Nonlin. Processes Geophys., 15, 221–232, 2008
www.nonlin-processes-geophys.net/15/221/2008/

Nonlinear Processes in Geophysics

Robust nonlinear canonical correlation analysis: application to seasonal climate forecasting

A. J. Cannon[1] **and W. W. Hsieh**[2]

[1]Meteorological Service of Canada, Environment Canada, 201-401 Burrard Street, Vancouver, BC V6C 3S5, Canada
[2]Dept. of Earth and Ocean Sciences, University of British Columbia, 6339 Stores Road, Vancouver, BC V6T 1Z4, Canada

Received: 3 September 2007 – Revised: 6 December 2007 – Accepted: 6 December 2007 – Published: 27 February 2008

Abstract. Robust variants of nonlinear canonical correlation analysis (NLCCA) are introduced to improve performance on datasets with low signal-to-noise ratios, for example those encountered when making seasonal climate forecasts. The neural network model architecture of standard NLCCA is kept intact, but the cost functions used to set the model parameters are replaced with more robust variants. The Pearson product-moment correlation in the double-barreled network is replaced by the biweight midcorrelation, and the mean squared error (mse) in the inverse mapping networks can be replaced by the mean absolute error (mae).

Robust variants of NLCCA are demonstrated on a synthetic dataset and are used to forecast sea surface temperatures in the tropical Pacific Ocean based on the sea level pressure field. Results suggest that adoption of the biweight midcorrelation can lead to improved performance, especially when a strong, common event exists in both predictor/predictand datasets. Replacing the mse by the mae leads to improved performance on the synthetic dataset, but not on the climate dataset except at the longest lead time, which suggests that the appropriate cost function for the inverse mapping networks is more problem dependent.

1 Introduction

Canonical correlation analysis (CCA) is a multivariate linear model used to find the modes of maximum correlation between two sets of variables (von Storch and Zwiers, 1999). CCA was first popularized as a tool for prediction in the atmospheric sciences by Glahn (1968) and has since been used extensively in climatology, particularly for seasonal forecasting (Barnett and Preisendorfer, 1987; Barnston and Ropelewski, 1992; Shabbar and Barnston, 1996).

Correspondence to: A. J. Cannon
(alex.cannon@ec.gc.ca)

CCA is a linear model and is thus unable to describe nonlinear relationships between datasets. To nonlinearly generalize CCA, various approaches, based on artificial neural network and kernel methods, have been proposed (Lai and Fyfe, 1999; Hsieh, 2000; Lai and Fyfe, 2000; Suykens et al., 2002; Melzer et al., 2003; Shawe-Taylor and Cristianini, 2004). For instance, Hsieh (2000) used three feed-forward (multi-layer perceptron) neural network mappings to perform nonlinear CCA (NLCCA). This method has been applied to climate research, for analyzing the structure of the El Niño-Southern Oscillation (ENSO) (Hsieh, 2001; Wu and Hsieh, 2002) and its interdecadal changes (Wu and Hsieh, 2003), and for determining the midlatitude atmospheric response to tropical Pacific sea surface temperature (SST) variability (Wu et al., 2003). Operational NLCCA forecasts of SST in the equatorial Pacific Ocean are also made available by the Climate Prediction Group of the University of British Columbia (see http://www.ocgy.ubc.ca/projects/clim.pred/ for more details).

While able to describe coupled nonlinear variability, this rather complicated NLCCA model is prone to overfitting (i.e., fitting to the noise rather than the signal), particularly when applied to the short, noisy datasets common in climate studies. This prompted the development of simpler multivariate nonlinear models such as nonlinear projection (Wu and Hsieh, 2004), which maps a univariate predictor to a multivariate predictand dataset, and nonlinear principal predictor analysis (Cannon, 2006), which maps a multivariate predictor dataset to a multivariate predictand dataset. While mitigating the influence of short, noisy datasets on model overfitting by reducing the number of neural networks in the model, neither nonlinear projection nor nonlinear principal predictor analysis are as general as NLCCA.

The main goal of this paper is the development of a robust version of NLCCA that can successfully operate on datasets with low signal-to-noise-ratios. The basic model architecture chosen by Hsieh (2000) is kept intact. Instead, the cost

Published by Copernicus Publications on behalf of the European Geosciences Union and the American Geophysical Union.

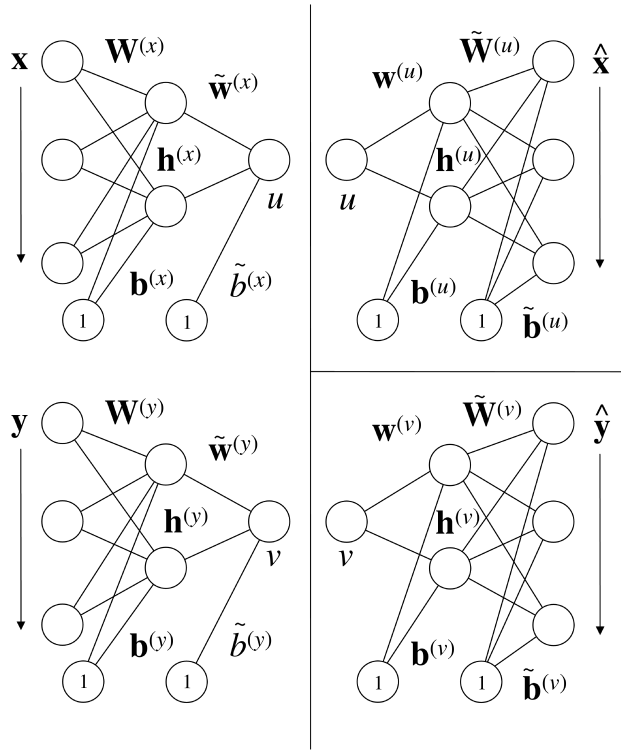

Fig. 1. Neural network architecture used to perform NLCCA.

functions used to set the model parameters are replaced with more robust versions. A cost function based on the biweight midcorrelation replaces one based on the Pearson (product-moment) correlation and cost functions based on the L_1-norm (i.e., mean absolute error, mae) replace ones based on the L_2-norm (i.e., mean squared error, mse). Robust variants of NLCCA are demonstrated on a synthetic dataset and are used to forecast SSTs in the tropical Pacific Ocean based on sea level pressure (SLP) data.

2 Method

2.1 NLCCA

Consider a dataset $\{x_i(t)\}$ with i variables and another dataset $\{y_j(t)\}$ with j variables, where each dataset has t=1, ..., N samples. The variables $\{x_i(t)\}$ can be grouped to form the vector $\boldsymbol{x}(t)$ and the variables $\{y_j(t)\}$ can be grouped to form the vector $\boldsymbol{y}(t)$. CCA looks for the linear combinations

$$u(t) = \boldsymbol{a} \cdot \boldsymbol{x}(t), \;\; v(t) = \boldsymbol{b} \cdot \boldsymbol{y}(t) \tag{1}$$

such that the Pearson correlation between the canonical variates u and v, i.e., cor(u, v), is maximized. If, for example, $\boldsymbol{x}$ is a gridded SLP dataset and $\boldsymbol{y}$ is a gridded SST dataset, then the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ represent correlated spatial patterns corresponding to the SLP and SST fields respectively. Unlike linear regression, which looks for relationships between a predictor dataset (e.g., $\boldsymbol{x}$) and each of the predictands (e.g., y_j) separately, CCA takes a holistic approach and looks for relationships between each of the sets of variables in their entirety. No distinction is made between the two fields; each can act interchangeably as predictors or predictands.

In NLCCA, the nonlinear analog of linear CCA, the linear mappings in Eq. (1) are replaced with nonlinear mappings performed by neural networks. The neural network architecture for NLCCA is shown in Fig. 1. The double-barreled network on the left-hand side nonlinearly maps $\boldsymbol{x}$ to u and $\boldsymbol{y}$ to v by

$$\begin{aligned} h_k^{(x)} &= \tanh[(\mathbf{W}^{(x)}\boldsymbol{x} + \boldsymbol{b}^{(x)})_k], \;\; u = \widetilde{\boldsymbol{w}}^{(x)} \cdot \boldsymbol{h}^{(x)} + \widetilde{b}^{(x)} \\ h_l^{(y)} &= \tanh[(\mathbf{W}^{(y)}\boldsymbol{y} + \boldsymbol{b}^{(y)})_l], \;\; v = \widetilde{\boldsymbol{w}}^{(y)} \cdot \boldsymbol{h}^{(y)} + \widetilde{b}^{(y)} \end{aligned} \tag{2}$$

where $h_k^{(x)}$ and $h_l^{(y)}$ are the hidden-layer nodes; $\tanh(\cdot)$ is the hyperbolic tangent function; $\mathbf{W}^{(x)}$ and $\mathbf{W}^{(y)}$ are the hidden-layer weight matrices; $\boldsymbol{b}^{(x)}$ and $\boldsymbol{b}^{(y)}$ are the hidden-layer bias vectors; $\widetilde{\boldsymbol{w}}^{(x)}$ and $\widetilde{\boldsymbol{w}}^{(y)}$ are the output-layer weight vectors; $\widetilde{b}^{(x)}$ and $\widetilde{b}^{(y)}$ are the output-layer biases; and k and l are indices of the vector elements. The number of hidden-layer nodes controls the overall complexity of the network; the hidden-layer must contain more than one node (k=1, ..., K, $K \geq 2$ and l=1, .., L, $L \geq 2$) to obtain a nonlinear solution (Hsieh, 2001).

Weight and bias parameters in the double-barreled network are set by minimizing the cost function

$$\begin{aligned} C_1 = &-\mathrm{cor}(u,\; v) + \langle u \rangle^2 + \langle v \rangle^2 + \left(\left\langle u^2 \right\rangle^{\frac{1}{2}} - 1 \right)^2 + \\ &\left(\left\langle v^2 \right\rangle^{\frac{1}{2}} - 1 \right)^2 + P_1 \left[\sum_{ki} \left(W_{ki}^{(x)} \right)^2 + \sum_{lj} \left(W_{lj}^{(y)} \right)^2 \right] \end{aligned} \tag{3}$$

where $\langle \cdot \rangle$ denotes the sample or temporal mean. The first term maximizes the correlation between the canonical variates u and v; the second, third, fourth, and fifth terms are normalization constraints that force u and v to have zero mean and unit variance; the sixth term is a weight penalty whose relative magnitude is controlled by the parameter P_1. Larger values of P_1 lead to smaller weights (i.e., fewer effective model parameters), which results in a more linear model. If $\tanh(\cdot)$ is replaced by the identity function, then Eq. (2) reduces to Eq. (1) and the network performs linear CCA.

Once the canonical variates u and v have been found, the inverse mappings to $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{y}}$ are given by the two neural networks on the right-hand side of Fig. 1:

$$\boldsymbol{h}_k^{(u)} = \tanh[(\mathbf{w}^{(u)}u + \boldsymbol{b}^{(u)})_k], \;\; \widehat{\boldsymbol{x}} = \widetilde{\mathbf{W}}^{(u)}\boldsymbol{h}^{(u)} + \widetilde{\boldsymbol{b}}^{(u)} \tag{4}$$

$$\boldsymbol{h}_l^{(v)} = \tanh[(\mathbf{w}^{(v)}v + \boldsymbol{b}^{(v)})_l], \;\; \widehat{\boldsymbol{y}} = \widetilde{\mathbf{W}}^{(v)}\boldsymbol{h}^{(v)} + \widetilde{\boldsymbol{b}}^{(v)}. \tag{5}$$

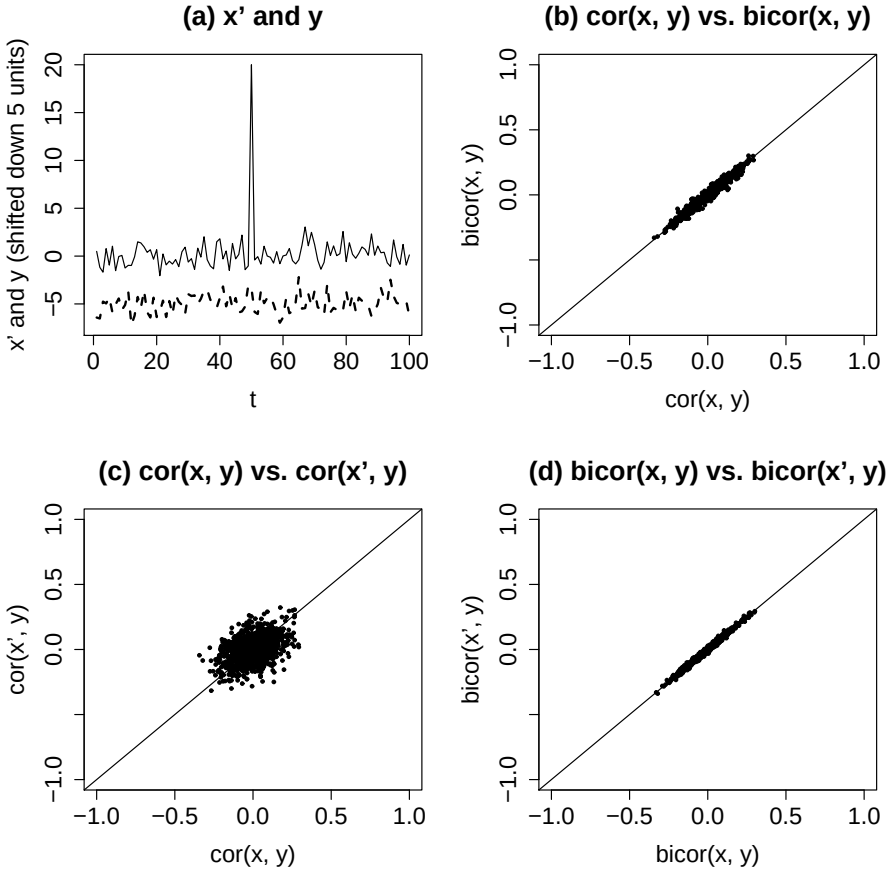

Fig. 2. Empirical comparison between the Pearson correlation (cor) and the biweight midcorrelation (bicor) on random variables x and y, each with samples drawn from a standard normal distribution, and x' and y, where x' is the same as x but with one case replaced with an outlier. Panel **(a)** shows sample time series of x' (solid) and y (dashed); **(b)** compares cor$(x,\ y)$ and bicor$(x,\ y)$; **(c)** compares cor$(x,\ y)$ and cor$(x',\ y)$; and **(d)** compares bicor$(x,\ y)$ and bicor$(x',\ y)$. Plots are for 1000 randomly generated datasets.

Weight and bias parameters in these two networks are found by minimizing the cost functions

$$C_2 = \left\langle \|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|^2 \right\rangle + P_2 \sum_k \left(w_k^{(u)} \right)^2 \tag{6}$$

$$C_3 = \left\langle \|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|^2 \right\rangle + P_3 \sum_l \left(w_l^{(v)} \right)^2 \tag{7}$$

respectively, where $\|\cdot\|^2$ is the square of the L_2-norm, with the L_p-norm given by

$$L_p(\boldsymbol{e}) = \left(\|\boldsymbol{e}\|^p \right)^{1/p} = \left(\sum_i |e_i|^p \right)^{1/p} . \tag{8}$$

C_2 and C_3 thus give the mse between the neural network predictions and the observed $\boldsymbol{x}$ and $\boldsymbol{y}$ variables subject to weight penalty terms whose magnitudes are controlled by the parameters P_2 and P_3. Once the first mode has been extracted from the data, the next leading mode can be extracted from the model residuals, and so on for higher modes.

For seasonal climate prediction tasks, where the goal is to predict values of a multivariate predictand dataset from a multivariate predictor dataset, e.g., $\widehat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x})$, values of the canonical variate v must be predicted from values of the canonical variate u. For canonical variates normalized to unit variance and zero mean, the linear least-squares regression solution is given by

$$\widehat{v} = u \ \mathrm{cor}(u,\ v) \tag{9}$$

(von Storch and Zwiers, 1999, pg. 325).

2.2 Biweight midcorrelation

The Pearson correlation is not a robust measure of association between two variables, as its estimates can be affected by the presence of a single outlier (Wilcox, 2004). For short, noisy datasets the cost function C_1 Eq. (3) in the NLCCA

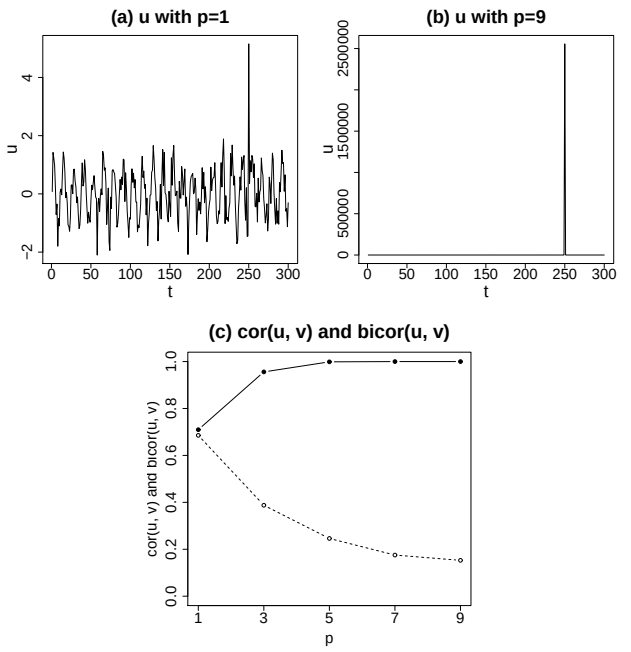

Fig. 3. Time series of u with **(a)** p=1 and **(b)** p=9; outliers in v occur at the same time as those in u. **(c)** The effect on cor(u, v) (solid line) and bicor(u, v) (dashed line) from increasing the separation between common outlier and non-outlier points by increasing p.

model may lead to overfitting as the model attempts to maximize the correlation between the canonical variates by generating mappings between $\boldsymbol{x}$ and u and $\boldsymbol{y}$ and v that are more complicated than necessary due to outliers. Rather than using the Pearson correlation, a more robust measure of association could instead be adopted in the cost function to avoid this problem.

Robust correlation coefficients, including commonly used functions like the Spearman rank correlation, are reviewed by Wilcox (2004). Trials with the Spearman rank correlation resulted in models with poor convergence properties and inconsistent performance on real-world climate datasets. Instead, the biweight midcorrelation (Wilcox, 2004) was selected as a robust alternative to the Pearson correlation. The biweight midcorrelation is calculated in the same manner as the Pearson correlation coefficient, except non-robust measures (the mean, expected deviation from the mean, and covariance) are replaced by robust measures. The biweight midcorrelation can also be used to predict v from u (and vice versa) in a manner similar to Eq. (9) for the standard NLCCA model, which is not possible with the Spearman rank correlation.

To calculate the biweight midcorrelation function bicor(x,y), first rescale x and y by

$$p = \frac{x - M_x}{9\,\mathrm{MAD}_x}, \quad q = \frac{y - M_y}{9\,\mathrm{MAD}_y} \tag{10}$$

where M_x and M_y are the median values of x and y respectively and MAD_x and MAD_y are the median values of $|x - M_x|$ and $|y - M_y|$ respectively. Next, the sample biweight midcovariance is given by

$$\mathrm{bicov}(x,\ y) = \frac{N \sum_t a(t)b(t)c(t)^2 d(t)^2 (x(t) - M_x)(y(t) - M_y)}{(\sum_t a(t)c(t)(1 - 5p(t)^2))(\sum_t b(t)d(t)(1 - 5q(t)^2))} \tag{11}$$

where $a(t)$=1 if $-1 \le p(t) \le 1$, otherwise $a(t)$=0; $b(t)$=1 if $-1 \le q(t) \le 1$, otherwise $b(t)$=0; $c(t)=(1-p(t)^2)$; and $d(t)=(1-q(t)^2)$. The biweight midcorrelation is then given by

$$\mathrm{bicor}(x,\ y) = \frac{\mathrm{bicov}(x,\ y)}{\sqrt{\mathrm{bicov}(x,\ x)\,\mathrm{bicov}(y,\ y)}}. \tag{12}$$

The biweight midcorrelation, like the Pearson correlation, ranges from -1 (negative association) to +1 (positive association).

Figure 2 shows estimates of the Pearson correlation and the biweight midcorrelation between normally distributed random variables $x \sim N(0, 1)$ and $y \sim N(0, 1)$ and between x' and y, where x' is the same as x but with one case replaced by an outlier (Fig. 2a). On the outlier-free dataset, both cor(x, y) and bicor(x, y) give approximately equal estimates of the strength of association between the variables (Fig. 2b). Estimates of cor(x', y) are strongly affected by the outlier, showing almost no association between values calculated with and without the outlying data point (Fig. 2c), whereas estimates of bicor(x', y) are essentially unaffected by the outlier (Fig. 2d).

NLCCA with the Pearson correlation cost function may fail when outliers occur simultaneously in both datasets. To illustrate, consider two identical sinusoidal series, each with a common outlier

$$x(t) = y(t) = \sin(0.5t) + \delta(t), \text{ where } \delta(t) = \begin{cases} 6 \text{ at } t = 150 \\ 0 \text{ elsewhere} \end{cases} \tag{13}$$

where $t = 1, 2, ..., 300$. Next, create new series x' and y' by adding noise drawn from $N(0, 0.5)$ to x and y. The expected values of cor(x', y') and bicor(x', y') are found to be 0.71 and 0.68 respectively. Now, consider values of cor(u, v) and bicor(u, v), where $u = x'^p$, $v = y'^p$, and p is an odd integer (Fig. 3a). Increasing the value of p effectively increases the separation between the outlier and the non-outliers (Fig. 3b). Values of cor(u, v) and bicor(u, v) for values of p from 1 to 9 are shown in Fig. 3c. The Pearson correlation can be increased simply by increasing p, whereas the biweight midcorrelation decreases as p increases. This case illustrates how increasing the nonlinearity of the mapping functions u and v (by increasing p) can lead to very high Pearson correlation. In the context of NLCCA, spuriously high values of cor(u, v) can be found by the double-barreled network

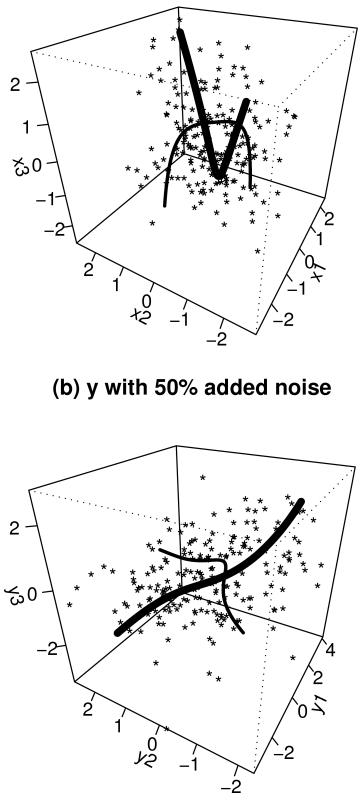

Fig. 4. Synthetic test datasets used to evaluate the performance of the standard and robust NLCCA models. The first mode is shown by the thick black curve and the second mode is shown by the thin black curve. Test samples with added noise are shown as asterisks.

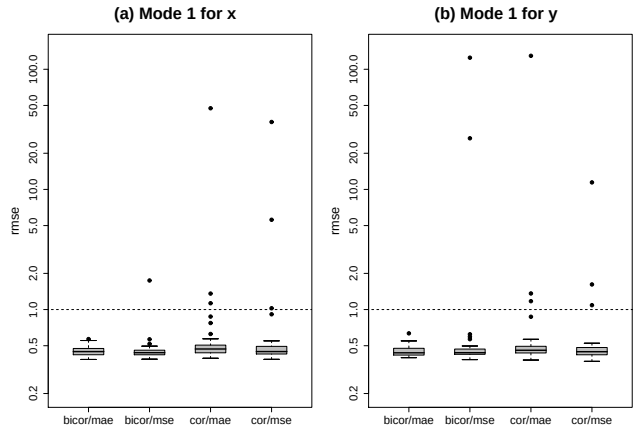

Fig. 5. Boxplots showing the distribution of rmse between the first synthetic mode and the first mode extracted by NLCCA models for **(a)** $\boldsymbol{x}$ and **(b)** $\boldsymbol{y}$ with different combinations of non-robust and robust cost functions over 50 trials. Boxes extend from the 25th to 75th percentiles, with the line indicating the median. Whiskers represent the most extreme data within ± 1.5 times the interquartile range (i.e., the box height); values outside this range are plotted as dots. The dashed line indicates a rmse equal to one. The ordinate is log-scaled to accomodate the large range in rmse.

when the nonlinear neural network mapping greatly magnifies an outlier in both x and y. This artifact can be particularly dangerous when NLCCA is applied to datasets that are affected by strong, concurrent climate signals, for example those with large El Niño or La Niña anomalies, as shown by Hsieh (2001). NLCCA performed worse than CCA when weight penalty terms were not used to reduce the nonlinearity of the double-barreled network. Based on results shown in Fig. 3, adopting bicor in the cost function should prevent this artifact.

When NLCCA models are used for multivariate prediction, a regression model is needed to estimate v from u (and vice versa). For the standard NLCCA model, the linear least squares estimate for the regression coefficient is given by Eq. (9). Similarly, the biweight midcorrelation is associated with a robust regression model that can be used to predict values of one canonical variate from the other. Following Lax (1985) and Hou and Koh (2004), the biweight midregression solution is given by

$$\widehat{v} = u\,\mathrm{bicor}(u,\ v) \tag{14}$$

for canonical variates normalized to unit variance and zero mean.

2.3 L_p-norm

Now consider the inverse maping from u and v back to $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{y}}$ (i.e., the networks on the right hand side of Fig. 1). The L_p-norm given in Eq. (8) forms the basis for a class of cost functions used in regression models (Bishop, 1995). Of these, the L_2-norm, which leads to the mse cost function, is commonly used in statistical models. Models that minimize the mse are optimal if the data are generated from a deterministic function corrupted by a normally distributed noise process with constant variance. However, a potential problem exists with cost functions based on the L_2-norm (e.g., C_2 and C_3 defined in Eqs. 6 and 7). Samples with the greatest errors exert disproportionately large influence on the cost function. Thus, a small number of outliers can come to dominate the solution. Adopting the L_1-norm, which leads to the mae cost function, reduces the influence of outliers.

Bishop (1995, Sects. 6.1–6.2) showed that in the limit of infinite samples and with a flexible enough model (e.g., a neural network with enough hidden nodes), the model converges to the conditional mean if the mse is used and the conditional median if the mae is used. The median is robust to outliers whereas the mean is not.

2.4 Robust NLCCA

Robust variants of NLCCA use the model architecture shown in Fig. 1 but with the cost functions C_1, C_2, and C_3 replaced by the robust versions described in Sect. 2.2 and 2.3. The biweight midcorrelation replaces the Pearson correlation in C_1 and the mae replaces the mse in C_2 and C_3.

3 Synthetic test problem

3.1 Data

To illustrate the effect of the changes to the NLCCA cost functions, consider the three dimensional synthetic test problem used by Hsieh (2000) to introduce the standard NLCCA model. The first correlated mode (**x** and **y**) is given by

$$x_1 = t - 0.3t^2, \quad x_2 = t + 0.3t^2, \quad x_3 = t^2 \tag{15}$$

$$y_1 = t^3, \quad y_2 = -t + 0.3t^3, \quad y_3 = t + 0.3t^2 \tag{16}$$

where t is a uniformly distributed random number in $[-1, 1]$. The second correlated mode (x' and y') is given by

$$x_1' = -s - 0.3s^2, \quad x_2' = s - 0.3s^3, \quad x_3' = -s^4 \tag{17}$$

$$y_1' = \mathrm{sech}(4s), \quad y_2' = s + 0.3s^3, \quad y_3' = s - 0.3s^2 \tag{18}$$

where s is a uniformly distributed random number in $[-1, 1]$. The shapes described by $\boldsymbol{x}$ and $\boldsymbol{x}'$ are shown in Fig. 4a and those described by $\boldsymbol{y}$ and $\boldsymbol{y}'$ are shown in Fig. 4b.

To test the performance of the NLCCA models, 50 training and test datasets, each with 500 samples, were randomly generated from Eqs. (15–18). The signal in each dataset was produced by adding the second mode to the first mode, with the variance of the second equal to one third that of the first. Normally distributed random noise with standard deviation equal to 50% of the signal standard deviation was added to the data. The variables were then standardized to zero mean and unit standard deviation.

3.2 Training and testing procedure

NLCCA models with different combinations of the non-robust (cor and mse) and robust (bicor and mae) cost functions were developed on the training datasets and applied to the test datasets. Following Hsieh (2000), all neural networks had three nodes in their hidden-layers and were trained without weight penalty terms. A quasi-Newton nonlinear optimization scheme with finite-difference approximations of the gradient and Hessian was used to minimize the cost functions. While the L_1 norm is not continuously differentiable, convergence problems were not noted during optimization. The L_1 norm can, however, be approximated by the Huber norm, which is continuously differentiable, if issues with convergence are found (Panayiotis et al., 2006). To avoid local minima in the error surface, each network in Fig. 1 was trained 30 times, each time starting from a different randomly selected set of initial weights and biases. The network with the lowest value of its associated cost function was then selected for use and applied to the test data.

3.3 Model performance

Root mse (rmse) values between the first synthetic mode and the first mode extracted by NLCCA models with different combinations of non-robust and robust cost functions are shown in Fig. 5 for the 50 test datasets. On average, all models performed approximately the same, although, for the leading NLCCA mode of the $\boldsymbol{x}$ dataset, NLCCA with bicor/mse cost functions yielded the lowest median rmse (0.44), followed by NLCCA with bicor/mae (0.45) and NLCCA with cor/mse (0.45). NLCCA with cor/mae performed worst with a median rmse of 0.47. Median rmse values and relative rankings of the models were the same for the leading NLCCA mode of the $\boldsymbol{y}$ dataset.

Of the four models, NLCCA with the robust cost functions (bicor/mae) was the most stable. No trial yielded an rmse in excess of the series standard deviation of one, with the maximum value under 0.6 for the **x** mode. The other models had at least one trial with an rmse value greater than one, which is indicative of severe overfitting. Maximum values for the $\boldsymbol{x}$ mode ranged from 1.8 for NLCCA with bicor/mse, to 47.4 for NLCCA with cor/mse, and 49.6 for cor/mae. NLCCA with bicor/mae performed similarly for the $\boldsymbol{y}$ mode, although two trials with rmse greater than 20 were found for NLCCA with bicor/mse cost functions.

Overall, results for the synthetic dataset suggest that replacing the cor/mse cost functions in NLCCA with bicor/mae cost functions leads to a more stable model that was less susceptible to overfitting and poor test performance. All models were run without weight penalty terms in this comparison. In practice, the non-robust models will need weight penalty terms to reduce overfitting, as is done in our next test, where NLCCA models are applied to a real-world climate prediction problem.

4 Seasonal prediction

4.1 Data

As the primary goal of the study is to investigate the effect of the robust cost functions on performance, and not to build an operational forecast model, predictor/predictand fields were selected as in Hsieh (2001).

SST data were obtained from the second version of the NOAA Extended Reconstructed SST (ERSST.v2) dataset (Smith and Reynolds, 2004). Monthly data on a $2^\circ \times 2^\circ$ grid were extracted for a spatial domain covering the tropical Pacific Ocean (22°S-22°N, 122°E-288°E) for the time period

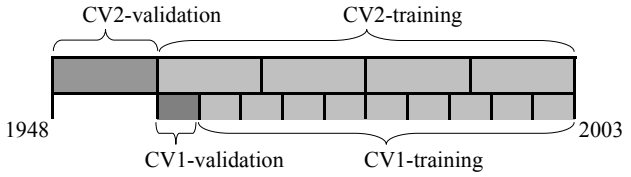

Fig. 6. Diagram showing how data were split into training (light gray) and validation (dark gray) segments for the first (CV1) and second (CV2) cross-validation procedures.

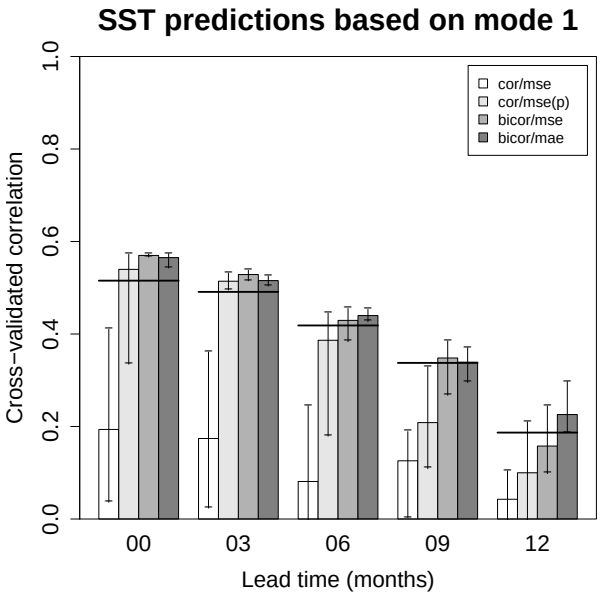

Fig. 7. Cross-validated correlation skill for NLCCA models trained with cor/mse, bicor/mse, and bicor/mae cost functions. Weight penalty was applied to the model denoted cor/mse(p). Bars show the mean correlation over the spatial domain, averaged over the 10 trials. Vertical lines extend from the minimum to the maximum spatial mean correlation from the 10 trials. Horizontal lines show correlation skill from the CCA model. The ordinate is limited to showing positive cross-validated skill.

1948 to 2003. The climatological seasonal cycle was removed, data were detrended, and a 3-month running mean filter was applied. Principal component analysis (PCA) was applied to the data; the first 6 modes accounting for 73% of the total SST variance were retained for further analysis. PC scores (i.e., the time series for the leading PCA modes) were scaled according to the amount of variance explained by each mode.

SLP data from the NCEP/NCAR Reanalysis (Kalnay et al., 1996) were obtained for the same region and period. Data on a 2.5°×2.5° grid had the climatological seasonal cycle removed, the data were detrended, and then smoothed by a 3-month running mean filter. PCA was applied to the data and the first 6 modes, accounting for 80% of the total SLP variance, were retained for further analysis.

4.2 Training and testing procedure

Three variants of the NLCCA model were applied to the SST and SLP datasets. The first, representing the standard NLCCA model, incorporated both non-robust cost functions (cor/mse). The second and third used the bicor cost function to train the double-barreled network and either the mae or mse cost function to train the inverse mapping networks. For brevity, the model with cor/mae cost functions was dropped from consideration.

To assess the usefulness of the three variants of NLCCA for seasonal forecasting, models were validated on the basis of their forecast performance. PC scores from the 6 leading PCs of the SLP dataset were used to predict PC scores from the 6 leadings PCs of the SST dataset at lead times of 0, 3, 6, 9, and 12-months. (Lead times are defined as the number of months from the predictor observation to the predictand observation, e.g., a forecast with a 3-month lead time from January would be for April.) Taking $\boldsymbol{x}$ to be historical values of the SLP PC scores and $\boldsymbol{y}$ to be historical values of the SST PC scores, forecasts for a new case $\widehat{\boldsymbol{y}}(t_n)$ at time t_n were made as follows. First, the double-barreled network was trained with $\boldsymbol{x}$ and $\boldsymbol{y}$ as inputs and the resulting values of u and v were used to train the inverse mapping networks. Given a new SLP data point $\boldsymbol{x}(t_n)$, a new value of the canonical variate $u(t_n)$ was obtained from the double-barreled network. Regression equations (Eq. 9 or Eq. 14) were then used to predict a new value of $\widehat{v}(t_n)$. Finally, $\widehat{v}(t_n)$ was entered into the appropriate inverse mapping network to give $\widehat{\boldsymbol{y}}(t_n)$. For the second and higher NLCCA modes, the same procedure was followed using residuals from the previous mode as inputs. Following Hsieh (2001), neural networks were trained both with and without weight penalty terms using two hidden-layer nodes. A two-stage cross-validation procedure was used to set the weight penalty coefficients and to estimate forecast performance. For reference, a schematic diagram showing how data were split into training/validation segments is shown in Fig. 6.

To avoid overfitting in models trained with weight penalty, values of the coefficients P_1, P_2, and P_3 in Eqs. 3, 6, and 7 were determined via 10-fold cross-validation on the training dataset (CV1 in Fig. 6). The training record was split into 10 contiguous segments. Models were trained on 9 of the 10 segments using weight penalties from the set $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Forecasts on the remaining segment were then recorded for each weight penalty coefficient. While fixing the weight penalties, these steps were repeated 9 times, each time making forecasts on a different segment. Finally, forecasts for all 10 segments were combined and validated against observations. Weight penalties that minimized the aggregated cross-validation error were recorded, neural networks were retrained on all 10 segments

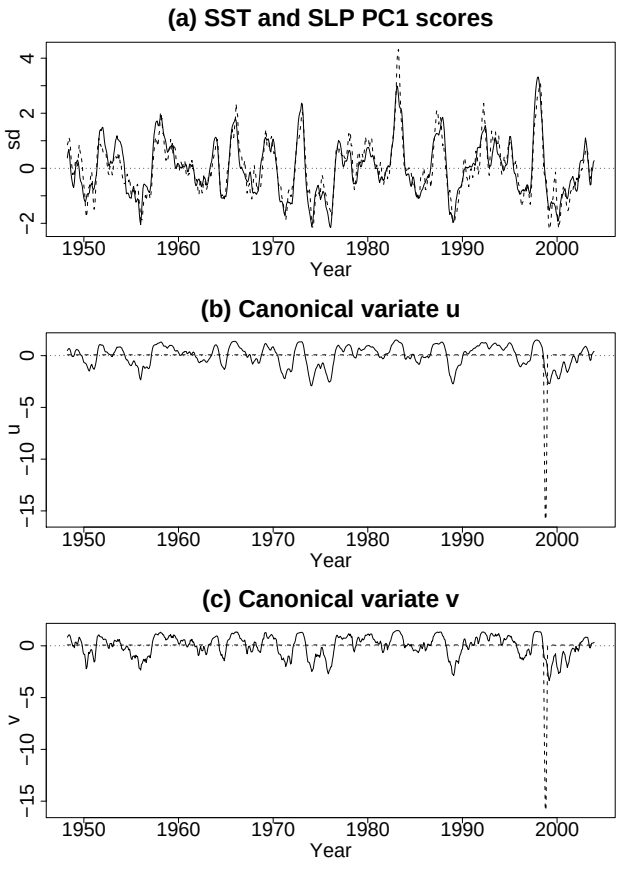

Fig. 8. Plots of **(a)** PC scores from the leading SST (solid line) and SLP (dashed line) PCA modes; **(b)** the canonical variate u for the leading NLCCA mode from a model with cor/mse cost functions (dashed line) and one with bicor/mse cost functions (solid line); and **(c)** canonical variate v for the leading NLCCA mode from a model with cor/mse cost functions (dashed line) and bicor/mse cost functions (solid line).

combined using these penalties. Ten models were trained in this manner to assess sensitivity to initial weights and biases.

A second round of cross-validation was used to estimate out-of-sample forecast performance of the models (CV2 in Fig. 6). The historical record was split into 5 contiguous segments (each approximately 11 years in length). Models were trained on 4 of the 5 segments using the cross-validation procedure outlined above. Forecasts on the remaining segment were then recorded. These steps were repeated 4 times, each time making forecasts on a different segment. Finally, forecasts for all 5 segments were combined and compared with observations.

4.3 Skill for models with one mode

Results from NLCCA models with one mode are shown in Fig. 7. For reference, results from linear CCA models are also shown. Cross-validated Pearson correlation skill is averaged over the entire domain following reconstruction of the SST field from the predicted SST PC scores. Values of rmse were also calculated, but are not shown as relative performance between models was the same as for correlation skill. Results with weight penalty are only given for the NLCCA model with cor/mse cost functions as the addition of penalty terms to models with the bicor cost function did not generally lead to significant changes in skill.

Without weight penalty, the NLCCA model with cor/mse cost functions performed poorly, exhibiting mean skills worse than CCA at all lead times. Even with concurrent predictor/predictand fields, the mean correlation skill was lower than 0.2. NLCCA with bicor/mse cost functions and bicor/mae cost functions performed much better, with mean correlation skills exceeding 0.5 at the 0-month lead time. Over the 10 trials, minimum skills from models incorporating the bicor cost function were higher than maximum skills from the corresponding cor/mse models without weight penalty.

For NLCCA with cor/mse cost functions, minimum correlations were lower than zero (i.e., no cross-validation skill) for 6, 9, and 12-month lead times. All NLCCA models with bicor/mse and bicor/mae cost functions, even those at a 12-month lead time, showed positive skill. In general, NLCCA models with bicor exhibited the least variability in skill between repeated trials. In no case was the range between minimum and maximum skill greater than 0.2. For NLCCA with cor/mse cost functions, the range in skill exceeded 0.2 at all lead times, indicating a very unstable model.

Little difference in skill was evident between bicor/mse and bicor/mae models, which suggests that the switch from cor to bicor in the double-barreled network cost function was responsible for most of the increase in skill relative to the standard NLCCA model. Inspection of the canonical variates shows that this was due to the insensitivity of the bicor cost function to the common outlier artifact described in Sect. 2.2 and illustrated in Fig. 3.

Plots of the canonical variates u and v for the first mode of NLCCA models with cor/mse and bicor/mse cost functions at the 0-month lead time are shown in Fig. 8 along with PC scores from the leading SST and SLP PCA modes. For these series, values of $\mathrm{cor}(u, v)$ and $\mathrm{bicor}(u, v)$ were 1.00 and 0.96 respectively. The high correlation between u and v for the NLCCA model with the cor cost function was driven almost exclusively by the common outliers present during 1997–1998. With the 1997–1998 outliers removed, $\mathrm{cor}(u, v)$ dropped to 0.28. On the other hand, the high correlation between u and v for the NLCCA model with the bicor cost function was indicative of the strong relationship between the SLP and SST series, as evidenced by the Pearson correlation of 0.91 between the leading SST and SLP PCs.

Results discussed to this point have been for NLCCA models without weight penalty. Hsieh (2001) found that the addition of weight penalty to the standard NLCCA model lead to improvements in performance, due in part to the avoidance

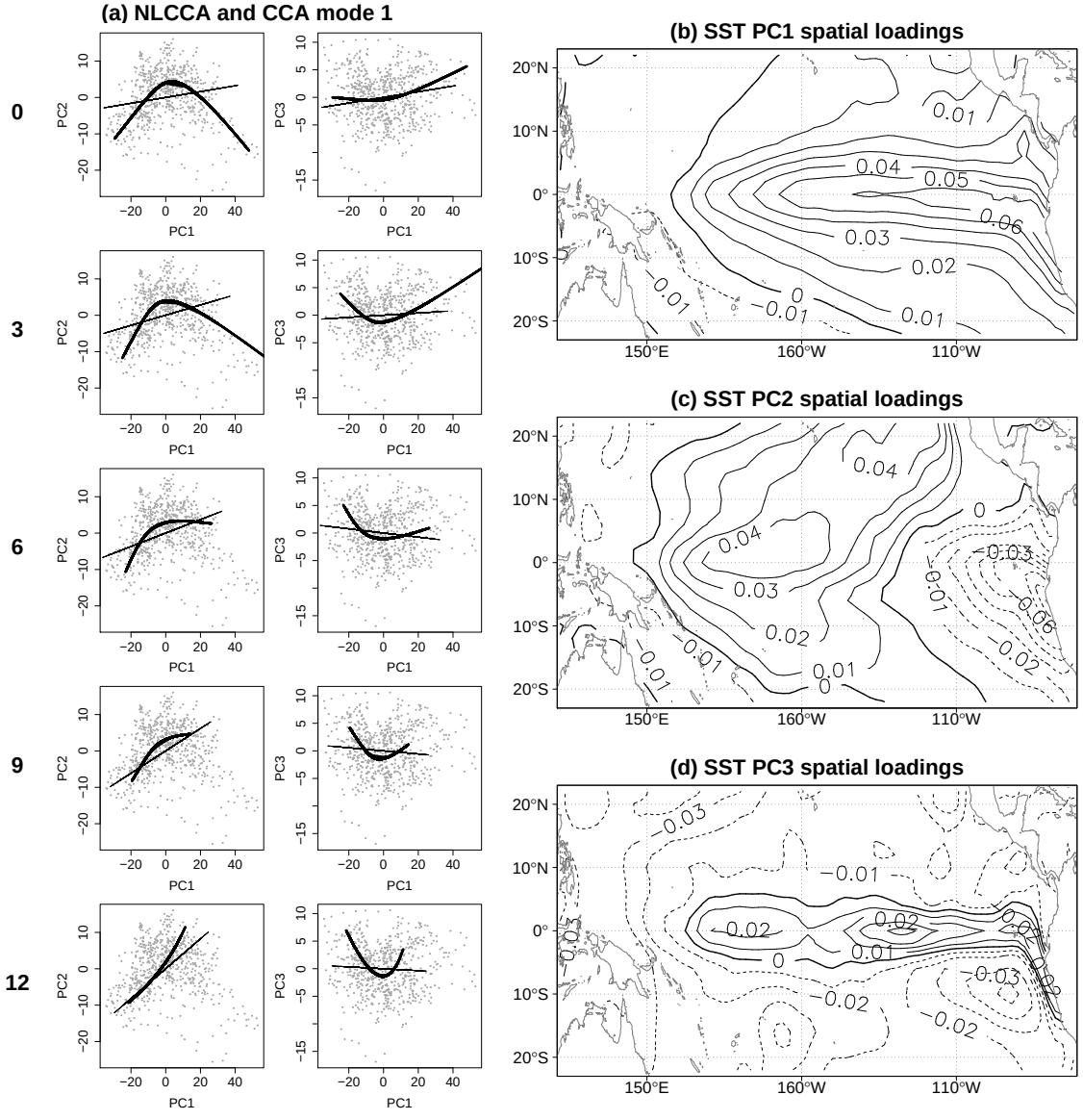

Fig. 9. **(a)** Plots of the first SST mode for CCA (thin line) and NLCCA with bicor/mae cost functions (thick line) in the planes of the PC1-PC2 and PC1-PC3 scores at 0, 3, 6, 9, and 12-month lead times. Spatial patterns for **(b)** PC1, **(c)** PC2, and **(d)** PC3, all normalized to unit norm.

of the common outlier artifact. Addition of weight penalty to the standard NLCCA model resulted in improvements in mean correlation skill, although performance still lagged behind NLCCA with the bicor cost function at 9 and 12-month lead times. At 0, 3, and 6-month lead times, maximum skill over the 10 trials did, however, exceed the mean level of skill of the bicor-based models, which suggests that an appropriate amount of weight penalty can result in a good performing model. Inspection of the time series of u and v for the best performing runs suggests that improved performance was due to avoidance of the common outlier artifact. However, the wide range in performance over the 10 trials (e.g., at 0 and 6-month lead times) reflects the instability of the training and cross-validation steps needed to choose the weight penalty coefficients. In practice, it may be difficult to consistently reach the performance level of the robust model by relying solely on weight penalty to control overfitting of the standard NLCCA model.

Returning to the NLCCA models with bicor/mse and bicor/mae cost functions, little difference in skill between the models is apparent from Fig. 7. At short lead times (0 and 3-months), when the signal is strongest, the bicor/mse model performed slightly better than the bicor/mae model, whereas at the longest lead time (12-months), when the signal is

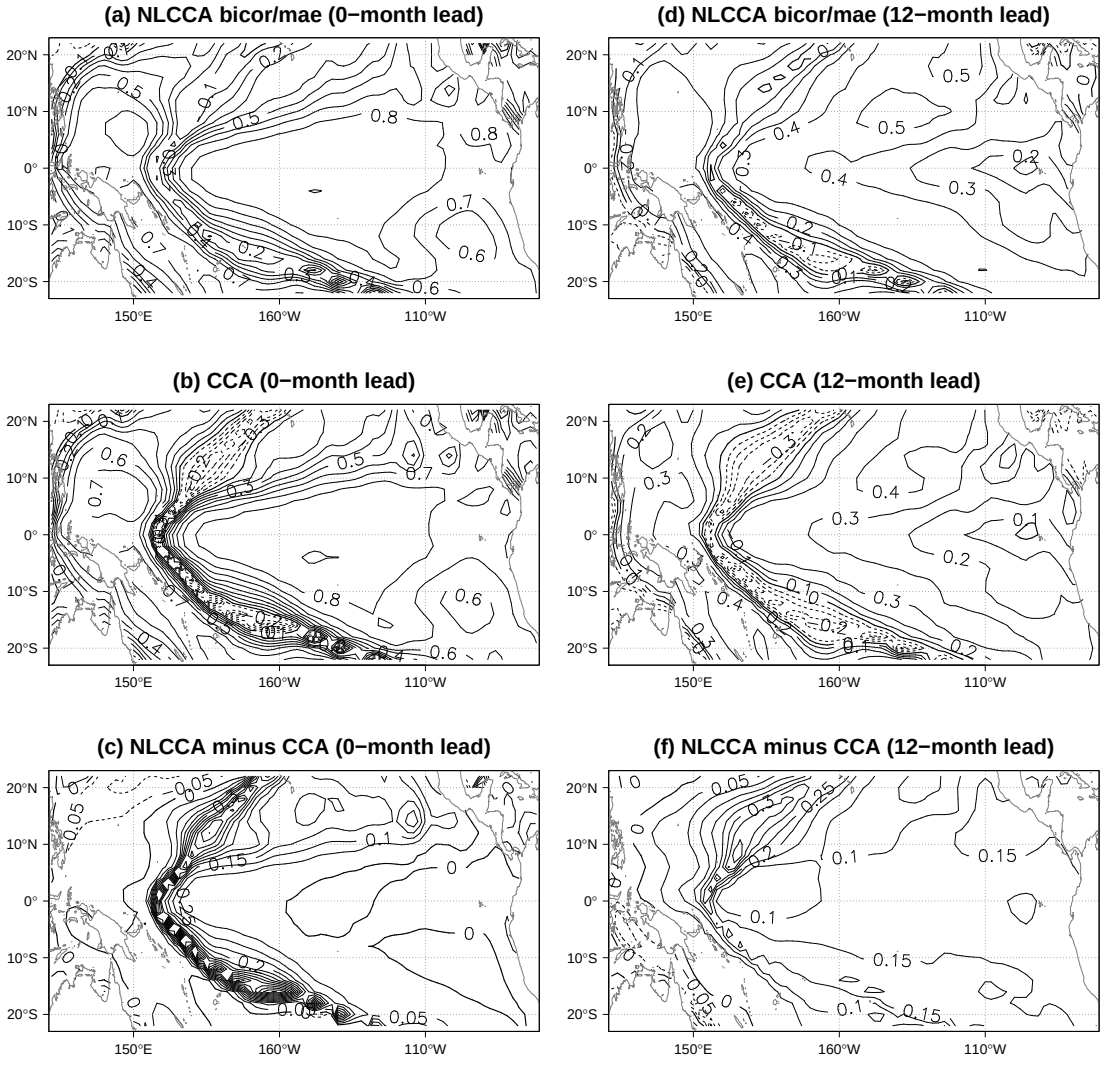

Fig. 10. Spatial correlation skill at 0-month lead time for **(a)** NLCCA with bicor/mae cost functions and **(b)** CCA. Panel **(c)** shows NLCCA skill minus CCA skill. Panels **(d)** to **(f)** as in (a) to (c) but for 12-month lead time

weakest, the bicor/mae model performed best (and with less run-to-run variability in skill).

NLCCA models with the bicor/mse and bicor/mae cost functions tended to perform slightly better than CCA. For the bicor/mae model, the small improvement in performance was significant (i.e., minimum skill over the 10 trials exceeded CCA skill) at 0, 3, 6, and 12-month lead times, while the same was true of the bicor/mse model at 0 and 3-month lead times.

To investigate the differences between the linear and nonlinear models, plots of the first CCA and NLCCA (bicor/mae) SST modes projected onto the PC1-PC2 and PC1-PC3 planes are shown in Fig. 9a. Spatial loading patterns associated with each PC are shown in Figs. 9b-d. For the NLCCA mode at short lead times, a quadratic response was present in the PC1-PC2 plane. Negative values of PC2 occurred when values of PC1 were both negative and positive, which, from the spatial loading patterns, means that the predicted SST response at the minimum/maximum values of v (which, at left/right extremes of the curve, correspond to La Niña/El Niño states respectively) exhibited east/west asymmetry. The curve rotated counter clockwise and straightened with increasing lead time. At short lead times, the leading CCA mode was driven mainly by PC1.

Conversely, the NLCCA curve in the PC1-PC3 plane displayed increased nonlinearity with lead time. Predicted values of PC3 were typically positive when values of PC1 were both negative and positive, which, from the spatial loading pattern of PC3, indicates differences in the contrast between predicted SST anomalies along the equator and off the equa-

tor during La Niña and El Niño states. Observed asymmetry in the spatial patterns and magnitudes of SST anomalies associated with La Niña and El Niño are present in the observational record and have previously been detected by nonlinear methods (Hsieh, 2001; Monahan, 2001; Wu and Hsieh, 2002).

To this point, reported skills have been averaged over the entire spatial domain. For reference, Fig. 10 shows spatial patterns of correlation skill for NLCCA models with bicor/mae cost functions at lead times of 0 and 12-months respectively. For comparison, correlation skills from CCA models are also plotted. Spatially at 0-month lead time, skill was highest in the central equatorial Pacific Ocean, with a secondary maximum to the northeast of Papua New Guinea and east of Australia. Somewhat similar spatial patterns are seen at the other lead times. Differences in skill between NLCCA and CCA are generally small, with the largest improvements by NLCCA occurring along the boundary separating the two skill maxima.

4.4 Skill for models with two modes

Results reported in the previous section were from NLCCA models with a single nonlinear mode. Inclusion of the second NLCCA mode may improve forecast performance in the tropical Pacific Ocean (Wu and Hsieh, 2002). To investigate, results from NLCCA models with two modes are shown in Fig. 11.

Model skill with two modes improved relative to NLCCA with a single mode at short lead times. For instance, mean correlation skill for the NLCCA model with bicor/mae went from 0.55 with one mode to 0.65 with two modes at a 0-month lead time, and from 0.52 to 0.59 at a 3-month lead time. At longer lead times performance dropped, even to a level below CCA at 6-months, which is indicative of overfitting. However, the same was also true of the CCA model where, at 9 and 12-month lead times, skill decreased relative to the model with a single mode. Results are somewhat at odds with those reported by Wu and Hsieh (2002), who found the largest improvements in model performance to occur at longer lead times. However, cross-validation was not employed by Wu and Hsieh (2002), which means that overfitting may have caused inflated skill estimates at these lead times.

As pointed out by Hsieh (2001), nonlinearity in the tropical Pacific is strongest in the leading SST mode and is much weaker (or even not evident) in higher modes. As a result, using a nonlinear model, even one that can be estimated robustly, to extract the second or higher modes may not be warranted and could lead to poor forecast performance. When the skill improvement of NLCCA over CCA is minimal, as is the case here even at short lead times, it may be more appropriate to apply CCA to residuals from the first NLCCA mode. This approach is currently used in operational NLCCA forecast models run by the Climate Prediction Group at the University of British Columbia (A. Wu, 2007, personal communication).

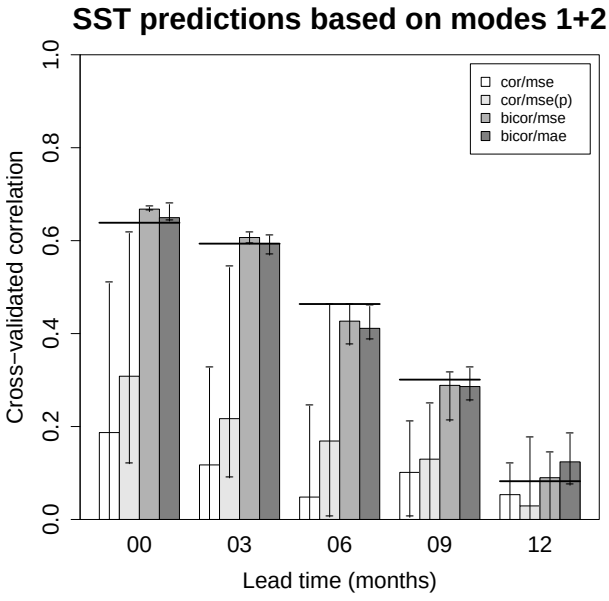

Fig. 11. As in Fig. 7, but for NLCCA models with two modes.

5 Conclusions

NLCCA based on multi-layer perceptron neural networks is a flexible model capable of nonlinearly generalizing linear CCA (Hsieh, 2000). However, the complicated model architecture and use of non-robust cost functions means that overfitting is difficult to avoid, particularly when dealing with the short, noisy datasets that are common in seasonal climate forecasting problems. To make NLCCA more robust, non-robust cost functions in the model are replaced by robust cost functions: the Pearson correlation in the double-barreled network is replaced by the biweight midcorrelation, while the mse in the inverse mapping network can be replaced by the mae.

Through analysis of a synthetic dataset and a real-world climate dataset, adoption of the biweight midcorrelation is shown to result in large improvements in model stability, mainly by avoiding the common outlier artifact noted by Hsieh (2001). Replacing the mse by the mae leads to improved performance on the synthetic dataset, but little improvement on the climate dataset, except at the longest lead time where the signal-to-noise ratio is smallest.

Based on these results, it is recommended that the biweight midcorrelation replace the Pearson correlation in the NLCCA model. Choosing the mse or mae cost function appears to be more problem dependent, and should be considered as part of the model selection process. Other cost functions, for example those based on the L_p norm with $1<p<2$

(Hanson and Burr, 1988), might also be viable, depending on the prediction task. More research is needed to determine the most appropriate cost function for the inverse mapping networks.

Development of a robust NLCCA model for operational prediction of SSTs in the equatorial Pacific Ocean is currently underway. To maximize skill, additional predictors, for example lagged SSTs (Wu et al., 2006), upper ocean heat content, and Madden-Julian oscillation indices (McPhaden et al., 2006), are being investigated. Model performance may also be improved by specifying corrections on predictions when the model extrapolates beyond the limits of the training data, as suggested by Wu et al. (2008[1]).

Acknowledgements. W. Hsieh is supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada, and a project grant from the Canadian Foundation for Climate and Atmospheric Sciences.

Edited by: H. A. Dijkstra
Reviewed by: V. Krasnopolsky and another anonymous referee

References

Barnett, T. P. and Preisendorfer, R.: 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, 1987.

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

Bishop, C. M.: Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 504 pp., 1995.

Cannon, A. J.: Nonlinear principal predictor analysis: Application to the Lorenz system, J. Climate, 19, 579–589, 2006.

Glahn, H. R.: Canonical correlation and its relationship to discriminant analysis and multiple regression, J. Atmos. Sci., 25, 23–31, 1968.

Hanson, S. J. and Burr, D. J.: Minkowski-r back-propagation: Learning in connectionist models with non-Euclidean error signals, Neural Information Processing Systems, American Institute of Physics, 348–357, 1988.

Hou, Z. and Koh, T. S.: Image denoising using robust regression, IEEE Signal Proc. Lett., 11, 243–246, 2004.

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

Hsieh, W. W.: Nonlinear canonical correlation analysis of the tropical Pacific climate variability using a neural network approach, J. Climate, 14, 2528–2539, 2001.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-year reanalysis project, B. Am. Meteorol. Soc., 77, 437–471, 1996.

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

Lai, P. L. and Fyfe, C.: Kernel and nonlinear canonical correlation analysis, Int. J. Neural. Syst., 10, 365–377, 2000.

Lax, D. A.: Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions, J. Am. Stat. Assoc., 80, 736–741, 1985.

McPhaden, M. J., Zhang, X. B., Hendon, H. H., and Wheeler, M. C.: Large scale dynamics and MJO forcing of ENSO variability, Geophys. Res. Lett., 33, L16702, doi:10.1029/2006GL026786, 2006.

Melzer, T., Reiter, M., and Bischof, H.: Appearance models based on kernel canonical correlation analysis, Pattern Recogn., 36, 1961–1971, 2003.

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

Panayiotis, C. A., Charalambous, C., and Martzoukos, S. H.: Robust artificial neural networks for pricing of European options, Computational Economics, 27, 329–351, 2006.

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

Shawe-Taylor, J. and Cristianini, N.: Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.

Smith, T. M. and Reynolds, R. W.: Improved extended reconstruction of SST (1854–1997), J. Climate, 17, 2466–2477, 2004.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J.: Least Squares Support Vector Machines, World Scientific, Singapore, 318 pp., 2002.

von Storch, H. and Zwiers, F. W.: Statistical Analysis in Climate Research, Cambridge University Press, 484 pp., 1999.

Wilcox, R. R.: Introduction to Robust Estimation and Hypothesis Testing, 2nd. Ed., Academic Press, 608 pp., 2004.

Wu, A. and Hsieh, W. W.: Nonlinear canonical correlation analysis of the tropical Pacific wind stress and sea surface temperature, Clim. Dynam., 19, 713–722, 2002.

Wu, A. and Hsieh, W. W.: Nonlinear interdecadal changes of the El Niño-Southern Oscillation, Clim. Dynam., 21, 719–730, 2003.

Wu, A. M. and Hsieh, W. W.: The nonlinear northern hemisphere winter atmospheric response to ENSO, Geophys. Res. Lett., 31, L02203, doi:10.1029/2003GL018885, 2004.

Wu, A. M., Hsieh, W. W., and Zwiers, F. W.: Nonlinear modes of North American winter climate variability derived from a general circulation model simulation, J. Climate, 16, 2325–2339, 2003.

Wu, A. M., Hsieh, W. W., and Tang, B. Y.: Neural network forecasts of the tropical Pacific sea surface temperatures, Neural Networks, 19, 145–154, 2006.

[1]Wu, A., Hsieh, W. W., Cannon, A. J., and Shabbar, A.: Improving neural network predictions of North American seasonal climate by outlier correction, Nonlin. Processes Geophys., under revision, 2008.