reveals that most of the information is present in just three variables, in comparison with more complicated models. Relevant predictions for short lead times can be made using a low dimensional system, despite its simplicity (Gámez et al., 2006a). The analysis of the SST data suggests that nonlinear

interaction between the oscillations is small, and that noise plays a secondary role in the fundamental dynamics of the oscillations.

6.1 Final remarks

6.1.1 On nonlinear dimensionality reduction of climate data

In Chapter 4, we suggested that nonlinear dimensionality reduction methods provide a useful way of analysing and modeling high dimensional data when nonlinear interactions are present. If the physical process can be embedded into a low dimensional manifold, the reduction of the relevant components is better achieved by nonlinear methods than by linear ones. In particular, Isomap (Tenenbaum et al., 2000) provides a physically appealing method of decomposing the data, as it substitutes the euclidean distances in the manifold by an approximation of the geodesic distances. We expect that this method could be successfully applied to other oscillatory extended systems and, in particular, to meteorological phenomena. Reduction allows to model the physical process by a low dimensional nonlinear dynamical system strictly based on data in order to make predictions of future states.

6.1.2 On the prediction of El Niño and the Annual Cycle

In Chapter 5, a well known non-parametric regression method was applied to the Annual Cycle and ENSO oscillations, as obtained by a particular method of dimensionality reduction, Isomap. The backfitting algorithm allowed us to build a model for the joint system of Annual Cycle and ENSO. We observed that, although the amount of data was low, we could predict future behaviours of the coupled ENSO-Annual Cycle system with an accuracy of less than six months, although the constructed system presented several drawbacks: few data points to input in the backfitting algorithm, untrained model, lack of forcing with external data and simplification using a close system. Moreover, ensemble prediction techniques show that the prediction skills of the three dimensional time series are as good as those found in much more complex models. This suggests that the climatological system in the tropics is essentially explained by the ocean dynamics, while the atmosphere plays a

secondary role in the physics of the process (Patil et al., 2001). For the lead times of six months or less, the prediction skills are similar to more complicated models found in literature (Chen et al. (2004), Tang and Hsieh (2002)). A second point of interest is the particular relationship between the Annual Cycle and ENSO. Observing that the particular functions found by regression in Chapter 5 are not unique in the sense that different adjusted polinomials give similar results, it is important to point that the interaction between ENSO and the Annual Cycle is apparently directed mainly from ENSO towards the Annual Cycle, but this could be an effect of mixing if the variables are not totally separated.

A global view of the work shows a general procedure to face modeling of climatological systems. First, we should find a suitable method of either linear or nonlinear dimensionality reduction. Then, low dimensional time series could be extracted out of the method applied. Finally, a low dimensional model could be found using a backfitting algorithm in order to predict future states of the system.

6.2 Outlook

A bunch of different working lines arise from this Thesis. Here, we are suggesting several of them.

• Mathematical development of the idea Isomap is constructed. A nonlinear reduction of dimensionality could be helpful when dealing with complex equations, as Navier-Stokes equations, the same way Fourier analysis (Franceschini, 1983) or PCA (Holmes et al. (1997), Duane and Tribbia (2001), Zoldi and Greenside (1997)) were successful in previous treatment of the problem. For example, the Kuramoto-Sivashinsky equation shows a remarkable difference in the dimensionality of the embedding subspace if turbulence is present (Meixner et al., 2000). Isomap could provide a better understanding of the transition via the relationship between correlation length and dimensionality (Gámez et al., 2006b).

• Exploration of the phase synchronisation scheme for the complex nonlinear system depicted in Chapter 5. Using the mathematical tools already at our disposal, we could achieve a mathematical proof of why El Niño peaks in December (Gámez et al., 2006c).

• Building a higher dimensional model of the interaction between the Annual Cycle and El Niño. Using the information we already know in the low dimensional space, we could model a high dimensional model of the joint Annual Cycle and El Niño system, considering not only the SST, but also the atmosphere, in order to achieve a better knowledge of the physics involving both oscillations as well as better prediction skills.

Acknowledgment

The road toward the completion of this Thesis was winding and very, very long. Fortunately, many people contributed to make the journey more pleasant and fruitful. In particular order, they are:

Prof. Dr. Jürgen Kurths, my supervisor, for his support. It has been a great experience working under his supervision.

Dr. Changsong Zhou, for spending hours of discussion and for his kindness and patience. His keenness and wisdom were essential to this work.

Dr. Markus Abel, who participated in the second half of the Thesis.

Dr. Axel Timmermann, who introduced us to the problem this Thesis is about, warmly hosted me at Kiel University and helped a lot from the climatologist point of view.

Now, in no particular order, my friends and colleagues at Potsdam and nearby: Birgit Voigt, for her helping hand, Jöerg-Uwe Tessmer, for helping me to win my fight against computers; Mamen Romano and Marco Thiel, who were always willing to help me; Douglas Maraun, who shared with me his knowledge of El Niño problem; Frank Spahn; Jürgen Schmidt, Bernd Blasius, Niels Wessel, Isao Tokuda, Keyser Soze, Udo Schwartz, Gustavo Rodrigues, Ralf Steuer, Miodrag Sremčevič, Daniel Kubas and the rest of the band. I would like to thank Murillo Baptista for being the brasilian master of chaos and my friend. Moltes gràcies, Ernest Montbrió i Fairen per tot. Vielen Dank, Jörg Wichard für alles. Gracias, Victor Sánchez Corbacho, por todo.

In the scientific part, I would like to thank the visitors and hosts who shared time and some of them even office with me and made useful suggestions: Celso Grebogi, Martin Hasler, Manuel Matías, Grigory Osipov, Itamar

Procaccia, Philip Rosenau and specially Ulrich Parlitz, who was very helpful in the first half of the work.

Finally, I would like to thanks my parents, family and friends for supporting me in the distance. Thanks Berlin...

... and María.

Bibliography

Abel, M. Nonparametric modeling and spatiotemporal dynamical systems. *Int. J. Bif. Chaos.* **14**(6), 2004.

Abel, M., Ahnert, K., Kurths, J. and Mandelj, S. Additive nonparametric reconstruction of dynamical systems from time series *Phys. Rev. E.* **71**, 015203(R), 2005.

An, S. I. and Wang, B. Mechanisms of locking of the El Niño and La Niña mature phases to boreal winter. *J. Climate.* **14**, 2164-2176, 2001.

Borg, I. and Groenen, P. *Modern Multidimensional Scaling: Theory and Applications.* Springer-Verlag, 1997.

Breiman, L., Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.*, **80**(391), 580-598, 1985.

Chang, P., Ji, L., Wang, B. and Li, T. Interactions between the seasonal cycle and El Niño-Southern Oscillation in an intermediate coupled ocean-atmosphere model. *J. Atmos. Sci.*, **52**, 2353-2372, 1994.

Chen, D., Cane, M. A., Kaplan, A., Zebiak, S. E. and Huang, D. Predictability of El Niño over the past 148 years. *Nature* **428**, 733-736, 2004.

Cobb, K. M., Charles, C. D., Cheng, H. and Edwards, R. L. El Niño/Southern Oscillation and tropical Pacific climate during the last millennium. *Nature* **424**, 271-276, 2003.

Do Carmo, M. *Differential Geometry of Curves and Surfaces.* Prentice-Hall, 1976.

Duane, G. S. Tribbia, J. J. Synchronized chaos in geophysical fluid dynamics. *Phys. Rev. Lett.* **86**, 4298-4301, 2001.

Franceschini, V. Two models of truncated Navier-Stokes equations on a two-dimensional torus. *Phys. Fluids* **26**, 433-447, 1983.

Floyd, R. W. Algorithm 97: Shortest path. *Communications of the ACM*, v.5 n.6, p.345, 1962.

Galanti, E., Tziperman, E., Harrison, M., Rosati, A. and Sirkes, Z. A study of ENSO prediction using a hybrid coupled model and the adjoint method for data assimilation. *Mon. Weather Rev.*, **131**, 2748-2764, 2003.

Gámez, A. J., Timmermann, A., Zhou, C. S. and Kurths, J. Nonlinear dimensionality reduction in climate data. *Nonlinear Processes in Geophysics*, **11**, 393-398, 2004.

Gámez, A. J., Abel, M., Zhou, C. S. and Kurths, J. Modelling the coupled system El Niño/Southern Oscillation and the Annual Cycle using a low order dynamical system. *submitted*, 2006.

Gámez, A. J., Zhou, C. S. and Kurths, J. Description of the Kuramoto-Sivashinskii equation using nonlinear dimensionality reduction techniques. *in preparation*, 2006.

Gámez, A. J., Zhou, C. S. and Kurths, J. Locking of El Niño to the boreal winter explained in terms of phase synchronisation theory. *in preparation*, 2006.

Gill, A. E. Some simple solutions for heat-induced tropical circulation *Quart. J. Roy. Met. Soc.*, **106**, 447-462, 1980.

Gill, A. E. *Atmosphere-Ocean dynamics*. International Geophysics Series. Academic Press, New York, 1982.

Gower, J. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325-328, 1966.

Grieger, B. and Latif, M. Reconstruction of the El Niño attractor with neural networks. *Climate Dynamics* **10**, 267-276, 1994.

Härdle, W. and Hall, P. *On the backfitting algorithm for additive regression models.* Cambridge University Press, 1993.

Ham, J., Lee, D. D., Mika, S. and Schölkopf, B. A kernel view of the dimensionality reduction of manifolds. *Proceedings of the 21st International Conference on Machine Learning (ICML-2004)*, Banff, Canada, 2004.

Hastie, T., Tibshirani, R. and Friedman, J. *The elements of statistical learning.* Springer Series in Statistics, Springer-Verlag, 2001.

Holmes, P., Lumley, J. L., Berkooz, G., Mattingly, J. C. and Wittenberg, R. W. Low-dimensional models of coherent structures in turbulence. *Physics Reports* **287**(4), 337-384, 1997.

Hsieh, W. W. Nonlinear principal component analysis by neural networks. *Tellus* **53**A, 599-615, 2001.

Jin, F. F., Neelin, J. D. and Ghil, M. El Niño/Southern Oscillation and the annual cycle: subharmonic frequency-locking and aperiodicity. *Physica D*, **98**, 442-465, 1996.

Jin, F. F. An equatorial recharge paradigm for ENSO. Part I: Conceptual model. *J. Atmos. Sci.*, **54**, 811-829, 1997.

Joliffe, I. T. *Principal Component Analysis.* Springer Verlag, 1986.

Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrica*, **23**, 187-200, 1958.

Kantz, H. and Schreiber, T. *Nonlinear time series analysis.* Cambridge University Press, 1997.

Kaplan, A., Cane M., Kushnir, Y., Clement, A., Blumenthal, M. and Rajagopalan, B. Analyses of global sea surface temperature 1856-1991. *J. Geophys. Res.*, **103**(C9), 18, 567-589, 1998.

Karhunen, K. Zur Spektraltheorie stochastischer Prozesse. *Ann. Acad. Sci. Fennicae*, Ser. A1, 34, 1946.

Li, T. and Philander, S. G. Decadal variability in ENSO predictability and prediction. *J. Climate*, **11**, 2804-2822, 1998.

Kovats, R. S., Bouma, M. J., Hajat, S., Worrall, E. and Haines, A. El Niño and health *The Lancet*, **362**, 1481-1489, 2003.

Krogh, A. and Sollich, P. Statistical mechanics of ensemble learning. *Phys. Rev. E*, **55**, 811825, 1997.

Latif, M., Barnett, T., Cane, M., Flügel, M., Graham, N. E., von Storch, H., Xu, J. S. and Zebiak, S. E. A review of ENSO prediction studies. *Climate Dynamics* **9**, 164-179, 1994.

Latif, M., Anderson, D., Barnett, T., Cane, M., Kleeman, R., Leetmaa, A., O'Brien, J., Rosati, A. and Schneider, E. A review of the predictability and prediction of ENSO. *J. Geophys. Res.*, **103**(C7), 14375-14393, 1998.

Li, T. and Philander, S. G. On the Annual Cycle of the eastern equatorial Pacific. *J. Climate*, **9**, 2986-2998, 1996.

Loève, M. Functions alèatoire de second ordre. *Comptes Rendus Acad. Sci.* Paris, 220, 1945.

McNames, J. A nearest trajectory strategy for time series prediction *Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, K. U. Leuven, Belgium, 1998.

Meixner, M. and Zoldi, S. M., Bose, S. and Schöll, E. Karhunen-Love local characterization of spatiotemporal chaos in a reaction-diffusion system *Phys. Rev. E*, **61**, 1382-1385, 2000.

Monahan, A. H. Nonlinear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure. *J. Climate*, **14**, 219-233, 2001.

Patil, D. J., Hunt, B. R., Kalnay, E., Yorke, J. A. and Ott, E. Local low dimensionality of atmospheric dynamics. *Phys. Rev. Lett.*, **86**, 5878-5881, 2001.

Philander, G. S. *El Niño, La Niña and the Southern Oscillation.* San Diego: Academic Press, 1990.

Pikovsky, A., Rosenblum, M. and Kurths, J. *Synchronization: A universal concept in nonlinear sciences.* Cambridge University Press, 2001.

Poggio, T and Girosi, F. Networks for approximation and learning. *Proc. IEEE*, **78**(9), 1481-1497, 1990.

Preisendorfer, R. W. *Principal Component Analysis in Meteorology and Oceanography.* Elsevier, 1988.

Reynolds, R. W. and Smith, T. M. A high resolution global sea surface temperature climatology. *J. Climate*, **8**, 1571-1583, 1995.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323-2326, 2000.

Seung, H. S. and Lee, D. D. The manifold ways of perception *Science*, **290**, 2268-2269, 2000.

de Silva, V. and Tenenbaum, J. B. Global versus local methods in nonlinear dimensionality reduction. Advances in Neural Information Processing Systems 15, 705-712. S. Becker, S. Thrun, K. Obermayer (eds.), *MIT Press*, Cambridge, 2003.

von Storch, H. and F. W. Zwiers. *Statistical Analysis in Climate Research.* Cambridge University Press, 1999.

Suarez, M. J., and Schopf, P. S. A delayed action oscillator for ENSO. *J. Atmos. Sci.*, **45**, 3283-3287, 1988.

Tang, Y. and Hsieh, W. W. Hybrid coupled models of the tropical Pacific: II ENSO prediction. *Climate Dynamics* **19**, 343-353, 2002.

Tang, Y., Kleeman, R. and Moore, A. M. Reliability of ENSO dynamical predictions. *J. Atmos. Sci.*, **62**, 1770-1791, 2005.

Tenenbaum, J. B., de Silva, V. and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319-2323, 2000.

Timmermann, A. and Jin, F. F. A nonlinear mechanism for decadal El Niño amplitude changes. *Geophys. Res. Lett.*, **29**, 1-4, 2002.

Timmermann, A., Jin, F. F. and Abshagen, J. A nonlinear theory for El Niño bursting. *J. Atmos. Sci.*, **60**, 152-165, 2003.

Tozuka, T. and Yamagata, T. Annual ENSO. *J. Phys. Oceanogr.*, **33**, 1564-1578, 2003.

Tziperman, E., Stone, L., Cane, M. A. and Jarosh, H. El Niño chaos: overlapping of resonances between the seasonal cycle and the Pacific ocean-atmosphere oscillator. *Science*, **264**, 72-74, 1994.

Tziperman, E., Cane, M. A., Zebiak, S. E., Xue, Y. and Blumenthal, B. Locking of El Niño's Peak Time to the End of the Calendar Year in the Delayed Oscillator Picture of ENSO. *J. Climate*, **11**, 2191-2199, 1998.

Vitushkin, A. G. On representation of functions by means of superpositions and related topics. *Enseignement Math.* 2 **23**, 255-320, 1977.

Voss, H., Bünner, M. J. and Abel, M. Identification of continuous spatiotemporal systems. *Phys. Rev. E*, **57**, 2820-2823, 1998.

Wichard, J. and Ogorzalek, M. Time series prediction with ensemble models *Proceedings of the IJCNN 2004*, 1625-1629, Budapest.

Xie, S. P. On the Genesis of the equatorial Annual Cycle. *J. Climate*, **7**, 2008-2013, 1994.

Xie, S. P. Interaction between the annual and interannual variations in the Equatorial Pacific. *J. Phys. Oceanogr.* 25, 1930?1941, 1995.

Xie, S. P. Effects of seasonal solar forcing on latitudinal asymmetry of the ITCZ. *J. Climate*, **9**, 2945-2950, 1996.

Xie, S. P. Westward propagation of latitudinal asymmetry in a coupled ocean-atmosphere model. *J. Atmos. Sci.*, **53**, 3236-3250, 1996.

Zebiak, S. E. Atmospheric convergence feedback in a simple model for El Niño. *Mon. Weather Rev.*, **114**, 1263-1271, 1986.

Zebiak, S. E. and Cane, M. A. A model El Niño/Southern Oscillation. *Mon. Weather Rev.*, **115**, 2262-2278, 1987.

Zha, H. and Zhang, Z. Isometric embedding and continuum ISOMAP. *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, 2003.

Zoldi, S. M. and Greenside, H. S. Karhunen-Loève decomposition of extensive chaos. *Phys. Rev. Lett.*, **78**, 1687-1690, 1997.