Modelling multivariate data systems: application to El Niño and the Annual Cycle

A J Gámez
Institut für Physik, Universität Potsdam, Postfach 601553, D-14415 Potsdam, Germany
E-mail: gamez@agnld.uni-potsdam.de and ajgamez@alephs.net

New Journal of Physics 9 (2007) 432
Received 24 August 2007
Published 5 December 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/12/432

Abstract. We propose a general procedure for the analysis of noisy high dimensional dynamical systems, potentially with time delay. This procedure combines novel dimensionality reduction techniques, nonlinear time series analysis for dynamical systems and nonparametric statistical estimation of functional dependencies. To check the feasibility of our method, we apply it to the sea surface temperature (SST) field in the tropical Pacific Ocean, in order to build a model for the interaction of El Niño/Southern Oscillation (ENSO) and the Annual Cycle coupled system. This dynamical representation is shown to be reducible to three dimensions by applying Isomap, a recent method of dimensionality reduction. From the resulting time series, we construct a stochastic dynamical system by using nonparametric estimation. This dynamical system is numerically integrated and compared with measured data a posteriori. The use for prediction of the joint system of ENSO and the Annual Cycle is discussed.
1. Introduction

In this paper, we propose a general framework for the analysis of noisy high-dimensional dynamical systems with time delay. More precisely, we reduce a high-dimensional system to a lower one that captures the essential dynamics of the physical process. Then, we build an approximate system of equations that represents the reduced data. Finally, in order to check the validity of our approach, we run simulations of the dynamical system and make predictions of future states.

Our procedure is composed of the following steps:

1. Reduction of the dimensionality of the dataset by using complex nonlinear methods.
2. Construction of a dynamical system that reproduces the data resulting from step 1 by using nonparametric regression. Eventually, knowledge of the physics involved in the process can improve this construction.
3. Prediction of new states of the system using the dynamical system found in step 2 and further reconstruction from the predicted solutions.

As an example, we apply our procedure to observed sea surface temperature (SST) data in the tropical Pacific Ocean. The SST evolution in the equatorial Pacific is characterized by two different oscillatory phenomena: El Niño/Southern Oscillation (ENSO) and the Annual Cycle. For a brief review of ENSO and Annual Cycle physics, see appendix B. Attempts to reconstruct ENSO’s attractor have been made using different linear (Penland and Sardeshmukh 1995) or nonlinear (Grieger and Latif 1994; Monahan 2001) methods. In fact, there is a debate as to whether ENSO could be explained by linear dynamics or not. In favour of the nonlinear point of view, some ENSO researchers support the fact that several features of ENSO, such as time asymmetry or skewness, cannot be explained by linear dynamics with added gaussian noise alone (An and Jin 2004; Hannachi et al 2003). Favouring the linear approximation, other authors have proven that linear dynamics can provide skewness to the SST distribution if the dynamical equations contain multiplicative and additive noise that are correlated (Sura and Sardeshmukh 2007). In any case, the method we apply in this work will not exclude the possibility that the dynamical equations could be linear, as we will explain later.

Most of the published analyses of ENSO dynamics are based on SST anomalies which are obtained by subtracting a mean Annual Cycle from the monthly averaged SST data. Therefore, the amplitude of the Annual Cycle is approximated to be constant. This approximation is
rough, but the reason for using this procedure lies in the fact that extracting a time-varying Annual Cycle and an ENSO mode in a multivariate way from SST data is not simple and the results depend on the assumptions used by different methods. This traditional approximation excludes two facts: that the Annual Cycle has different amplitude each year and that ENSO and the Annual Cycle are interacting oscillations because they exhibit in some sense a joint synchronized behaviour (ENSO amplitude is enhanced 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 coupling is of great importance for understanding the variability of ENSO. In the last few years, several papers discussing how ENSO and the Annual Cycle interact in the tropical Pacific Ocean have been published (An and Wang 2001; Chang et al 1995; Jin et al 1996; Li and Hogan 1999; Xie 1995).

The possibly nonlinear interaction may lead to erroneous conclusions when subtracting an Annual Cycle of constant amplitude from the SST data because the state space cannot be decomposed into a sum of linear subspaces spanned by independent variables. Hence, the separation of the SST into physically independent modes is not trivial, and could even be impossible. Therefore, the ENSO–Annual Cycle coupled system is a perfect candidate to test the power of our method.

This paper is organized as follows. In section 2, we briefly explain how to find a lower dimensional embedding for our data, in order to use it for building the model. In section 3, a stochastic nonlinear system is constructed. In section 4, we analyse the model in order to predict future events. The potentials and limits of the model are also discussed in section 5.

2. Nonlinear dimensionality reduction of the multivariate data

Multivariate data obtained from complex systems are usually difficult to model. Correlations between the variables may result in an excess of information. For this reason, reducing the dimensionality helps to handle the information offered by data in order to construct a suitable model that represents the system appropriately. However, there are many different linear and nonlinear methods for reducing dimensionality. A first step is, then, to choose the best method for our particular problem. We consider the best method to be the one that best captures the variability of the data by using the lowest number of dimensions. A concise but detailed description of the methods we use throughout this work can be found in appendix A. In the particular example we are treating in this work, we have considered the SST as a suitable observable of the ENSO phenomenon. The SST data has been taken from two public databases: the Reynolds–Smith (RS) database (Reynolds and Smith 1995) and the Kaplan extended (KE) database (Kaplan et al 1998). The spatial resolution of the RS database is 25 times better than the KE database, which has a resolution of five degrees. The KE database covers the period from January 1856 to October 1981 with a time resolution of one month. The RS database covers the November 1981 to October 2002 time range and was used to complement the KE database for that period. We have analysed the data with principal component analysis (PCA) (Joliffe 1986), multidimensional scaling (MDS) (Borg and Groenen 1997) and Isomap (Tenenbaum et al 2000). The first two methods are linear methods, the last is a nonlinear method of dimensionality reduction based on MDS. We would like to stress that PCA is equivalent to MDS if the data is normalized by the statistical variance (Borg and Groenen 1997). Compared to the linear methods, Isomap offers a clear cut-off in the number of dimensions to which the data can be reduced (see figure 1). In figure 1, we observe that 15 is the number of dimensions.
Figure 1. The residual statistical variances $r_p$ (definition in appendix A) plotted against the number of dimensions $p$ considered for PCA (triangles) and Isomap (circles). For the KE dataset, the dimensionality shown by Isomap is three, as the residual variance ceases to decrease significantly if more dimensions are added. The dimensionality shown by PCA is not well defined because of the slow convergence of $r_p$ (see appendix A). In the inner box, the three dimensional (3D) representation of the resulting variables is plotted.

needed for PCA to explain the variance that Isomap covers in only three (Gámez et al 2004). Considering that adding more than 3D does not capture more variance of the data as the residual variance converges, we set three as the lowest number of dimensions that our model should have. It is important to note that this number is consistent throughout different time intervals and that it is independent of the database used. Fortunately, three dimensions allow one to plot the reduced system, as seen in figure 1. A proper selection of coordinates isolates a 12 months rotation in two components, while the third one is an irregular oscillation that describes ENSO (figure 2). There is no annual component in it, as Fourier decomposition shows. We consider that the first coordinate represents the variation of SST due to the Annual Cycle and that the third coordinate represents the variation due to ENSO. These coordinates are not independent, though. The interaction between the modes, represented in the second coordinate, shows two important facts: firstly, that it is impossible to separate the two oscillations independently because of the interaction between them and, secondly, that considering the Annual Cycle as constant (also known as climatology) does not reflect the true variability of the data. It is important to stress that the coordinates are no longer orthogonal, as in the case of PCA, for example. This means that we cannot project the data on to the ‘principal components’ obtained by this method to obtain time invariant spatial patterns, but it is possible to obtain spatial patterns for particular values of the variables by constructing a function that associates each 3D point in the reduced space to a spatiotemporal point in the original data space (Poggio and Girosi 1990). For example, two patterns, one for a condition where the Annual Cycle is maximum and ENSO is zero and one for a condition where ENSO is maximum and the Annual Cycle is zero are shown in figure 3. We stress again that these patterns do not form a orthogonal basis in this space: in nonlinear spaces the associative property does not hold in general.
Figure 2. (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.

Figure 3. (a) SST pattern in degrees Celsius for a condition where ENSO is nearly zero and the Annual Cycle is maximum. (b) SST pattern for maximum ENSO with zero Annual Cycle.
In summary, we take these three time series as a minimal input for modelling a dynamical system that could represent the whole dataset. The most general system of dynamical equations,

\[ \begin{align*}
\dot{x}_1 &= f_1(x_1, x_2, x_3), \\
\dot{x}_2 &= f_2(x_1, x_2, x_3), \\
\dot{x}_3 &= f_3(x_1, x_2, x_3)
\end{align*} \tag{1} \]

is our first ansatz, as typical way of embedding (Sauer et al 1991).

3. Nonparametric estimation of a stochastic model

In this section, we describe the method to estimate the functions defined in the system (1). Consider each point \( \vec{x}(t) = (x_1, x_2, x_3) \in \mathbb{R}^3 \) of the Isomap output as a realization of the flow in phase space. If we assume equal dimensionality between the dynamical space and the embedding space, we can cast equation (1) in the general form,

\[ \dot{\vec{x}} = \Phi(\vec{x}(t)), \tag{2} \]

where \( \Phi : \mathbb{R}^3 \to \mathbb{R}^3 \). For each component \( \Phi_i \) of \( \Phi \), one obtains as a best estimator of (2) in the least-square sense (Kantz and Schreiber 1997),

\[ \Phi_i = E[\dot{x}_i | x_1, \ldots, x_m], \tag{3} \]

where \( E[\cdot | \cdot] \) is the conditional expectation value (CEV) operator. In practice, the CEV is calculated by using the neighbours of \( \vec{x} \), as in most nearest neighbour methods. Extracting analytical models from this formula is very hard, and visualization is obviously impossible for \( m > 2 \). To counter this drawback, we start building the model from the simplest version in terms of estimation and cast the rhs of equation (2) into an additive model, which is a well understood form for nonparametric estimation (Hastie and Tibshirani 1990):

\[ \Phi_i = \sum_{j=1}^{m} \phi_{ij}(x_j). \tag{4} \]

After having estimated the components \( \phi_{ij}(x_j) \), we can easily visualize the functions and try analytical formulae for the \( \phi_{ij} \). With (4), we have a less general model than (2), 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 (Kolmogorov 1963), later refined by Vitushkin (Vitushkin 1977). If it turns out that the strictly additive model is not sufficient, successively more complicated dependencies can be taken into account. The optimal estimate for the \( \phi_{ij} \) is calculated by a backfitting algorithm (Hastie and Tibshirani 1990). It works by alternatively applying the CEV operator to projections of \( \Phi_i \) on to the coordinates: \( \phi_{ij}(x_j) = E[\dot{x}_i - \sum_{k \neq j} \phi_{ik} | x_j] \), 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 computed numerically by running mean, local linear fits, polynomials, etc. In Härdle 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 or other techniques 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 \( C_{ij} \) between the
shows that a value of approximately

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

where \( C \) is the correlation function. A correlation close to one means a good model, close to zero means a bad one. It is worth noting the geometrical aspect of our approach: equation (2) defines a hypersurface in \( \mathbb{R}^m \), approximated by the sum of the functions \( \phi_{ij} \) in (4). 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. Starting from equation (1), we consider \( f_1, f_2 \) and \( f_3 \) to be measurable linear or nonlinear functions of the variables \( x_1, x_2 \) and \( x_3 \), which are the three Isomap coordinates. On the other hand, it is known that delayed variables play an important role in the low-dimensional description of ENSO physics (Tziperman et al 1994) (appendix B). Therefore, we also allow delayed terms related to the second and third variables, \( x_4 = x_3(t - \tau) \) and \( x_5 = x_3(t - \tau) \). In any case, if the delay is not affecting the data evolution, the variables \( \dot{x}_1, \dot{x}_2 \) and \( \dot{x}_3 \) do not depend on \( x_4 \) or \( x_5 \), and there is no delayed term in the resulting equations. In the same way, if the system is purely linear, the functions will show a clearly linear behaviour. This is automatically analysed by the modelling technique described in the next section. As we will see later, the first variable is not directly affected by delayed variables, as \( x_1 \) is essentially capturing the effect of the Annual Cycle. As a second ansatz that includes the delayed terms, we write,

\[ \begin{align*}
\dot{x}_1 &= f_1(x_1, x_2, x_3), \quad \dot{x}_2 = f_2(x_1, x_2, x_3, x_4, x_5), \quad \dot{x}_3 = f_3(x_1, x_2, x_3, x_4, x_5).
\end{align*} \]  

Taking the new contributions into account, the model (6) is too complicated to be parametrized. Thus, more heuristic reasoning is required to simplify the equations. Obviously, a period of 12 months has to appear in the solution of variables \( x_1, x_2 \). Therefore, we require 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 has a linear component. As explained in the context of equation (4), we restrict the model to be additive. We also add a stochastic part \( \mu_1 \) of the measurement to represent the model uncertainty. Including this information we finally obtain the model:

\[ \begin{align*}
\dot{x}_1 &= -c_1 x_2 + f_{12}(x_2) + f_{13}(x_3) + \mu_1, \\
\dot{x}_2 &= c_2 x_1 + f_{21}(x_1) + f_{23}(x_3) + f_{24}(x_4) + f_{25}(x_5) + \mu_2, \\
\dot{x}_3 &= f_{31}(x_1) + f_{32}(x_2) + f_{33}(x_3) + f_{34}(x_4) + f_{35}(x_5) + \mu_3,
\end{align*} \]

where we renamed \( x_4 = x_3(t - \tau) \) and \( x_5 = x_3(t - \tau) \) to simplify the notation. The fundamental frequency of the Annual Cycle, represented by the first and second equation, is \( \omega = \sqrt{c_1 c_2} \). The functions \( f_{ij} \) where \( i = 1, 2, 3 \) and \( j = 1, 2, 3, 4, 5 \) are nonlinear corrections, not necessarily polynomial. This is the input to the backfitting algorithm we are applying.

As a pre-processing, we have to calculate the derivative from the Isomap data series. We use a spectral estimator, which turns out to be numerically well behaved and fast. For details see Abel et al (2005). For the delay, figure 4 shows that a value of approximately \( \tau = 6 \) months yields an optimal correlation for the second and third variable, which is consistent with the characteristic time of ocean wave dynamics in ENSO evolution.
Figure 4. Variation of the correlation $C_{ij}$ (equation (5)) for variable delay. An optimal delay of six months is found for $x_5$ (left) and $x_4$ (right).

Figure 5. (a) Plot of $\dot{x}_1$ versus $x_2$. A straight line of slope $-1$ is a good approximation of the nonlinear function fit (red line). (b) Plot of $\dot{x}_2$ versus $x_1$. A straight line of slope $0.27416$ approximates the nonlinear function fit (red line).

Now, we take a closer look at the first equation. A scatter plot of $\dot{x}_1$ versus $x_2$ yields figure 5(a). A slope of value $c_1 = -1$ approximates quite well the optimal straight line through the data. A fully nonlinear analysis yields the red line in figure 5(a), which, up to the usual effects at the boundaries (Abel 2004), follows a linear behaviour. For the second equation, we
Table 1. Definition of the functions in equations (7)–(9) according to polynomial regression.

| Function | Regression |
|----------|------------|
| \( c_1 \) | -1         |
| \( c_2 \) | 0.27416    |
| \( f_{12} \) | 0         |
| \( f_{13} \) | -0.01 + 0.45x_3 |
| \( f_{21} \) | 0         |
| \( f_{24} \) | -0.13x_4 |
| \( f_{31} \) | 0         |
| \( f_{33} \) | 0.01 - 1.33x_3 + 2.35x_3^2 - 0.75x_3^5 |
| \( f_{34} \) | -0.02 + 0.85x_4 |
| \( f_{35} \) | 0.26x_5 - 0.01x_5^2 - 0.65x_5^3 |

fix \( c_2 = 0.27416 \) in order to get a main oscillation of annual period. This value is consistent with the error bars of the estimated slope. The corresponding plot is shown in figure 5(b): a linear fitting approximates well the nonlinear function, even though the data are few and show very inhomogeneous scatter. So we can set the functions \( f_{12} \) and \( f_{21} \) to be zero.

The most relevant equation in terms of the system and also in terms of the modelling difficulties is the third equation, as it contains most of the ENSO physics. It turns out that the \( x_1 \) and \( x_5 \) components are close to zero for central values. The deviation at the boundaries is again mainly due to the inhomogeneous data distribution. The correlations, however, are high. Therefore, there are two options for the modelling: the first could consider that the contribution of these components is not relevant, the 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 \( f_{12} \) is not affecting the stability of the system. Therefore, we assume that \( f_{31} = 0 \). For the \( x_3 \) and \( x_4 \) components, the relevant contributions are similar to those of the delayed oscillator (DO) model (Tziperman et al 1994) with opposite signs. This could be due to the fact that the variable \( x_5 \) is strongly correlated to \( x_3 \) and \( x_4 \).

Summarizing all these considerations, table 1 gives the regression of the functions represented in figure 6. The first equation is purely linear. It contains a term that develops a regular Annual Cycle and a linear contribution of ENSO 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 DO paradigm with a nonlinear term and a nonlinear contribution from the interaction variable \( x_5 \). We would like to stress here that, as we pointed out before, some equations are linear and some are not, and the method is perfectly capable of distinguishing both behaviours without any previous assumption. This is the model that we are going to integrate in the next section.

4. Simulation of the model for prediction: potentials and limits of the model

To check the validity of the model, we test its predictive power by a posteriori comparison with real data. In this section, we sketch the procedure and make simulations without a detailed study of noise distribution or extensive averaging. We integrate our system numerically
by using a classical Euler algorithm with a time step \( \Delta t = 0.001 \). We note that a thousand integrations are needed to make the system evolve one month. To model the stochastic part in equations (7)–(9), \( \mu_1, \mu_2 \) and \( \mu_3 \), we fit different functional forms of noise to the residuals of the nonparametric regression algorithm. We find that a good approximation is the selection of gaussian white noise, where the variances \( \sigma_i \) are directly calculated from the fitting. The selection of functional forms that are different from the gaussian does not significantly change the results. In particular, we found similar results for the exponential power, Cauchy, Rayleigh and Landau distributions. It is interesting to analyse in which interval the noise variance is physically relevant. 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, the dynamical behaviour is much richer. As we see in figure 7, in the absence of noise, there is a relaxation to a fixed point with a transient time of approximately one year, which is the typical time of an ENSO oscillation. As noise increases (\( \sigma_3 \in \{0.1, 0.5, 1.0, 5.0\} \)), the system undergoes irregular oscillations at low (\( \sigma_3 = 0.1 \)) variance, then regular oscillations are created (\( \sigma_3 \in \{0.5, 1.0\} \)) that are smoothly destroyed when noise increases to reach irregular oscillations. We observe that a noise level of \( \sigma_3 = 5.0 \) creates irregular oscillations that are, in a way, partially phase-locked to rational multiples of six months. Increasing the noise to values greater than 5.0 results in non-physical solutions that we are not considering, as the amplitude and frequency of the third variable are much higher than observed in real data. In fact, the observed solutions are similar, in the long run, to the solutions found for noise around \( \sigma_3 = 0.1 \). The variances computed by fitting the residuals of the backfitting algorithm to gaussian distributions are \( \sigma_1 = 0.142, \sigma_2 = 0.123 \) and \( \sigma_3 = 0.136 \). These values are physically consistent with our previous analysis. Concerning the particular form of the equations, we note that increasing coupling, therefore increasing \( f_{13}, f_{24} \) and \( f_{35} \), changes the regime of oscillations to regular oscillations around a different fixed point. The amplitude of the Annual Cycle also increases. Importantly, if \( f_{35} \) is set to zero, ending the

**Figure 6.** From left to right, the polynomial functions (red lines) representing the three different functional dependencies in the \( x_3 \) equation: \( x_3, x_4 \) and \( x_5 \).
coupling, we find that the oscillation found in the system (figure 7, first row) disappears. This means that coupling is fundamental in this model for ENSO burst and its locking. The particular form of the polynomials shows that the coupling between the Annual Cycle and ENSO can be mainly represented by linear coupling with a small nonlinear contribution of higher order terms. This result complements some stochastic models where the Annual Cycle is represented as a stochastic forcing with periodic statistics (Penland 1996). In our case, the stochastic forcing is substituted by noise and an oscillatory variable, which we associate with the Annual Cycle. This approach is more appealing to our eyes, because it has the advantage that the skewness of the time series that we see in the data is better represented, meaning that the positive oscillations of ENSO are bigger than the negative ones.

Having all the ingredients for the model depicted in equations (7)–(9), we are now ready to use it for forecasting future states of the system. We make calculations for prediction times ranging from one month to one year. It is worth pointing out, as we commented before, that the predicted points in the low-dimensional system can be converted into approximate spatio-temporal SST values by parametrizing a function that transforms the spatio-temporal data in the Isomap points and its inverse (see, for example, Poggio and Girosi (1990)). We run 50 realizations of noise in order to make a mean value for each predicted point. To measure the predictive power of the model, we compute the statistical correlation between the predicted and the real time series, as well as their root mean square (rms) error. In figure 8, we observe

**Figure 7.** From top to bottom, sample simulation of the $\dot{x}_3$ equation for noise variances $\sigma_3 = 0, 0.1, 0.5, 1.0, 5.0$, where $\sigma_1$ and $\sigma_2$ are set to zero. The approximate value for $\sigma_3$ taken from the nonparametric regression algorithm is equal to 0.136.
that, as expected, the prediction capabilities of the model worsens when the prediction lead time is increased.

It is noteworthy to say that the system works worse after certain conditions of high El Niño events. For this reason, if we measure the correlation for different decades, we observe big differences in the predictability of the system. In summary, the model faithfully represents the first and second variables for lead times under one year, while the third variable shows lower predictability for lead times greater than five months. It is important to say that the dynamical system proposed here does not achieve the prediction capabilities for the third variable of some high-dimensional ENSO models found in the literature (see Latif et al (1994); Latif et al (1998) for reviews), but it is better than the DO model for particular decades (Kirtman and Schopf 1998). There are several points to consider in our case: firstly, there are few data points to input in the nonparametric regression algorithm; secondly, the system is not trained; thirdly, it is not artificially pumped with external data; fourthly, it does not consider effects out of the boundaries of the tropics and is not sustained by the atmosphere; fifth, it predicts future states of the joint system ENSO–Annual Cycle where previous models only studied the system partially. Anyway, we expect that the dynamical system could be improved by considering the atmospheric influence, which we believe is secondary to the ocean dynamics in this particular system (Patil et al 2001).

5. Conclusions

In this paper, a general procedure for the study of high-dimensional dynamical systems with time delay has been presented. In particular, we have studied the coupled system composed of ENSO and the Annual Cycle, which are the two main oscillations that interact in the tropical Pacific Ocean on human timescales. A nonparametric backfitting algorithm

Figure 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.
has been used to build a nonlinear dynamical system for representing the reduced SST variables obtained by a modern method of dimensionality reduction, the Isomap. Therefore, our procedure favours the existence of nonlinearities to explain the ENSO bursting, the decadal variability and the skewness (Timmermann et al 2003). Using surrogate data associated with different indexes, such as the Southern Oscillation, shows that the amount of data available could be not sufficient to favour linear or nonlinear dynamics (Schreiber and Schmitz 2000). Anyway, we observe that, although the amount of data is little, we can predict future states of the coupled ENSO–Annual Cycle system with significant accuracy. One major advantage in our approach is that the constructed system is neither trained in any way nor is it forced by real conditions such as wind stress (Blanke et al 1997). This suggests that the climatological system in the tropics is driven by the ocean dynamics, while the atmosphere plays a secondary, although relevant role in the physics of the process (Patil et al 2001). Interestingly, for lead times of six months or less, the prediction skills are similar to more complicated models found in the literature (Chen et al 2004, Tang and Hsieh 2002). This suggests that including the interaction between ENSO and the Annual Cycle significantly improves the prediction capabilities of ENSO models. A second point of interest is the particular relationship between the Annual Cycle and ENSO. Considering that the particular functions found by fitting the output of the nonparametric regression algorithm are not unique (different adjusted polynomials give similar results), it is important to stress that the interaction between ENSO and the Annual Cycle seems to be directed mainly from ENSO towards the Annual Cycle. In the future, this could help in building more advanced theoretical models for the interaction between ENSO and the Annual Cycle and, moreover, we expect that the method presented here could improve the analysis of complex spatiotemporal systems.

Acknowledgments

We thank M Abel, A Timmermann and J Wichard for useful discussions. We also thank the referees for their comments. The author was partially supported by the European Training Network COSYC of SENS (contract number HPRN-CT-2000-00158).

Appendix A

Let us define a matrix $T_{n \times m}$ of data, in such a way that $T_{ij}$ is the SST at time $t_i$ in a certain spatial point $x_j$. With this notation, the matrix is built up of $n$ points $\{T_i\}$ in a $\mathbb{R}^m$ space, and these points belong to a trajectory as they are ordered in time. In this space, for every two points $\{T_i, T_j\}$ in $\mathbb{R}^m$, the metric or distance $d_{ij} = d(T_i, T_j)$ can be defined as a function onto nonnegative real numbers that obey the following rules: $d_{ij} = 0$ iff $i = j$, $d_{ij} = d_{ji}$ and the triangle inequality, which states that $d_{ij} + d_{jk} \geq d_{ik}$.

The distance can be interpreted as a measure of similarity: if $d_{ij} < d_{ik}$ we say that the physical state represented by $T_i$ is closer to $T_j$ than $T_k$. The MDS approach makes use of the matrix of squared distances

$$\left( D^{(2)} \right)_{ij} = d_{ij}^2 \quad \text{(A.1)}$$

and applies the following procedure (Borg and Groenen 1997). First, $D^{(2)}$ is transformed via an operation called double-centring. More precisely, using a matrix $J$ where $J_{ij} = \delta_{ij} - 1/n$, where
\( \delta_{ij} \) is the Kronecker delta, we define the matrix
\[
Z^{(2)} = -\frac{1}{2} J D^{(2)} J. 
\] (A.2)

Then, \( Z^{(2)} \) is decomposed into its eigenvalues and eigenvectors \( Z^{(2)} p_i = \lambda_i p_i \). Defining \( \Lambda_{ij} = \lambda_i \delta_{ij} \) with \( \lambda_i > \lambda_{i+1} \) and \( P_{ij} = p_{ij} \) the component \( j \) of vector \( p_i \), then \( Z^{(2)} = P \Lambda P \). If the set of eigenvalues have some leading components \( \{ \lambda_1, \ldots, \lambda_p \} \) with \( p < m \), and the rest decay very fast, meaning \( \lambda_p \gg \lambda_{p+1} \), we could approximate \( Z^{(2)} \) to its projection into the subspace spanned by the eigenvectors \( \{ p_1, \ldots, p_p \} \). In this way we can reduce the dimensionality of the data by taking \( P_i \Lambda_i^{1/2} \) from \( i = 1 \) to some value \( p < m \), which could be \( p \ll m \). Therefore, 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 residual variance \( r_p \), of dimension \( p \) is defined using \( r_p = 1 - \frac{\sum \lambda_i}{\text{Tr}(Z)} \), and takes the value \( r_p = 1 \) if no statistical variance is explained, and \( r_p = 0 \) if all the variability is taken into account.

The main difference between the MDS and Isomap comes from the particular definition of distance used in equation (A.1). If the data points belong to a nonlinear manifold, the orthonormal projection spreads contributions to the variance on to 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. To obtain a better representation of similarities between data points, we use the natural metric for nonlinear manifolds, the geodesic distance (Do Carmo 1976), which can be defined as follows: let us have an euclidean space of temperatures \( T \) of coordinates \( \{ T_1, \ldots, T_m \} \). Suppose there is a manifold \( \Theta \subseteq T \) represented by the coordinates \( \{ \Theta_1, \ldots, \Theta_p \} \). The metric of \( \Theta \) is the matrix \( g \) with elements defined by,
\[
g_{ij} = \sum_{k=1}^{m} \frac{\partial T_k}{\partial \Theta_i} \frac{\partial T_k}{\partial \Theta_j}. \quad (A.3)
\]

Generally speaking, we can define the distance between two data points \( \theta_1(T_1) \) and \( \theta_2(T_2) \) as,
\[
d(\theta_1, \theta_2) = ||\theta_1 - \theta_2|| = \int_{\theta_1}^{\theta_2} \sqrt{\sum_{i,j=1}^{m} g_{ij} d\Theta_i d\Theta_j}. \quad (A.4)
\]

In the case of a euclidean manifold, \( g_{ij} = \delta_{ij} \) and the distance between two points \( T_i \) and \( T_j \) is
\[
d_{ij} = \|T_i - T_j\| = \sqrt{\sum_{k=1}^{m} (T_{ik} - T_{jk})^2} \quad (A.5)
\]
which is the well-known expression of euclidean distance.

Summarizing, PCA can be regarded as an euclidean MDS for normalized data. We would like to stress that PCA is a linear method of decomposition, where the data are projected into orthonormal linear subspaces. 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. Details about how to calculate a geodesic squared distance matrix for a particular manifold can be found in Tenenbaum et al (2000) and references therein.

![New Journal of Physics 9 (2007) 432 (http://www.njp.org/)](http://www.njp.org/)
Appendix B

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 at the Equator. Due to the conservation of angular momentum, the mean winds are easterly (i.e. westward) 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 easterly 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 characterized 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°N in August and September, while it moves south to the vicinity of the Equator in March and April. Concerning the SST, the moist air rises where the SST is highest while dry air subsides where waters are cold. For these reasons, 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. Considering all these processes, the resulting 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.

Under this scheme, ENSO could be seen as a perturbation of the mean state described by the Annual Cycle. 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. The wind stress perturbation creates a Kelvin wave that moves to the east and 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 reinforce the instability growth in the east but do not travel back as they 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. This dynamics can be described by reduced Navier–Stokes equations on a rotating sphere with a equatorial beta plane approximation (Gill 1982), in what is called the shallow water equations, plus the thermodynamical equations and atmospheric wind stress. As this model is too complicated to be solved even numerically, several approximations are necessary.

There are several models that describe the phenomenon mathematically, the DO model (Suarez and Schopf 1988; Tziperman et al 1994) being the one that comes closer to our approach in terms of dimensionality and definition of variables. The DO model is a low-order chaotic system driven by a external force associated with 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 heuristic equation of the temperature deviations $\hat{T}$ from the mean profile is,

$$\frac{d\hat{T}(t)}{dt} = a\hat{T} - b\hat{T}^3 - c\hat{T}(t - \delta),$$

where $a$, $b$, $c$ are parameters and $\delta$ is the delay that accounts for the time needed by the Kelvin wave to travel to the eastern boundary and get back as Rossby waves. The linear terms consider the effect of the oceanic waves, while the nonlinearity is necessary to limit the growth of the wave amplitude.

References

Abel M 2004 Nonparametric modeling and spatiotemporal dynamical systems Int. J. Bifurcat. Chaos 14 2027–39
Abel M, Ahnert K, Kurths J and Mandelj S 2005 Additive nonparametric reconstruction of dynamical systems from time series Phys. Rev. E 71 015203
An S I and Wang B 2001 Mechanisms of locking of the El Niño and La Niña mature phases to boreal winter J. Clim. 14 2164–76
An S I and Jin F F 2004 Nonlinearity and asymmetry of ENSO J. Clim. 17 2399–412
Blanke B, Neelin J D and Gutzler D 1997 Estimating the effect of stochastic wind stress forcing on ENSO irregularity J. Clim. 10 1473–86
Borg I and Groenen P 1997 Modern Multidimensional Scaling: Theory and Applications (Berlin: Springer)
Chang P, Ji L, Wang B and Li T 1994 Interactions between the seasonal cycle and El Niño-Southern Oscillation in an intermediate coupled ocean-atmosphere model J. Atmos. Sci. 52 2353–72
Chen D, Cane M A, Kaplan A, Zebiak S E and Huang D 2004 Predictability of El Niño over the past 148 years Nature 428 733–6
Do Carmo M 1976 Differential Geometry of Curves and Surfaces (Englewood Cliffs, NJ: Prentice-Hall)
Gámez A J, Timmermann A, Zhou C S and Kurths J 2004 Nonlinear dimensionality reduction in climate data Nonlinear Processes Geophys. 11 393–8
Gill A E 1982 Atmosphere-Ocean dynamics (International Geophysics Series) (New York: Academic)
Grieger B and Latif M 1994 Reconstruction of the El Niño attractor with neural networks Clim. Dyn. 10 267–76
Härdle W and Hall P 1993 On the Nonparametric Regression Algorithm for Additive Regression Models (Cambridge: Cambridge University Press)
Hannachi A, Stephenson D B and Sperber K R 2003 Probability-based methods for quantifying nonlinearity in the ENSO Clim. Dyn. 20 241–56
Hastie T and Tibshirani R 1990 Generalized Additive Models (London: Chapman and Hall)
Jin F F, Neelin J D and Ghil M 1996 El Niño/Southern Oscillation and the Annual Cycle: subharmonic frequency-locking and aperiodicity Physica D 98 442–65
Joliffe I T 1986 Principal Component Analysis (Berlin: Springer)
Kantz H and Schreiber T 1997 Nonlinear Time Series Analysis (Cambridge: Cambridge University Press)
Kaplan A, Cane M, Kushner Y, Clement A, Blumenthal M and Rajagopalan B 1998 Analyses of global sea surface temperature 1856–1991 J. Geophys. Res. 103 18567–89
Li T and Philander S G 1998 Decadal variability in ENSO predictability and prediction J. Clim. 11 2804–22
Kolmogorov A N 1963 On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition AMS Transl. 28 55–9
Latif M, Barnett T, Cane M, Flügel M, Graham N E, von Storch H, Xu J S and Zebiak S E 1994 A review of ENSO prediction studies Clim. Dyn. 9 164–79
Latif M, Anderson D, Barnett T, Cane M, Kleeman R, Leetmaa A, O’Brien J, Rosati A and Schneider E 1998 A review of the predictability and prediction of ENSO J. Geophys. Res. 103 14375–93

New Journal of Physics 9 (2007) 432 (http://www.njp.org/)
Li T and Hogan T F 1999 The role of the annual-mean climate on seasonal and interannual variability of the tropical Pacific in a coupled GCM J. Clim. 12 780–92
Li T and Philander S G 1996 On the Annual Cycle of the eastern equatorial Pacific J. Clim. 9 2986–98
Monahan A H 2001 Nonlinear principal component analysis: tropical Indo-Pacific sea surface temperature and sea level pressure J. Clim. 14 219–33
Patil D J, Hunt B R, Kalhay E, Yorke J A and Ott E 2001 Local low dimensionality of atmospheric dynamics Phys. Rev. Lett. 86 5878–81
Penland C and Sardeshmukh P D 1995 The optimal growth of tropical sea surface temperature anomalies J. Clim. 8 1999–2024
Penland C 1996 A stochastic model of IndoPacific sea surface temperature anomalies Physica D 98 534–58
Pikovsky A, Rosenblum M and Kurths J 2001 Synchronization: A Universal Concept in Nonlinear Sciences (Cambridge: Cambridge University Press)
Poggio T and Girosi F 1990 Networks for approximation and learning Proc. IEEE 78 1481–97
Reynolds R W and Smith T M 1995 A high resolution global sea surface temperature climatology J. Clim. 8 1571–83
Sauer T, Yorke J and Casdagli M 1991 Embeddology J. Stat. Phys. 65 579–616
Schreiber T and Schmitz A 2000 Surrogate time series Physica D 142 346–82
Suarez M J and Schopf P S 1988 A delayed action oscillator for ENSO J. Atmos. Sci. 45 3283–7
Sura P and Sardeshmukh P D 2007 A global view of non-gaussian SST variability J. Phys. Oceanogr. at press
Tang Y and Hsieh W W 2002 Hybrid coupled models of the tropical Pacific: II ENSO prediction Clim. Dyn. 19 343–53
Tenenbaum J B, de Silva V and Langford J C 2000 A global geometric framework for nonlinear dimensionality reduction Science 290 2319–23
Timmermann A, Jin F F and Abshagen J 2003 A nonlinear theory for El Niño bursting J. Atmos. Sci. 60 152–65
Tziperman E, Stone L, Cane M A and Jarosh H 1994 El Niño chaos: overlapping of resonances between the seasonal cycle and the Pacific ocean-atmosphere oscillator Science 264 72–4
Vitushkin A G 1977 On representation of functions by means of superpositions and related topics Enseignement Math. 23 255–320
Voss H, Bünner M J and Abel M 1998 Identification of continuous spatiotemporal systems Phys. Rev. E 57 2820–23
Xie S P 1994 On the genesis of the equatorial Annual Cycle J. Clim. 7 2008–13
Xie S P 1995 Interaction between the annual and interannual variations in the Equatorial Pacific J. Phys. Oceanogr. 25 1930–41