NONLINEAR MULTIVARIATE AND TIME SERIES ANALYSIS BY NEURAL NETWORK METHODS

William W. Hsieh
Department of Earth and Ocean Sciences
University of British Columbia
Vancouver, British Columbia, Canada

Received 29 March 2002; revised 6 November 2003; accepted 12 November 2003; published 18 March 2004.

[1] Methods in multivariate statistical analysis are essential for working with large amounts of geophysical data, data from observational arrays, from satellites, or from numerical model output. In classical multivariate statistical analysis, there is a hierarchy of methods, starting with linear regression at the base, followed by principal component analysis (PCA) and finally canonical correlation analysis (CCA). A multivariate time series method, the singular spectrum analysis (SSA), has been a fruitful extension of the PCA technique. The common drawback of these classical methods is that only linear structures can be correctly extracted from the data. Since the late 1980s, neural network methods have become popular for performing nonlinear regression and classification. More recently, neural network methods have been extended to perform nonlinear PCA (NLPCA), nonlinear CCA (NLCCA), and nonlinear SSA (NLSSA). This paper presents a unified view of the NLPCA, NLCCA, and NLSSA techniques and their applications to various data sets of the atmosphere and the ocean (especially for the El Niño-Southern Oscillation and the stratospheric quasi-biennial oscillation). These data sets reveal that the linear methods are often too simplistic to describe real-world systems, with a tendency to scatter a single oscillatory phenomenon into numerous unphysical modes or higher harmonics, which can be largely alleviated in the new nonlinear paradigm. *INDEX TERMS:* 3299 Mathematical Geophysics: General or miscellaneous; 3394 Meteorology and Atmospheric Dynamics: Instruments and techniques; 4294 Oceanography: General: Instruments and techniques; 4522 Oceanography: Physical: El Niño; *KEYWORDS:* neural networks, principal component analysis, canonical correlation analysis, singular spectrum analysis, El Niño.

Citation: Hsieh, W. W. (2004), Nonlinear multivariate and time series analysis by neural network methods, *Rev. Geophys.*, *42*, RG1003, doi:10.1029/2002RG000112.

1. INTRODUCTION

[2] In a standard text on classical multivariate statistical analysis [e.g., *Mardia et al.*, 1979] the chapters typically proceed from linear regression to principal component analysis and then to canonical correlation analysis. In regression one tries to find how the response variable y is linearly affected by the predictor variables $\mathbf{x} \equiv [x_1, \ldots, x_l]$, i.e.,

$$y = \mathbf{r} \cdot \mathbf{x} + r_0 + \epsilon, \tag{1}$$

where ϵ is the error (or residual) and the regression coefficients $\mathbf{r}$ and r_0 are found by minimizing the mean of ϵ^2.

1.1. Principal Component Analysis

[3] However, in many data sets one cannot separate variables into predictor and response variables. For instance, one may have a data set of the monthly sea surface temperatures (SST) collected at 1000 grid locations over several decades; that is, the data set is of the form $\mathbf{x}(t) = [x_1, \ldots, x_l]$, where each variable x_i ($i = 1, \ldots, l$) has N samples labeled by the index t. Very often, t is simply the time, and each x_i is a time series containing N observations. Principal component analysis (PCA), also known as empirical orthogonal function (EOF) analysis, looks for u, a linear combination of the x_i, and an associated vector $\mathbf{a}$, with

$$u(t) = \mathbf{a} \cdot \mathbf{x}(t), \tag{2}$$

so that

$$\langle \|\mathbf{x}(t) - \mathbf{a}u(t)\|^2 \rangle$$

is minimized, where $\langle\ \rangle$ denotes a sample or time mean. Here u, called the first principal component (PC) (or score), is often a time series, while $\mathbf{a}$, called the first eigenvector (also called an EOF or loading), is the first eigenvector of the data covariance matrix, and $\mathbf{a}$ often describes a spatially standing oscillation pattern. Together u and $\mathbf{a}$ make up the first PCA mode. In essence, a given data set is approximated by a straight line (oriented in the direction of $\mathbf{a}$), which accounts for the maximum amount of variance in the data; pictorially, in a scatterplot of the data the straight line found by PCA passes through the "middle" of the data set. From the residual, $\mathbf{x} - \mathbf{a}u$, the second PCA mode can similarly be extracted and so on for the higher modes. In practice, the

8755-1209/04/2002RG000112$15.00
Paper number 2002RG000112

common algorithms for PCA extract all modes simultaneously [*Jolliffe*, 2002; *Preisendorfer*, 1988]. By retaining only the leading modes, PCA has been commonly used to reduce the dimensionality of the data set and to extract the main patterns from the data set. PCA has also been extended to the singular spectrum analysis (SSA) technique for time series analysis [*Elsner and Tsonis*, 1996; *von Storch and Zwiers*, 1999; *Ghil et al.*, 2002].

1.2. Canonical Correlation Analysis

[4] Next consider two data sets $\{x_i(t)\}$ and $\{y_j(t)\}$, each with N samples. We group the $\{x_i(t)\}$ variables to form the vector $\mathbf{x}(t)$ and the $\{y_j(t)\}$ variables to form the vector $\mathbf{y}(t)$. Canonical correlation analysis (CCA) [*Mardia et al.*, 1979; *Bretherton et al.*, 1992; *von Storch and Zwiers*, 1999] looks for linear combinations

$$u(t) = \mathbf{a} \cdot \mathbf{x}(t) \qquad v(t) = \mathbf{b} \cdot \mathbf{y}(t), \tag{3}$$

where the canonical variates u and v have maximum correlation; that is, the weight vectors $\mathbf{a}$ and $\mathbf{b}$ are chosen such that cor(u, v), the Pearson correlation coefficient between u and v, is maximized. For instance, if $\mathbf{x}(t)$ is the sea level pressure (SLP) field and $\mathbf{y}(t)$ is the SST field, then CCA can be used to extract the correlated spatial patterns $\mathbf{a}$ and $\mathbf{b}$ in the SLP and SST fields. Unlike regression, which tries to study how each y_j is related to the $\mathbf{x}$ variables, CCA examines how the entire $\mathbf{y}$ field is related to the $\mathbf{x}$ field. This holistic view has made CCA popular [*Barnett and Preisendorfer*, 1987; *Barnston and Ropelewski*, 1992; *Shabbar and Barnston*, 1996].

[5] In the environmental sciences, researchers have to work with large data sets from satellite images of the Earth's surface and global climate data to voluminous output from large numerical models. Multivariate techniques such as PCA and CCA have become indispensable in extracting essential information from these massive data sets [*von Storch and Zwiers*, 1999]. However, the restriction of finding only linear relations means that nonlinear relations are either missed or misinterpreted by these methods. The introduction of nonlinear multivariate and time series techniques is crucial to further advancement in the environmental sciences.

1.3. Feed Forward Neural Network Models

[6] The nonlinear neural network (NN) models originated from research trying to understand how the brain functions with its networks of interconnected neurons [*McCulloch and Pitts*, 1943]. There are many types of NN models; some are only of interest to neurological researchers, while others are general nonlinear data techniques. There are now many good textbooks on NN models [*Bishop*, 1995; *Rojas*, 1996; *Ripley*, 1996; *Cherkassky and Mulier*, 1998; *Haykin*, 1999].

[7] The most widely used NN models are the feed forward NNs, also called multilayer perceptrons [*Rumelhart et al.*, 1986], which perform nonlinear regression and classification. The basic architecture (Figure 1) consists of a layer of input neurons x_i (a "neuron" is simply a variable in NN jargon) linked to a layer or more of "hidden" neurons, which are, in turn, linked to a layer of output neurons y_j. In Figure 1, there is only one layer of hidden neurons h_k. A transfer function (an "activation" function in NN jargon) maps from the inputs to the hidden neurons. There is a variety of choices for the transfer function, the hyperbolic tangent function being a common one, i.e.,

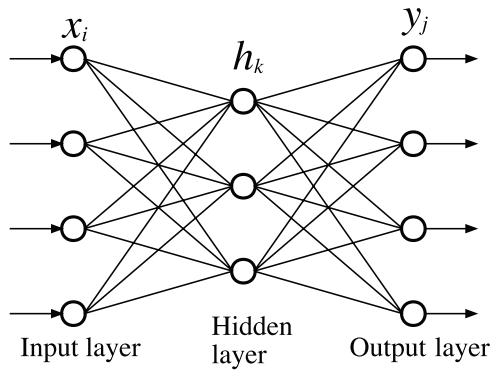

Figure 1. Schematic diagram of the feed forward neural network (NN) model, with one "hidden" layer of neurons (i.e., variables) (denoted by circles) sandwiched between the input layer and the output layer. In the feed forward NN model the information only flows forward starting from the input neurons. Increasing the number of hidden neurons increases the number of model parameters. Adapted from *Hsieh and Tang* [1998].

$$h_k = \tanh\left(\sum_i w_{ki} x_i + b_k\right), \tag{4}$$

where w_{ki} and b_k are the weight and bias parameters, respectively. The tanh(z) function is a sigmoidal-shaped function, where its two asymptotic values of ± 1 as $z \to \pm\infty$ can be viewed as representing the two states of a neuron (at rest or activated), depending on the strength of the excitation z. (If there is more than one hidden layer, then equations of the same form as equation (4) are used to calculate the values of the next layer of the hidden neurons from the current layer of neurons). When the feed forward NN is used for nonlinear regression, the output neurons y_j are usually calculated by a linear combination of the neurons in the preceding layer, i.e.,

$$y_j = \sum_k \tilde{w}_{jk} h_k + \tilde{b}_j. \tag{5}$$

[8] Given observed data y_{oj}, the optimal values for the weight and bias parameters (w_{ki}, $\tilde{w}_{jk}$, b_k, and $\tilde{b}_j$) are found by "training" the NN, i.e., performing a nonlinear optimization, where the cost function or objective function

$$J = \left\langle \sum_j \left(y_j - y_{oj}\right)^2 \right\rangle \tag{6}$$

is minimized, with J simply being the mean squared error (MSE) of the output. The NN has found a set of nonlinear regression relations $y_j = f_j(\mathbf{x})$. To approximate a set of

continuous functions f_j, only one layer of hidden neurons is enough, provided enough hidden neurons are used in that layer [*Hornik et al.*, 1989; *Cybenko*, 1989]. The NN with one hidden layer is commonly called a two-layer NN, as there are two layers of mapping (equations (4) and (5)) going from input to output; however, there are other conventions for counting the number of layers, and some authors refer to our two-layer NN as a three-layer NN since there are three layers of neurons.

1.4. Local Minima and Overfitting

[9] The main difficulty of the NN method is that the nonlinear optimization often encounters multiple local minima in the cost function. This means that starting from different initial guesses for the parameters, the optimization algorithm may converge to different local minima. Many approaches have been proposed to alleviate this problem [*Bishop*, 1995; *Hsieh and Tang*, 1998]; a common approach involves multiple optimization runs starting from different random initial parameters so that, hopefully, not all runs will be stranded at shallow local minima.

[10] Another pitfall with the NN method is overfitting, i.e., fitting to the noise in the data, because of the tremendous flexibility of the NN to fit the data. With enough hidden neurons the NN can fit the data, including the noise, to arbitrary accuracy. Thus, for a network with many parameters, reaching the global minimum may mean nothing more than finding a badly overfitted solution. Usually, only a portion of the data record is used to train (i.e., fit) the NN model; the other is reserved to validate the model. If too many hidden neurons are used, then the NN model fit to the training data will be excellent, but the model fit to the validation data will be poor, thereby allowing the researchers to gauge the appropriate number of hidden neurons. During the optimization process it is also common to monitor the MSE over the training data and over the validation data separately. As the number of iterations of the optimization algorithm increased, the MSE calculated over the training data would decrease; however, beyond a certain number of iterations the MSE over the validation data would begin to increase, indicating the start of overfitting and hence the appropriate time to stop the optimization process. Another approach to avoid overfitting is to add weight penalty terms to the cost function, as discussed in Appendix A. Yet another approach is to compute an ensemble of NN models starting from different random initial parameters. The mean of the ensemble of NN solutions tends to give a smoother solution than the individual NN solutions.

[11] If forecast skills are to be estimated, then another unused part of the data record will have to be reserved as independent test data for estimating the forecast skills, as the validation data have already been used to determine the model architecture. Some authors interchange the terminology for "validation" data and "test" data; the terminology here follows *Bishop* [1995]. For poor quality data sets (e.g., short, noisy data records) the problems of local minima and overfitting could render nonlinear NN methods incapable of offering any advantage over linear methods.

[12] The feed forward NN has been applied to a variety of nonlinear regression and classification problems in environmental sciences such as meteorology and oceanography and has been reviewed by *Gardner and Dorling* [1998] and *Hsieh and Tang* [1998]. Some examples of recent applications using NN include the following: tornado diagnosis [*Marzban*, 2000], efficient radiative transfer computation in atmospheric general circulation models [*Chevallier et al.*, 2000], multiparameter satellite retrievals from the Special Sensor Microwave/Imager [*Gemmill and Krasnopolsky*, 1999], wind retrieval from scatterometer [*Richaume et al.*, 2000], adaptive nonlinear model output statistics [*Yuval and Hsieh*, 2003], efficient computation of sea water density or salinity from a nonlinear equation of state [*Krasnopolsky et al.*, 2000], tropical Pacific sea surface temperature prediction [*Tang et al.*, 2000; *Yuval*, 2001], and an empirical atmosphere in a hybrid coupled atmosphere-ocean model of the tropical Pacific [*Tang and Hsieh*, 2002]. For NN applications in geophysics (seismic exploration, well log lithology determination, electromagnetic exploration, and earthquake seismology), see *Sandham and Leggett* [2003].

[13] To keep within the scope of a review paper, I have to omit reviewing the numerous fine papers on using NN for nonlinear regression and classification and focus on the topic of how the feed forward NN can be extended from its original role as nonlinear regression to nonlinear PCA (section 2), nonlinear CCA (section 3), and nonlinear SSA (section 4), illustrated by examples from the tropical Pacific atmosphere-ocean interactions and the equatorial stratospheric wind variations. These examples reveal various disadvantages of the linear methods; the most common one is the tendency to scatter a single oscillatory phenomenon into numerous modes or higher harmonics.

2. NONLINEAR PRINCIPAL COMPONENT ANALYSIS (NLPCA)

2.1. Open Curves

[14] As PCA finds a straight line which passes through the "middle" of the data cluster, the obvious next step is to generalize the straight line to a curve. *Kramer* [1991] proposed a neural network-based NLPCA model where the straight line is replaced by a continuous open curve for approximating the data.

[15] The fundamental difference between NLPCA and PCA is that PCA only allows a linear mapping (equation (2)) between $\mathbf{x}$ and the PC u, while NLPCA allows a nonlinear mapping. To perform NLPCA, the feed forward NN in Figure 2a contains three hidden layers of neurons between the input and output layers of variables. The NLPCA is basically a standard feed forward NN with four layers of transfer functions mapping from the inputs to the outputs. One can view the NLPCA network as composed of two standard two-layer feed forward NNs placed one after the other. The first two-layer network maps from the inputs $\mathbf{x}$ through a hidden layer to the bottleneck layer with only

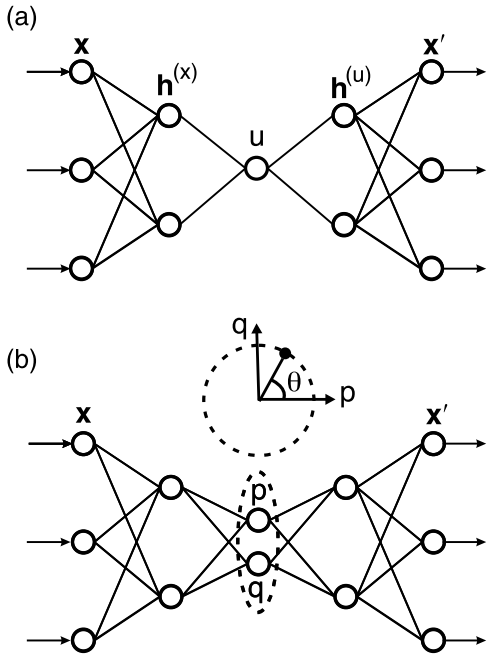

Figure 2. (a) Schematic diagram of the NN model for calculating the nonlinear principal component analysis (NLPCA). There are three layers of hidden neurons sandwiched between the input layer $\mathbf{x}$ on the left and the output layer $\mathbf{x}'$on the right. Next to the input layer is the encoding layer, followed by the "bottleneck" layer (with a single neuron u), which is then followed by the decoding layer. A nonlinear function maps from the higher-dimension input space to the one-dimensional bottleneck space, followed by an inverse transform mapping from the bottleneck space back to the original space represented by the outputs, which are to be as close to the inputs as possible by minimizing the cost function $J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle$. Data compression is achieved by the bottleneck, with the bottleneck neuron giving u, the nonlinear principal component (NLPC). (b) Schematic diagram of the NN model for calculating the NLPCA with a circular node at the bottleneck (NLPCA(cir)). Instead of having one bottleneck neuron u, there are now two neurons p and q constrained to lie on a unit circle in the p-q plane, so there is only one free angular variable θ, the NLPC. This network is suited for extracting a closed curve solution. From *Hsieh* [2001a], reprinted with permission from Blackwell Science, Oxford.

one neuron u, i.e., a nonlinear mapping $u = f(\mathbf{x})$. The next two-layer feed forward NN inversely maps from the nonlinear PC (NLPC) u back to the original higher-dimensional $\mathbf{x}$ space, with the objective that the outputs $\mathbf{x}' = \mathbf{g}(u)$ be as close as possible to the inputs $\mathbf{x}$ (thus the NN is said to be autoassociative). Note $\mathbf{g}(u)$ nonlinearly generates a curve in the $\mathbf{x}$ space and hence a one-dimensional (1-D) approximation of the original data. To minimize the MSE of this approximation, the cost function $J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle$ is minimized to solve for the weight and bias parameters of the NN. Squeezing the input information through a bottleneck layer with only one neuron accomplishes the dimensional reduction. Details of the NLPCA are given in Appendix A.

[16] In effect, the linear relation (2) in PCA is now generalized to $u = f(\mathbf{x})$, where f can be any nonlinear continuous function representable by a feed forward NN mapping from the input layer to the bottleneck layer; and instead of $\langle \|\mathbf{x}(t) - \mathbf{a}u(t)\|^2 \rangle$, $\langle \|\mathbf{x} - \mathbf{g}(u)\|^2 \rangle$ is minimized. The residual, $\mathbf{x} - \mathbf{g}(u)$, can be input into the same network to extract the second NLPCA mode and so on for the higher modes.

[17] That the classical PCA is indeed a linear version of this NLPCA can be readily seen by replacing all the transfer functions with the identity function, thereby removing the nonlinear modeling capability of the NLPCA. Then the forward map to u involves only a linear combination of the original variables as in the PCA.

[18] The NLPCA has been applied to the radiometric inversion of atmospheric profiles [*Del Frate and Schiavon*, 1999] and to the *Lorenz* [1963] three-component chaotic system [*Monahan*, 2000; *Hsieh*, 2001a]. For the tropical Pacific climate variability the NLPCA has been used to study the SST field [*Monahan*, 2001; *Hsieh*, 2001a] and the SLP field [*Monahan*, 2001]. The Northern Hemisphere atmospheric variability [*Monahan et al.*, 2000, 2001], the Canadian surface air temperature [*Wu et al.*, 2002], and the subsurface thermal structure of the Pacific Ocean [*Tang and Hsieh*, 2003] have also been investigated by the NLPCA.

[19] In the classical linear approach, there is a well-known dichotomy between PCA and rotated PCA (RPCA) [*Richman*, 1986]. In PCA the linear mode that accounts for the most variance of the data set is sought. However, as illustrated by *Preisendorfer* [1988, Figure 7.3], the resulting eigenvectors may not align close to local data clusters, so the eigenvectors may not represent actual physical states well. One application of RPCA methods is to rotate the PCA eigenvectors so they point closer to the local clusters of data points [*Preisendorfer*, 1988]. Thus the rotated eigenvectors may bear greater resemblance to actual physical states (though they account for less variance) than the unrotated eigenvectors; hence RPCA is also widely used [*Richman*, 1986; *von Storch and Zwiers*, 1999]. As there are many possible criteria for rotation, there are many RPCA schemes, among which the varimax [*Kaiser*, 1958] scheme is perhaps the most popular.

[20] The tropical Pacific climate system contains the famous interannual variability known as the El Niño-Southern Oscillation (ENSO), a coupled atmosphere-ocean interaction involving the oceanic phenomenon El Niño and the associated atmospheric phenomenon the Southern Oscillation. The coupled interaction results in anomalously warm SST in the eastern equatorial Pacific during El Niño episodes and cool SST in the central equatorial Pacific during La Niña episodes [*Philander*, 1990; *Diaz and Markgraf*, 2000]. ENSO is an irregular oscillation, but spectral analysis does reveal a broad spectral peak at the 4- to 5-year period. *Hsieh* [2001a] used the tropical Pacific SST data (1950–1999) to make a three-way comparison between NLPCA, RPCA, and PCA. The tropical Pacific SST anomaly (SSTA) data (i.e., the SST data with the climatological seasonal cycle removed) were prefiltered by

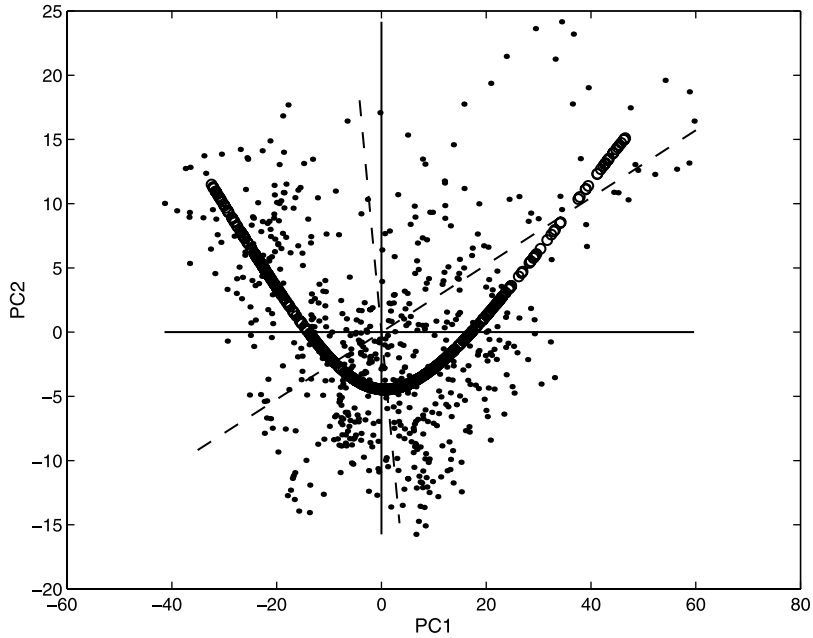

Figure 3. Scatterplot of the sea surface temperatures (SST) anomaly (SSTA) data (shown as dots) in the principal component (PC)1-PC2 plane, with the El Niño states lying in the top right corner and the La Niña states lying in the top left corner. The PC2 axis is stretched relative to the PC1 axis for better visualization. The first-mode NLPCA approximation to the data is shown by the (overlapping) small circles, which traced out a U-shaped curve. The first principal component analysis (PCA) eigenvector is oriented along the horizontal line, and the second PCA is oriented along the vertical line. The varimax method rotates the two PCA eigenvectors in a counterclockwise direction, as the rotated PCA (RPCA) eigenvectors are oriented along the dashed lines. (As the varimax method generates an orthogonal rotation, the angle between the two RPCA eigenvectors is 90° in the three-dimensional PC1-PC2-PC3 space). From *Hsieh* [2001a], reprinted with permission from Blackwell Science, Oxford.

PCA, with only the three leading modes retained. PCA modes 1, 2, and 3 accounted for 51.4%, 10.1%, and 7.2%, respectively, of the variance in the SSTA data. The first three PCs (PC1, PC2, and PC3) were used as the input **x** for the NLPCA network.

[21] The data are shown as dots in a scatterplot in the PC1-PC2 plane (Figure 3), where the cool La Niña states lie in the top left corner and the warm El Niño states lie in the top right corner. The NLPCA solution is a U-shaped curve linking the La Niña states at one end (low u) to the El Niño states at the other end (high u), similar to the solution found originally by *Monahan* [2001]. In contrast, the first PCA eigenvector lies along the horizontal line, and the second PCA lies along the vertical line (Figure 3), neither of which would come close to the El Niño states in the top right corner nor the La Niña states in the top left corner, thus demonstrating the inadequacy of PCA. For comparison, a varimax rotation [*Kaiser*, 1958; *Preisendorfer*, 1988] was applied to the first three PCA eigenvectors. (The varimax criterion can be applied to either the loadings or the PCs depending on one's objectives [*Richman*, 1986; *Preisendorfer*, 1988]; here it is applied to the PCs.) The resulting first RPCA eigenvector, shown as a dashed line in Figure 3, spears through the cluster of El Niño states in the top right corner, thereby yielding a more accurate description of the El Niño anomalies (Figure 4c) than the first PCA mode (Figure 4a), which did not fully represent the intense warming of Peruvian waters. The second RPCA eigenvector, also shown as a dashed line in Figure 3, did not improve much on the second PCA mode, with the PCA spatial pattern shown in Figure 4b and the RPCA pattern shown in Figure 4d. In terms of variance explained, the first NLPCA mode explained 56.6% of the variance versus 51.4% explained by the first PCA mode and 47.2% explained by the first RPCA mode.

[22] With the NLPCA, for a given value of the NLPC u, one can map from u to the three PCs. This is done by assigning the value u to the bottleneck neuron and mapping forward using the second half of the network in Figure 2a. Each of the three PCs can be multiplied by its associated PCA (spatial) eigenvector, and the three can be added together to yield the spatial pattern for that particular value of u. Unlike PCA, which gives the same spatial anomaly pattern except for changes in the amplitude as the PC varies, the NLPCA spatial pattern generally varies continuously as the NLPC changes. Figures 4e and 4f show the spatial anomaly patterns when u has its maximum value (corresponding to the strongest El Niño) and when u has its minimum value (strongest La Niña), respectively. Clearly, the asymmetry between El Niño and La Niña (i.e., the cool anomalies during La Niña episodes (Figure 4f) are observed to center much farther west of the warm anomalies during El Niño (Figure 4e) [*Hoerling et al.*, 1997]) is well captured by the first NLPCA mode; in contrast, the PCA mode 1

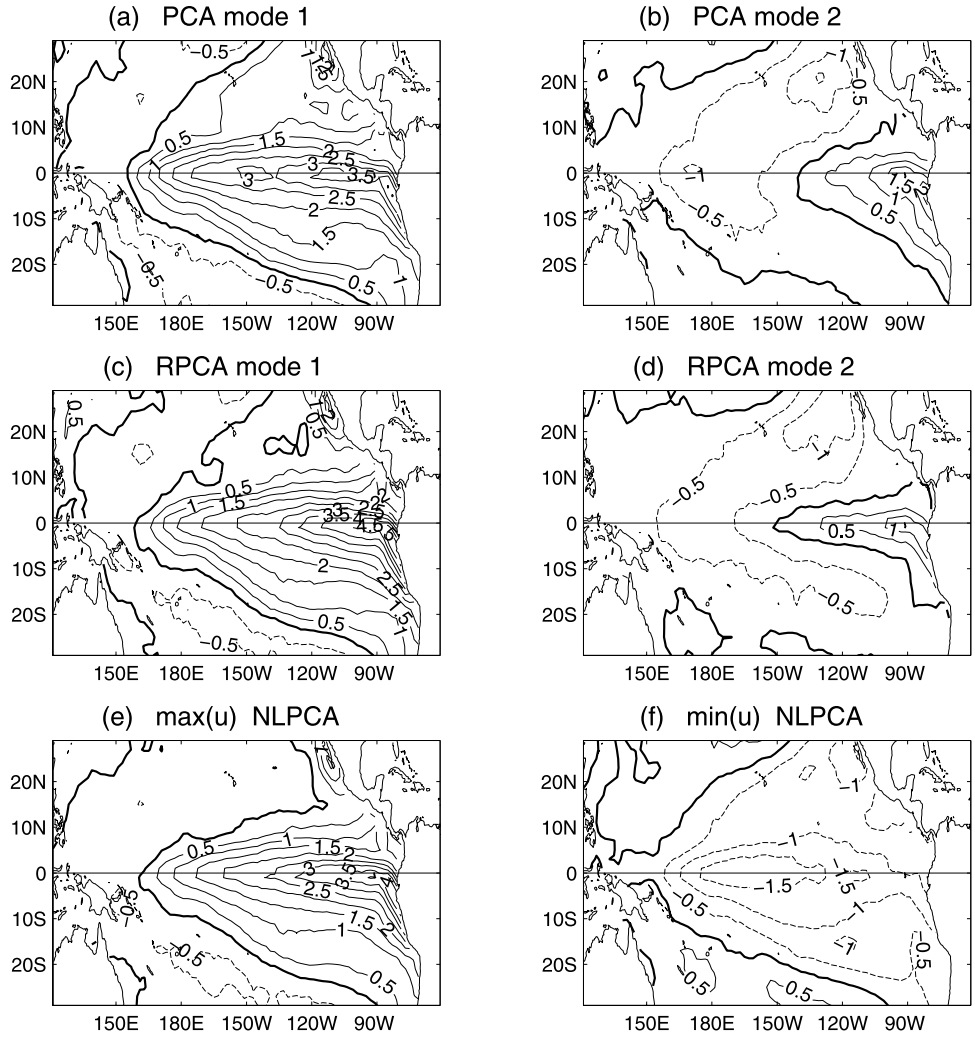

Figure 4. SSTA patterns (in °C) of the PCA, RPCA, and the NLPCA. (a) First and (b) second PCA spatial modes (both with their corresponding PCs at maximum value). (c) First and (d) second varimax RPCA spatial modes (both with their corresponding RPCs at maximum value). Anomaly pattern as the NLPC u of the first NLPCA mode varying from (e) maximum (strong El Niño) to (f) its minimum (strong La Niña). With a contour interval of 0.5°C the positive contours are shown as solid curves, negative contours are shown as dashed curves, and the zero contour is shown as a thick curve. Adapted from *Hsieh* [2001a], reprinted with permission from Blackwell Science, Oxford.

gives a La Niña that is simply the mirror image of the El Niño (Figure 4a). While El Niño has been known by Peruvian fishermen for many centuries because of its strong SSTA off the coast of Peru and its devastation of the Peruvian fishery, La Niña, with its weak manifestation in the Peruvian waters, was not appreciated until the last 2 decades of the twentieth century.

[23] In summary, PCA is used for two main purposes: (1) to reduce the dimensionality of the data set and (2) to extract features or recognize patterns from the data set. It is the second purpose where PCA can be improved upon. Both RPCA and NLPCA take the PCs from PCA as input. However, instead of multiplying the PCs by a fixed orthonormal rotational matrix, as performed in the varimax RPCA approach, NLPCA performs a nonlinear mapping of the PCs. RPCA sacrifices on the amount of variance explained, but by rotating the PCA eigenvectors, RPCA eigenvectors tend to point more toward local data clusters and are therefore more representative of physical states than the PCA eigenvectors.

[24] With a linear approach it is generally impossible to have a solution simultaneously (1) explaining maximum global variance of the data set and (2) approaching local data clusters hence the dichotomy between PCA and RPCA, with PCA aiming for objective 1 and RPCA aiming for objective 2. *Hsieh* [2001a] pointed out that with the more flexible NLPCA method both objectives 1 and 2 may be attained together; thus the nonlinearity in NLPCA unifies the PCA and RPCA approaches. It is easy to see why the dichotomy between PCA and RPCA in the linear approach automatically vanishes in the nonlinear approach. By increasing m, the number of hidden neurons in the encoding layer (and the decoding layer), the solution is capable of going through all local data clusters while maximizing the global variance explained. (In fact, for large enough m, NLPCA can pass through all data points,

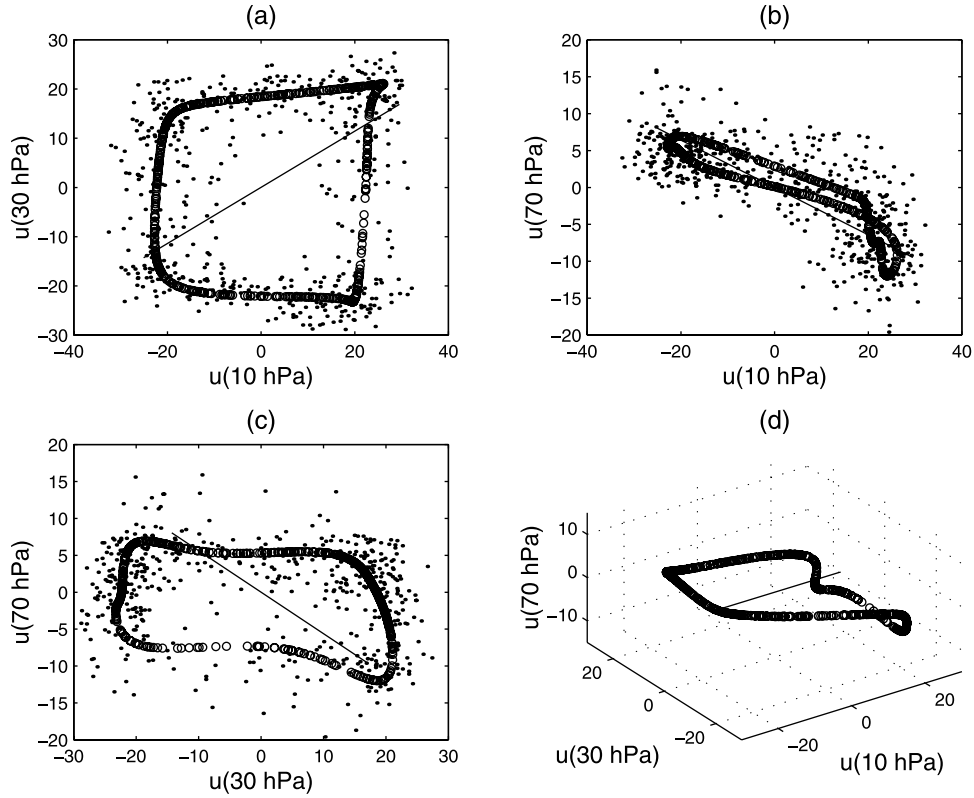

Figure 5. NLPCA(cir) mode 1 solution for the equatorial stratospheric zonal wind shown by the (overlapping) circles, with the data shown as dots. For comparison, the PCA mode 1 solution is shown as a thin straight line. Only three out of seven dimensions are shown, namely, u at the top, middle, and bottom levels (10, 30, and 70 hPa). (a)–(c) Two-dimensional views. (d) Three-dimensional view. From *Hamilton and Hsieh* [2002].

although this will in general give an undesirable, overfitted solution).

[25] The tropical Pacific SST example illustrates that with a complicated oscillation like the El Niño-La Niña phenomenon, using a linear method such as PCA results in the nonlinear mode being scattered into several linear modes. (In fact, all three leading PCA modes are related to this phenomenon.) This brings to mind the famous parable of the three blind men and their disparate descriptions of an elephant. Thus we see the importance of NLPCA as a unifier of the separate linear modes. In the study of climate variability the wide use of PCA methods has created the somewhat misleading view that our climate is dominated by a number of spatially fixed oscillatory patterns, which is, in fact, due to the limitation of the linear method. Applying NLPCA to the tropical Pacific SSTA, we found no spatially fixed oscillatory patterns but an oscillation evolving in space as well as in time.

2.2. Closed Curves

[26] The NLPCA is capable of finding a continuous open curve solution, but there are many geophysical phenomena involving waves or quasiperiodic fluctuations, which call for a continuous closed curve solution. *Kirby and Miranda* [1996] introduced a NLPCA with a circular node at the network bottleneck (henceforth referred to as the NLPCA (cir)), so that the NLPC as represented by the circular node is an angular variable θ, and the NLPCA(cir) is capable of approximating the data by a closed continuous curve. Figure 2b shows the NLPCA(cir) network, which is almost identical to the NLPCA of Figure 2a, except at the bottleneck, where there are now two neurons p and q constrained to lie on a unit circle in the p-q plane, so there is only one free angular variable θ, the NLPC. Details of the NLPCA(cir) are given in Appendix B.

[27] Applications of the NLPCA(cir) have been made to the tropical Pacific SST [*Hsieh*, 2001a] and to the equatorial stratospheric zonal wind (i.e., the east-west component of the wind) for the quasi-biennial oscillation (QBO) [*Hamilton and Hsieh*, 2002]. The QBO dominates over the annual cycle or other variations in the equatorial stratosphere, with the period of oscillation varying roughly between 22 and 32 months, with a mean of about 28 months. After the 45-year means were removed, the zonal wind u at seven vertical levels in the stratosphere became the seven inputs to the NLPCA(cir) network. The NLPCA(cir) mode 1 solution gives a closed curve in a seven-dimensional space. The system goes around the closed curve once, as the NLPC θ varies through one cycle of the QBO. Figure 5 shows the solution in three of the seven dimensions, namely, the wind anomalies at the 70-, 30-, and 10-hPa pressure levels (corresponding to elevations ranging roughly between 20 and 30 km above sea level). The NLPCA(cir) mode 1 explains 94.8% of the variance. For comparison, the linear

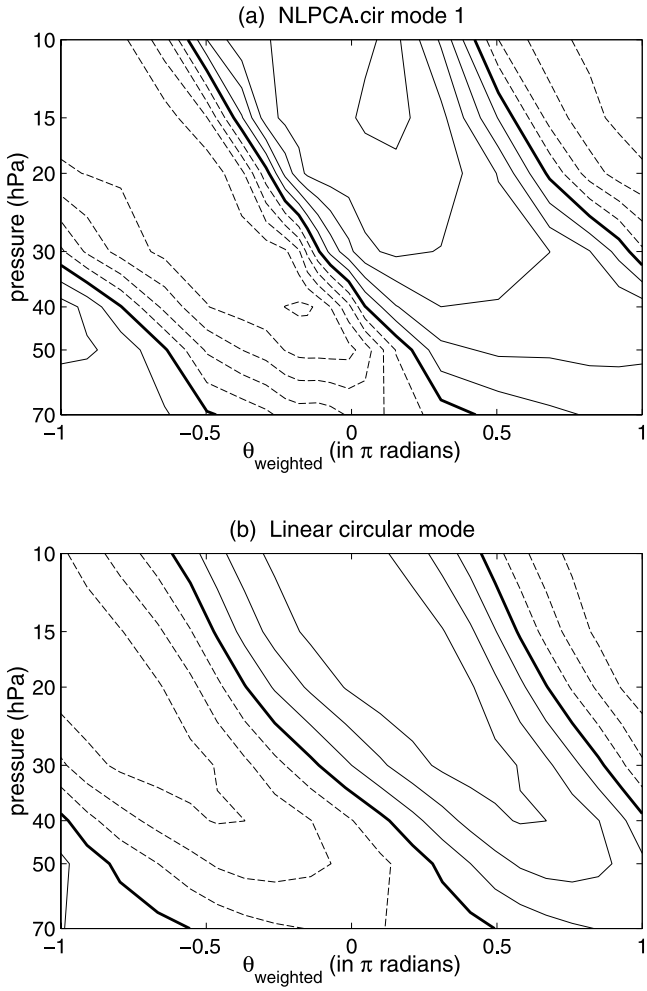

Figure 6. (a) Contour plot of the NLPCA(cir) mode 1 zonal wind anomalies as a function of pressure and $\theta_{weighted}$, where $\theta_{weighted}$ is θ weighted by the histogram distribution of θ. Thus $\theta_{weighted}$ is more representative of actual time during a cycle than θ. Contour interval is 5 m s^{-1}, with westerly winds indicated by solid lines, easterlies indicated by dashed lines, and zero contours indicated by thick lines. (b) Similar plot for a linear circular model of $\theta_{weighted}$. From *Hamilton and Hsieh* [2002].

PCA yields seven modes explaining 57.8, 35.4, 3.1, 2.1, 0.8, 0.5, and 0.3% of the variance, respectively. To compare with the NLPCA(cir) mode 1, *Hamilton and Hsieh* [2002] constructed a linear model of θ. In the plane spanned by PC1 and PC2 (each normalized by its standard deviation), an angle θ can be defined as the arctangent of the ratio of the two normalized PCA coefficients. This linear model accounts for 83.0% of the variance in the zonal wind, considerably less than the 94.8% accounted for by the NLPCA(cir) mode 1. The QBO as θ varies over one cycle is shown in Figure 6 for the NLPCA(cir) mode 1 and for the linear model. The observed strong asymmetries between the easterly and westerly phases of the QBO [*Hamilton*, 1998; *Baldwin et al.*, 2001] are captured by the nonlinear mode but not by the linear mode.

[28] The actual time series of the wind measured at a particular height level is somewhat noisy, and it is often desirable to have a smoother representation of the QBO time series which captures the essential features at all vertical levels. Also, the reversal of the wind from westerly to easterly and vice versa occurs at different times for different height levels, rendering it difficult to define the phase of the QBO. *Hamilton and Hsieh* [2002] found that the phase of the QBO as defined by the NLPC θ is more accurate than previous attempts to characterize the phase, leading to a stronger link between the QBO and Northern Hemisphere polar stratospheric temperatures in winter (the Holton-Tan effect) [*Holton and Tan*, 1980] than previously found.

2.3. Other Approaches (Principal Curves and Self-Organizing Maps)

[29] Besides the autoassociative NN, there have been several other approaches developed to generalize PCA [*Cherkassy and Mulier*, 1998]. The principal curve method [*Hastie and Stuetzle*, 1989; *Hastie et al.*, 2001] finds a nonlinear curve which passes through the middle of the data points. Developed originally in the statistics community, this method does not appear to have been applied to the environmental sciences or geophysics. There is a subtle but important difference between NLPCA (by autoassociative NN) and principal curves. In the principal curve approach each point in the data space is projected to a point on the principal curve, where the distance between the two is the shortest. In the NLPCA approach, while the mean square error (hence distance) between the data point and the projected point is minimized, it is only the mean which is minimized. There is no guarantee for an individual data point that it will be mapped to the closest point on the curve found by NLPCA. Hence, unlike the projection in principal curves, the projection used in NLPCA is suboptimal [*Malthouse*, 1998]. However, NLPCA has an advantage over the principal curve method in that its NN architecture provides a continuous (and differentiable) mapping function.

[30] *Newbigging et al.* [2003] used the principal curve projection concept to improve the NLPCA solution. *Malthouse* [1998] made a comparison between principal curves and the NLPCA model by autoassociative NN. Unfortunately, when testing a closed curve solution, he used NLPCA instead of NLPCA(cir) (which would have extracted the closed curve easily), thereby ending up with the conclusion that the NLPCA was not satisfactory for extracting the closed curve solution.

[31] Another popular NN method is the self-organizing map (SOM) [*Kohonen*, 1982, 2001], used widely for clustering. Since this approach fits a grid (usually a 1-D or 2-D grid) to a data set, it can be thought of as a discrete version of nonlinear PCA [*Cherkassky and Mulier*, 1998]. SOM has been applied to the clustering of winter daily precipitation data [*Cavazos*, 1999], to satellite ocean color classification [*Yacoub et al.*, 2001], and to high-dimensional hyperspectral Airborne Visible/Infrared Imaging Spectrometer data to classify the geology of the land surface [*Villmann et al.*, 2003]. For seismic data, SOM has been used to identify and classify multiple events [*Essenreiter et al.*, 2001] and used in well log calibration [*Taner et al.*, 2001].

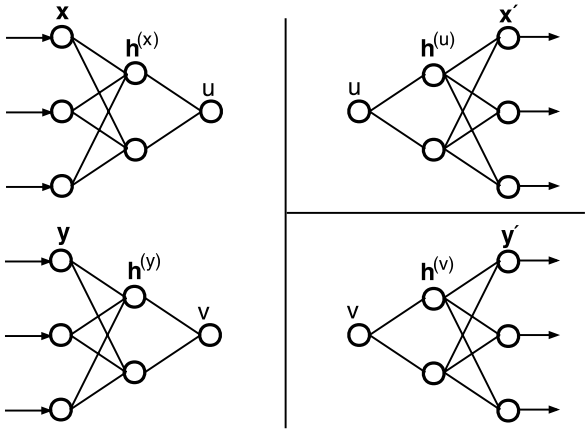

Figure 7. Three feed forward NNs used to perform nonlinear canonical correlation analysis (NLCCA). The double-barreled NN on the left maps from the inputs **x** and **y** to the canonical variates u and v, respectively. The cost function J forces the correlation between u and v to be maximized. On the right side the top NN maps from u to the output layer $\mathbf{x}'$. The cost function J_1 basically minimizes the mean-square error (MSE) of $\mathbf{x}'$ relative to **x**. The third NN maps from v to the output layer $\mathbf{y}'$. The cost function J_2 basically minimizes the MSE of $\mathbf{y}'$ relative to **y**. From *Hsieh* [2001b], reprinted with permission from the American Meteorological Society.

[32] Another way to generalize PCA is via independent component analysis (ICA) [*Comon*, 1994; *Hyvärinen et al.*, 2001], which was developed from information theory and has been applied to study the tropical Pacific SST variability by *Aires et al.* [2000]. Since ICA uses higher-order statistics (e.g., kurtosis, which is very sensitive to outliers), it may not be robust enough for the noisy data sets commonly encountered in climate or seismic studies (T. Ulrych, personal communication, 2003).

3. NONLINEAR CANONICAL CORRELATION ANALYSIS (NLCCA)

[33] While many techniques have been developed for nonlinearly generalizing PCA, there has been much less activity in developing nonlinear CCA. A number of different approaches have recently been proposed to nonlinearly generalize CCA [*Lai and Fyfe*, 1999, 2000; *Hsieh*, 2000; *Melzer et al.*, 2003]. *Hsieh* [2000] proposed using three feed forward NNs to accomplish NLCCA, where the linear mappings in equation (3) for the CCA are replaced by nonlinear mapping functions using two-layer feed forward NNs. The mappings from **x** to u and **y** to v are represented by the double-barreled NN on the left-hand side of Figure 7. By minimizing the cost function $J = -\mathrm{cor}(u, v)$, one finds the parameters that maximize the correlation $\mathrm{cor}(u, v)$. After the forward mapping with the double-barreled NN has been solved, inverse mappings from the canonical variates u and v to the original variables, as represented by the two standard feed forward NNs on the right side of Figure 7, are to be solved, where the MSE of their outputs $\mathbf{x}'$ and $\mathbf{y}'$ are minimized with respect to **x** and **y**, respectively. For details, see Appendix C.

[34] Consider the following test problem from *Hsieh* [2000]. Let

$$X_1 = t - 0.3t^2, \quad X_2 = t + 0.3t^3, \quad X_3 = t^2, \tag{7}$$

$$Y_1 = \tilde{t}^3, \quad Y_2 = -\tilde{t} + 0.3\tilde{t}^3, \quad Y_3 = \tilde{t} + 0.3\tilde{t}^2, \tag{8}$$

where t and $\tilde{t}$ are uniformly distributed random numbers in $[-1, 1]$. Also let

$$X_1' = -s - 0.3s^2, \quad X_2' = s - 0.3s^3, \quad X_3' = -s^4, \tag{9}$$

$$Y_1' = \mathrm{sech}(4s), \quad Y_2' = s + 0.3s^3, \quad Y_3' = s - 0.3s^2, \tag{10}$$

where s is a uniformly distributed random number in $[-1, 1]$. The shapes described by the $\mathbf{X}$ and $\mathbf{X}'$ vector functions are displayed in Figure 8, and those described by $\mathbf{Y}$ and $\mathbf{Y}'$ are displayed in Figure 9. To lowest order, equation (7) for $\mathbf{X}$ describes a quadratic curve, and equation (9) for $\mathbf{X}'$ describes a quartic. Similarly, to lowest order, $\mathbf{Y}$ is a cubic, and $\mathbf{Y}'$ is a hyperbolic secant. The signal in the test data was produced by adding the second mode ($\mathbf{X}'$, $\mathbf{Y}'$) to the first mode ($\mathbf{X}$, $\mathbf{Y}$), with the variance of the second mode being 1/3 that of the first mode. A small amount of Gaussian random noise, with standard deviation equal to 10% of the signal standard deviation, was also added to the data set. The data set of $N = 500$ points was then standardized (i.e., each variable with mean removed was normalized by the standard deviation). Note that different sequences of random numbers t_n and $\tilde{t}_n$ ($n = 1, \ldots, N$) were used to generate the first modes $\mathbf{X}$ and $\mathbf{Y}$, respectively. Hence these two dominant modes in the **x** space and the **y** space are unrelated. In contrast, as $\mathbf{X}'$ and $\mathbf{Y}'$ were generated from the same sequence of random numbers s_n, they are strongly related. The NLCCA was applied to the data, and the first NLCCA mode retrieved (Figures 10 and 11) resembles the expected theoretical mode ($\mathbf{X}'$, $\mathbf{Y}'$). This is quite remarkable considering that $\mathbf{X}'$ and $\mathbf{Y}'$ have only 1/3 the variance of $\mathbf{X}$ and $\mathbf{Y}$; that is, the NLCCA ignores the large variance of $\mathbf{X}$ and $\mathbf{Y}$ and succeeded in detecting the nonlinear correlated mode ($\mathbf{X}'$, $\mathbf{Y}'$). In contrast, if the NLPCA is applied to **x** and **y** separately, then the first NLPCA mode retrieved from **x** will be $\mathbf{X}$, and the first mode from **y** will be $\mathbf{Y}$. This illustrates the essential difference between NLPCA and NLCCA.

[35] The NLCCA has been applied to analyze the tropical Pacific sea level pressure anomaly (SLPA) and SSTA fields [*Hsieh*, 2001b], where the six leading PCs of the SLPA and the six PCs of the SSTA during 1950–2000 were inputs to an NLCCA model. The first NLCCA mode is plotted in the PC spaces of the SLPA and the SSTA (Figure 12), where only the three leading PCs are shown. For the SLPA (Figure 12a) in the PC1-PC2 plane the La Niña states are in the left corner (corresponding to low u values), while the El Niño states are in the top right corner (high u values). The CCA solutions are shown as thin straight lines. For the

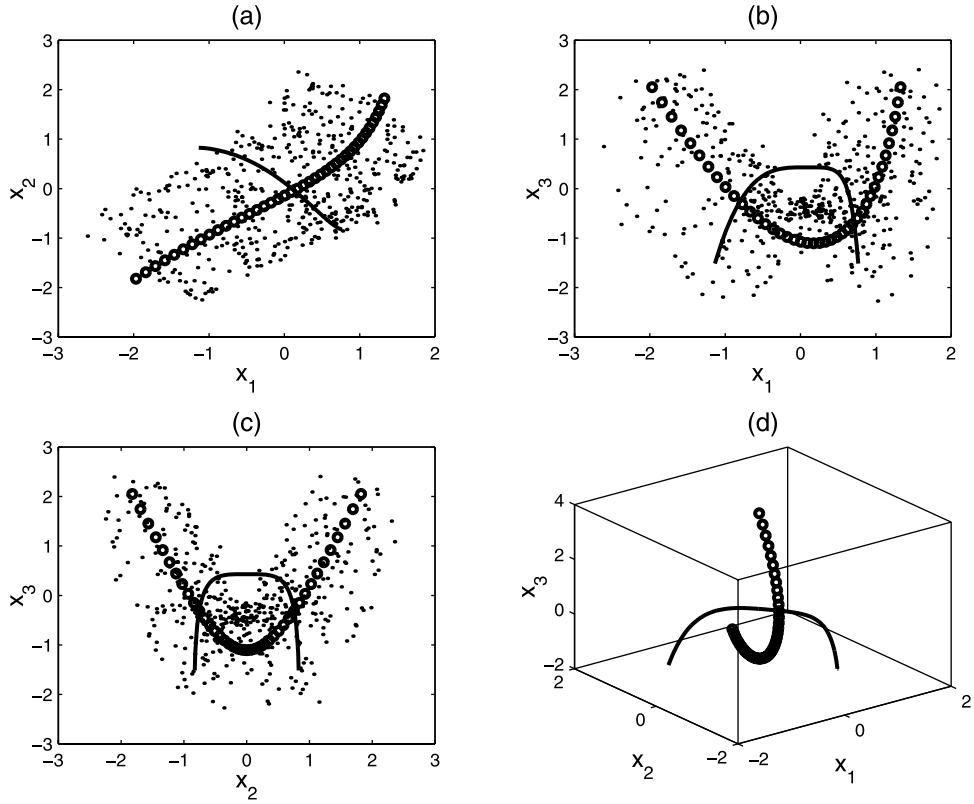

Figure 8. First theoretical mode **X** generated from equation (7) shown by curve made up of small circles. The solid curve shows the second mode $\mathbf{X}'$ generated from equation (9). A projection in the (a) x_1-x_2, (b) x_1-x_3, and (c) x_2-x_3 planes and a (d) three-dimensional plot. The actual data set of 500 points (shown by dots) was generated by adding mode 2 to mode 1 (with mode 2 having 1/3 the variance of mode 1) and adding a small amount of Gaussian noise. Follows *Hsieh* [2000], with permission from Elsevier Science.

SSTA (Figure 12b) in the PC1-PC2 plane the first NLCCA mode is a U-shaped curve linking the La Niña states in the top left corner (low v values) to the El Niño states in the top right corner (high v values). In general, the nonlinearity is greater in the SSTA than in the SLPA, as the difference between the CCA mode and the NLCCA mode is greater in Figure 12b than in Figure 12a.

[36] The MSE of the NLCCA divided by the MSE of the CCA is a useful measure on how different the nonlinear solution is relative to the linear solution; a smaller ratio means greater nonlinearity, while a ratio of 1 means the NLCCA can only find a linear solution. This ratio is 0.951 for the SLPA and 0.935 for the SSTA, confirming that the mapping for the SSTA was more nonlinear than that for the SLPA. When the data record was divided into two halves (1950–1975 and 1976–1999) to be separately analyzed by the NLCCA, *Hsieh* [2001b] found that this ratio decreased for the second half, implying an increase in the nonlinearity of ENSO during the more recent period.

[37] For the NLCCA mode 1, as u varies from its minimum value to its maximum value, the SLPA field varies from the strong La Niña phase to the strong El Niño phase (Figure 13). The zero contour is farther west during La Niña (Figure 13a) than during strong El Niño (Figure 13b). Similarly, as v varies from its minimum to its maximum, the SSTA field varies from strong La Niña to strong El Niño (Figure 14), revealing that the SST anomalies during La Niña are centered farther west of the anomalies during El Niño.

[38] *Wu and Hsieh* [2002, 2003] studied the relation between the tropical Pacific wind stress anomaly (WSA) and SSTA fields using the NLCCA. *Wu and Hsieh* [2003] found notable interdecadal changes of ENSO behavior before and after the mid-1970s climate regime shift, with greater nonlinearity found during 1981–1999 than during 1961–1975. Spatial asymmetry (for both SSTA and WSA) between El Niño and La Niña episodes was significantly enhanced in the later period. During 1981–1999 the location of the equatorial easterly WSA in the NLCCA solution during La Niña was unchanged from the earlier period, but during El Niño the westerly WSA was shifted eastward by up to 30°. From dynamical considerations based on the delay oscillator theory for ENSO (where the farther east the location of the WSA, the longer is the duration of the resulting SSTA in the eastern equatorial Pacific), *Wu and Hsieh* [2003] concluded that this interdecadal change would lengthen the duration of the ENSO warm phase but leave the duration of the cool phase unchanged, which was confirmed with numerical model experiments. This is an example of a nonlinear data analysis detecting a feature missed by previous studies using linear techniques, which, in turn, leads to new dynamical insight.

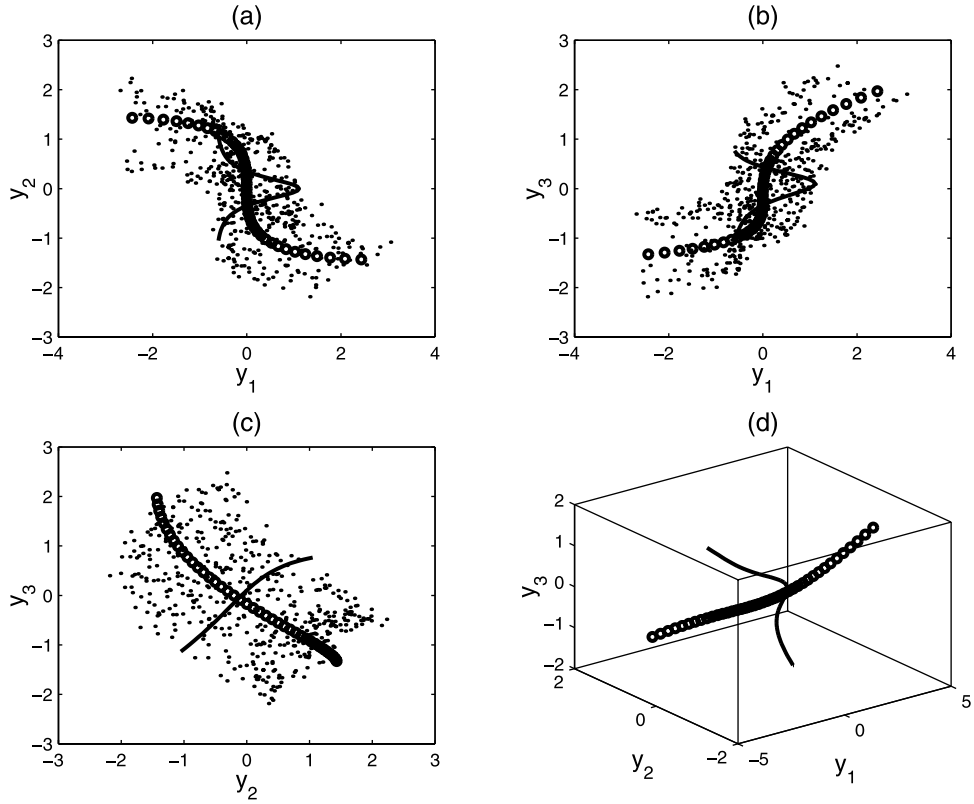

Figure 9. First theoretical mode **Y** generated from equation (8) shown by curve made up of small circles. The solid curve shows the second mode $\mathbf{Y}'$ generated from equation (10). A projection in the (a) y_1-y_2 plane, (b) y_1-y_3 plane, and (c) y_2-y_3 plane and a (d) three-dimensional plot. The data set of 500 points was generated by adding mode 2 to mode 1 (with mode 2 having 1/3 the variance of mode 1) and adding a small amount of Gaussian noise. Follows *Hsieh* [2000], with permission from Elsevier Science.

[39] The NLCCA has also been applied to study the relation between the tropical Pacific SSTA and the Northern Hemisphere midlatitude winter atmospheric variability (500 mbar geopotential height and North American surface air temperature) simulated in an atmospheric general circulation model (GCM), demonstrating the value of NLCCA as a nonlinear diagnostic tool for GCMs [*Wu et al.*, 2003].

4. NONLINEAR SINGULAR SPECTRUM ANALYSIS (NLSSA)

[40] By the 1980s, interest in chaos theory and dynamical systems led to further extension of the PCA method to singular spectrum analysis [*Elsner and Tsonis*, 1996; *Golyandina et al.*, 2001; *Ghil et al.*, 2002]. Given a time series $y_j = y(t_j)$ ($j = 1, \ldots, N$), lagged copies of the time series are stacked to form the augmented matrix **Y**,

$$\mathbf{Y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{N-L+1} \\ y_2 & y_3 & \cdots & y_{N-L+2} \\ \vdots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & \cdots & y_N \end{bmatrix}. \tag{11}$$

This matrix has the same form as the data matrix produced by L variables, each being a time series of length $n = N - L + 1$. **Y** can also be viewed as composed of its column vectors $\mathbf{y}_j$, forming a vector time series $\mathbf{y}(t_j)$ ($j = 1, \ldots, n$). The standard PCA can be performed on the augmented data matrix **Y**, resulting in

$$\mathbf{y}(t_j) = \sum_i x_i(t_j)\mathbf{e}_i, \tag{12}$$

where x_i is the ith PC, a time series of length n, and $\mathbf{e}_i$ is the ith eigenvector (or loading vector) of length L. Together x_i and $\mathbf{e}_i$ represent the ith SSA mode. This resulting method is the SSA with window L.

[41] In the multivariate case, with M variables $y_k(t_j) \equiv y_{kj}$ ($k = 1, \ldots, M$; $j = 1, \ldots, N$), the augmented matrix can be formed by letting

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1,N-L+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_{M1} & y_{M2} & \cdots & y_{M,N-L+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_{1L} & y_{1,L+1} & \cdots & y_{1N} \\ \vdots & \vdots & \vdots & \vdots \\ y_{ML} & y_{M,L+1} & \cdots & y_{MN} \end{bmatrix}. \tag{13}$$

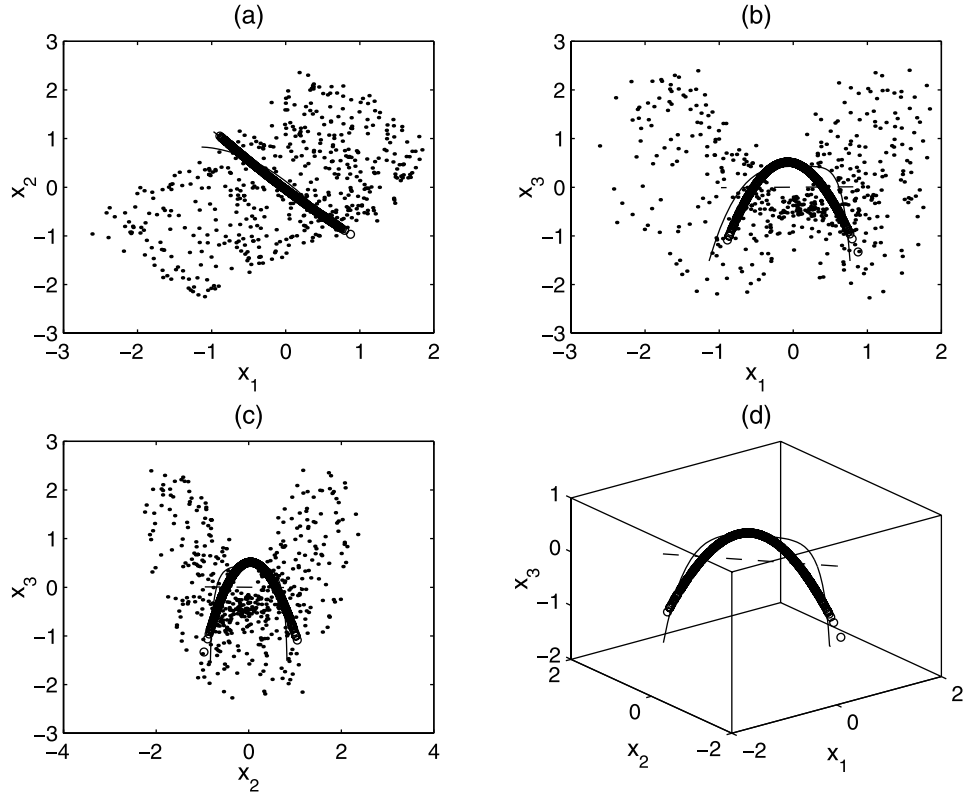

Figure 10. Nonlinear canonical correlation analysis (NLCCA) mode 1 in **x** space shown as a string of (densely overlapping) small circles. The theoretical mode $\mathbf{X}'$ is shown as a thin solid curve, and the linear canonical correlation analysis (CCA) mode is shown as a thin dashed line. The dots display the 500 data points. The number of hidden neurons (see Appendix C) used is $l_2 = m_2 = 3$. Follows *Hsieh* [2000], with permission from Elsevier Science.

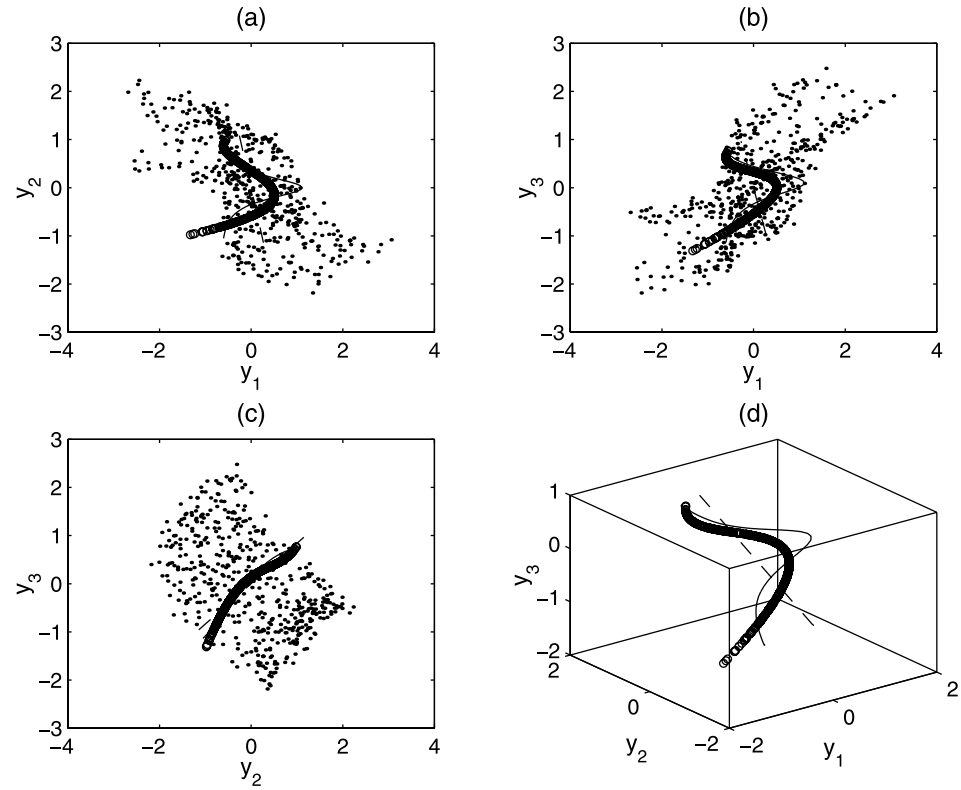

Figure 11. NLCCA mode 1 in **y** space shown as a string of overlapping small circles. The thin solid curve is the theoretical mode $\mathbf{Y}'$, and the thin dashed line is the CCA mode. Follows *Hsieh* [2000], with permission from Elsevier Science.

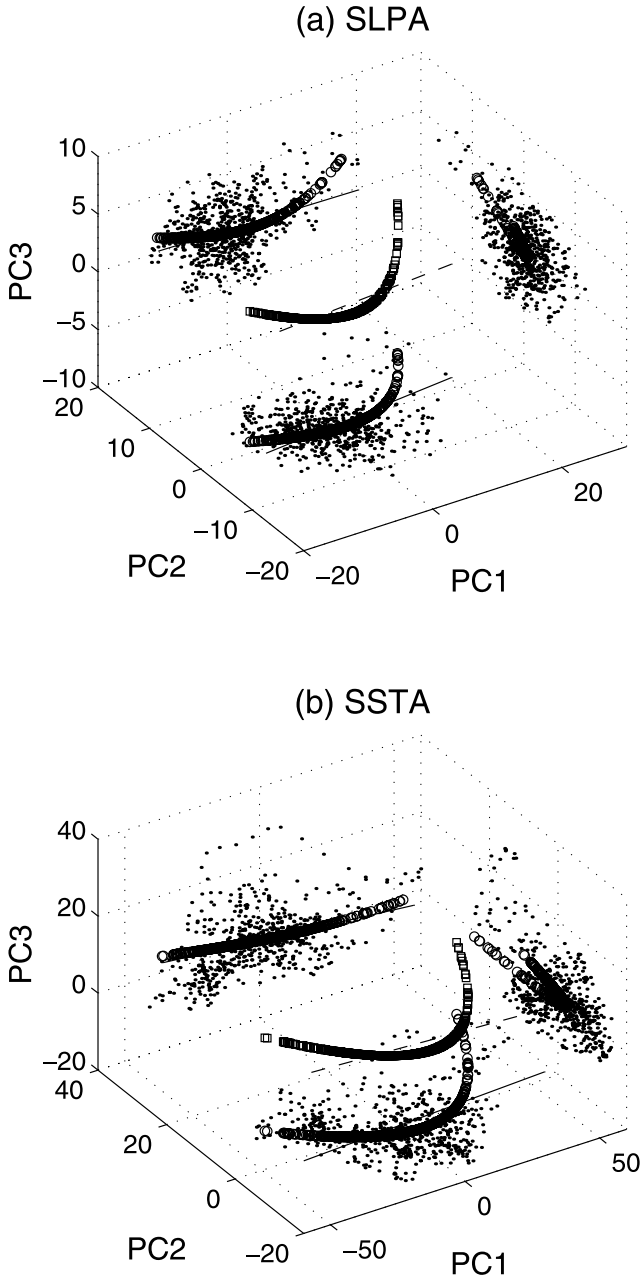

Figure 12. NLCCA mode 1 between the tropical Pacific (a) sea level pressure anomaly (SLPA) and (b) SSTA, plotted as (overlapping) squares in the PC_1-PC_2-PC_3 three-dimensional (3-D) space. The linear (CCA) mode is shown as a dashed line. The NLCCA mode and the CCA mode are also projected onto the PC_1-PC_2 plane, the PC_1-PC_3 plane, and the PC_2-PC_3 plane, where the projected NLCCA is indicated by (overlapping) circles, and the CCA is indicated by thin solid lines, and the projected data points (during 1950–2000) are shown by the scattered dots. There is no time lag between the SLPA and the corresponding SSTA data. The NLCCA solution was obtained with the number of hidden neurons $l_2 = m_2 = 2$; with $l_2 = m_2 = 1$, only a linear solution can be found. Adapted from *Hsieh* [2001b].

PCA can again be applied to **Y** to get the SSA modes, resulting in the multichannel SSA (MSSA) method, also called the space-time PCA method or the extended EOF (EEOF) method (though in typical EEOF applications, only a small number of lags are used). For brevity, we will use the term SSA to denote both SSA and MSSA. Commonly used in the meteorological and oceanographic communities [*Ghil et al.*, 2002], SSA has also been used to analyze solar activity [*Watari*, 1996; *Rangarajan and Barreto*, 2000] and storms on Mars [*Hollingsworth et al.*, 1997].

[42] *Hsieh and Wu* [2002] proposed the NLSSA method: Assume SSA has been applied to the data set, and after discarding the higher modes, we have retained the leading PCs, $\mathbf{x}(t) = [x_1, \ldots, x_l]$, where each variable x_i ($i = 1, \ldots, l$) is a time series of length n. The variables $\mathbf{x}$ are the inputs to the NLPCA(cir) network (Figure 2b). The NLPCA(cir), with its ability to extract closed curve solutions, is particularly ideal for extracting periodic or wave modes in the data. In SSA it is common to encounter periodic modes, each of which has to be split into a pair of SSA modes [*Elsner and Tsonis*, 1996], as the underlying PCA technique is not capable of modeling a periodic mode (a closed curve) by a single mode (a straight line). Thus two (or more) SSA modes can easily be combined by NLPCA(cir) into one NLSSA mode, taking the shape of a closed curve. When implementing NLPCA(cir), *Hsieh* [2001a] found that there were two possible configurations, a restricted configuration

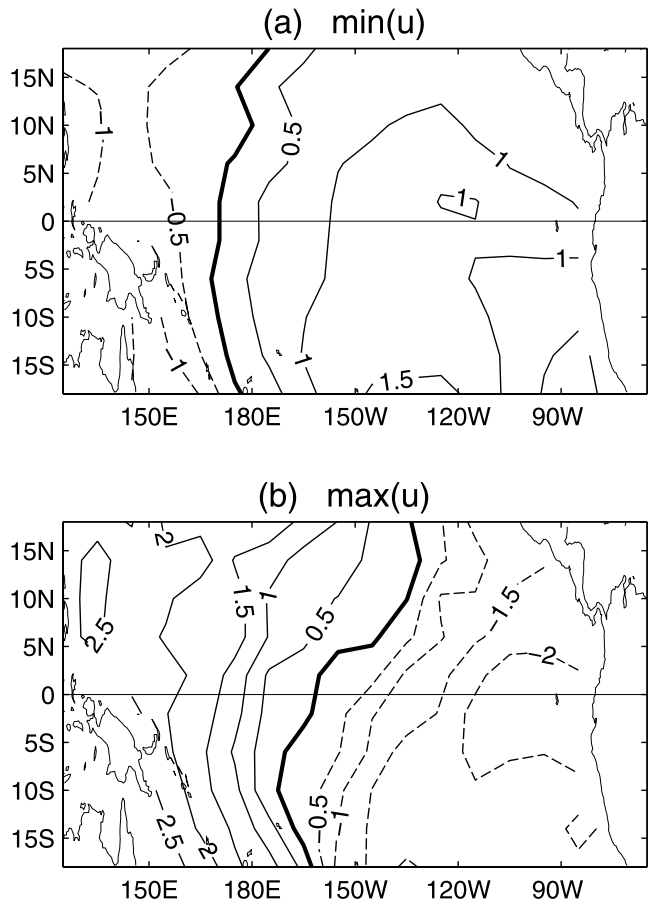

Figure 13. SLPA field when the canonical variate u of the NLCCA mode 1 is at (a) its minimum (strong La Niña) and (b) its maximum (strong El Niño). Contour interval is 0.5 mbar. Reprinted from *Hsieh* [2001b], with permission from the American Meteorological Society.

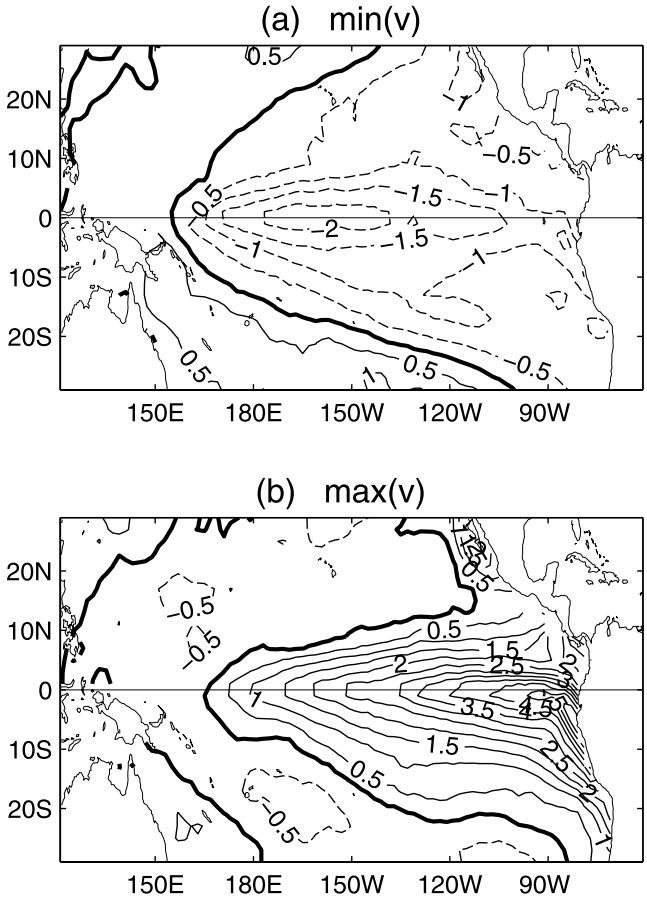

Figure 14. SSTA field when the canonical variate v is at (a) its minimum (strong La Niña) and (b) its maximum (strong El Niño). Contour interval is 0.5°C. Reprinted from *Hsieh* [2001b], with permission from the American Meteorological Society.

and a general configuration (see Appendix B). We will use the general configuration here. After the first NLSSA mode has been extracted, it can be subtracted from **x** to get the residual, which can be input again into the same network to extract the second NLSSA mode, and so forth for the higher modes.

[43] To illustrate the difference between the NLSSA and the SSA, consider a test problem with a nonsinusoidal wave of the form

$$f(t) = \begin{cases} 3 & t = 1, \ldots, 7 \\ -1 & t = 8, \ldots, 28 \end{cases} \tag{14}$$

and periodic thereafter. This is a square wave with the peak stretched to be 3 times as tall but only 1/3 as broad as the trough; it has a period of 28. Gaussian noise with twice the standard deviation as this signal was added, and the time series was normalized to unit standard deviation (Figure 15). The time series has 600 data points.

[44] SSA with window $L = 50$ was applied to this time series, with the first eight eigenvectors shown in Figure 16. The first eight modes individually accounted for 6.3, 5.6, 4.6, 4.3, 3.3, 3.3, 3.2, and 3.1% of the variance of the augmented time series **y**. The leading pair of modes displays oscillations of period 28, while the next pair manifests oscillations at a period of 14, i.e., the first harmonic. The nonsinusoidal nature of the SSA eigenvectors can be seen in mode 2 (Figure 16), where the trough is broader and shallower than the peak but nowhere as intense as in the original stretched square wave signal. The PCs for modes 1–4 are also shown in Figure 17. Both the eigenvectors (Figure 16) and the PCs (Figure 17) tend to appear in pairs, each member of the pair having similar appearance except for the quadrature phase difference.

[45] The first eight PC time series were served as inputs to the NLPCA(cir) network, with m (the number of hidden neurons in the encoding layer) ranging from 2 to 8 (and the weight penalty parameter $P = 1$, see Appendix B). The MSE dropped with increasing m, until $m = 5$, beyond which the MSE showed no further improvement. The resulting NLSSA mode 1 (with $m = 5$) is shown in Figure 18. Not surprisingly, the PC1 and PC2 are united by the approximately circular curve. What is more surprising are the Lissajous-like curves found in the PC1-PC3 plane (Figure 18b) and in the PC1-PC4 plane (Figure 18c), indicating relations between the first SSA mode and the higher modes 3 and 4. (It is well known that for two sinusoidal time series $z_1(t)$ and $z_2(t)$ oscillating at frequencies ω_1 and ω_2, respectively, a plot of the trajectory in the z_1-z_2 plane reveals a closed Lissajous curve if and only if ω_2/ω_1 is a rational number.) There was no relation found between PC1 and PC5, as PC5 appeared independent of PC1 (Figure 18d). However, with less noise in the input, relations can be found between PC1 and PC5 and even higher PCs.

[46] The NLSSA reconstructed component 1 (NLRC1) is the approximation of the original time series by the NLSSA mode 1. The neural network output $\mathbf{x}'$ are the NLSSA mode 1 approximation for the eight leading PCs. Multiplying these approximated PCs by their corresponding SSA eigenvectors and summing over the eight modes allows the reconstruction of the time series from the NLSSA mode 1. As each eigenvector contains the loading over a range of lags, each value in the reconstructed time series at time t_j also involves averaging over the contributions at t_j from the various lags.

[47] In Figure 15, NLRC1 (curve f) from NLSSA is to be compared with the reconstructed component (RC) from SSA mode 1 (RC1) (curve c). The nonsinusoidal nature of the oscillations is not revealed by the RC1 but is clearly manifested in the NLRC1, where each strong narrow peak is followed by a weak broad trough, similar to the original stretched square wave. Also, the wave amplitude is more steady in the NLRC1 than in the RC1. Using contributions from the first two SSA modes, RC1-2 (not shown) is rather similar to RC1 in appearance except for a larger amplitude.

[48] In Figure 15, curves d and e show the RC from SSA using the first three modes and the first eight modes, respectively. These curves, referred to as RC1-3 and RC1-8, respectively, show increasing noise as more modes are used. Among the RCs, with respect to the stretched square

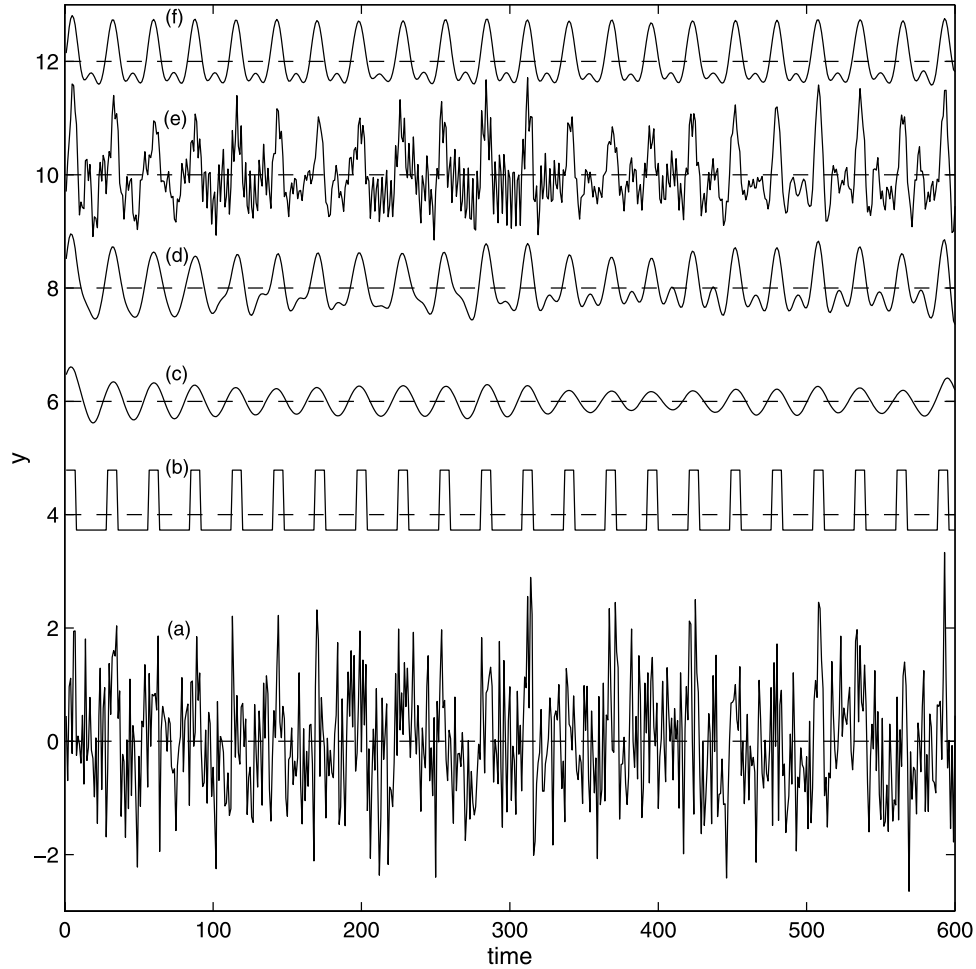

Figure 15. Noisy time series y containing a stretched square wave signal shown by curve a. Curve b shows the stretched square wave signal, which we will try to extract from the noisy time series. Curves c, d, and e are the reconstructed components (RC) from SSA leading modes, using one, three, and eight modes, respectively. Curve f is the NLSSA mode 1 RC (NLRC1). The dashed lines show the means of the various curves, which have been vertically shifted for better visualization.

wave time series (curve b), RC1-3 attains the most favorable correlation (0.849) and root-mean-square error (RMSE) (0.245) but remains behind the NLRC1, with correlation (0.875) and RMSE (0.225).

[49] The stretched square wave signal accounted for only 22.6% of the variance in the noisy data. For comparison, NLRC1 accounted for 17.9%, RC1 accounted for 9.4%, and RC1–2 accounted for 14.1% of the variance. With more modes the RCs account for increasingly more variance, but beyond RC1-3 the increased variance is only from fitting to the noise in the data.

[50] When classical Fourier spectral analysis was performed, the most energetic bands were the sine and cosine at a period of 14, the two together accounting for 7.0% of the variance. In this case the strong scattering of energy to higher harmonics by the Fourier technique has actually assigned 38% more energy to the first harmonic (at period 14) than to the fundamental period of 28. Next the data record was slightly shortened from 600 to 588 points, so the data record is exactly 21 times the fundamental period of our known signal; this is to avoid violating the periodicity assumption of Fourier analysis and the resulting spurious energy scatter into higher spectral bands. The most energetic Fourier bands were the sine and cosine at the fundamental period of 28, the two together accounting for 9.8% of the variance, compared with 14.1% of the variance accounted for by the first two SSA modes. Thus, even with great care, the Fourier method scatters the spectral energy considerably more than the SSA method.

[51] The SSA has also been applied to the multivariate case by *Hsieh and Wu* [2002]. The tropical Pacific monthly SLPA data [*Woodruff et al.*, 1987] during 1950–2000 were used. The first eight SSA modes of the SLPA accounted for 7.9, 7.1, 5.0, 4.9, 4.0, 3.1, 2.5, and 1.9%, respectively, of the total variance of the augmented data. In Figure 19 the first two modes displayed the Southern Oscillation (SO), the east-west seesaw of SLPA at around the 50-month period, while the higher modes displayed fluctuations at around the QBO [*Hamilton*, 1998] average period of 28 months.

[52] The eight leading PCs of the SSA were then used as inputs, $x_1, \ldots, x_8$, to the NLPCA(cir) network, yielding the NLSSA mode 1 for the SLPA. This mode accounts for

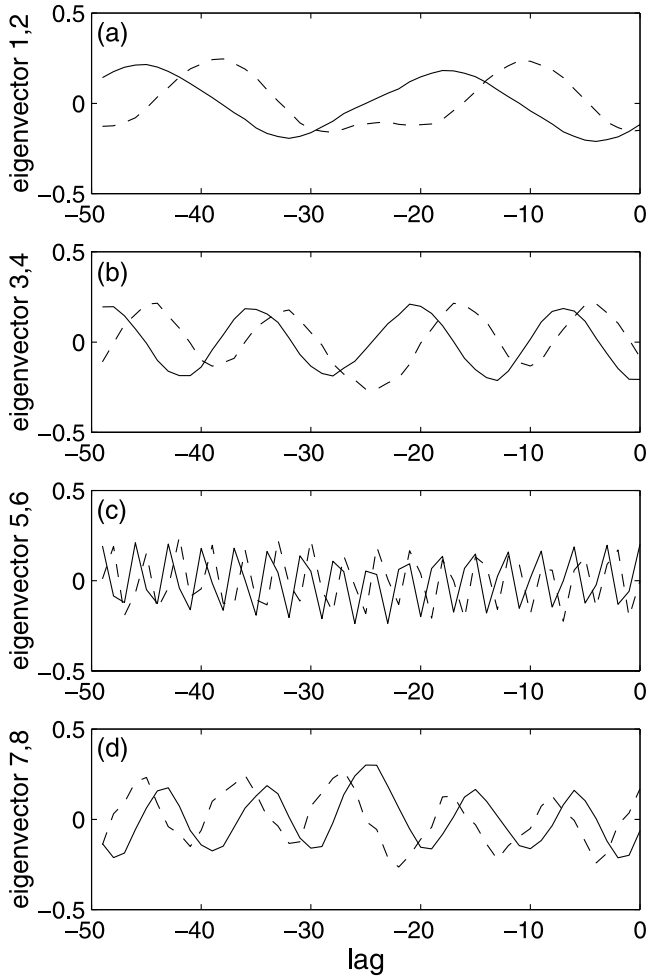

Figure 16. First eight SSA eigenvectors as a function of time lag: (a) mode 1 (solid curve) and mode 2 (dashed curve), (b) mode 3 (solid curve) and mode 4 (dashed curve), (c) mode 5 (solid curve) and mode 6 (dashed curve), and (d) mode 7 (solid curve) and mode 8 (dashed curve).

17.1% of the variance of the augmented data, more than the variance explained by the first two SSA modes (15.0%). This is not surprising as the NLSSA mode did more than just combine the SSA modes 1 and 2: It also connects the SSA mode 3 to the SSA modes 1 and 2 (Figure 20). In the x_1-x_3 plane the bowl-shaped projected solution implies that PC3 tends to be positive when PC1 takes on either large positive or large negative values. Similarly, in the x_2-x_3 plane the hill-shaped projected solution indicates that PC3 tends to be negative when PC2 takes on large positive or negative values. These curves reveal interactions between the longer-timescale (50 months) SSA modes 1 and 2 and the shorter-timescale (28 months) SSA mode 3.

[53] In the linear case of PCA or SSA, as the PC varies, the loading pattern is unchanged except for scaling by the PC. In the nonlinear case, as the NLPC varies, the loading pattern changes as it does not generally lie along a fixed eigenvector. The space-time loading patterns for the NLSSA mode 1 at various values of the NLPC θ (Figure 21) manifest prominently the growth and decay of the negative phase of the SO (i.e., negative SLPA in the eastern equatorial Pacific and positive SLPA in the west) as time progresses. The negative phase of the SO here is much shorter and more intense than the positive phase, in agreement with observations and in contrast to the SSA modes 1 and 2 (Figures 19a and 19b), where the negative and positive phases of the SO are about equal in duration and magnitude.

[54] The tropical Pacific SSTA field was also analyzed by the NLSSA method by *Hsieh and Wu* [2002]. Comparing the NLSSA mode 1 loading patterns with the patterns from the first two SSA modes of the SSTA, *Hsieh and Wu* [2002] found three notable differences: (1) The presence of warm anomalies for 24 months followed by cool anomalies for 24 months in the first two SSA modes is replaced in the NLSSA mode 1 by warm anomalies for 18 months followed by cool anomalies for about 33 months; although the cool anomalies can be quite mild for long periods, they can develop into full La Niña cool episodes. (2) The El Niño warm episodes are strongest near the eastern boundary, while the La Niña episodes are strongest near the central equatorial Pacific in the NLSSA mode 1, an asymmetry not found in the individual SSA modes. (3) The magnitude of the peak positive anomalies is significantly larger than that of the peak negative anomalies in the NLSSA mode 1, again an asymmetry not found in the individual SSA modes. All three differences indicate that the NLSSA mode 1 is much closer to the observed ENSO properties than the first two SSA modes are.

[55] Furthermore, from the residual the NLSSA mode 2 has been extracted by *Hsieh and Wu* [2002] for the SLPA field and for the SSTA field. For both variables the NLSSA mode 2 has a 39-month period, considerably longer than the QBO periods typically reported by previous studies using linear techniques [*Ghil et al.*, 2002]. Intriguingly, the coupling between the SLPA and the SSTA fields for the second nonlinear mode of a 39-month period was found to be considerably stronger than their coupling for the first nonlinear "ENSO" mode of a 51-month period [*Hsieh and Wu*, 2002]. The NLSSA technique has also been used to study the stratospheric equatorial winds for the QBO phenomenon [*Hsieh and Hamilton*, 2003].

5. SUMMARY AND CONCLUSIONS

[56] This paper has reviewed the recent extension of the feed forward NN from its original role for nonlinear regression and classification to nonlinear PCA (for open and closed curves), nonlinear CCA, and nonlinear SSA. With examples from the atmosphere and the ocean, notably the ENSO and the stratospheric QBO phenomena, these NN methods can be seen to advance our understanding of geophysical phenomena. To highlight only a few of the many new findings by the nonlinear techniques, we note that the nonlinearity in the tropical Pacific interannual variability has been found to have increased in recent decades [*Hsieh*, 2001b; *Wu and Hsieh*, 2003]; that besides the main coupling at the ENSO timescale of about 51 months the strongest coupling between the tropical

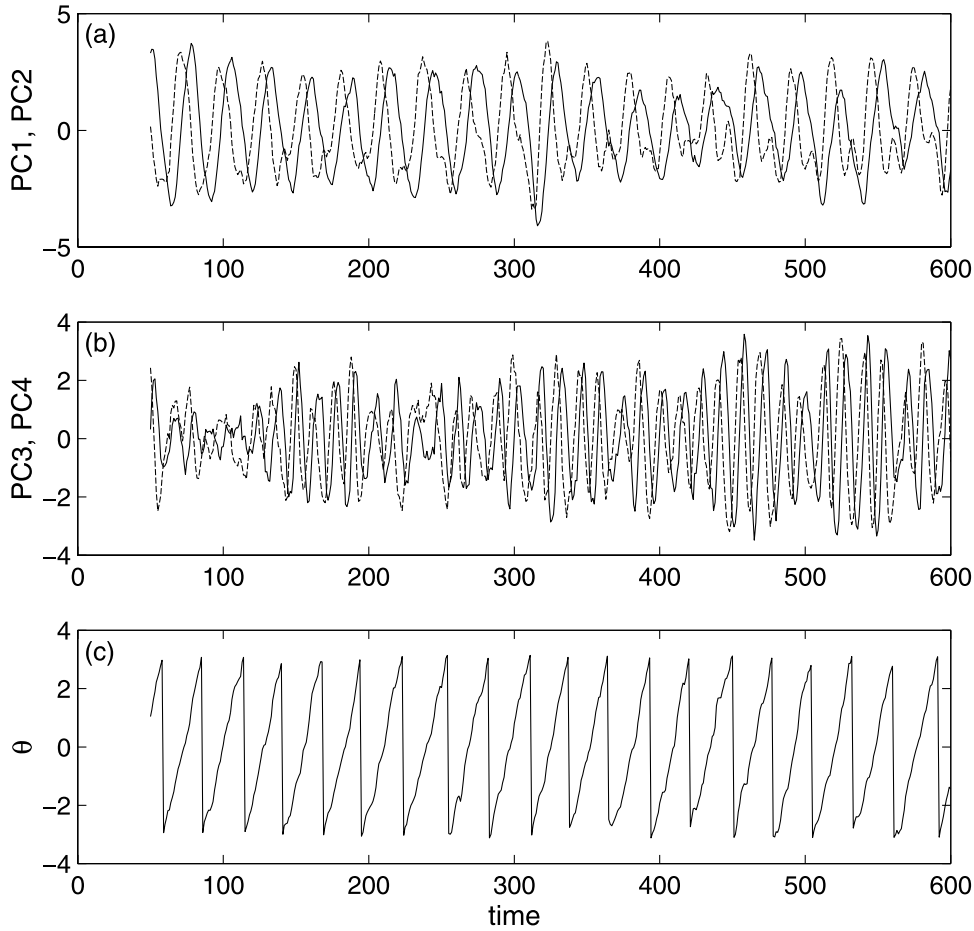

Figure 17. PC time series of SSA (a) mode 1 (solid curve) and mode 2 (dashed curve), (b) mode 3 (solid curve) and mode 4 (dashed curve), and (c) θ, the nonlinear PC from NLSSA mode 1. Note θ is periodic, here bounded between $-\pi$ and π radians.

Pacific SLP and SST has been identified at a second nonlinear mode of a 39-month period, this unusual period itself arising from the interaction between the linear modes with ENSO and QBO timescales [*Hsieh and Wu*, 2002]; and that the phase of the stratospheric QBO can be better defined, resulting in an enhancement of the Holton-Tan effect [*Hamilton and Hsieh*, 2002]. The nonlinear PCA, CCA, and SSA codes (written in MATLAB® are freely downloadable from the author's web site (http://www.ocgy.ubc.ca/projects/clim.pred).

[57] PCA is widely used for two main purposes: (1) to reduce the dimensionality of the data set and (2) to extract features or recognize patterns from the data set. It is purpose 2 where PCA can be improved upon. Rotated PCA (RPCA) sacrifices on the amount of variance explained, but by rotating the PCA eigenvectors, RPCA eigenvectors can point more toward local data clusters and can therefore be more representative of physical states than the PCA eigenvectors. With the tropical Pacific SST as an example it was shown that RPCA represented El Niño states better than PCA, but neither method represented La Niña states well. In contrast, nonlinear PCA (NLPCA) passed through both the clusters of El Niño and La Niña states, thus representing both well within a single mode; the NLPCA first mode also accounted for more variance of the data set than the first mode of PCA or RPCA.

[58] With PCA the straight line explaining the maximum variance of the data is found. With NLPCA the straight line is replaced by a continuous, open curve. NLPCA(cir) (NLPCA with a circular node at the bottleneck) replaces the open curve with a closed curve, so periodic or wave solutions can be modeled. When dealing with data containing a nonlinear or periodic structure, the linear methods scatter the energy into multiple modes, which is usually prevented when the nonlinear methods are used.

[59] With two data fields **x** and **y** the classical CCA method finds the canonical variate u (from a linear combination of the **x** variables) and its partner v (from the **y** variables), so that the correlation between u and v is maximized. CCA finds a line in the **x** space, where fluctuations of the **x** data projected onto this line are most highly correlated with fluctuations of **y** data projected onto another line in the **y** space. NN can perform nonlinear CCA (NLCCA), where u and v can be nonlinear functions of the **x** and **y** variables, respectively. NLCCA finds a curve in the **x** space where fluctuations of the **x** data projected onto this curve are most highly correlated with fluctuations of **y** data projected onto another curve in the **y** space.

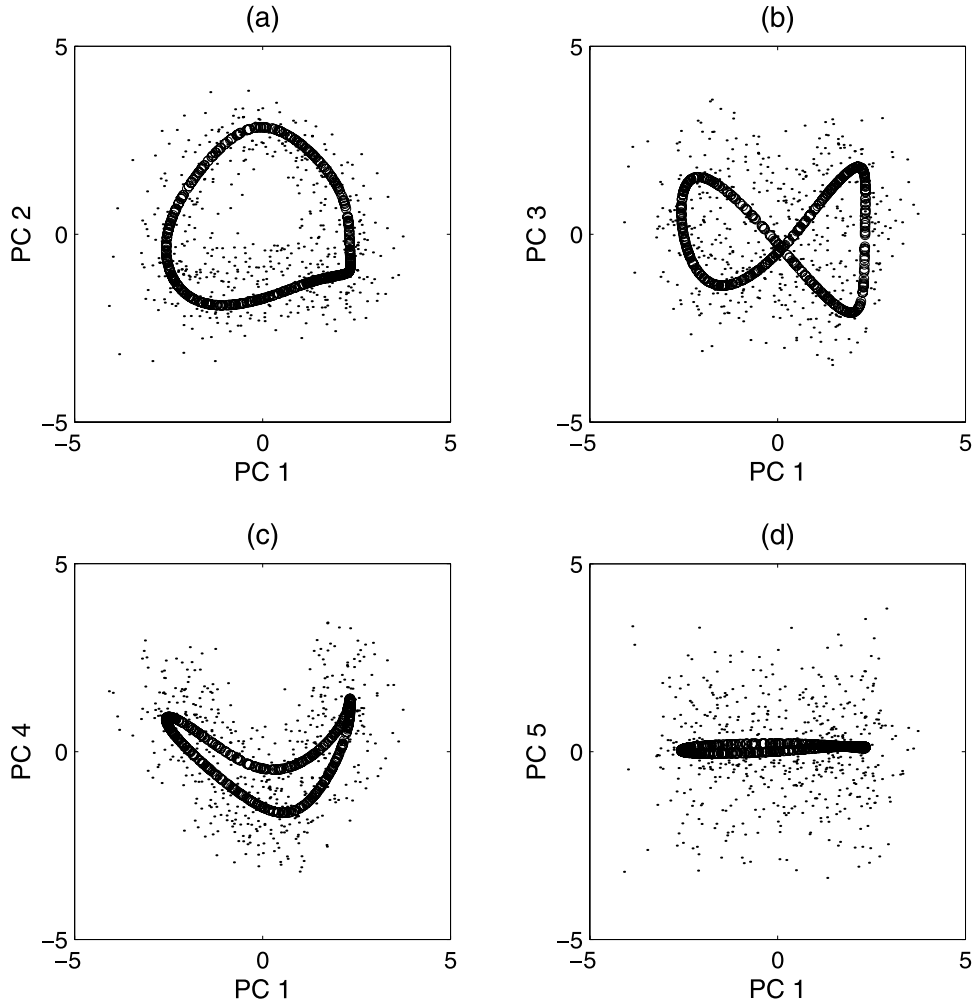

Figure 18. First NLSSA mode indicated by the (overlapping) small circles, with the input data shown as dots. The input data were the first eight PCs from the SSA of the time series y containing the stretched square wave. The NLSSA solution is a closed curve in an eight-dimensional PC space. The NLSSA solution projected onto (a) the PC1-PC2 plane, (b) the PC1-PC3 plane, (c) the PC1-PC4 plane, and (d) the PC1-PC5 plane.

[60] For univariate and multivariate time series analysis the PCA method has been extended to the SSA technique. NN can also be used to perform nonlinear SSA (NLSSA): The data set is first condensed by the SSA; then several leading PCs from the SSA are chosen as inputs to the NLPCA(cir) network, which extracts the NLSSA mode by nonlinearly combining the various SSA modes.

[61] In general, NLSSA has several advantages over SSA: (1) The PCs from different SSA modes are linearly uncorrelated; however, they may have relationships that can be detected by the NLSSA. (2) Although the SSA modes are not restricted to sinusoidal oscillations in time like the Fourier spectral components, in practice, they are inefficient in modeling nonsinusoidal periodic signals (e.g., the stretched square wave in section 4), scattering the signal energy into many SSA modes, similar to the way Fourier spectral analysis scatters the energy of a nonsinusoidal wave to its higher harmonics. The NLSSA recombines the SSA modes to extract the nonsinusoidal signal, alleviating the spurious transfer of energy to higher frequencies. In the tropical Pacific the NLSSA mode 2 of both the SSTA field and the SLPA field yielded a 39-month signal, considerably lower in frequency than the QBO frequency signals found by linear methods.

Figure 19. (a–f) SSA modes 1–6 for the tropical Pacific sea level pressure anomalies (SLPA), respectively. The contour plots display the SSA space-time eigenvectors (loading patterns), showing the SLPA along the equator as a function of the lag. Solid contours indicate positive anomalies, and dashed contours indicate negative anomalies, with the zero contour indicated by the thick solid curve. In a separate graph beneath each contour plot the PC of each SSA mode is also plotted as a time series (where each tick mark on the abscissa indicates the start of a year). The time of the PC is synchronized to the lag time of 0 month in the space-time eigenvector. Figure 19 courtesy of A. Wu.

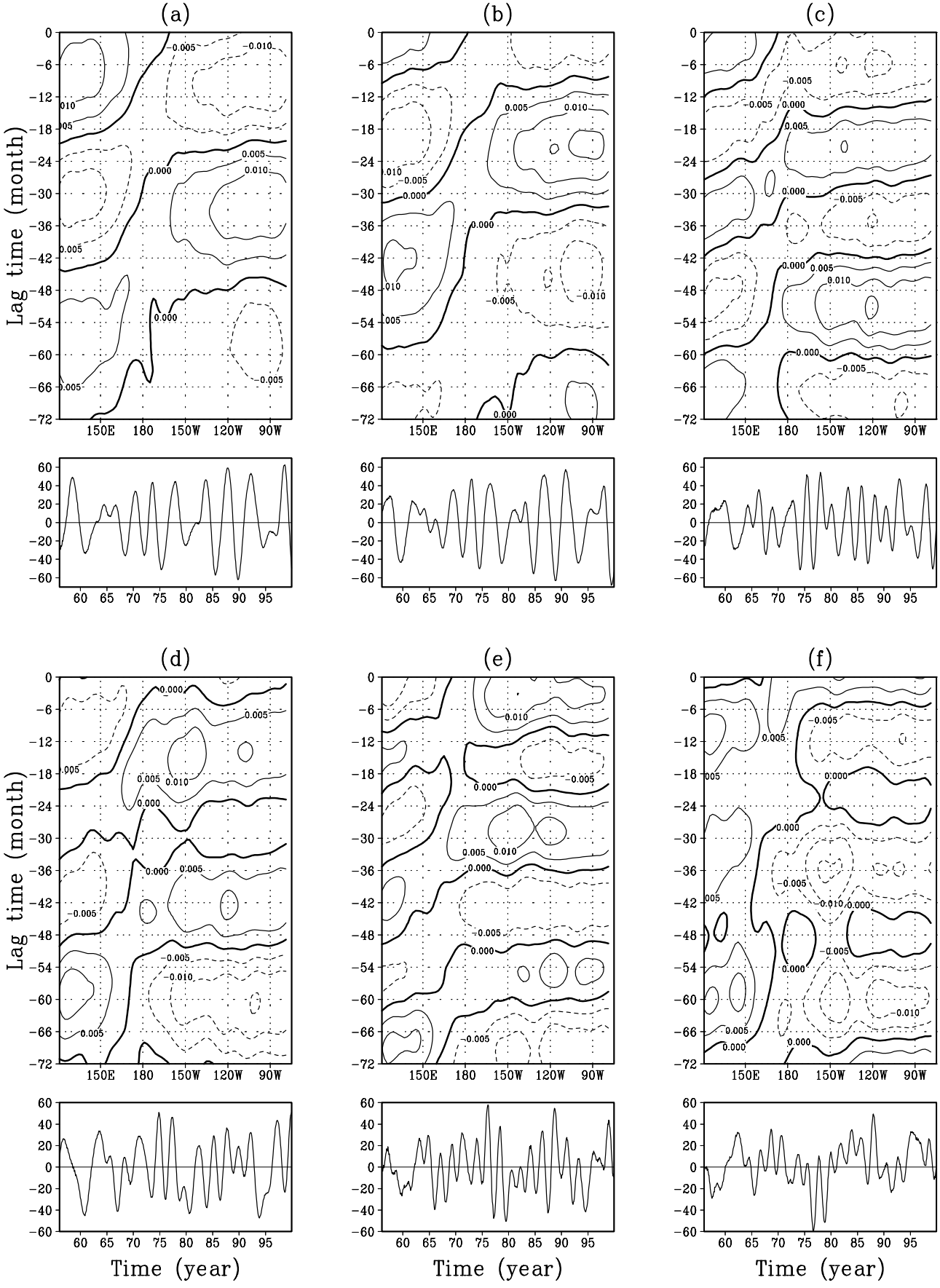
(a)
(b)
(c)
(d)
(e)
(f)
Lag time (month)
Time (year)
150E 180 150W 120W 90W
60 65 70 75 80 85 90 95

Figure 19

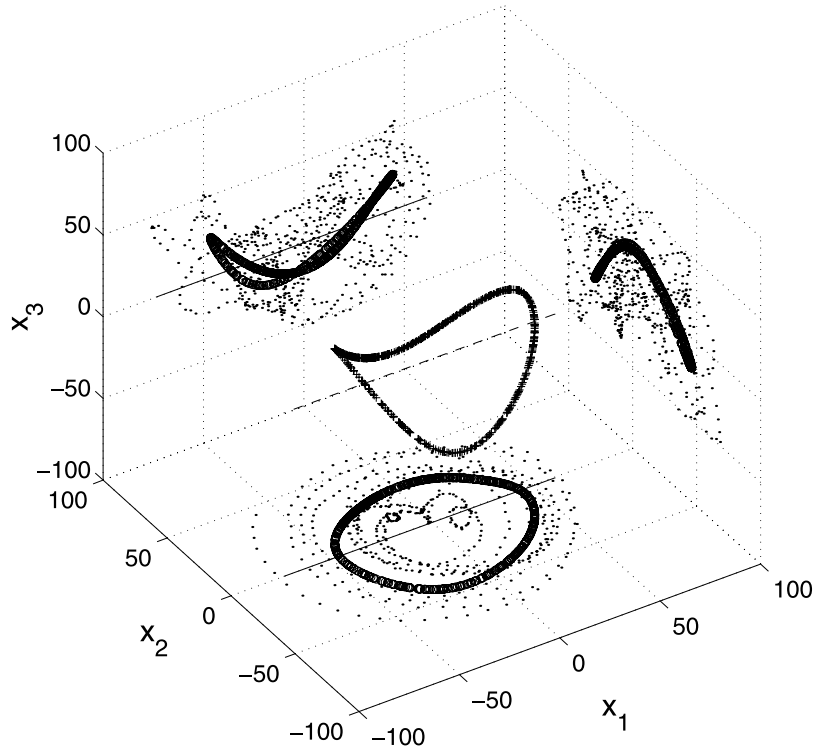

Figure 20. NLSSA mode 1 for the tropical Pacific SLPA. The PCs of SSA modes 1–8 were used as inputs $x_1, \ldots, x_8$ to the NLPCA(cir) network, with the resulting NLSSA mode 1 shown as (densely overlapping) crosses in the x_1-x_2-x_3 3-D PC space. The projections of this mode onto the x_1-x_2, x_1-x_3, and x_2-x_3 planes are denoted by the (densely overlapping) small circles, and the projected data are shown by dots. For comparison, the linear SSA mode 1 is shown by the dashed line in the 3-D space and by the projected solid lines on the 2-D planes. From *Hsieh and Wu* [2002].

[62] In summary, the linear methods currently used are often too simplistic to describe complicated real-world systems, resulting in a tendency to scatter a single oscillatory phenomenon into numerous modes or higher harmonics. This would introduce unphysical spatially standing patterns or spurious high-frequency energy. These problems are shown to be largely alleviated by the use of nonlinear methods.

[63] The main disadvantage of NN methods compared with the linear methods lies in their instability or nonuniqueness; with local minima in the cost function, optimizations started from different initial parameters often end up at different minima for the NN approach. A number of optimization runs starting from different random initial parameters is needed, where the best run is chosen as the solution; even then, there is no guarantee that the global minimum has been found. Proper scaling of the input data is essential to avoid having the nonlinear optimization algorithm searching for parameters with a wide range of magnitudes. Regularization by adding weight penalty terms to the cost functions generally improved the stability of the NN methods. Nevertheless, for short records with noisy data one may not be able to find a reliable nonlinear solution, and the linear solution may be the best one can extract from ~~of~~ the data. The time averaging of data (e.g., averaging daily data to yield monthly data) may also, through the central limit theorem, severely reduce the nonlinearity which can be detected [*Yuval and Hsieh*, 2002].

[64] Hence, whether the nonlinear approach has a significant advantage over the linear approach is highly dependent on the data set: The nonlinear approach is generally ineffective if the data record is short and noisy or the underlying physics is essentially linear. For the Earth's climate, tropical variability such as ENSO and the stratospheric QBO have strong signal-to-noise ratio and are handled well by the nonlinear methods; in contrast, in the middle and high latitudes the signal-to-noise ratio is much weaker, rendering the nonlinear methods less effective. Presently, the number of hidden neurons in the NN and the weight penalty parameters are often determined by a trial and error approach; adopting techniques such as generalized cross validation [*Yuval*, 2000] and information criterion [*Burnham and Anderson*, 1998] may help in the future to provide more guidance on the choice of the most appropriate NN architecture. While NN has been widely used as the main workhorse in nonlinear multivariate and time series analysis, new emerging techniques such as kernel-based methods [*Vapnik*, 1998] may play an increasingly important role in the future.

APPENDIX A: NLPCA MODEL

[65] In Figure 2a the transfer function f_1 maps from $\mathbf{x}$, the input column vector of length l, to the first hidden layer (the

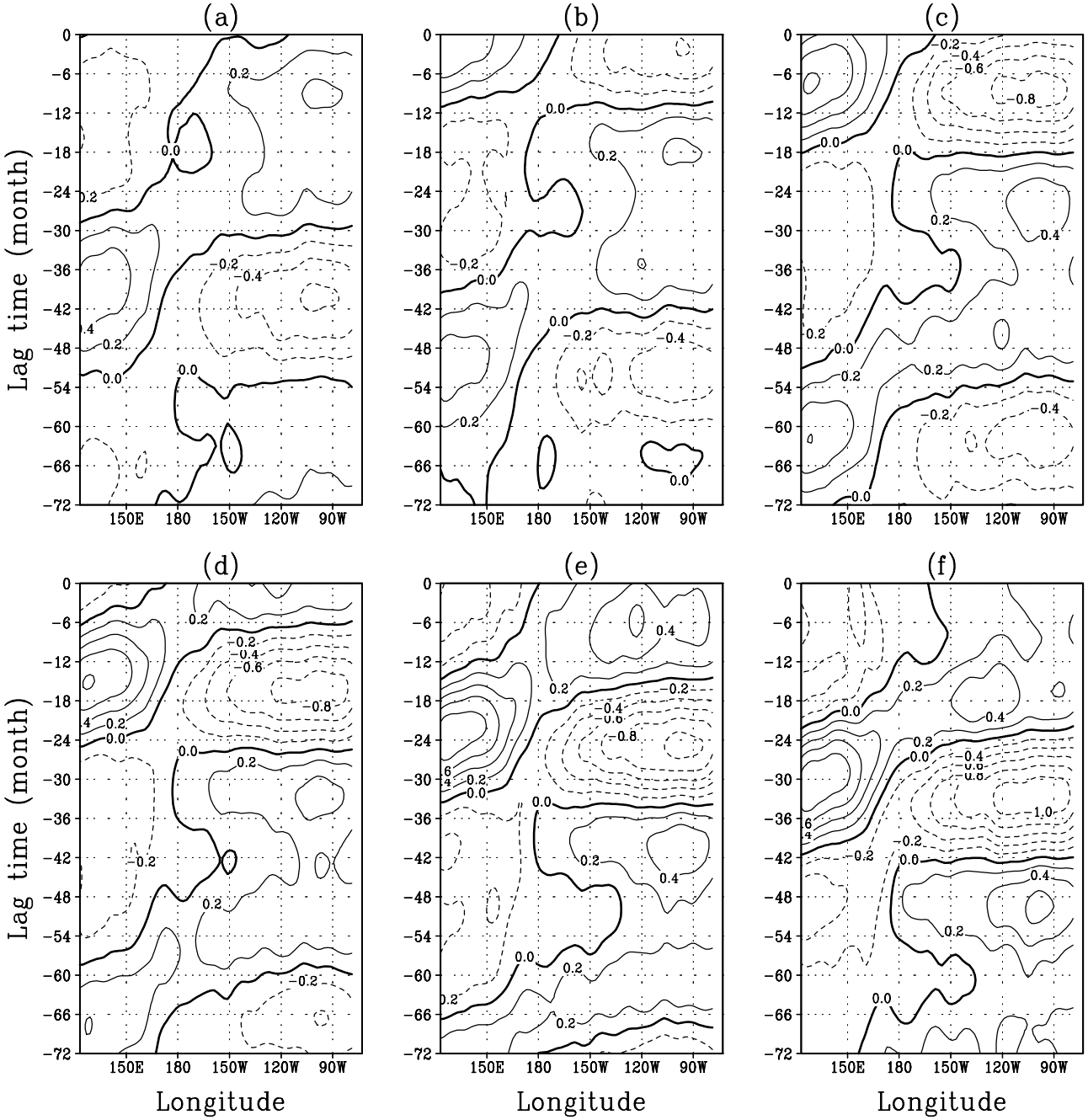

Figure 21. SLPA NLSSA mode 1 space-time loading patterns for various values of the NLPC θ: (a) θ = 0°, (b) θ = 60°, (c) θ = 120°, (d) θ = 180°, (e) θ = 240°, and (f) θ = 300°. The contour plots display the SLPA along the equator as a function of the lag time. Contour interval is 0.2 mbar. Figure 21 courtesy of A. Wu.

encoding layer), represented by $\mathbf{h}^{(x)}$, a column vector of length m, with elements

$$h_k^{(x)} = f_1\left[\left(\mathbf{W}^{(x)}\mathbf{x} + \mathbf{b}^{(x)}\right)_k\right], \tag{A1}$$

where (with the capital bold font reserved for matrices and the small bold font for vectors), $\mathbf{W}^{(x)}$ is an $m \times l$ weight matrix, $\mathbf{b}^{(x)}$ is a column vector of length m containing the bias parameters, and $k = 1, \ldots, m$. Similarly, a second transfer function f_2 maps from the encoding layer to the bottleneck layer containing a single neuron, which represents the nonlinear principal component u,

$$u = f_2\left(\mathbf{w}^{(x)} \cdot \mathbf{h}^{(x)} + \overline{b}^{(x)}\right). \tag{A2}$$

The transfer function f_1 is generally nonlinear (usually the hyperbolic tangent or the sigmoidal function, though the exact form is not critical), while f_2 is usually taken to be the identity function.

[66] Next a transfer function f_3 maps from u to the final hidden layer (the decoding layer) $\mathbf{h}^{(u)}$,

$$h_k^{(u)} = f_3\left[\left(\mathbf{w}^{(u)}u + \mathbf{b}^{(u)}\right)_k\right], \tag{A3}$$

($k = 1, \ldots, m$), followed by f_4 mapping from $\mathbf{h}^{(u)}$ to $\mathbf{x}'$, the output column vector of length l, with

$$x'_i = f_4\left[\left(\mathbf{W}^{(u)}\mathbf{h}^{(u)} + \overline{\mathbf{b}}^{(u)}\right)_i\right]. \tag{A4}$$

[67] The cost function $J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle$ is minimized by finding the optimal values of $\mathbf{W}^{(x)}$, $\mathbf{b}^{(x)}$, $\mathbf{w}^{(x)}$, $\overline{b}^{(x)}$, $\mathbf{w}^{(u)}$, $\mathbf{b}^{(u)}$, $\mathbf{W}^{(u)}$, and $\overline{\mathbf{b}}^{(u)}$. The mean-square error (MSE) between the NN output $\mathbf{x}'$ and the original data $\mathbf{x}$ is thus minimized. The NLPCA was implemented using the hyperbolic tangent function for f_1 and f_3 and the identity function for f_2 and f_4, so that

$$u = \mathbf{w}^{(x)} \cdot \mathbf{h}^{(x)} + \overline{b}^{(x)} \tag{A5}$$

$$x'_i = \left(\mathbf{W}^{(u)}\mathbf{h}^{(u)} + \overline{\mathbf{b}}^{(u)}\right)_i. \tag{A6}$$

[68] Furthermore, we adopt the normalization conditions that $\langle u \rangle = 0$ and $\langle u^2 \rangle = 1$. These conditions are approximately satisfied by modifying the cost function to

$$J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle + \langle u \rangle^2 + \left(\langle u^2 \rangle - 1\right)^2. \tag{A7}$$

The total number of (weight and bias) parameters used by the NLPCA is $2lm + 4m + l + 1$, though the number of effectively free parameters is two less because of the constraints on $\langle u \rangle$ and $\langle u^2 \rangle$.

[69] The choice of m, the number of hidden neurons in both the encoding and decoding layers, follows a general principle of parsimony. A larger m increases the nonlinear modeling capability of the network but could also lead to overfitted solutions (i.e., wiggly solutions which fit to the noise in the data). If f_4 is the identity function and $m = 1$, then equation (A6) implies that all x'_i are linearly related to a single hidden neuron; hence there can only be a linear relation between the x'_i variables. Thus, for nonlinear solutions we need to look at $m \geq 2$. It is also possible to have more than one neuron at the bottleneck layer. For instance, with two bottleneck neurons the mode extracted will span a 2-D surface instead of a 1-D curve.

[70] The nonlinear optimization was carried out by the MATLAB function "fminu," a quasi-Newton algorithm. Because of local minima in the cost function, there is no guarantee that the optimization algorithm reaches the global minimum. Hence a number of runs with random initial weights and bias parameters was made. Also, 20% of the data were randomly selected as validation data and withheld from the training of the NNs. Runs where the MSE was larger for the validation data set than for the training data set were rejected to avoid overfitted solutions. Then the run with the smallest MSE was selected as the solution.

[71] In general, the most serious problem with NLPCA is the presence of local minima in the cost function. As a result, optimizations started from different initial parameters often converge to different minima, rendering the solution unstable or nonunique. Regularization of the cost function by adding weight penalty terms is an answer.

[72] The purpose of the weight penalty terms is to limit the nonlinear power of the NLPCA, which came from the nonlinear transfer functions in the network. The transfer function tanh has the property that given x in the interval $[-L, L]$, one can find a small enough weight w, so that $\tanh(wx) \approx wx$; that is, the transfer function is almost linear. Similarly, one can choose a large enough w so that tanh approaches a step function, thus yielding Z-shaped solutions. If we can penalize the use of excessive weights, we can limit the degree of nonlinearity in the NLPCA solution. This is achieved with a modified cost function

$$J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle + \langle u \rangle^2 + \left(\langle u^2 \rangle - 1\right)^2 + P \sum_{ki} \left(W_{ki}^{(x)}\right)^2, \tag{A8}$$

where P is the weight penalty parameter. A large P increases the concavity of the cost function and forces the weights $\mathbf{W}^{(x)}$ to be small in magnitude, thereby yielding smoother and less nonlinear solutions than when P is small or zero. Hence increasing P also reduces the number of effectively free parameters of the model. The percentage of the variance explained by the NLPCA mode is simply

$$100\ \% \times \left(1 - \frac{\langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle}{\langle \|\mathbf{x} - \overline{\mathbf{x}}\|^2 \rangle}\right), \tag{A9}$$

with $\overline{\mathbf{x}}$ being the mean of $\mathbf{x}$.

APPENDIX B: NLPCA(cir) MODEL

[73] At the bottleneck in Figure 2b, analogous to u in equation (A5), we calculate the prestates p_o and q_o by

$$p_o = \mathbf{w}^{(x)} \cdot \mathbf{h}^{(x)} + \overline{b}^{(x)} \qquad q_o = \tilde{\mathbf{w}}^{(x)} \cdot \mathbf{h}^{(x)} + \tilde{b}^{(x)}, \tag{B1}$$

where $\mathbf{w}^{(x)}$ and $\tilde{\mathbf{w}}^{(x)}$ are weight parameter vectors and $\overline{b}^{(x)}$ and $\tilde{b}^{(x)}$ are bias parameters. Let

$$r = \left(p_o^2 + q_o^2\right)^{1/2}, \tag{B2}$$

then the circular node is defined with

$$p = p_o/r \qquad q = q_o/r, \tag{B3}$$

satisfying the unit circle equation $p^2 + q^2 = 1$. Thus, even though there are two variables p and q at the bottleneck, there is only one angular degree of freedom from θ (Figure 2b) because of the circle constraint. The mapping from the bottleneck to the output proceeds as in Appendix A, with equation (A3) replaced by

$$h_k^{(u)} = \tanh\left[\left(\mathbf{w}^{(u)}p + \tilde{\mathbf{w}}^{(u)}q + \mathbf{b}^{(u)}\right)_k\right]. \tag{B4}$$

[74] When implementing NLPCA(cir), *Hsieh* [2001a] found that there are actually two possible configurations: (1) A restricted configuration where the constraints $\langle p \rangle = 0 = \langle q \rangle$ are applied and (2) a general configuration without

the constraints. With configuration 1 the constraints can be satisfied approximately by adding the extra terms $\langle p \rangle^2$ and $\langle q \rangle^2$ to the cost function. If a closed curve solution is sought, then configuration 1 is better than configuration 2 as it has effectively two fewer parameters. However, configuration 2, being more general than configuration 1, can more readily model open curve solutions like a regular NLPCA. The reason is that if the input data mapped onto the *p*-*q* plane cover only a segment of the unit circle instead of the whole circle, then the inverse mapping from the *p*-*q* space to the output space will yield a solution resembling an open curve. Hence, given a data set, configuration 2 may yield either a closed curve or an open curve solution. Its generality comes with a price, namely, that there may be more local minima to contend with. The number of parameters is $2lm + 6m + l + 2$; though under configuration 1 the number of effectively free parameters is two less because of the imposed constraints. Unlike NLPCA, which reduces to PCA when only linear transfer functions are used, NLPCA (cir) does not appear to have a linear counterpart.

APPENDIX C: NLCCA MODEL

[75] In Figure 7 the inputs $\mathbf{x}$ and $\mathbf{y}$ are mapped to the neurons in the hidden layer:

$$h_k^{(x)} = \tanh\left[\left(\mathbf{W}^{(x)}\mathbf{x} + \mathbf{b}^{(x)}\right)_k\right]$$
$$h_n^{(y)} = \tanh\left[\left(\mathbf{W}^{(y)}\mathbf{y} + \mathbf{b}^{(y)}\right)_n\right], \tag{C1}$$

where $\mathbf{W}^{(x)}$ and $\mathbf{W}^{(y)}$ are weight matrices and $\mathbf{b}^{(x)}$ and $\mathbf{b}^{(y)}$ are bias parameter vectors. The dimensions of $\mathbf{x}$, $\mathbf{y}$, $\mathbf{h}^{(x)}$, and $\mathbf{h}^{(y)}$ are l_1, m_1, l_2, and m_2, respectively.

[76] The canonical variate neurons u and v are calculated from a linear combination of the hidden neurons $\mathbf{h}^{(x)}$ and $\mathbf{h}^{(y)}$, respectively, with

$$u = \mathbf{w}^{(x)} \cdot \mathbf{h}^{(x)} + \overline{b}^{(x)} \qquad v = \mathbf{w}^{(y)} \cdot \mathbf{h}^{(y)} + \overline{b}^{(y)}. \tag{C2}$$

These mappings are standard feed forward NNs and are capable of representing any continuous functions mapping from $\mathbf{x}$ to u and from $\mathbf{y}$ to v to any given accuracy, provided large enough l_2 and m_2 are used.

[77] To maximize cor(u, v), the cost function $J = -\,\mathrm{cor}(u, v)$ is minimized by finding the optimal values of $\mathbf{W}^{(x)}$, $\mathbf{W}^{(y)}$, $\mathbf{b}^{(x)}$, $\mathbf{b}^{(y)}$, $\mathbf{w}^{(x)}$, $\mathbf{w}^{(y)}$, $\overline{b}^{(x)}$, and $\overline{b}^{(y)}$. We also adopt the constraints $\langle u \rangle = 0 = \langle v \rangle$ and $\langle u^2 \rangle = 1 = \langle v^2 \rangle$, which are approximately satisfied by modifying the cost function to

$$J = -\mathrm{cor}(u, v) + \langle u \rangle^2 + \langle v \rangle^2 + \left(\langle u^2 \rangle^{1/2} - 1\right)^2 + \left(\langle v^2 \rangle^{1/2} - 1\right)^2. \tag{C3}$$

[78] On the right side of Figure 7 the top NN (a standard feed forward NN) maps from u to $\mathbf{x}'$ in two steps:

$$h_k^{(u)} = \tanh\left[\left(\mathbf{w}^{(u)}u + \mathbf{b}^{(u)}\right)_k\right] \qquad \mathbf{x}' = \mathbf{W}^{(u)}\mathbf{h}^{(u)} + \overline{\mathbf{b}}^{(u)}. \tag{C4}$$

The cost function $J_1 = \langle \|\mathbf{x}' - \mathbf{x}\|^2 \rangle$ is minimized by finding the optimal values of $\mathbf{w}^{(u)}$, $\mathbf{b}^{(u)}$, $\mathbf{W}^{(u)}$, and $\overline{\mathbf{b}}^{(u)}$. The MSE between the NN output $\mathbf{x}'$ and the original data $\mathbf{x}$ is thus minimized.

[79] Similarly, the bottom NN on the right side of Figure 7 maps from v to $\mathbf{y}'$:

$$h_n^{(v)} = \tanh\left[\left(\mathbf{w}^{(v)}v + \mathbf{b}^{(v)}\right)_n\right] \qquad \mathbf{y}' = \mathbf{W}^{(v)}\mathbf{h}^{(v)} + \overline{\mathbf{b}}^{(v)}, \tag{C5}$$

with the cost function $J_2 = \langle \|\mathbf{y}' - \mathbf{y}\|^2 \rangle$ minimized. The total number of parameters used by the NLCCA is $2(l_1 l_2 + m_1 m_2) + 4(l_2 + m_2) + l_1 + m_1 + 2$, though the number of effectively free parameters is four less because of the constraints on $\langle u \rangle$, $\langle v \rangle$, $\langle u^2 \rangle$, and $\langle v^2 \rangle$.

[80] A number of runs mapping from ($\mathbf{x}$, $\mathbf{y}$) to (u, v), using random initial parameters, were performed. The run attaining the highest cor(u, v) was selected as the solution. Next a number of runs (mapping from u to $\mathbf{x}'$) was used to find the solution with the smallest MSE in $\mathbf{x}'$. Finally, a number of runs were used to find the solution yielding the smallest MSE in $\mathbf{y}'$. After the first NLCCA mode has been retrieved from the data, the method can be applied again to the residual to extract the second mode and so forth.

[81] That the CCA is indeed a linear version of this NLCCA can be readily seen by replacing the hyperbolic tangent transfer functions in equations (C1), (C4), and (C5) with the identity function, thereby removing the nonlinear modeling capability of the NLCCA. Then the forward maps to u and v involve only a linear combination of the original variables $\mathbf{x}$ and $\mathbf{y}$, as in the CCA.

[82] With three NNs in NLCCA, overfitting can occur in any of the three networks. With noisy data the three cost functions are modified to

$$J = -\mathrm{cor}(u, v) + \langle u \rangle^2 + \langle v \rangle^2 + \left(\langle u^2 \rangle^{1/2} - 1\right)^2 + \left(\langle v^2 \rangle^{1/2} - 1\right)^2 + P\left[\sum_{ki}\left(W_{ki}^{(x)}\right)^2 + \sum_{nj}\left(W_{nj}^{(y)}\right)^2\right], \tag{C6}$$

$$J_1 = \langle \|\mathbf{x}' - \mathbf{x}\|^2 \rangle + P_1 \sum_k \left(w_k^{(u)}\right)^2, \tag{C7}$$

$$J_2 = \langle \|\mathbf{y}' - \mathbf{y}\|^2 \rangle + P_2 \sum_n \left(w_n^{(v)}\right)^2, \tag{C8}$$

where P, P_1, and P_2 are nonnegative weight penalty parameters. Since the nonlinearity of a network is controlled by the weights in the hyperbolic tangent transfer function, only those weights are penalized.

[83] **ACKNOWLEDGMENTS.** I would like to thank Aiming Wu, who supplied Figures 19 and 21. The American Meteorological Society, Elsevier Science, and Blackwell Science kindly granted permission to reproduce or adapt figures from their publications. Support through strategic and research grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Foundation for Climate and Atmospheric Sciences is gratefully acknowledged.

[84] Kendal McGuffie was the Editor responsible for this paper. He thanks one technical reviewer and one cross-disciplinary reviewer.

REFERENCES

Aires, F., A. Chédin, and J. P. Nadal (2000), Independent component analysis of multivariate time series: Application to the tropical SST variability, *J. Geophys. Res.*, *105*(D13), 17,437–17,455.

Baldwin, M., et al. (2001), The quasi-biennial oscillation, *Rev. Geophys.*, *39*, 179–229.

Barnett, T. P., and R. Preisendorfer (1987), Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis, *Mon. Weather Rev.*, *115*, 1825–1850.

Barnston, A. G., and C. F. Ropelewski (1992), Prediction of ENSO episodes using canonical correlation analysis, *J. Clim.*, *5*, 1316–1345.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, 482 pp., Clarendon, Oxford, U.K.

Bretherton, C. S., C. Smith, and J. M. Wallace (1992), An intercomparison of methods for finding coupled patterns in climate data, *J. Clim.*, *5*, 541–560.

Burnham, K. P., and D. R. Anderson (1998), *Model Selection and Inference*, 353 pp., Springer-Verlag, New York.

Cavazos, T. (1999), Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas, *J. Clim.*, *12*, 1506–1523.

Cherkassky, V., and F. Mulier (1998), *Learning From Data*, 441 pp., John Wiley, Hoboken, N. J.

Chevallier, F., J. J. Morcrette, F. Cheruy, and N. A. Scott (2000), Use of a neural-network-based long-wave radiative-transfer scheme in the ECMWF atmospheric model, *Q. J. R. Meteorol. Soc.*, *126*, 761–776.

Comon, P. (1994), Independent component analysis—A new concept?, *Signal Process.*, *36*, 287–314.

Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function, *Math. Control Signals Syst.*, *2*, 303–314.

Del Frate, F., and G. Schiavon (1999), Nonlinear principal component analysis for the radiometric inversion of atmospheric profiles by using neural networks, *IEEE Trans. Geosci. Remote Sens.*, *37*, 2335–2342.

Diaz, H. F., and V. Markgraf (2000), *El Niño and the Southern Oscillation: Multiscale Variability and Global and Regional Impacts*, 496 pp., Cambridge Univ. Press, New York.

Elsner, J. B., and A. A. Tsonis (1996), *Singular Spectrum Analysis*, 164 pp., Plenum, New York.

Essenreiter, R., M. Karrenbach, and S. Treitel (2001), Identification and classification of multiple reflections with self-organizing maps, *Geophys. Prospect.*, *49*, 341–352.

Gardner, M. W., and S. R. Dorling (1998), Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences, *Atmos. Environ.*, *32*, 2627–2636.

Gemmill, W. H., and V. M. Krasnopolsky (1999), The use of SSM/I data in operational marine analysis, *Weather Forecasting*, *14*, 789–800.

Ghil, M., et al. (2002), Advanced spectral methods for climatic time series, *Rev. Geophys.*, *40*(1), 1003, doi:10.1029/2000RG000092.

Golyandina, N. E., V. V. Nekrutin, and A. A. Zhigljavsky (2001), *Analysis of Time Series Structure*, *SSA and Related Techniques*, 320 pp., Chapman and Hall, New York.

Hamilton, K. (1998), Dynamics of the tropical middle atmosphere: A tutorial review, *Atmos. Ocean*, *36*, 319–354.

Hamilton, K., and W. W. Hsieh (2002), Representation of the QBO in the tropical stratospheric wind by nonlinear principal component analysis, *J. Geophys. Res.*, *107*(D15), 4232, doi:10.1029/2001JD001250.

Hastie, T., and W. Stuetzle (1989), Principal curves, *JASA J. Am. Stat. Assoc.*, *84*, 502–516.

Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning*, 552 pp., Springer-Verlag, New York.

Haykin, S. (1999), *Neural Networks: A Comprehensive Foundation*, 842 pp., Prentice-Hall, Old Tappan, N. J.

Hoerling, M. P., A. Kumar, and M. Zhong (1997), El Niño, La Niña and the nonlinearity of their teleconnections, *J. Clim.*, *10*, 1769–1786.

Hollingsworth, J. L., R. M. Haberle, and J. Schaeffer (1997), Seasonal variations of storm zones on Mars, *Adv. Space Res.*, *19*, 1237–1240.

Holton, J. R., and H.-C. Tan (1980), The influence of the equatorial quasi-biennial oscillation on the global circulation at 50 mb, *J. Atmos. Sci.*, *37*, 2200–2208.

Hornik, K., M. Stinchcombe, and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, *2*, 359–366.

Hsieh, W. W. (2000), Nonlinear canonical correlation analysis by neural networks, *Neural Networks*, *13*, 1095–1105.

Hsieh, W. W. (2001a), Nonlinear principal component analysis by neural networks, *Tellus, Ser. A*, *53*, 599–615.

Hsieh, W. W. (2001b), Nonlinear canonical correlation analysis of the tropical Pacific climate variability using a neural network approach, *J. Clim.*, *14*, 2528–2539.

Hsieh, W. W., and K. Hamilton (2003), Nonlinear singular spectrum analysis of the tropical stratospheric wind, *Q. J. R. Meteorol. Soc.*, *129*, 2367–2382.

Hsieh, W. W., and B. Tang (1998), Applying neural network models to prediction and data analysis in meteorology and oceanography, *Bull. Am. Meteorol. Soc.*, *79*, 1855–1870.

Hsieh, W. W., and A. Wu (2002), Nonlinear multichannel singular spectrum analysis of the tropical Pacific climate variability using a neural network approach, *J. Geophys. Res.*, *107*(C7), 3076, doi:10.1029/2001JC000957.

Hyvärinen, A., J. Karhunen, and E. Oja (2001), *Independent Component Analysis*, 504 pp., John Wiley, Hoboken, N. J.

Jolliffe, I. T. (2002), *Principal Component Analysis*, 502 pp., Springer-Verlag, New York.

Kaiser, H. F. (1958), The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, *23*, 187–200.

Kirby, M. J., and R. Miranda (1996), Circular nodes in neural networks, *Neural Comp.*, *8*, 390–402.

Kohonen, T. (1982), Self-organized formation of topologically correct feature maps, *Biol. Cybernetics*, *43*, 59–69.

Kohonen, T. (2001), *Self-Organizing Maps*, 501 pp., Springer-Verlag, New York.

Kramer, M. A. (1991), Nonlinear principal component analysis using autoassociative neural networks, *AIChE J.*, *37*, 233–243.

Krasnopolsky, V. M., D. Chalikov, L. C. Breaker, and D. Rao (2000), Application of neural networks for efficient calculation of sea water density or salinity from the UNESCO equation of state, in *Proceedings of the Second Conference on Artificial Intelligence*, pp. 27–30, Am. Math. Soc., Providence, R. I.

Lai, P. L., and C. Fyfe (1999), A neural implementation of canonical correlation analysis, *Neural Networks*, *12*, 1391–1397.

Lai, P. L., and C. Fyfe (2000), Kernel and non-linear canonical correlation analysis, *Int. J. Neural Syst.*, *10*, 365–377.

Lorenz, E. N. (1963), Deterministic nonperiodic flow, *J. Atmos. Sci.*, *20*, 130–141.

Malthouse, E. C. (1998), Limitations of nonlinear PCA as performed with generic neural networks, *IEEE Trans. Neural Networks*, *9*, 165–173.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979), *Multivariate Analysis*, 518 pp., Academic, San Diego, Calif.

Marzban, C. (2000), A neural network for tornado diagnosis: Managing local minima, *Neural Comput. Appl.*, *9*, 133–141.

McCulloch, W. S., and W. Pitts (1943), A logical calculus of the ideas immanent in neural nets, *Bull. Math. Biophys.*, *5*, 115–137.

Melzer, T., M. Reiter, and H. Bischof (2003), Appearance models based on kernel canonical correlation analysis, *Pattern Recognit.*, *36*, 1961–1971.

Monahan, A. H. (2000), Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz system, *J. Clim.*, *13*, 821–835.

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

Monahan, A. H., 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.

Monahan, A. H., L. Pandolfo, and J. C. Fyfe (2001), The preferred structure of variability of the Northern Hemisphere atmospheric circulation, *Geophys. Res. Lett.*, *28*, 1019–1022.

Newbigging, S. C., L. A. Mysak, and W. W. Hsieh (2003), Improvements to the nonlinear principal component analysis method, with applications to ENSO and QBO, *Atmos. Ocean*, *41*, 291–299.

Philander, S. G. (1990), *El Niño*, *La Niña, and the Southern Oscillation*, 293 pp., Academic, San Diego, Calif.

Preisendorfer, R. W. (1988), *Principal Component Analysis in Meteorology and Oceanography*, 425 pp., Elsevier Sci., New York.

Rangarajan, G. K., and L. M. Barreto (2000), Long term variability in solar wind velocity and IMF intensity and the relationship between solar wind parameters and geomagnetic activity, *Earth Planets Space*, *52*, 121–132.

Richaume, P., F. Badran, M. Crepon, C. Mejía, H. Roquet, and S. Thiria (2000), Neural network wind retrieval from ERS-1 scatterometer data, *J. Geophys. Res.*, *105*(C4), 8737–8751.

Richman, M. B. (1986), Rotation of principal components, *J. Clim.*, *6*, 293–335.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, 403 pp., Cambridge Univ. Press, New York.

Rojas, R. (1996), *Neural Networks—A Systematic Introduction*, 502 pp., Springer-Verlag, New York.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986), Learning internal representations by error propagation, in *Parallel Distributed Processing*, vol. 1, edited by D. E. Rumelhart, J. L. McClelland, and P. R. Group, pp. 318–362, MIT Press, Cambridge, Mass.

Sandham, W., and M. Leggett (2003), *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*, 348 pp., Kluwer Acad., Norwell, Mass.

Shabbar, A., and A. G. Barnston (1996), Skill of seasonal climate forecasts in Canada using canonical correlation analysis, *Mon. Weather Rev.*, *124*, 2370–2385.

Taner, M. T., T. Berge, J. A. Walls, M. Smith, G. Taylor, D. Dumas, and M. B. Carr (2001), Well log calibration of Kohonen-classified seismic attributes using Bayesian logic, *J. Pet. Geol.*, *24*, 405–416.

Tang, B. Y., W. W. Hsieh, A. H. Monahan, and F. T. Tangang (2000), Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial Pacific sea surface temperatures, *J. Clim.*, *13*, 287–293.

Tang, Y., and W. W. Hsieh (2002), Hybrid coupled models of the tropical Pacific-ENSO prediction, *Clim. Dyn.*, *19*, 343–353.

Tang, Y., and W. W. Hsieh (2003), Nonlinear modes of decadal and interannual variability of the subsurface thermal structure in the Pacific Ocean, *J. Geophys. Res.*, *108*(C3), 3084, doi:10.1029/2001JC001236.

Vapnik, V. N. (1998), *Statistical Learning Theory*, 736 pp., John Wiley, Hoboken, N. J.

Villmann, T., E. Merenyi, and B. Hammer (2003), Neural maps in remote sensing image analysis, *Neural Networks*, *16*, 389–403.

von Storch, H., and F. W. Zwiers (1999), *Statistical Analysis in Climate Research*, 484 pp., Cambridge Univ. Press, New York.

Watari, S. (1996), Separation of periodic, chaotic, and random components in solar activity, *Sol. Phys.*, *168*, 413–422.

Woodruff, S. D., R. J. Slutz, R. L. Jenne, and P. M. Steurer (1987), A comprehensive ocean-atmosphere data set, *Bull. Am. Meteorol. Soc.*, *68*, 1239–1250.

Wu, A., and W. W. Hsieh (2002), Nonlinear canonical correlation analysis of the tropical Pacific wind stress and sea surface temperature, *Clim. Dyn.*, *19*, 713–722, doi:10.1007/s00382-002-0262-8.

Wu, A., and W. W. Hsieh (2003), Nonlinear interdecadal changes of the El Niño-Southern Oscillation, *Clim. Dyn.*, *21*, 719–730, doi:10.1007/s00382-003-0361-1.

Wu, A., W. W. Hsieh, and A. Shabbar (2002), Nonlinear characteristics of the surface air temperature over Canada, *J. Geophys. Res.*, *107*(D21), 4571, doi:10.1029/2001JD001090.

Wu, A., W. W. Hsieh, and F. W. Zwiers (2003), Nonlinear modes of North American winter climate variability detected from a general circulation model, *J. Clim.*, *16*, 2325–2339.

Yacoub, M., F. Badran, and S. Thiria (2001), A topological hierarchical clustering: Application to ocean color classification, in *Artificial Neural Networks—ICANN 2001, International Conference, Vienna, Austria, August 21–25, 2001: Proceedings*, *Lect. Notes Comput. Sci.*, vol. 2130, edited by G. Dorffner, H. Bischof, and K. Hornik, pp. 492–499, Springer-Verlag, New York.

Yuval, (2000), Neural network training for prediction of climatological time series, regularized by minimization of the generalized cross validation function, *Mon. Weather Rev.*, *128*, 1456–1473.

Yuval (2001), Enhancement and error estimation of neural network prediction of Niño 3.4 SST anomalies, *J. Clim.*, *14*, 2150–2163.

Yuval, and W. W. Hsieh (2002), The impact of time-averaging on the detectability of nonlinear empirical relations, *Q. J. R. Meteorol. Soc.*, *128*, 1609–1622.

Yuval, and W. W. Hsieh (2003), An adaptive nonlinear MOS scheme for precipitation forecasts using neural networks, *Weather Forecasting*, *18*, 303–310.

W. W. Hsieh, Department of Earth and Ocean Sciences, University of British Columbia, 6339 Stores Road, Vancouver, BC V6T 1Z4, Canada. (whsieh@eos.ubc.ca)