Nonlinear canonical correlation analysis of the tropical Pacific climate variability using a neural network approach

William W. Hsieh

Oceanography/EOS, University of British Columbia,
Vancouver, B.C. V6T 1Z4, Canada
tel: (604) 822-2821, fax: (604) 822-6091
email: whsieh@eos.ubc.ca

J. Climate (in press)

Abstract

Recent advances in neural network modeling have led to the nonlinear generalization of classical multivariate analysis techniques such as principal component analysis and canonical correlation analysis (CCA). The nonlinear canonical correlation analysis (NLCCA) method is used to study the relationship between the tropical Pacific sea level pressure (SLP) and sea surface temperature (SST) fields. The first mode extracted is a nonlinear El Niño-Southern Oscillation (ENSO) mode, showing the asymmetry between the warm El Niño states and the cool La Niña states. The nonlinearity of the first NLCCA mode is found to increase gradually with time. During 1950-1975, the SLP showed no nonlinearity, while the SST revealed weak nonlinearity. During 1976-1999, the SLP displayed weak nonlinearity, while the weak nonlinearity in the SST was further enhanced. The second NLCCA mode displays longer time-scale fluctuations, again with weak but noticeable nonlinearity in the SST but not in the SLP.

1 Introduction

Classical multivariate statistical methods, e.g. principal component analysis (PCA) and canonical correlation analysis (CCA) are widely used for data analysis in many fields, including meteorology and oceanography (von Storch and Zwiers 1999). For a set of variables $\{x_i\}$, PCA (also known as empirical orthogonal function analysis) extracts the eigenmodes of the data covariance matrix. It is used to (i) reduce the dimensionality of the dataset, and (ii) extract features from the dataset. When there are two sets of variables $\{x_i\}$ and $\{y_j\}$, CCA finds the modes of maximum correlation between $\{x_i\}$ and $\{y_j\}$, rendering CCA a standard tool for discovering relations between two fields (Barnett and Preisendorfer, 1987).

Recent advances in neural network (NN) modeling have led to the nonlinear generalization of PCA and CCA. Nonlinear principal component analysis (NLPCA) using NN was first introduced by Kramer (1991) in the chemical engineering literature, and is now used in many fields. Nonlinear canonical correlation analysis (NLCCA) was recently introduced by Hsieh (2000) using an NN approach. While NLCCA has been demonstrated with synthetic data, it has yet to be applied to real data. This paper examines the problems encountered when applying NLCCA to real data— the tropical Pacific sea level pressure (SLP) and sea surface temperature (SST) fields— and extracts from the data, the nonlinear coupled modes, including the famous tropical Pacific climate variability known as the El Niño-Southern Oscillation (ENSO). ENSO has warm El Niño states and cool La Niña states, with changes found not only in the SST, but also in the SLP. The previous NN studies of the tropical Pacific by our group used NN only for nonlinear regression (Tangang et al. 1997; Tangang et al. 1998a; Tangang et al. 1998b; Tang et al. 2000, Yuval, 2000, Tang et al. 2001), and for NLPCA (Monahan, 2001 and Hsieh, 2001).

This paper is organized as follows: The theory of NLCCA is presented in Sec.2. While the theory largely follows Hsieh(2000), there are several improvements for working with real data, e.g. scaling of the input variables, normalization of the canonical variates, and the addition of weight penalty terms to the cost functions to avoid overfitting. The NLCCA is applied to the tropical Pacific SLP and SST fields to extract the first mode in Sec.3, and the second mode in Sec.4.

2 Theory of nonlinear canonical correlation analysis

Consider two data sets $\{x_i(t)\}$ and $\{y_j(t)\}$, where t is the time, or simply a label of the particular sample. Assume there is a total of N samples in t for each variable $x_i(t)$ and $y_j(t)$. We group the

$\{x_i(t)\}$ variables to form the vector $\mathbf{x}(t)$, and $\{y_j(t)\}$ to $\mathbf{y}(t)$. CCA looks for linear combinations

$$u(t) = \mathbf{a} \cdot \mathbf{x}(t)\,, \quad \text{and} \quad v(t) = \mathbf{b} \cdot \mathbf{y}(t)\,, \tag{1}$$

where the canonical variates u and v have maximum correlation, i.e. the weight vectors $\mathbf{a}$ and $\mathbf{b}$ are chosen such that $\mathrm{cor}(u, v)$, the Pearson correlation coefficient between u and v, is maximized. CCA is widely used in meteorological/oceanographic studies (Barnett and Preisendorfer 1987; Barnston and Ropelewski 1992; Shabbar and Barnston 1996).

In NLCCA, we follow the same procedure as in CCA, except that the linear mappings in (1) are replaced by nonlinear mapping functions using NNs. The mappings from $\mathbf{x}$ to u and $\mathbf{y}$ to v are represented by the double-barreled NN on the left hand side of Fig.1. The inputs $\mathbf{x}$ and $\mathbf{y}$ are mapped to the neurons (i.e. variables) in the hidden layer:

$$h_k^{(x)} = \tanh((\mathbf{W}^{(x)}\mathbf{x} + \mathbf{b}^{(x)})_k), \quad h_n^{(y)} = \tanh((\mathbf{W}^{(y)}\mathbf{y} + \mathbf{b}^{(y)})_n), \tag{2}$$

where $\mathbf{W}^{(x)}$ and $\mathbf{W}^{(y)}$ are weight matrices, $\mathbf{b}^{(x)}$ and $\mathbf{b}^{(y)}$, bias parameter vectors, and k, n, indices of the vector elements (with the capital bold font reserved for matrices and the small bold font for vectors). The hyperbolic tangent function is used as the transfer function (see Bishop, 1995, Sec.4.3 for a discussion on the choice of transfer functions). 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.

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

$$u = \mathbf{w}^{(x)} \cdot \mathbf{h}^{(x)} + \overline{b}^{(x)}\,, \quad v = \mathbf{w}^{(y)} \cdot \mathbf{h}^{(y)} + \overline{b}^{(y)}\,. \tag{3}$$

These mappings are standard feedforward 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 (Bishop, 1995, Sec.4.3; Cybenko, 1989).

To maximize $\mathrm{cor}(u, v)$, 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 $\overline{b}^{(y)}$. Without loss of generality, u and v are required to have zero mean; then $\overline{b}^{(x)}$ and $\overline{b}^{(y)}$ are no longer free parameters, with

$$\overline{b}^{(x)} = -\langle \mathbf{w}^{(x)} \cdot \mathbf{h}^{(x)} \rangle \quad \text{and} \quad \overline{b}^{(y)} = -\langle \mathbf{w}^{(y)} \cdot \mathbf{h}^{(y)} \rangle\,, \tag{4}$$

where $\langle \ldots \rangle$ denotes the sample or time mean. We also adopt the normalization conditions $\langle u^2 \rangle = \langle v^2 \rangle = 1$, which are approximately satisfied by modifying the cost function to

$$J = -\mathrm{cor}(u, v) + (\langle u^2 \rangle^{1/2} - 1)^2 + (\langle v^2 \rangle^{1/2} - 1)^2\,. \tag{5}$$

Incidentally, compared to $(\langle u^2 \rangle^{1/2} - 1)^2$, the simpler normalization term $(\langle u^2 \rangle - 1)^2$ is not always effective, as it yielded very small values of $\langle u^2 \rangle$ for some runs. The gradient of the function $f(u) = (u^2 - 1)^2$, at $u = 0$ is 0. In contrast, the gradient of $((u^2)^{1/2} - 1)^2$ is nonzero at $u = 0$, thus forcing the solution to move away from $u = 0$.

There is a variant of CCA known as 'SVD' (Singular Value Decomposition) (Bretherton et al., 1992, Newman and Sardeshmukh, 1995). Noting that the name 'SVD' is used to denote both a statistical method and a matrix decomposition method, von Storch and Zwiers (1999) proposed using the less confusing name MCA (Maximum Covariance Analysis) for the statistical method. The difference between MCA and CCA lies in the fact that MCA maximizes the covariance $\mathrm{cov}(u, v)$, while CCA maximizes the correlation. Using the cost function $J = -\mathrm{cov}(u, v)$ yielded unbounded values for u

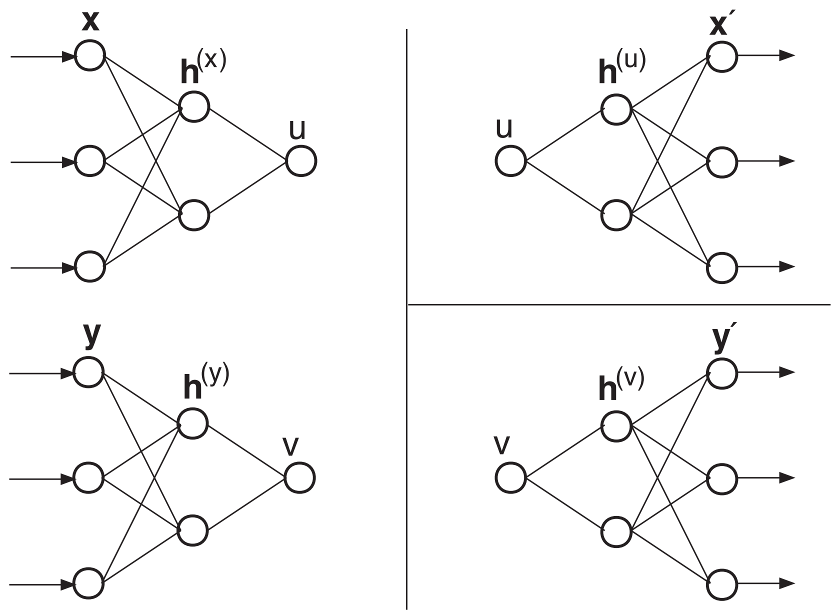

Figure 1: The three NNs used to perform NLCCA. The double-barreled NN on the left maps from the inputs $\mathbf{x}$ and $\mathbf{y}$ to the canonical variates u and v. Starting from the left, there are l_1 input $\mathbf{x}$ variables ('neurons' in NN jargon), denoted by circles. The information is then mapped to the next layer (to the right)— a 'hidden' layer $\mathbf{h}^{(x)}$ (with l_2 neurons). For input $\mathbf{y}$, there are m_1 neurons, followed by a hidden layer $\mathbf{h}^{(y)}$ (with m_2 neurons). The mappings continued onto u and v. 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 a hidden layer $\mathbf{h}^{(u)}$ (with l_2 neurons), followed by the output layer $\mathbf{x}'$ (with l_1 neurons). The cost function J_1 minimizes the MSE (mean square error) of $\mathbf{x}'$ relative to $\mathbf{x}$. The third NN maps from v to a hidden layer $\mathbf{h}^{(v)}$ (with m_2 neurons), followed by the output layer $\mathbf{y}'$ (with m_1 neurons). The cost function J_2 minimizes the MSE of $\mathbf{y}'$ relative to $\mathbf{y}$. When applied to the tropical Pacific, the inputs $\mathbf{x}$ were the first 6 PCs of the SLP field, and the inputs $\mathbf{y}$, the first 6 PCs of the SST. For some runs, instead of 6 pairs of PCs, only 3 pairs were used as inputs.

and v, as the network provided no constraints on the magnitudes of u and v when $\mathrm{cov}(u, v)$ was being maximized; in contrast, $J = -\mathrm{cor}(u, v)$ was well-behaved. With the normalization terms added to the cost function in (5), replacing $\mathrm{cor}(u, v)$ by $\mathrm{cov}(u, v)$ does not lead to significantly different results. Thus, it does not seem possible to have a nonlinear MCA method distinctly different from NLCCA.

After the forward mapping with the double-barreled NN has been solved, inverse mappings from the canonical variates to the original variables must then be found. On the right side of Fig.1, the top NN (a standard feed-forward NN) maps from u to $\mathbf{x}'$ in two steps:

$$h_k^{(u)} = \tanh((\mathbf{w}^{(u)} u + \mathbf{b}^{(u)})_k)\,, \quad \text{and} \quad \mathbf{x}' = \mathbf{W}^{(u)}\mathbf{h}^{(u)} + \overline{\mathbf{b}}^{(u)}. \tag{6}$$

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 (mean square error) between the NN output $\mathbf{x}'$ and the original data $\mathbf{x}$ is thus minimized.

Similarly, the bottom NN on the right side of Fig.1 maps from v to $\mathbf{y}'$:

$$h_n^{(v)} = \tanh((\mathbf{w}^{(v)} v + \mathbf{b}^{(v)})_n)\,, \quad \text{and} \quad \mathbf{y}' = \mathbf{W}^{(v)}\mathbf{h}^{(v)} + \overline{\mathbf{b}}^{(v)}, \tag{7}$$

with the cost function $J_2 = \langle \|\mathbf{y}' - \mathbf{y}\|^2 \rangle$ minimized. The total number of free parameters used by the NLCCA is $2(l_1 l_2 + m_1 m_2) + 4(l_2 + m_2) + l_1 + m_1$. For most datasets in meteorology and oceanography, a pre-filter of the datasets using PCA is needed, i.e. the first few principal component (PC) time series for $\mathbf{x}$ and the first few PCs for $\mathbf{y}$ are used as inputs to the NLCCA.

The nonlinear optimizations for the three NNs were all carried out by the function 'fminu' in the MATLAB Optimization Toolbox. An ensemble of 30 NNs mapping from $(\mathbf{x}, \mathbf{y})$ to (u, v), using random initial parameters, was run. The NN attaining the highest $\mathrm{cor}(u, v)$ was selected as the solution. Next a random ensemble of 30 NNs (mapping from u to $\mathbf{x}'$) was used to find the solution with the smallest MSE in $\mathbf{x}'$. Finally, another ensemble of 30 was used to find the NN yielding the smallest MSE in $\mathbf{y}'$. For noisy data, overfitting can be avoided by reserving some data as test data, and rejecting ensemble members which perform poorer on the test data than on the training data. 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.

That the CCA is indeed a linear version of this NLCCA can be readily seen by replacing the hyperbolic tangent transfer functions in (2), (6) and (7) 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.

The nonlinear optimization algorithm is inaccurate if the optimal parameters to be determined have a wide range of magnitudes. Hsieh (2001) examined what would happen to NLPCA if the input variables $\mathbf{x}$ were scaled by a factor α. If the original weight parameters are all of order 1, replacing $\mathbf{x}$ by $\alpha\mathbf{x}$ would cause some parameters to scale by α and some by α^{-1}, resulting in an increase of α^2 in the range of magnitudes. With NLCCA, because separate networks are used for the forward mapping and the inverse mappings, a similar argument as in Hsieh (2001) would show that, if $\mathbf{x}$ and $\mathbf{y}$ were replaced by $\alpha\mathbf{x}$ and $\alpha\mathbf{y}$, there would only be an increase of α in the range of magnitudes during a network optimization. Even though NLCCA is less sensitive to scaling than NLPCA, it is nevertheless a good idea to scale the input variables appropriately. One possibility is to standardize all the input variables, i.e. for each variable, remove its mean and divide by it standard deviation. If the input variables are themselves the leading principal component (PC) time series (i.e. PCA has been used to compact the dataset), then it would be appropriate to normalize each input variable x_i by subtracting its mean and dividing by the standard deviation of the first PC of $\mathbf{x}$; and similarly for the y_j variables.

The use of excessive nonlinearity to arrive at an 'overfitted' solution (i.e. a wiggly solution fitted to the noise in the data) is a well-known problem in NN modeling (Hsieh and Tang, 1998). With three NNs in NLCCA, overfitting can occur in any of the three networks. Regularization by adding weight penalty terms to the cost function is a common way to prevent overfitting in NNs (Bishop, 1995, Sec.5.4). The three cost functions are modified to:

$$J = -\mathrm{cor}(u, v) + (\langle u^2 \rangle^{1/2} - 1)^2 + (\langle v^2 \rangle^{1/2} - 1)^2 + p \left[\sum_{ki} (W_{ki}^{(x)})^2 + \sum_{nj} (W_{nj}^{(y)})^2 \right] , \tag{8}$$

$$J_1 = \langle \|\mathbf{x}' - \mathbf{x}\|^2 \rangle + p_1 \sum_k (w_k^{(u)})^2 , \tag{9}$$

$$J_2 = \langle \|\mathbf{y}' - \mathbf{y}\|^2 \rangle + p_2 \sum_n (w_n^{(v)})^2 , \tag{10}$$

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. The function tanh has the property that given x in the interval $[-L, L]$, one can find a small enough weight w, so that $\tanh(wx) \approx wx$, i.e. an almost linear transfer function results from using a small enough weight w. By increasing the weight penalty parameter, one can suppress the use of weights with large magnitudes, thereby preventing excessively nonlinear solutions. Using a weight penalty also reduces the sensitivity of the solution to the number of hidden neurons— the reason is that with a large number of hidden neurons, the abundant parameters lead to excessively nonlinear solutions, unless a weight penalty suppresses the excessive nonlinear capability of the network.

3 First NLCCA mode for the tropical Pacific

The tropical Pacific monthly SLP data from COADS (Comprehensive Ocean-Atmosphere Data Set) (Woodruff et al. 1987) for Jan., 1950 to June, 2000 were used. The 2° by 2° resolution data were combined into 10° longitude by 4° latitude gridded data, with climatological seasonal cycle removed, and smoothed by a 3-month running average. PCA of the data resulted in the first six modes accounting for 29.5%, 16.5%, 4.8%, 3.4%, 2.5% and 2.2%, respectively, of the total variance.

The tropical Pacific monthly SST data from NOAA (Smith et al. 1996) for the same period (where the original 2° by 2° resolution data had been combined into 4° by 4° gridded data, with climatological seasonal cycle removed, and smoothed by a 3-month running average) were used. PCA resulted in the first six modes accounting for 51.8%, 10.1%, 7.3%, 4.3%, 3.5% and 3.1%, respectively, of the total SST variance.

For the NLCCA, the inputs $\mathbf{x}$ are the first six PCs of the SLP field, while the inputs $\mathbf{y}$ are the first six PCs of the SST field, i.e. the NLCCA architecture has $l_1 = m_1 = 6$. For the number of hidden neurons, some experimentation with l_2 and m_2 was needed. For $l_2 = m_2 = 1$, the solutions were found to be essentially the same as the linear CCA solutions. To get nonlinear solutions, l_2 and m_2 both have to be at least 2. Following the principle of parsimony, I chose $l_2 = m_2 = 2$. The penalty parameters are generally chosen to be as small as possible while avoiding overfitting. The values $p = p_1 = p_2 = 0.1$ were used.

The first NLCCA mode is shown in Fig.2 for the SLP and in Fig.3 for the SST. For SLP, in the PC1-PC2 plane (Fig.2a), the El Niño states are in the upper right corner (high u values), while the La Niña states are in the left corner (low u values). Fig.2d offers a full 3-D view of the first NLCCA mode in the PC1-PC2-PC3 space. For comparison, the linear solution (CCA mode 1) is shown as a thin straight line.

For the SST, in the PC1-PC2 plane (Fig.3a), the first NLCCA mode is a U-shaped curve linking the La Niña states in the upper left corner (low v values) to the El Niño states in the upper right

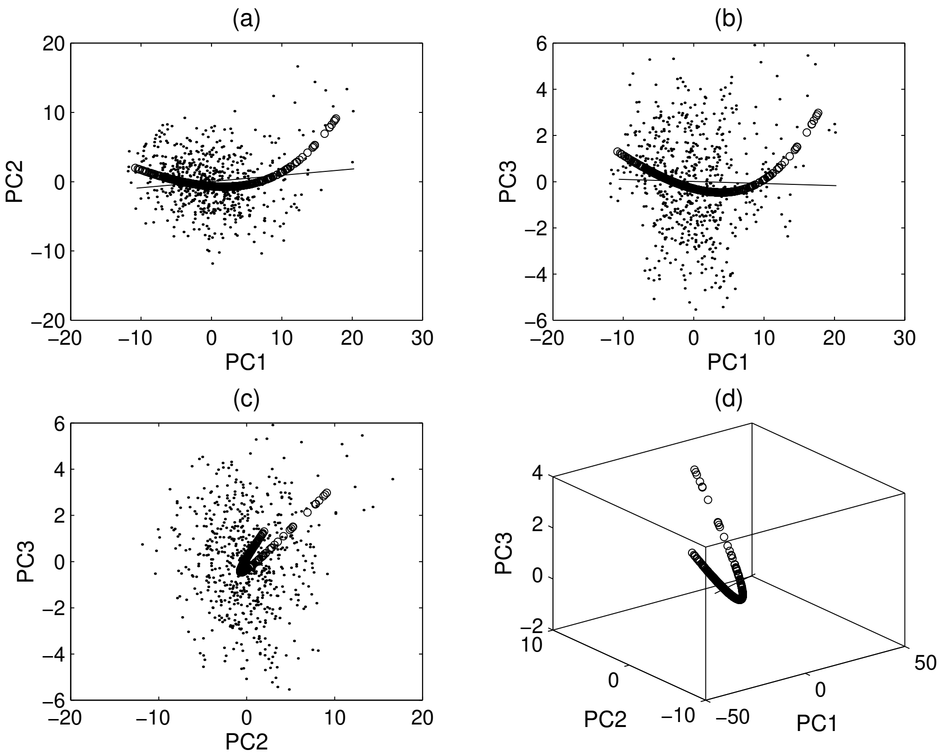

Figure 2: The NLCCA mode 1 in the SLP PC-space (where only PC1, PC2 and PC3 of the 6 SLP PCs are shown). Four panels are used to show the (a) PC1-PC2 plane, (b) the PC1-PC3 plane, (c) the PC2-PC3 plane, and (d) the 3-dimensional PC1-PC2-PC3 space. The data are shown as dots, and the projection of the data onto the first NLCCA mode gives the string of small (overlapping) circles. As the canonical variate u moves from its minimum to its maximum, the SLP system moves from one end of the string to the other, i.e. from La Niña to El Niño. In panel (a), the La Niña states are to the left corner, and the El Niño states to the upper right corner. The CCA mode 1 is shown as the thin straight line.

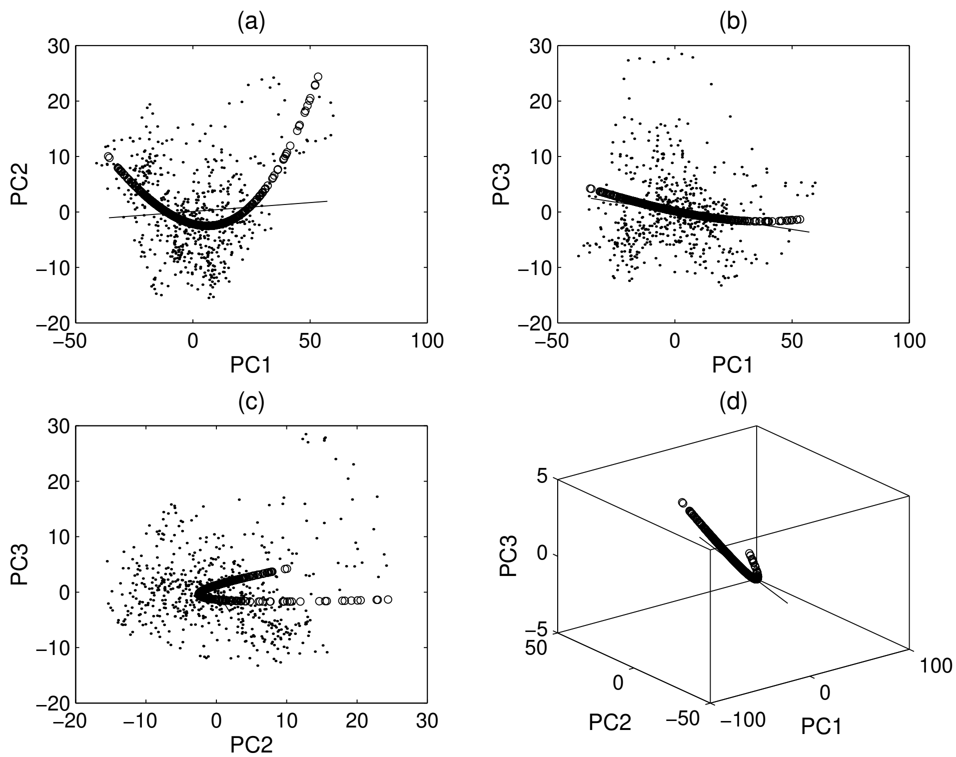

Figure 3: Similar to Fig.2, but for the SST data (where PC1, PC2 and PC3 of the 6 PCs are shown). As the canonical variate v moves from its minimum to its maximum, the SST system moves from one end of the string to the other, i.e. from La Niña states at the upper left corner to El Niño states at the upper right corner of panel (a). The CCA mode 1 is shown as the thin straight line.

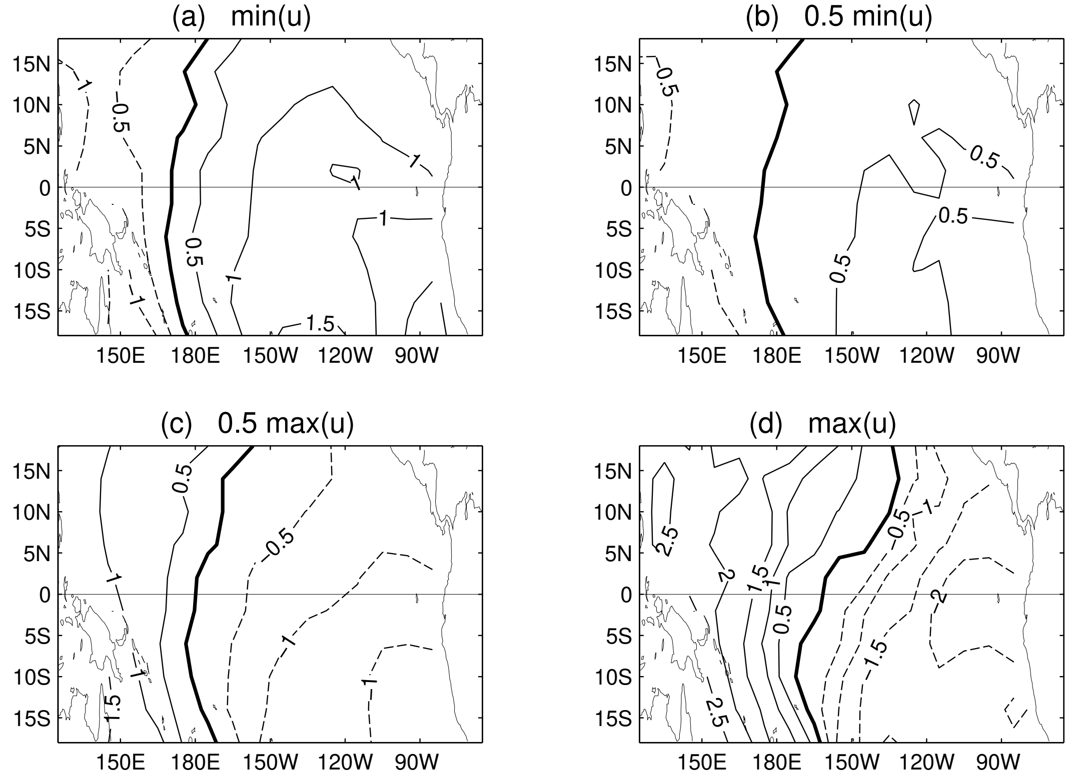

Figure 4: The SLP field as the canonical variate u of the first NLCCA mode varies from (a) its minimum (strong La Niña), to (b) half its minimum (weak La Niña), to (c) half its maximum (weak El Niño) and (d) its maximum (strong El Niño). Contour interval is 0.5 mb.

corner (high v values). In general, the nonlinearity is greater in the SST than in the SLP, as the difference between the CCA mode and the NLCCA mode is greater in Fig.3a than in Fig.2a. The MSE of the NLCCA divided by the MSE of the CCA is 0.951 for the SLP and 0.935 for the SST, confirming that the mapping for the SST was more nonlinear than that for the SLP.

For a given value of u, one can map from u to the 6 PCs of SLP. Each of the PCs can be multiplied by its associated PCA (spatial) eigenvector (also known as the empirical orthogonal function), and the six modes added together to yield the spatial anomaly pattern for that particular value of u. For the first NLCCA mode, as u varies from its minimum value to its maximum value, the SLP field varies from the strong La Niña phase of the Southern Oscillation to the strong El Niño phase of the Southern Oscillation (Fig.4). The zero contour is further west during La Niña (Fig.4a) than during strong El Niño (Fig.4d), though the eastward shift of the zero contour mainly occurs between Fig.4c and Fig.4d, i.e. between weak El Niño and strong El Niño. Similarly, as v varies from its minimum to its maximum, the SST field varies from strong La Niña to strong El Niño (Fig.5), revealing that the SST anomalies during La Niña are centered further west of the anomalies during El Niño.

For comparison, the CCA mode 1 spatial patterns for the SLP and SST are shown in Fig.6. The CCA first mode is not capable of simulating the asymmetry between El Niño and La Niña, as the mode involves a fixed pattern of SLP (Fig.6a) multiplied by $u(t)$ and a fixed pattern of SST (Fig.6b) multiplied by $v(t)$.

The first mode canonical variate time series u and v, for the NLCCA and for the CCA, reveal that

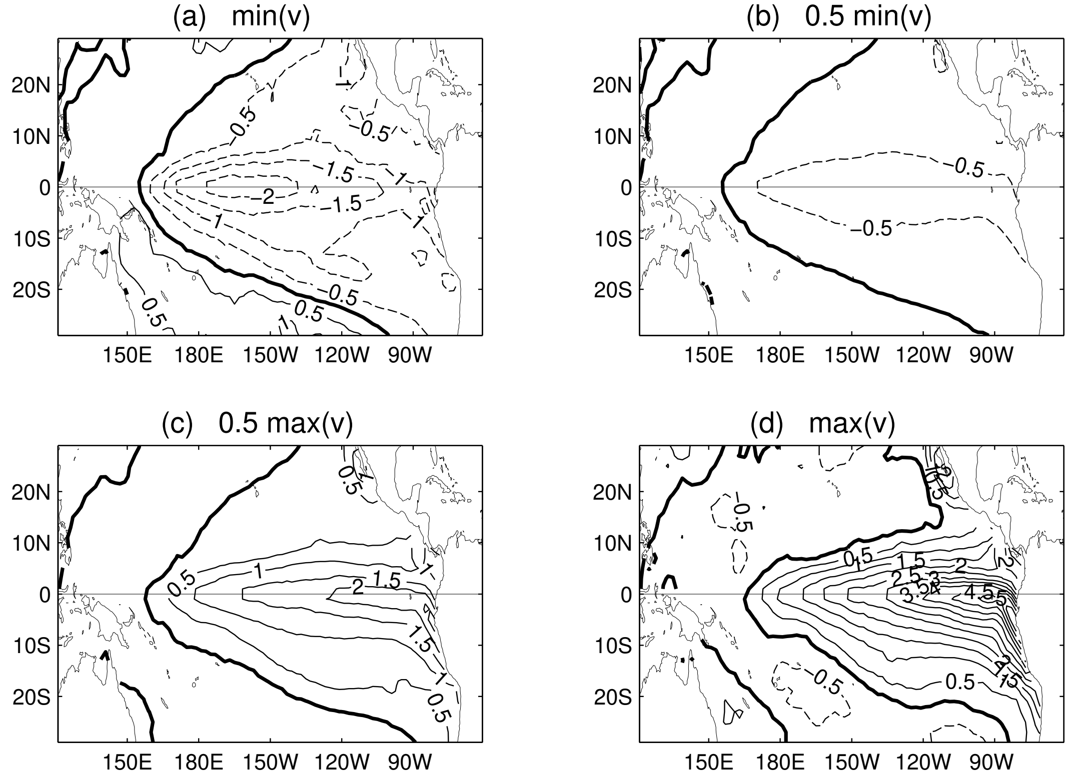

Figure 5: Similar to Fig.4 but for the SST field, as the canonical variate v varies from (a) its minimum (strong La Niña), to (b) half its minimum (weak La Niña), to (c) half its maximum (weak El Niño) and (d) its maximum (strong El Niño). Contour interval is 0.5°C.

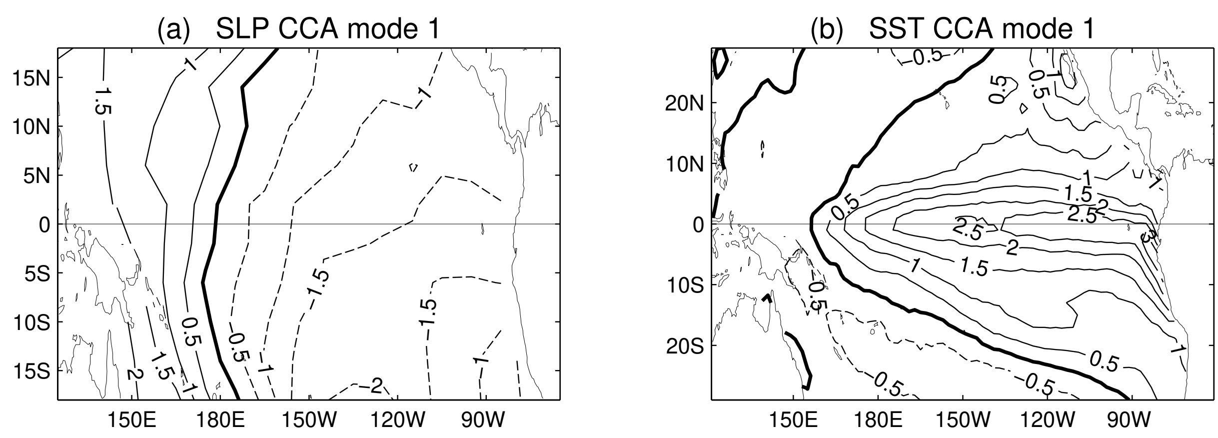

Figure 6: The CCA mode 1 for (a) the SLP and (b) the SST. The pattern in (a) is scaled by $u_1 = [\max(u) - \min(u)]/2$, and (b) by $v_1 = [\max(v) - \min(v)]/2$. Contour interval is 0.5 mb in (a) and 0.5°C in (b).

Table 1: Ratios between the MSE of the NLCCA and of the CCA for various periods. A ratio around 1 means that the nonlinear solution is almost identical to the linear solution, whereas a ratio much less than 1 implies the nonlinear solution has improved on the linear solution.

Period	MSE (SLP)	MSE (SST)
1950-75	1.027	0.847
1976-99	0.879	0.759
1950-59	1.012	0.987
1960-69	1.028	0.923
1970-79	1.011	0.846
1980-89	0.857	0.948
1990-99	0.862	0.613

they are all rather similar (Fig.7), in that El Niño events show up as peaks and La Niña events, as troughs. The correlation between u and v is 0.954 for the NLCCA versus 0.949 for the CCA.

Also shown in Fig.7 are the u and v values of an NLCCA run with all the weight penalty parameters set to zero. Serious overfitting is evident, in that u and v only depart significantly from their bottom values for three big El Niño events, while achieving an extremely high correlation of 0.995 (cf 0.954 for the penalty case). This superficially high correlation was costly, as the inverse maps fared poorly, attaining a ratio of MSE (relative to the CCA model) of 1.51 for SLP and 1.38 for SST — i.e. the NLCCA with no penalty did much worse than the CCA model, due to severe 'overfitting' in the forward mapping from the inputs to the canonical variates u and v. Hence there is a need for weight penalty terms in the cost functions to prevent the use of excessively nonlinear mapping functions for datasets with outliers.

Another run was made using only the first 3 PCs of the SLP and the first 3 PCs of the SST as inputs (with the number of hidden neurons unchanged). The ratios between the MSE of the NLCCA and of the CCA are: 0.912 for SLP and 0.835 for SST. The MSE ratios are farther from 1 than in the previous run with 6 pairs of PCs, indicating that the advantage of NLCCA over CCA occurs mainly over the first 3 pairs of PCs, than over the last 3 pairs. When Figs.2-5 were redrawn for the run with 3 pairs of input PCs, there were only minor changes (not shown). Thus, in this case, the NLCCA mode 1 can be adequately extracted with only 3 pairs of PCs as inputs.

To see if there are gradual changes in the nonlinearity of the data, we ran the NLCCA with 3 pairs of PCs as inputs for two 25-year periods— 1950-1975 and 1976-1999— with the results shown in Table 1. During 1950-1975, SLP was linear, while SST was weakly nonlinear. During 1976-1999, SLP became weakly nonlinear, while the nonlinearity in the SST was enhanced from the previous period. Calculations were repeated for individual decades in Table 1. While a decade is a very short record, the findings from the 25-year period studies were supported by the decadal studies. These results suggest that the nonlinearity in the tropical Pacific climate system, albeit weak, may be increasing gradually with time.

Two caveats: (1) There is actually no proven relation between the nonlinearity found in the data and the nonlinearity in the governing dynamical equations. In fact, even for linear systems, there are no established relations between the linear dynamical modes and the empirical PCA modes. (2) The poorer quality of the earlier data, where missing data were interpolated by principal component reconstructions, could have a linearizing effect on the data (F. Zwiers, 2000, personal communication). On smaller spatial scales, this effect can be major, but on a scale as large as ENSO, it is unlikely that this effect alone can account for the marked decline of nonlinearity in the data for the earlier decades. The most probable explanation remains that nonlinear processes may have become more important in the tropical Pacific in recent decades.

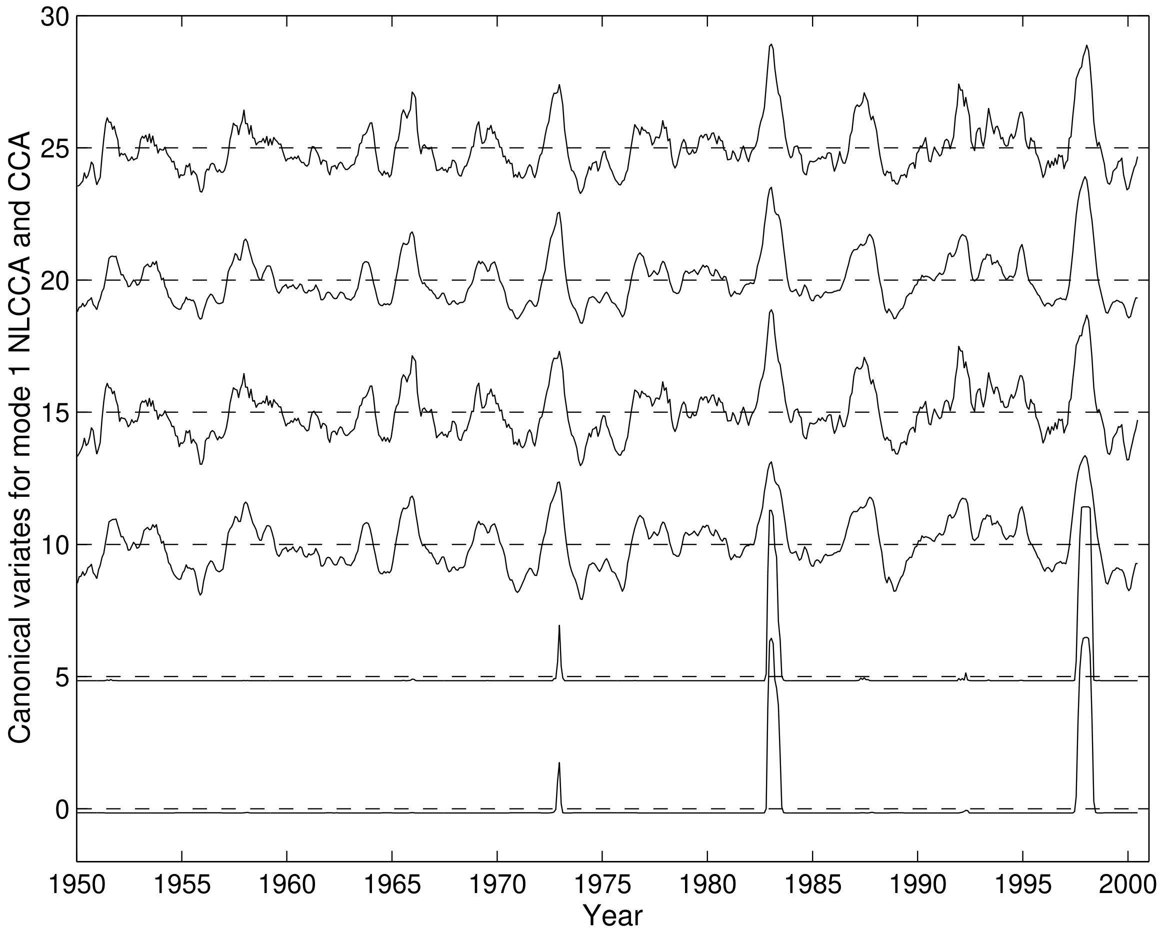

Figure 7: The canonical variates u and v plotted as a function of time from 1950 to 2000. The top curve is u (for SLP) from the first NLCCA mode, with v (for SST) immediately below it. Next, u and v from the first CCA mode are shown respectively, as the third and fourth curves from the top. The results of the first mode from the NLCCA model with no weight penalty terms is shown in the second curve from the bottom (u) and the bottom curve (v), illustrating serious overfitting when no penalty terms were used. For better visualization, the curves have been shifted by steps of 5 units from the bottom curve, with the dashed lines indicating the mean positions of the curves.

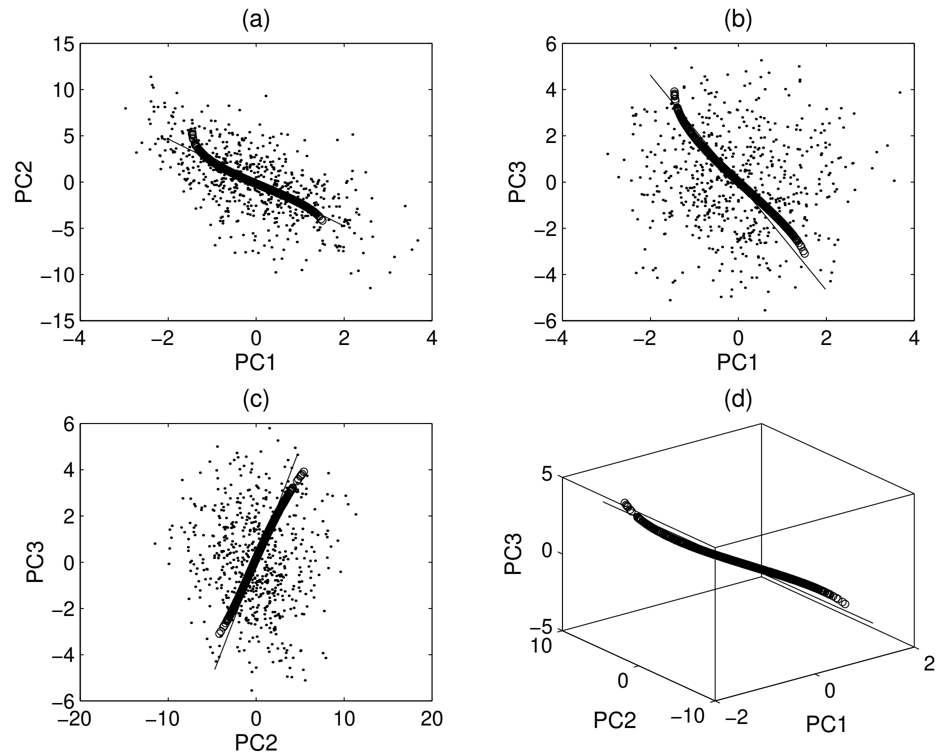

Figure 8: The NLCCA mode 2 (circles) in the SLP PC-space (where PC1, PC2 and PC3 of the 6 SLP PCs are shown). The dots show the residual data after the NLCCA mode 1 has been subtracted. The linear solution is shown as the thin straight line. (This linear solution to the dataset after NLCCA mode 1 has been removed is not the same as CCA mode 2, which is the linear solution to the dataset after CCA mode 1 has been removed.)

4 Second NLCCA mode for the tropical Pacific

After the first NLCCA mode had been extracted, the residual (i.e. the original data minus the first NLCCA mode) was served as input to the NLCCA network (with 6 pairs of input PCs) again to extract the second mode. The second mode in the SLP PC-space (Fig.8) shows a nearly linear solution. In contrast, nonlinearity can be seen in the SST PC-space (Fig.9). The canonical variates for NLCCA and CCA are plotted as a function of time in Fig.10, showing that mode 2 is 'noisier' than mode 1 (Fig.7), but it also has longer time-scale oscillations than the ENSO oscillations in mode 1. The low v state was found to last several years in the early 1950s, and from the late 1990s to the present (Fig.10). In fact, from 1976 to 1997, the ocean was in a generally high v state, with a sharp transition to the low v state in 1998, resembling a 'regime shift'.

The spatial patterns of the SLP associated with the NLCCA mode 2 (Fig.11) confirms the almost linear behaviour found in Fig.8. For instance, Fig.11a (at minimum u) is almost an exact mirror image of Fig.11d (at maximum u). The spatial patterns of the SST associated with the NLCCA mode 2 (Fig.12) reveal some differences between Fig.12a (at minimum v) and Fig.12d (at maximum v)— in Fig.12a, the cool anomalies in the central equatorial Pacific, the warm anomalies off Peru, and the warm anomalies around the northwest corner of the picture, are all more intense than the corresponding anomalies in Fig.12d. While it is always somewhat risky to infer the physics from empirical modes (Newman and Sardeshmukh, 1995), it seems plausible that mode 2 at low u and v, corresponds to an anomalous high pressure region centred north of the equator (Fig.11a), inducing

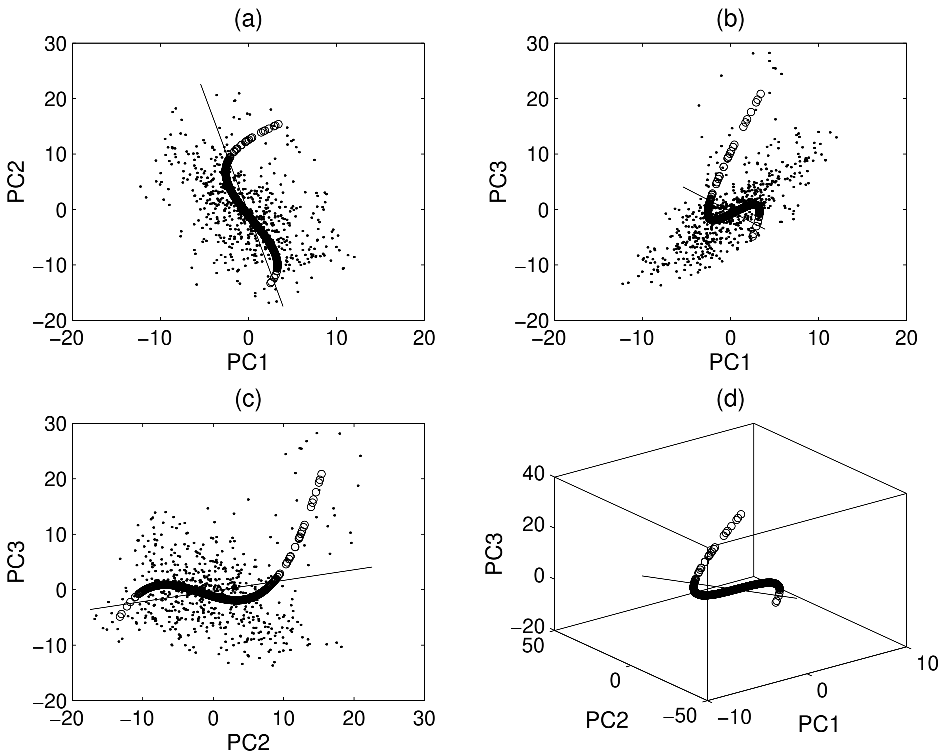

Figure 9: Similar to Fig.8, but for the SST data (where PC1, PC2 and PC3 of the 6 SST PCs are shown). The linear solution is shown as the thin straight line.

easterly winds and upwelling of cool waters along the equator (Fig.12a). For comparison, the CCA mode 2 is shown in Fig.13.

When only 3 pairs of PCs were used as inputs, the second NLCCA mode was quite different from that using 6 pairs of PCs. In particular, the longer time-scale variability is much weaker. Thus it appears that the higher PC modes are needed to extract the NLCCA mode 2 correctly. Another run with 8 pairs of PCs as input was performed; the resulting NLCCA mode 2 did not differ much from that obtained from 6 pairs of input PCs, thereby confirming that 6 pairs of input PCs were adequate for extracting the NLCCA mode 2.

5 Summary and discussion

In classical multivariate statistics, one has a hierarchy of tools: (1) multiple linear regression, (2) PCA, and (3) CCA. Neural network (NN) models have allowed the nonlinear generalization of all three: The standard feed-forward back-propagating model of Rumelhart et al. (1986), which was largely responsible for the surge of interest on neural networks in the late 1980s (Crick 1989), is a nonlinear regression model. PCA was nonlinearly generalized by the NLPCA model of Kramer (1991), and CCA by the NLCCA model of Hsieh (2000). (Codes for NLPCA and NLCCA are downloadable from the Web site http://www.ocgy.ubc.ca/projects/clim.pred).

When dealing with data from the tropical Pacific, it was found that the NLCCA needed proper scaling of the input variables, and regularization by adding weight penalty terms to the cost functions to prevent 'overfitting' (the use of excessively nonlinear mappings). In the tropical Pacific, the NLCCA applied to the SLP and SST fields found a nonlinear ENSO mode. The asymmetry between the warm

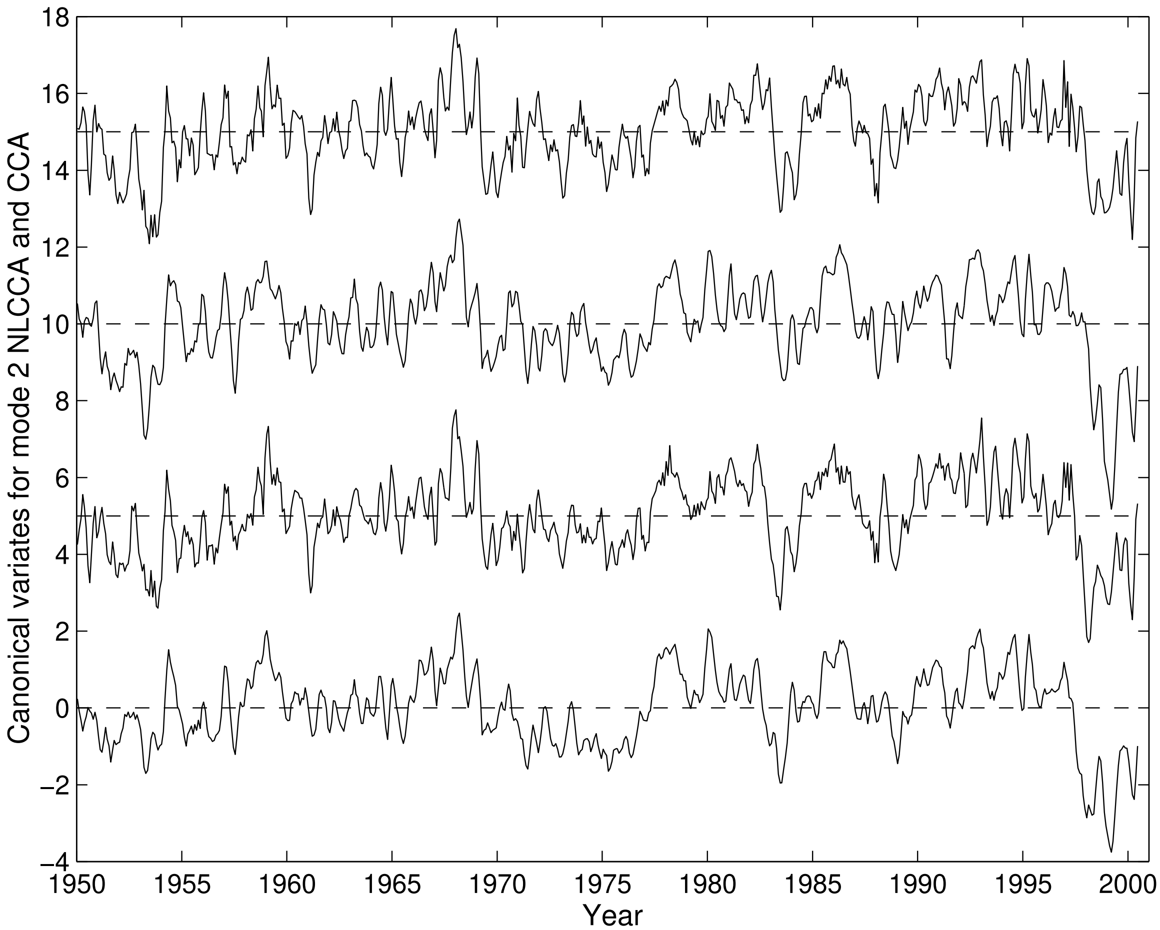

Figure 10: The canonical variates u and v plotted as a function of time. The top curve is u (for SLP) from the NLCCA mode 2, with v (for SST) immediately below it. Next, u and v from the CCA mode 2 are shown respectively, as the third and fourth curves from the top. For better visualization, the curves have been shifted by steps of 5 units.

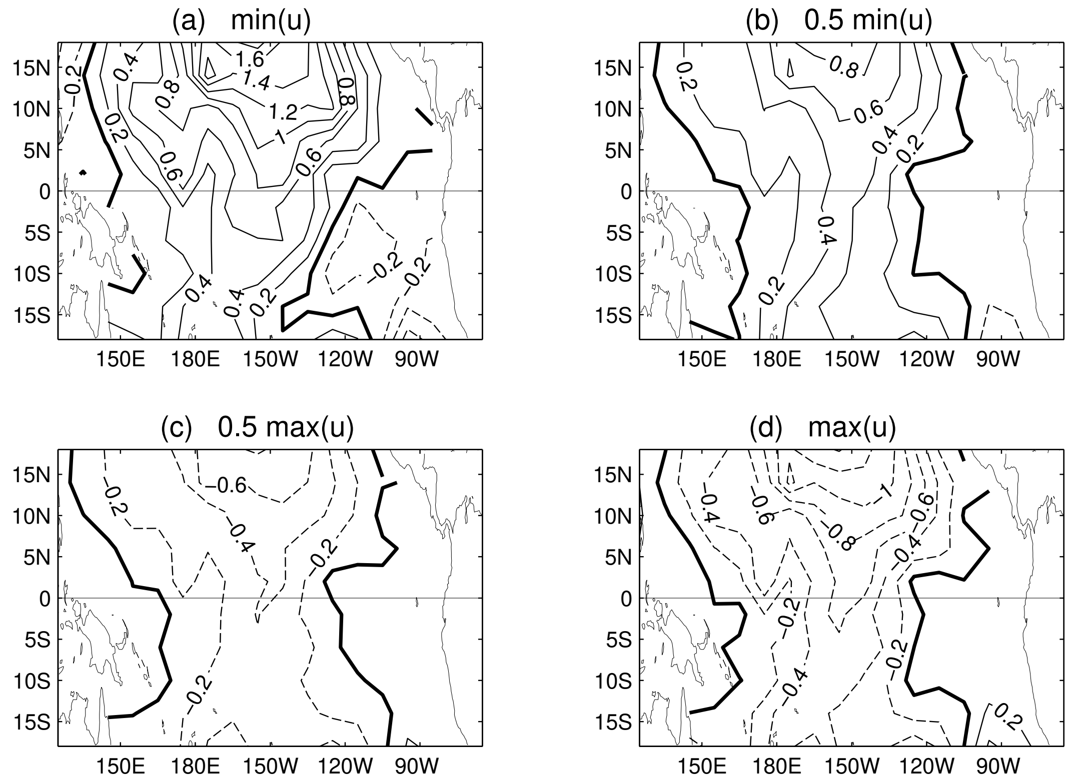

Figure 11: The SLP field as the canonical variate u of the NLCCA mode 2 varies from (a) its minimum, to (b) half its minimum, to (c) half its maximum and (d) its maximum. Contour interval is 0.2 mb.

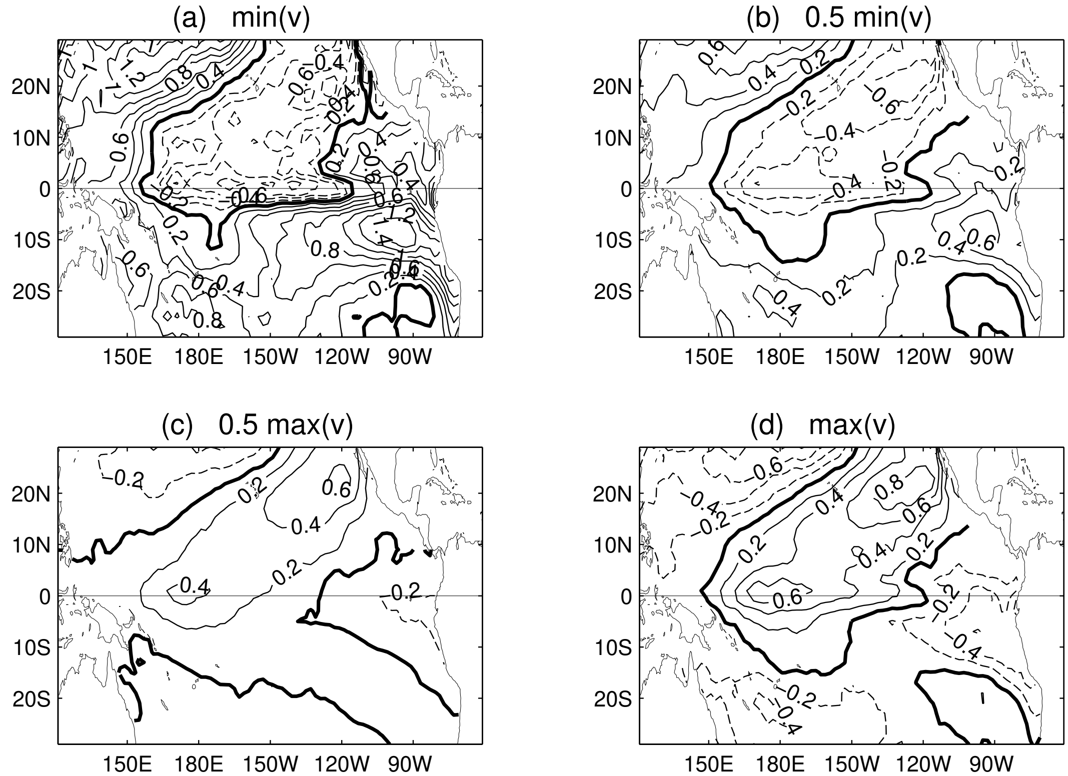

Figure 12: Similar to Fig.11 but for the SST field, as the canonical variate v varies from (a) its minimum, to (b) half its minimum, to (c) half its maximum, and (d) its maximum. Contour interval is 0.2°C.

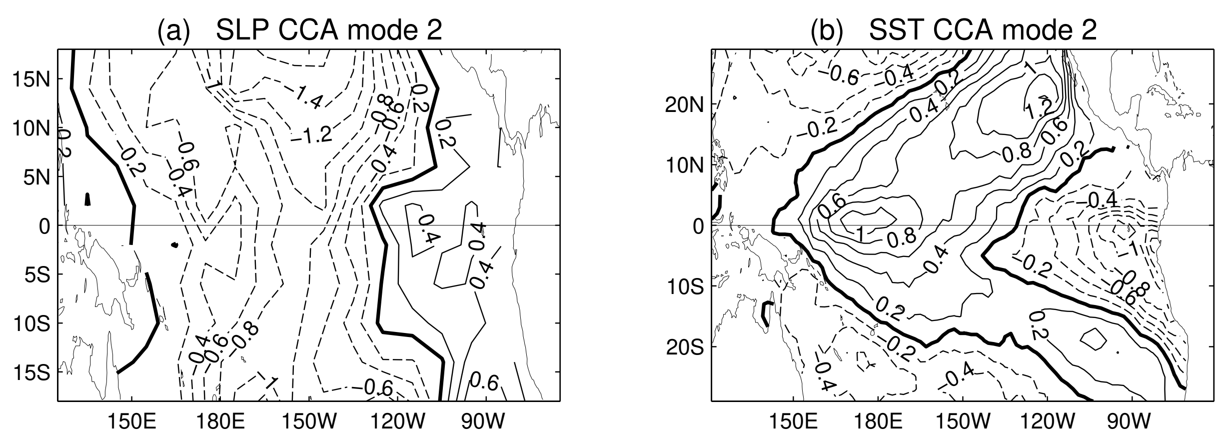

Figure 13: The CCA mode 2 for (a) the SLP and (b) the SST. Contour interval is 0.2 mb in (a) and 0.2°C in (b).

El Niño states and the cool La Niña states was well modeled by the NLCCA first mode, whereas the CCA first mode was incapable of modeling the asymmetry. Interestingly, even though the nonlinearity in the equatorial region is quite weak relative to the nonlinearity in the extratropical regions (Hoerling et al. 1997), the NLCCA managed to detect the weak nonlinearity, and showed that the SST is slightly more nonlinear than the SLP in the tropical Pacific. The nonlinearity of the first NLCCA mode was found to increase gradually with time. The second NLCCA mode showed longer time scale fluctuations, again with weak but noticeable nonlinearity in the SST but not in the SLP.

As the nonlinearity in the tropical Pacific is weak, the nonlinear power of the NLCCA and NLPCA is not very well demonstrated. Hsieh (2000 and 2001) provided examples with much stronger nonlinearity, and showed that the NN methods handled them well. Comparing the linear mappings used by PCA and CCA to the continuous (nonlinear) mappings by NLPCA and NLCCA, it is clear that NN techniques have considerably expanded our ability to empirically model multivariate datasets.

There are two disadvantages with nonlinear NN methods: (i) The presence of multiple minima in the cost functions. Even with an ensemble of optimization runs starting from random initial parameters, there is no guarantee that the best solution in the ensemble is close to the global minimum. However, this problem is greatly alleviated with the use of weight penalty regularization, which increases the concavity of the cost function, so that in theory a global minimum can always be found, provided a large enough weight penalty parameter is used. Of course, excessively large penalty parameters lead to linear solutions (and ultimately to constant solutions where the penalized weights are all zero). (ii) There are as yet no satisfactory ways to objectively determine the number of hidden neurons (l_2 and m_2) and the weight penalty parameters for the model. F. Zwiers (2000, personal communication) pointed out the potential of using Akaike Information Criterion (AIC) (von Storch and Zwiers, 1999) to choose the most appropriate values of l_2 and m_2. With 3 pairs of PCs as input to the NLCCA model, the AIC did choose $l_2 = m_2 = 2$ for mode 1, but with 6 pairs of PCs as input, AIC chose $l_2 = m_2 = 1$, i.e. an essentially linear solution— even though the $l_2 = m_2 = 2$ solution for 6 pairs of input PCs is very similar to that for 3 pairs of PCs. The appropriate choice for the number of input PCs to NLCCA is also not known, though from PCA theory, there are some guidelines (Preisendorfer, 1988). Hopefully, future research will provide more objective criteria for the network architecture.

Acknowlegments

Dr. Benyang Tang kindly sent me the SLP and SST datasets and his Matlab contouring package. Helpful comments were provided by Dr. Francis Zwiers of CCCma, and by Youmin Tang, Aiming Wu and Yuval of our research group. Support through research and strategic grants from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

References

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. Wea. Rev.*, **115,** 1825-1850.

Barnston, A.G., and C.F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate*, **5,** 1316-1345.

Bishop, C.M., 1995: *Neural Networks for Pattern Recognition.* Oxford, Clarendon Pr., 482 pp.

Bretherton, C.S., C. Smith, and J.M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate,* **5,** 541-560.

Crick, F., 1989: The recent excitement about neural networks. *Nature,* **337,** 129-132.

Cybenko, G., 1989: Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, **2,** 303-314.

Hoerling, M.P., A. Kumar, and M. Zhong, 1997: El Niño, La Niña and the nonlinearity of their teleconnections. *J. Climate*, **10,** 1769-1786.

Hsieh, W.W., 2000: Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, **13,** 1095-1105.

Hsieh, W.W., 2001: Nonlinear principal component analysis by neural networks. *Tellus*, (in press).

Hsieh, W.W., and B. Tang, 1998: Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Amer. Meteor. Soc.*, **79,** 1855-1870.

Kramer, M.A., 1991: Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, **37,** 233-243.

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

Newman, M., and P.D. Sardeshmukh, 1995: A caveat concerning singular value decomposition. *J. Climate*, **8,** 352-360.

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

Rumelhart, D.E., G.E. Hinton, and R.J. Williams, 1986: Learning internal representations by error propagation. *Parallel Distributed Processing*. D.E. Rumelhart, J.L. McClelland and PDP Research Group. Cambridge, MA, MIT Press. **1,** 318-362.

Shabbar, A., and A.G. Barnston, 1996: Skill of seasonal climate forecasts in Canada using Canonical Correlation Analysis. *Mon. Wea. Rev.*, **124,** 2370-2385.

Smith, T.M., R.W. Reynolds, R.E. Livezey, and D.C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, **9,** 1403-1420.

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. Climate*, **13,** 287-293.

Tang, Y., W.W. Hsieh, B. Tang, and K. Haines, 2001: A neural network atmospheric model for hybrid coupled modelling. *Clim. Dynam.*, **17,** 445-455.

Tangang, F.T., W.W. Hsieh, and B. Tang, 1997: Forecasting the equatorial Pacific sea surface temperatures by neural network models. *Clim. Dynam.*, **13,** 135-147.

Tangang, F.T., W.W. Hsieh, and B. Tang, 1998a: Forecasting the regional sea surface temperatures of the tropical Pacific by neural network models, with wind stress and sea level pressure as predictors. *J. Geophys. Res.*, **103,** 7511-7522.

Tangang, F.T., B. Tang, A.H. Monahan, and W.W. Hsieh, 1998b: Forecasting ENSO events— a neural network-extended EOF approach. *J. Climate*, **11,** 29-41.

von Storch, H., and F.W. Zwiers, 1999: *Statistical Analysis in Climate Research. Cambridge*, Cambridge Univ. Pr., 484 pp.

Woodruff, S.D., R.J. Slutz, R.L. Jenne, and P.M. Steurer, 1987: A Comprehensive Ocean-Atmosphere Data Set. *Bull. Amer. Meteor. Soc.*, **68,** 1239-1250.

Yuval, 2000: Neural network training for prediction of climatological time series; regularized by minimization of the Generalized Cross Validation function. *Mon. Wea. Rev.*, **128,** 1456- 1473.