Improvements to the Non-linear Principal Component Analysis Method, with Applications to ENSO and QBO

Stephen C. Newbigging[1], Lawrence A. Mysak[1] and William W. Hsieh[2*]

[1]*Department of Atmospheric and Oceanic Sciences, McGill University, Montreal QC*
[2]*Department of Earth and Ocean Sciences, University of British Columbia, Vancouver BC V6T 1Z4*

[Original manuscript received 24 July 2002; in revised form 15 May 2003]

ABSTRACT *Two improvements to the Non-linear Principal Component Analysis (NLPCA) method are presented. In the normal application of this method, a non-linear curve C is found that best fits the data. The method provides a projection function mapping from the data space to the curve C. However, this projection function is faulty in that points in the data space are generally not projected onto their closest neighbours on C. Here, a new projection function is introduced which ensures that the data points are projected onto their closest neighbours on C, resulting in an increase in the amount of variance explained by the NLPCA mode. This is illustrated by an analysis of the sea surface temperature anomaly data from the tropical Pacific, where the El Niño-Southern Oscillation (ENSO) phenomenon is manifested. A second shortcoming of the NLPCA method is that the curve C comes with a parametrization which is arbitrary and has no physical interpretation. Here, the curve is re-parametrized by arc length. This allows the computation of more meaningful time series, which we illustrate through an analysis of the Quasi-Biennial Oscillation (QBO) in the equatorial stratospheric zonal wind data.*

RÉSUMÉ *[traduit par la rédaction] On présente deux améliorations à la méthode d'analyse non linéaire des composantes principales (NLPCA). Normalement, avec cette méthode, on trouve une courbe non linéaire C qui satisfait au mieux les données. La méthode fournit une fonction de projection établissant une correspondance entre l'espace de données et la courbe C. Cependant, cette fonction de projection est imparfaite car les points dans l'espace de données ne sont généralement pas projetés sur leurs plus proches voisins sur la courbe C. Ici, on introduit une nouvelle fonction de projection qui assure que les points de données sont projetés vers leurs plus proches voisins sur C, ce qui aboutit à un accroissement dans la quantité de variance expliquée par le mode NLPCA. Ceci est montré par une analyse des données d'anomalie de température de la surface de la mer dans le Pacifique tropical, où le phénomène El Niño–oscillation australe (ENSO) se manifeste. Une deuxième lacune de la méthode NLPCA est que la courbe C vient avec une paramétrisation arbitraire et ne possède pas de signification physique. Ici, la courbe est reparamétrisée par longueur d'arc. Ceci permet le calcul de séries temporelles plus significatives, ce que nous illustrons au moyen d'une analyse de l'oscillation quasi biennale dans les données sur le vent zonal stratosphérique équatorial.*

1 Introduction

A number of papers have recently appeared which are concerned with the Non-linear Principal Component Analysis (NLPCA) method and its applications to climate data (Hsieh, 2001a; Hamilton and Hsieh, 2002; Monahan et al., 2000; Monahan, 2001). First proposed by Kramer (1991), NLPCA is a neural-net-based generalization of principal component analysis (PCA), using an autoassociative multi-layer perceptron network architecture. The purpose of this paper is to introduce two improvements to the NLPCA method that were devised following a re-examination of the El Niño-Southern Oscillation (ENSO) study of Monahan (2001; hereafter M2001) and the stratospheric Quasi-Biennial Oscillation (QBO) investigation of Hamilton and Hsieh (2002; hereafter HH2002). First, it will be shown that because of the nature of the mapping from the data space to the NLPCA curve used by M2001 in the NLPCA application to tropical Pacific sea surface temperature (SST) anomalies, the analysis underestimates the closeness with which the curve approximates the observed SST data. Second, we show that because of the nature of the curve parametrization used in HH2002, there are shortcomings in their QBO index time series, which can be removed by introducing an arc-length parametrization for the aforementioned curve.

Although the discussion and examples given here are in the context of NLPCA, the above shortcomings are also present in, and the improvements are valid for, a recent neural-net-based non-linear generalization of canonical correlation analysis (see Hsieh, 2001b). We also note that the existence

*Corresponding author's e-mail: whsieh@eos.ubc.ca

of the above two shortcomings were recognized in Malthouse (1998), hereafter M98. However, to the best of our knowledge, this is the first time that the particular consequences of these shortcomings have been examined.

It should be mentioned that Kramer's autoassociative neural network NLPCA is only one of several approaches to non-linearly generalizing PCA (Cherkassky and Mulier, 1998). Another common approach, the principal curves method (Hastie and Stuetzle, 1989), finds a non-linear curve which passes through the middle of the data points. Malthouse (1998) made a comparison between principal curves and Kramer's NLPCA model. Unfortunately, when testing a closed curve solution, he used Kramer's NLPCA (which is suitable only for open curves) instead of the version of NLPCA proposed by Kirby and Miranda (1996) (which handles closed curves well). There is no conclusive study that shows which approach is superior. While the neural network NLPCA method has the advantage of using analytical mapping functions, its projection function may be sub-optimal. In our study, we correct the problems in NLPCA by using concepts from principal curves, namely the projection index and arc-length parametrization.

With synthetic data, it is easy to demonstrate the superiority of non-linear techniques to linear techniques. However, with real data, non-linear techniques may not produce better results than linear techniques. This happens if the data record is short and noisy, or the underlying dynamics are linear. But for strong signals like ENSO and QBO, the advantage of non-linear techniques over linear techniques is easily seen (Hsieh, 2001a,b; HH 2002).

The remainder of this paper is structured as follows. In Section 2 a perspective of the NLPCA method is given. In Section 3 the first improvement, the projection index, is introduced and applied to ENSO data. In Section 4 the arc-length parametrization is described and applied to the QBO. The conclusions are given in Section 5, and an Appendix outlines the method by which we implement the projection index and the arc-length parametrization.

2 A perspective of NLPCA

The problem addressed by NLPCA is as follows: suppose a multivariate dataset clearly exhibits a non-linear structure but the functional form of the non-linearity is unclear. We would like to establish a procedure for producing a curve which passes through the middle of the dataset, and we would like the procedure itself to choose the functional form of the curve. In the case where we specify that the curve be a straight line, a widely used technique to fit such a dataset is PCA. PCA extracts eigenvectors representing spatial patterns (also called loadings), and associated time coefficients called principal components (PCs).

For an introduction to the NLPCA method, see the original paper by Kramer (1991), or the review paper by Hsieh (2003). Here we describe only those aspects of NLPCA which are relevant to the improvements introduced in this paper. NLPCA should be understood as a method for searching a non-parametric family of functions for the member which best satisfies a certain condition (to be defined shortly) known as the selection criterion. The family of functions being searched has the property that each of its members can be written naturally as the composition of two functions, $C : \mathcal{R} \to \mathcal{R}^n$ and $\mathcal{P} : \mathcal{R}^n \to \mathcal{R}$. C is therefore a parametrized curve in $\mathcal{R}^n$, and the composition $C \circ \mathcal{P}$ is a map from $\mathcal{R}^n$ to the curve parametrized by C. The condition to be satisfied is that the root mean square (rms) distance between the data points $x_i \in \mathcal{R}^n$ and their images under $C \circ \mathcal{P}$ (x_i) should be minimized. This is the same requirement used in iterative versions of PCA, where the first eigenvector gives the straight line which minimizes the rms distance from the data to the line.

The family of functions from which $C \circ \mathcal{P}$ is to be chosen must be specified in advance, and given any continuous function F, a family may be chosen so as to contain a member which approximates F as closely as desired (see Cybenko, 1989). However, once chosen, the family of candidate functions remains fixed during the search. Moreover, if the chosen family is too large, a problem known as over-fitting is encountered. Over-fitting occurs when the curve fits the data too closely so that it models noise in the data as well as any underlying relationships that may be present (in extreme cases, a zigzag curve could be chosen to pass through all data points). The usual method of avoiding over-fitting is to restrict heavily the family of curves from which C is to be drawn. However, this practice exacerbates the problems associated with the projection function which are discussed in the following section.

3 The projection index and application to ENSO

It should be realized that $C \circ \mathcal{P}$ is not a projection function in the usual sense. First, it is not idempotent, i.e., $C \circ \mathcal{P}$ does not map points on C to themselves, as the projection function was trained with only a finite sample, and with a finite number of neurons. Second, it does not in general map points $x_i \in \mathcal{X}$ to their closest neighbours on C. This problem is illustrated in Fig. 1, which shows the result of analysing a synthetic dataset with NLPCA. The points mark the elements in the dataset and the parabola-like curve C is the NLPCA approximation to the data. The curves which cut C are lines of constant projection under $C \circ \mathcal{P}$: all points along a given intersecting curve are mapped by $C \circ \mathcal{P}$ to the same point on C. As can be seen in Fig. 1, the errors are significant over large areas of the data-space, but are particularly severe near the middle of the parabola. In this region points lying close to the arms of the parabola are mapped far down the arm towards the centre of the curve.

This problem is an artefact of the method and not specific to this particular analysis. It arises from the fact that $C \circ \mathcal{P}$ is required to be a continuous map, and is exacerbated by the limitations placed on the family of candidate curves to avoid over-fitting. Briefly, to satisfy the selection criterion, $C \circ \mathcal{P}$ must project points from the data space to their closest neighbours on C. However, between the two arms of the parabola lies a line of 'ambiguity' points (Hastie and Stuetzle, 1989; M98), each

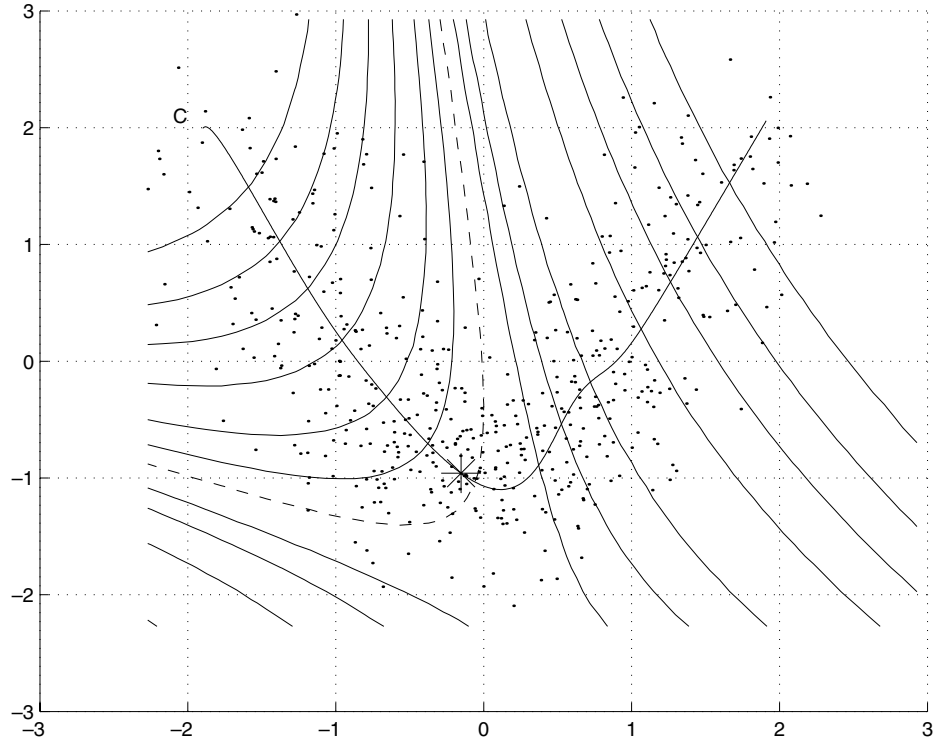

Fig. 1 Inefficiencies in $C \circ \mathcal{P}$. The scatter-plot is of the points in a synthetic bivariate dataset. The parabola-like curve C is the NLPCA approximation of the data, and the transecting curves are isopleths of $C \circ \mathcal{P}$. If T is one of the transecting curves, then each of the points along T is projected to the same point on C, and moreover that point is near, but not equal to, the point of intersection of T and C. For example, all points along the dashed curve are projected to the point marked with a star.

having two closest neighbours on the parabola (one on each arm). Points lying slightly to the left of this line should be mapped to the parabola's left arm, and points lying to the right should be mapped to the parabola's right arm. This implies a discontinuity in the projection map. The only way to approximate this discontinuous function with a continuous one is to map points around the line of ambiguity towards the middle of the parabola, implying a region of inaccuracy near the line of ambiguity, because points neighbouring the line of ambiguity must be fanned out by the mapping to cover the middle of the parabola. To see this behaviour, observe the dashed line and its two neighbours on either side in Fig. 1. At the top of the figure, these five lines lie close together in the centre of the parabola near the line of ambiguity. However, their intersections with C cover a large portion of the centre of the parabola. If the family of functions from which $C \circ \mathcal{P}$ is chosen is sufficiently large, this region of ambiguity can, in theory, be made as small as desired. However, by doing so the problem of over-fitting is again encountered, rendering this cure undesirable. Further discussion may be found in M98 where his Fig. 4 is particularly illuminating.

The function $\mathcal{P}$ is essential to the algorithm by which $C \circ \mathcal{P}$ (and therefore C) is selected. However, after $C \circ \mathcal{P}$ has been found, the curve C is known independently of $\mathcal{P}$. The errors in projection associated with the composite map $C \circ \mathcal{P}$ can therefore be avoided by discarding $C \circ \mathcal{P}$ in favour of the projection index $\mathcal{P}_r$, which is defined to be the function which sends points in the data space to their closest neighbours on C (Hastie and Stuetzle, 1989). It should be noted that algorithms to approximate $\mathcal{P}_r$ as closely as desired are easy to implement (see Appendix). The following text illustrates the advantage of doing so.

A measure of the extent to which a low-dimensional summary $\mathcal{L}$ (here $\mathcal{L}$ is either the NLPCA C or the first PCA eigenvector) of a dataset $\mathcal{X}$ captures the essential features of $\mathcal{X}$ is called the fraction of explained variance (*FEV*). The *FEV* is defined as follows:

$$FEV = 1 - \frac{\sum \left\| \mathcal{P}_r(x_i) - x_i \right\|^2}{\sum \left\| x_i \right\|^2}$$

where the x_i are the points in $\mathcal{X}$, the numerator is the sum of the squares of the distance of the data points from their projections onto $\mathcal{L}$, and the denominator is the total variance of the dataset, assuming without loss of generality that the dataset has zero mean.

M2001 found empirically that NLPCA shares the partition of variance property with PCA, i.e., that the variance of the dataset is equal to the sum of the variance of its projection onto C plus the variance of the residuals:

$$\sum \left\| x_i \right\|^2 = \sum \left\| \mathcal{P}_r(x_i) \right\|^2 + \sum \left\| x_i - \mathcal{P}_r(x_i) \right\|^2.$$

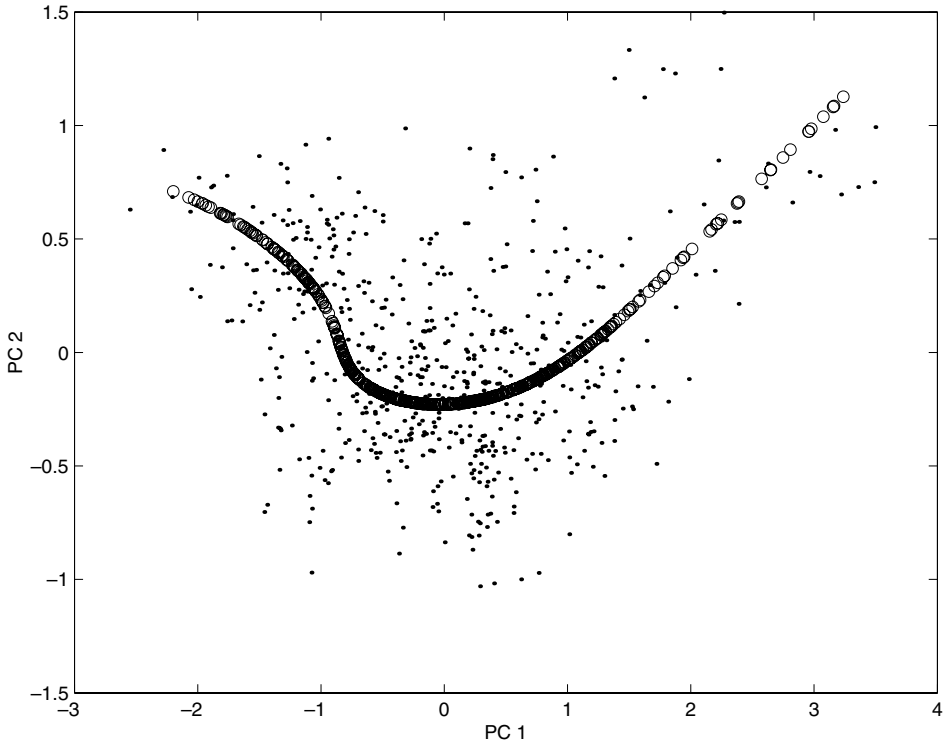

Fig. 2 Results of the NLPCA analysis of ENSO data. The scatter-plot points are the PC1-PC2 coordinates of the SSTA data, and the curve is the NLPC projected onto this plane. The NLPC explains 62.6% of the variance using the NLPCA projection function, but under the projection index this increases to 63.1%.

NLPCA has been judged to be superior to PCA because the *FEV* associated with C is generally greater than the *FEV* associated with the first PCA mode. Since $\mathcal{P}_r$ is by definition the map that minimizes the quantity $(\mathcal{P}_r(x_i) - x_i)^2$ for each point $x_i \in \mathcal{X}$, (thereby maximizing the *FEV*), $C \circ \mathcal{P} \neq \mathcal{P}_r$ implies that using $C \circ \mathcal{P}$ in place of $\mathcal{P}_r$ in the computation of the *FEV* underestimates the closeness with which NLPCA approximates the data.

We illustrate with an example. M2001 performed an NLPCA of the tropical Pacific SST. Here we performed NLPCA on the same data using code obtained from Hsieh (2001a). We then implemented a version of $\mathcal{P}_r$ and compared the NLPCA *FEV* using both $C \circ \mathcal{P}$ and $\mathcal{P}_r$. The family of curves from which $C \circ \mathcal{P}$ was chosen was the same as in M2001. The data analysed were the time series of the first ten PCs of the monthly Pacific SST anomalies from the National Oceanic and Atmospheric Administration for the period January 1950 to April 1999. The results of our analysis are shown in Fig. 2: the projections of the data onto the PC1–PC2 plane appear as points in a scatter plot, and the curve passing through the middle is the projection of the NLPCA solution onto this plane.

For us the NLPCA algorithm converged to a curve with an *FEV* of 62.6% as computed with $C \circ \mathcal{P}$, which is an improvement on the *FEV* = 57.7% associated with the first PCA mode. When the *FEV* was re-computed using the projections of the data onto C using $\mathcal{P}_r$, this same curve had an *FEV* of 63.1%, revealing that errors in the projection function had caused the *FEV* to be underestimated by half a percentage point. This half-percentage point difference is about 10% of the improvement in the *FEV* achieved by using NLPCA instead of PCA.

For this particular example, the difference using the improvement is small; we chose the example because it comes from the literature. In general, the more data points there are in regions where $C \circ \mathcal{P}$ is likely to be inaccurate (i.e., near the ambiguity points), the greater the difference our procedure is likely to make. To illustrate this, we turn our attention back to the artificial data from Fig. 1, where the data were generated so that the first PCA explains 50% of the variance. Here, the *FEV* associated with the non-linear PC is 86.5% under $C \circ \mathcal{P}$, whereas under $\mathcal{P}_r$ the *FEV* increases to 91.2%. This is an increase of almost 5%. The possibility therefore exists that unless the projection index is used, future applications of NLPCA will significantly underestimate the degree to which NLPCA characterizes the data.

4 Arc-length parametrization and application to the QBO

The function $\mathcal{P}_r$ associates with each data point x_i a number $\lambda_i = \mathcal{P}_r(x_i)$ which is called the score value of the data point. These score values have been used to produce time series of the projections of the data onto the NLPCA mode (M2001, HH2002). The score values thus produced serve as an index of the phenomenon being studied (ENSO in the case of M2001, the QBO in the case of HH2002), and the time series of these score values is a representation of how the phenomenon is evolving with time.

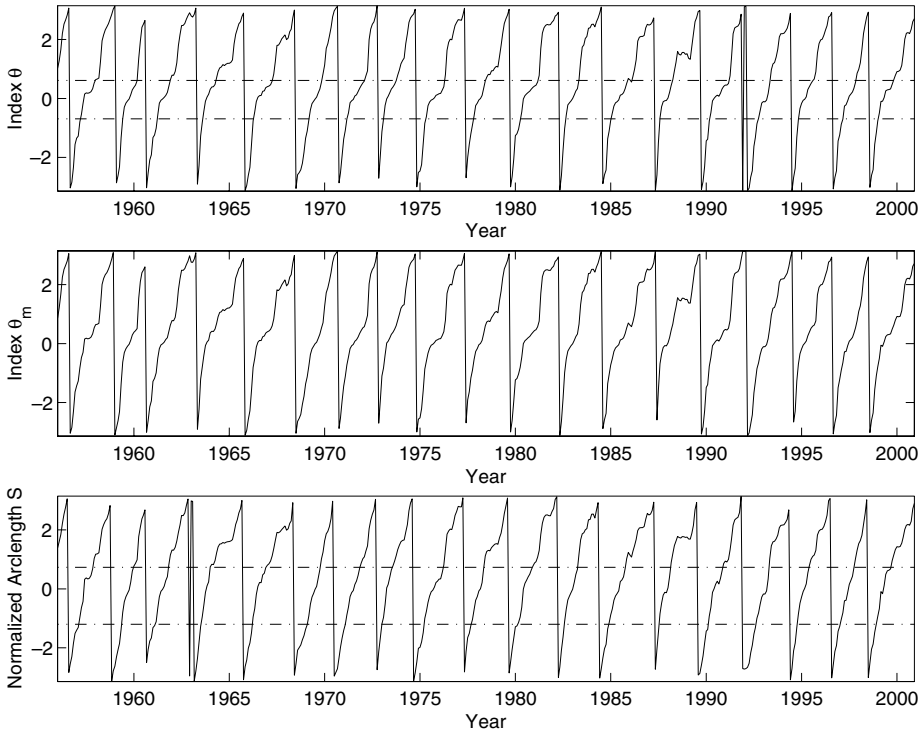

Fig. 3 Time series of the score values of the QBO. In the top panel the score values are computed with the NLPCA projection and parametrization θ. The middle panel plots the score values as computed using the projection index, but still relative to the NLPCA parametrization θ. The bottom panel is a plot of the time series as computed using both the projection index and the re-parametrization by arc length, *S*. Note that the ordinates are periodically bounded between (–π, π). For the meaning of the dot-dashed lines in the top and bottom panels, see text.

There are two shortcomings to using these time series. We have already seen in Section 3 that $C \circ \mathcal{P}$ inaccurately projects data points onto the NLPCA mode; this means that the score values of the data points will also be inaccurate. This source of inaccuracy is eliminated once $\mathcal{P}_r$ has been substituted for $C \circ \mathcal{P}$. The second shortcoming arises from the fact that C, as produced by the NLPCA algorithm, is not parametrized by arc length, hence is not easily interpretable. We will discuss the way in which a non-arc-length parametrization distorts time series in the context of the results of HH2002.

Using a version of the NLPCA method proposed by Kirby and Miranda (1996), HH2002 analyse the QBO in the equatorial stratospheric zonal wind data (Naujokat, 1986). The data are the monthly means of the zonal wind components taken at 70, 50, 40, 30, 20, 15 and 10 hPa, from September 1967 to December 2000 for a total of 540 monthly values. These data from the seven heights were treated as points in a seven-dimensional data space, and a version of NLPCA was performed in which the family of curves from which C is chosen was a family of closed curves. The curve C which results from the analysis is therefore parametrized by a cyclic variable of period 2π which HH2002 call θ. In this case, the original analysis was available to us (W.W. Hsieh, personal communication, 2001), and so we are in possession of the functions C and $\mathcal{P}$ resulting from their study.

The phase of the QBO has been difficult to define, since the wind change occurs at different times for different vertical levels. The NLPCA approach of HH2002 was successful in extracting the phase θ using data from all seven levels. The phase of the QBO is known to be related to the stratospheric polar winter temperature anomalies in the northern hemisphere (the Holton-Tan effect), and the phase from HH2002 identified a stronger effect than previous studies.

HH2002 constructed a time series of the QBO score values, which is replotted here in the top panel of Fig. 3. HH2002 noted that θ seems to progress systematically more quickly through some of its values than others. A goal in HH2002 is to construct a composite of the QBO by tracking the average progress of the QBO through phase space. This is done in two steps: first, an NLPCA of the zonal wind data is performed, resulting in the curve C. The average progress of the projections of the QBO onto C (where the average is taken over many QBO cycles) is then to be followed, taking equal time-interval snapshots of the vertical wind structure. In order to obtain equal time-interval snapshots, they cannot simply take snapshots at equal θ-intervals; rather, they must compensate for the fact that θ routinely advances through some values more quickly than through others. To do this, they divide C into a number of segments such that θ increases by a uniform amount over each segment. They then count how often the projection of the QBO onto C lands in each segment, and reason that the QBO must be spending more time near those segments where its projection is often found. When calculating the final composite, they therefore weigh each segment

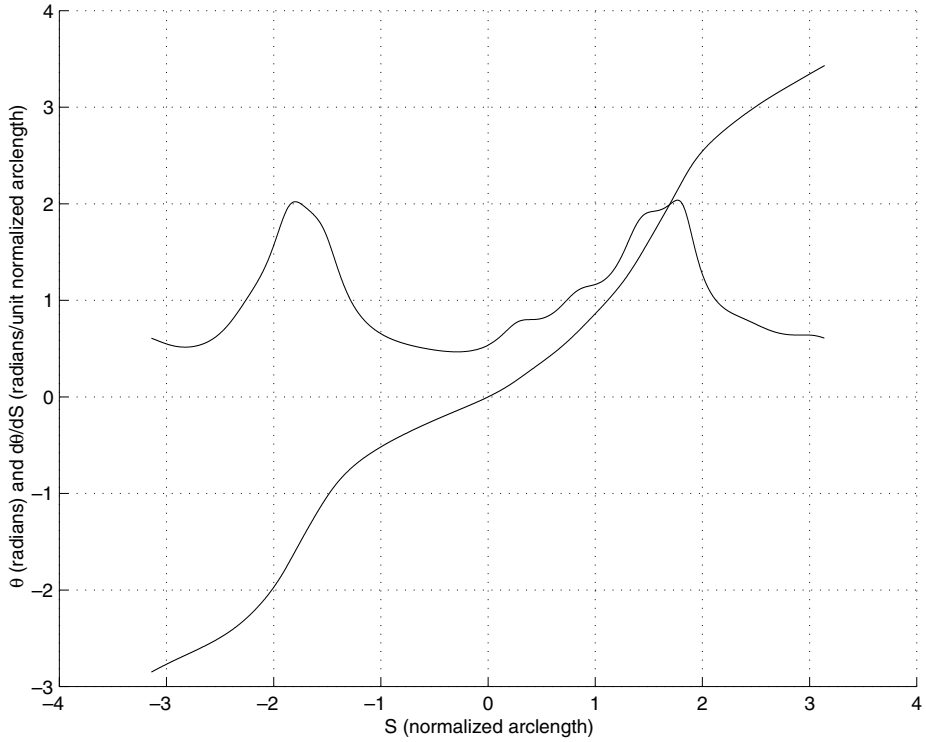

Fig. 4 Comparison of the NLPCA and arc length parametrizations for the NLPC of the QBO data as computed in HH2002. The monotonically increasing function is a plot of θ vs. S, and the curve showing local extrema is its derivative $d\theta/dS$. The rate of change of θ with S varies by a factor of four, indicating significant distortions in time series as computed with θ.

according to the frequency with which it contains the projection of the QBO.

We claim that part of the variability in the speed with which θ increases is a result of the fact that θ does not increase proportionally to the arc length along the curve; the remaining part of this variability is a representation of a genuine variation in the speed at which the QBO advances through phase space. Moreover, by re-parametrizing C by arc length, we show that we can separate these two sources of variability, thereby isolating the true variability with which the QBO progresses through phase space. This is interesting in its own right, and it may also be used to construct a composite QBO as in HH2002 where the systemic biases associated with the arbitrary parametrization have been entirely removed.

In order to see the manner in which the change in score value with time is influenced by the choice of parametrization, consider C under two different parametrizations. In what follows $C(\theta)$ will represent the curve C with the variable-speed parametrization supplied by the NLPCA algorithm, whereas $C(S)$ will denote the re-parametrization of C by arc length S.

In general, the values of S and θ associated with a given point on C will not be the same, and thus θ may be considered as a function of S, i.e., $\theta = \theta(S)$. The nature of NLPCA is such that $\theta(S)$ is a differentiable function, and since both θ and S are normalized cyclical variables of period 2π, the value of $d\theta/dS$ will average to one when integrated over one period. The true speed with which the projection of the data moves along C is, by definition, dS/dt. Differentiating θ with respect to time via the chain rule yields

$$\frac{d\theta}{dt} = \frac{d\theta}{dS}\frac{dS}{dt}$$

and so $d\theta/dt$ differs from the true rate of propagation by a factor of $d\theta/dS$. Thus $d\theta/dt$ will underestimate the true speed of progression where $d\theta/dS < 1$ and overestimate this speed where $d\theta/dS > 1$.

A comparison of θ and S is shown in Fig. 4. In this figure, the monotonically increasing line is the function $\theta(S)$, and the line with local extrema is $d\theta/dS$. It is evident that the NLPCA parametrization of C is far from being of unit speed, as the derivative $d\theta/dS$ varies from a minimum of 0.47 at $S = -0.29$ to a maximum of 2.04 at $S = 1.77$, which is a variation of a factor of four. The rate of change $d\theta/dt$ should therefore be expected to be a poor estimator of dS/dt.

That this is the case can be seen by examining Fig. 3. In the top panel, the values of θ were computed with $C \circ \mathcal{P}$, and therefore exhibit both types of shortcoming described above. The time series points in the middle panel are the projections of the data points onto the curve made with $\mathcal{P}_r$, and thus the θ values plotted correspond to slightly different points on C than do the values in the top panel. Interestingly, these two time series are virtually identical. This is due to the fact that the signal-to-noise ratio of the QBO is very high, and therefore

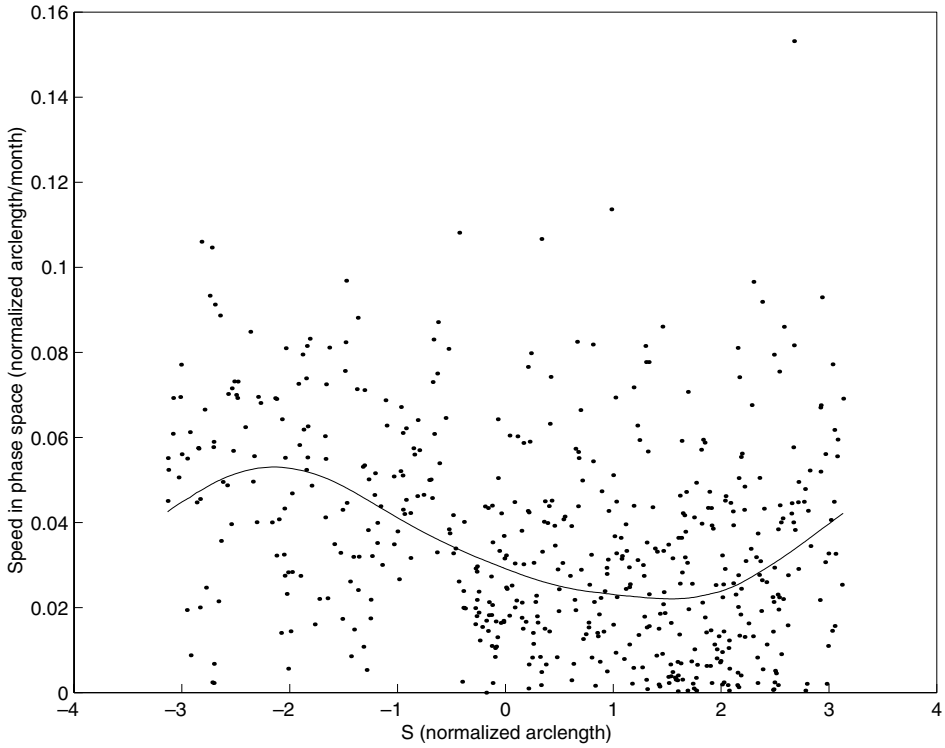

Fig. 5 Average rate of propagation of the QBO relative to its phase as measured by S. The curved line is a local average of the points with similar S values. The maximum at $S = -2.2$ corresponds to the middle of the shift from easterly to westerly wind patterns; the minimum at $S = 1.5$ corresponds to the middle of the shift from westerly to easterly winds.

few points are far from C. In this case, unlike the case of the ENSO anomalies analysed above, $C \circ \mathcal{P}$ is clearly a close approximation to $\mathcal{P}_r$. In order to quantify the accuracy, we note that the *FEV* as computed with $C \circ \mathcal{P}$ is 0.948, whereas the *FEV* computed using $\mathcal{P}_r$ is 0.949.

It is the bottom panel in Fig. 3 which makes clear the effect of the re-parametrization. This plot of the time series of S is an objective estimator of the position of the QBO in phase space. The region $-1.20 \le S \le 0.73$, lying between the dot-dashed lines in this panel corresponds to the region $-0.69 \le \theta \le 0.61$ lying between the dot-dashed lines in the top panel. Examination of this portion of the time series in the top panel reveals what seems to be a systematic slowing of the QBO for these values of θ. Examination of the corresponding portion of the time series parametrized by arc length shows little if any systematic slowing, indicating that the apparent slowing in the top panel was an artefact of the NLPCA parametrization, and not a property of the QBO.

We now use the time series in the top panel of Fig. 3 to obtain a true estimate of the speed at which the QBO moves through phase space. The speed of the QBO at any time t_0 may be approximated by a forward difference scheme, computing the distance in phase space between the projections of successive observations of the QBO onto C, and dividing by the time between observations (one month). A scatter plot of the results is shown in Fig. 5, where the abscissae are the score values measured by the arc length of the QBO observations, and the ordinates are the estimates of the QBO speeds in phase space. The curved line is a kind of local average of the data points, and is the result of a local averaging procedure called robust locally averaged scatter-plot smoothing (see Cleveland, 1979). This curve, which we denote $\langle dS/dt \rangle$, represents an average speed of propagation of the QBO and is a function of S. It confirms our above observations: the region $-1.20 \le S \le 0.73$ is characterized by an average dS/dt, and is not slow, contrary to what the top panel of Fig. 3 would have had us believe.

Note that the average speeds represented by the curve in Fig. 5 are not the same as the speed $d\theta/dS$ obtained from Fig. 4. The former is dS/dt and is a true estimate of the speed of progression of the QBO through phase space. The latter is a measure of the extent to which the true speed of the QBO will be distorted if it is measured by $d\theta/dt$.

The systematic speeding up of the QBO near $S = -2$ does seem to be genuine, as dS/dt (in dimensional units) reaches its absolute maximum of 16.5 m s^{-1} mo^{-1} at $S = -2.2$. This corresponds to the middle of the easterly to westerly transition. The absolute minimum of 6.6 m s^{-1} mo^{-1} is reached in the middle of the westerly to easterly transition at $S = 1.5$. This is in keeping with the well-known phenomenon that the easterly to westerly transition is more rapid than the westerly to easterly transition. In terms of the average speed of propagation through phase space, the one transition is about twice as fast as the other.

We now create a new composite QBO using the re-parametrized Non-linear Principal Component (NLPC) and the

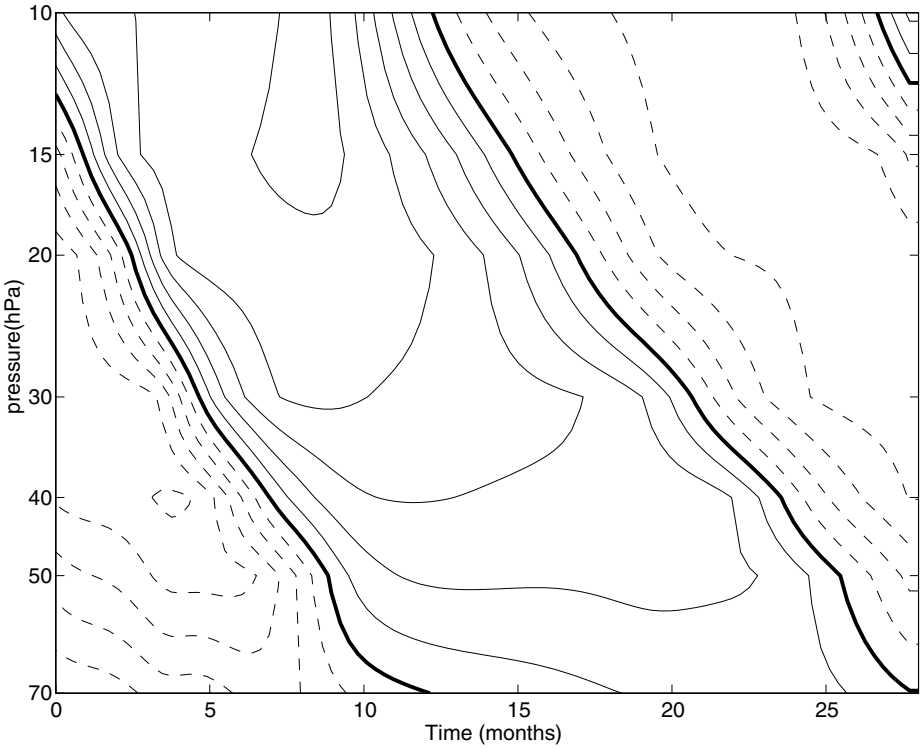

Fig. 6 A composite QBO, computed by following the NLPCA at its local average-speed $\langle dS/dt \rangle$ through phase space and taking equal time interval snapshots. The contour interval is 5 m s^{-1}, with positive (i.e., westerly) contours indicated with solid lines, negative contours by dashed lines, and zero contours by thick lines.

average speed $\langle dS/dt \rangle$ from Fig. 5 as follows: we chose $S = 0$ as an arbitrary starting point. We then followed S in the direction of QBO propagation, progressing at a speed equal to $\langle dS/dt \rangle$, so that we were able to take equal time interval snapshots of the reconstructed QBO. We re-emphasize the fact that $\langle dS/dt \rangle$ is a function of S: it is the local average-speed of the propagation of the QBO through phase space, the average being taken over many cycles. These snapshots were compiled into the composite shown in Fig. 6. In this way we used a true picture of the speed at which the QBO propagated and avoided the shortcomings associated with the smoothed-frequency histogram approach used in HH2002. Our approach is only possible once the re-parametrization has been done. The composite appearing in Fig. 6 can be compared with that shown in Fig. 5 of HH2002.

5 Conclusions

Neural-net-based generalizations of PCA and Canonical Correlation Analysis (CCA) have been developed which are capable of approximating low-dimensional structures present in complex datasets. However, the family of functions used as candidate data summaries for NLPCA has at least two undesirable properties. First, when used to project the data onto the NLPCA, the projection function does not project points to their nearest neighbours; one consequence is that the fraction of variance explained by the NLPC is underestimated.

In Section 3 we showed that the NLPCA approximation to the ENSO data showed an improvement in *FEV* over the first PC of 5% when $\mathcal{P}$ was used, but the gains were slightly larger when $\mathcal{P}_r$ was used instead, being of the order of 5.5%. We showed also that this effect has the potential to be quite significant by examining an artificial dataset where the discrepancy in *FEV* was over 5%. In Section 4 we demonstrated that the NLPC curves produced by NLPCA are not parametrized by arc length, and hence time series of the score values of the data are subject to arbitrary distortions. Since these distortions are systematic in nature, they can lead to misleading time series which display artificial structures, as seen in the QBO study of HH2002. The NLPCA codes used in Sections 3 and 4 are downloadable from our website http://www.ocgy.ubc.ca/projects/clim.pred/.

Acknowledgments

This work was supported by Canadian Natural Sciences and Engineering Research Council Research Grants awarded to L.A.M. and W.W.H.

Appendix: Implementation of the projection map and arc-length parametrization

In order to implement the projection index $\mathcal{P}_r$ and the arc-length parametrization S, we first approximate C with an open n-sided polygon C_r whose $n + 1$ vertices lie on C. This is done as follows.

We first choose the positions of the initial and final vertices of C_r. In the case where C_r is a closed curve, the initial vertex v_1 of C_r is chosen arbitrarily from the image of C, and the final

vertex v_{n+1} is set equal to v_1. When C is an open curve, we set $v_1 = \lim_{\theta \to -\infty} C(\theta)$, and $v_{n+1} = \lim_{\theta \to \infty} C(\theta)$. The existence of these limits (which we estimate numerically) is guaranteed by the functional form of C.

The positions of the internal vertices v_i, $i \in (2,\ldots, n)$ are chosen by computing points along C at $n-1$ evenly spaced values of its parameter θ and setting $v_i = C(\theta_i)$, where $i \in (2,\ldots, n)$. The interval between these θ-values is chosen small enough so that

$$\max \left\| C(\theta_{i+1}) - C(\theta_i) \right\| < \varepsilon, \qquad \forall i \in [1,\ldots, n+1].$$

Here ε is a positive user-defined parameter which sets the accuracy of the approximation. We note that because C is differentiable it is rectifiable (i.e., can be approximated by straight line segments), and so ε may be chosen small enough to make the maximum distance between C and C_r as small as desired. The rectifiability of C, together with the fact that its image has compact support, guarantees the existence of a finite set of θ values which satisfy Eq. (1).

The projection index and the arc-length parametrization are now easy to implement. To implement $\mathcal{P}_r$ we observe that if x is any point in the data space, its nearest neighbour on C_r may be found by computing its projection x onto each of the line segments comprising C_r and then setting $\mathcal{P}_r(x)$ equal to the closest of these points. To parametrize by arc length, we note that if x' is any point lying on C_r, it will lie between two vertices v_i and v_{i+1}. The distance as measured along C_r from v_1 to x' may be computed by summing the lengths of all line segments lying between v_1 and v_i, and adding the distance $\|x' - v_i\|$. At an ambiguity point, there are at least two closest points on C_r, the projection index chooses the point on C_r with the largest arc length.

References

CHERKASSKY, V. AND F. MULIER. 1998. *Learning from Data*. Wiley, 441 pp.

CLEVELAND, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Soc*. **74**: 829–836.

CYBENKO, G. 1989. Approximation by superpositions of a sigmoidal function. *Math. Contrib. Signals Syst.* **2**: 303–314.

HAMILTON, K. AND W.W. HSIEH. 2002. Representation of the QBO in the tropical stratospheric wind by non-linear principal component analysis. *J. Geophys. Res*. **107**(**D15**): 4232, DOI: 10.1029/2001JD001250.

HASTIE, T. AND W. STUETZLE. 1989. Principal curves. *J. Am. Stat. Soc.* **84**: 502–516.

HSIEH, W.W. 2001a. Non-linear principal component analysis by neural networks. *Tellus*, **53**: 599–615.

———. 2001b. Non-linear canonical correlation analysis of the tropical Pacific climate variability using a neural network approach. *J. Clim.* **14**: 2528–2539.

———. 2003. Non-linear multivariate and time series analysis by neural network methods. *Rev. Geophys.* (in press).

KIRBY, M.J. and R. MIRANDA. 1996. Circular nodes in neural networks. *Neural Comp*. **8**: 390–402.

KRAMER, M.A. 1991. Non-linear principal component analysis using autoassociative neural networks. *AIChE J.* **37**: 223–243.

MALTHOUSE, E.C. 1998. Limitations of non-linear PCA as performed with generic neural networks. *IEEE Trans. Neural Networks*, **9**: 165–173.

MONAHAN, A.H. 2001. Non-linear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure. *J. Clim.* **14**: 219–233.

———.; J.C. FYFE and G.M. FLATO. 2000. A regime view of northern hemisphere atmospheric variability and change under global warming. *Geophys. Res. Lett*. **27**: 1139–1142.

NAUJOKAT, B. 1986. An updat of the observed quasi-biennial ocillation of the stratospheric winds over the tropics. *J. Atmos. Sci.* **43**: 1873–1877.