Nonlinear post-processing of numerical seasonal climate forecasts

Joel Finnis*, Dept. of Geography, Memorial University of Newfoundland
William W. Hsieh, Dept. of Earth & Ocean Sciences, University of British Columbia
Hai Lin, Environment Canada
William Merryfield, Canadian Centre for Climate Modeling and Analysis, Environment Canada

Submitted to Atmosphere-Ocean (20 May 2011)

Corresponding Author: jfinnis@mun.ca; Phone: 709-864-8987
Dept. of Geography, Memorial University of Newfoundland, St. John's NL, A1B 3X9

Abstract

Although numerical models are increasingly being used to generate operational seasonal forecasts, the reliability of these products remains relatively low. Regression-based post-processing methods have proven useful in increasing forecast skill, but as yet efforts have focused on linear regression. Given the nonlinear nature of the climate system and sources of model error, nonlinear analogues of these post-processing methods may offer considerable improvements. The current study tests this hypothesis, applying both linear and nonlinear regression to the correction of climate hindcasts produced with general circulation models. Results indicate that nonlinear support vector regression is better able to extract indices of the Pacific-North American pattern and the North Atlantic Oscillation from coupled model output, while linear approaches are better suited to atmosphere-only model output. Statistically significant predictions are produced at lead times of up to nine months, and can be obtained from model output with no forecast skill prior to processing.

1. Introduction

Numerical seasonal climate forecasts are becoming increasingly common, as the research community experiments with new tools and techniques while various user groups explore their potential value. Building on improvements in climate models and computing power realized in recent years, several prominent forecasting centers have developed operational systems to predict the climate state at lead times of several months to several seasons (Derome et al. 2001; Livezey and Timofeyeva 2008; Saha et al. 2006). A common

approach has been to force an atmospheric general circulation model (GCM) with sea surface conditions observed at the beginning of a forecast period, recognizing that the ocean boundary captures various slow-varying processes with strong atmospheric influences. Several of these atmosphere-only (AGCM) forecasting systems have demonstrated modest skill at lead times of one month to a season, during which sea surface conditions are unlikely to change significantly. These early successes have stoked interest in systems based on coupled atmosphere-ocean GCMs (AOGCMs). The addition of an evolving ocean state has the potential to extend the useful lead-time of a forecast, if the coupled model responds to initial conditions in a realistic manner. In Canada, an AOGCM-based forecasting system has been developed through the Coupled Historical Forecasting Project (CHFP). This effort is a natural extension of the first and second Historical Forecasting Projects (HFP & HFP2), which formed the basis for Environment Canada's current AGCM-based forecasting system (Merryfield et al. 2010).

The finite inherent predictability of the climate system presents one of the primary challenges to seasonal forecasting. The influence of the slow-varying climate processes used in dynamical forecasting is often much smaller than the climate's internal variability; i.e. predictable signals are obscured by climate noise. The magnitude of this problem varies by region; for example, predictability is high in the tropics relative to the extratropics (Rowell 1998). Skillful extratropical forecasts are therefore often associated with accurate prediction of processes with the highest signal-to-noise ratios. Much of the seasonal forecast skill in North America can be attributed to the influence of the El Nino-Southern Oscillation (ENSO; e.g. Derome et al. 2001; Livezey and Timofeyeva 2008; Shukla et al. 2000), which generates a highly predictable signal over much of the

continent. Other teleconnection patterns have similar, if smaller, value as North American climate predictors, including the North Atlantic Oscillation (NAO; e.g. Hurrell and van Loon 1997) and the Pacific-North American Pattern (PNA; e.g. Leathers et al. 1991), among others.

Other key challenges in numerical forecasts involve the quality of the models being used. All models produce systematic error in their representation of the climate system and its integration from initial conditions. Limitations in resolution and model physics push the model integrations towards a biased interpretation of the true climate, and this drift obscures the simulated evolution from initial conditions. Furthermore, the response itself may deviate considerably from that of the true climate. Although a model may produce a predictable, systematic response to prescribed forcings, the spatial characteristics of this response may differ from that of the true climate. That is, a GCM's response may be statistically related to the true outcome, yet produce forecast fields with limited accuracy. These problems are exacerbated in coupled GCMs, as nonlinear interactions between component models can amplify any existing biases.

Much of the effort to improve numerical forecasts focuses on improving the models used. This work is necessary, but also expensive and time intensive. A complementary, and less resource intensive, approach involves the development of post-processing methods that identify and correct systematic model biases. Post-processing techniques typically use computationally efficient linear statistical analyses, such as regression, to relate GCM projections with true climate outcomes. This process has proven useful in improving temperature, precipitation, and teleconnection forecasts, (Jia et al. 2010; Lin et al. 2005; Lin et al. 2008). However, given that both the climate system and sources of GCM bias

involve nonlinear processes, it is possible that the use of linear statistics has limited the effectiveness of this approach. If this is the case, nonlinear analogues of the statistical methods applied in past studies should give superior results. In many cases, appropriate nonlinear methods exist, and have already proven useful in climate and weather applications. In particular, nonlinear regression techniques are well developed, and can be readily adapted to numerical climate forecast correction. Past studies have successfully applied nonlinear regression to seasonal forecasting of El Nino events (Wu et al. 2006), extreme precipitation (Zeng et al. 2011), and air temperatures (Wu et al. 2005), among other climate and weather applications.

The current study assesses the potential value of nonlinear regression in climate forecast post-processing, applying the method to HFP2 and CHFP hindcast data. Classical multivariate linear regression is compared against support vector regression, a machine learning method with nonlinear capabilities. The present study focuses on prediction of key teleconnection patterns influencing North American climate; these predictions are themselves of limited operational value, but are strongly related to forecast variables of value and provide a simple illustration of the post-processing methods' ability to extract a useful signal from forecast noise.

2. Data

Post-processing methods are tested here using numerical climate hindcasts produced for the first phase of Environment Canada's Coupled Historical Forecasting Project (CHFP; Merryfield et al. 2010) and second Historical Forecasting Project (HFP2; Kharin et al. 2009), downloaded from the Canadian Centre for Climate Modelling and Analysis

(CCCMA). The simulations examined were generated with version three of the CCCMA atmospheric GCM (AGCM3), a global spectral model run at T63 horizontal resolution with 32 vertical levels (Scinocca et al. 2008). The two prediction systems both generate ensembles of 10 parallel integrations for each hindcast period, facilitating comparison of their predictive skill. Brief descriptions of the initialization and ensemble generation schemes follow; a more detailed discussion is available in Merryfield et al. (2010).

The complete set of HFP2 forecasts includes output from four atmospheric models. For the purposes of the current student, only forecasts generated with AGCM3 have been used, in order to ease comparison with CHFP1. These hindcasts were generated by forcing AGCM3 with static sea surface temperature (SST) anomalies, observed during the month preceding a simulation's start date. Sea ice extents were also initialized to the previous month's observations, then relaxed to climatology over the first fifteen days of the simulation. Hindcasts extend four months from the start date, a period over which SST anomalies are expected to change relatively little. The ensemble is generated using different atmospheric initial conditions, taken from the NCEP-NCAR reanalysis (Kalnay et al. 1996) at 12 hr intervals preceding the simulation start date. Twelve hindcasts were generated for each year, beginning at the first of every month. Data cover the period 1969-2001 (Kharin et al. 2009).

CHFP1 uses the version 3 of the CCCMA Coupled GCM (CGCM3), which couples the AGCM3 with the National Center for Atmospheric Research Community Ocean Model (NCOM; Gent et al. 1998). The inclusion of a dynamic ocean boundary is intended to improve predictive skill at long lead times, and CHFP1 hindcasts have a twelve month duration. The 10 member ensemble was produced by combining two ocean

initializations, taken one and five days prior to the start date, and five atmospheric initializations, taken at 00:00Z from the five days preceding the simulation. These states were produced by a coupled model run in which SSTs were relaxed strongly to observed values leading up to the forecast, so that the atmospheric states contain the influence of observed SSTs. (Unlike the HFP2 initial states they do not contain details of the atmospheric state as represented by the NCEP-NCAR reanalysis at those times; this difference tends to degrade CHFP1 skills relative to HFP2 in the DJF period.) Four hindcasts were produced for each year, beginning on the first of December, March, June, and September respectively (Merryfield et al. 2010). Data cover 1972-2001.

The present study examines December (winter) runs valid for 1972-2001, covering the largest coincident period of the HFP2 and CHFP1 experiments. Winter runs were chosen due to the greater influence of synoptic weather systems and teleconnection indices on winter climate relative to other seasons. Stronger upper level westerlies at this time of year favor Rossby wave propagation, which in turn is influenced by teleconnections such as ENSO. This generally results in more skillful seasonal forecasts relative to other seasons, and renders it particularly well suited to numerical climate prediction. Results are evaluated against data from the NCEP-NCAR reanalysis and time series of prominent teleconnection patterns from the National Oceanographic and Atmospheric Administration's Climate Prediction Center (NOAA-CPC; http://www.cpc.ncep.noaa.gov). Loading patterns in the 500 hPa geopotential height (Z500) field for the North Atlantic Oscillation (NAO) and Pacific-North America (PNA) pattern were generated by regressing NAO and PNA time series from NOAA-CPC onto gridded anomaly fields from NCEP-NCAR reanalysis, producing regressed anomaly

("loading") maps that were subsequently normalized to a unit vector (Figures 1a and 2a). Separate loading patterns were calculated for overlapping three-month periods (e.g. DJF, JFM, FMA, etc) in order to reflect seasonal variations in the structure and intensity of the two teleconnections. CHFP1 and HFP2 predictions of the teleconnection indices were then calculated as the projection of the loading patterns onto the ensemble mean Z500 anomaly hindcast for each three-month period.

3. Methods

3.a Support Vector Regression:

Support vector regression (SVR) is an extension of the support vector machine method (developed originally for classification problems) to regression problems (Vapnik 1997). A brief description of the technique is described here; more detailed explanation can be found in Hsieh (2009).

Let **x** be a vector of predictor variables and y be a desired predictand. Linear relationships between y and **x** can be readily identified through classical linear regression, while nonlinear relationships cannot. However, a nonlinear mapping function, $\mathbf{\Phi}$, can be used to convert the nonlinear relationship into a linear regression problem, i.e.

$$f(\mathbf{x},\mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b, \qquad (1)$$

where <,> denotes the inner product, and **w** and b are the regression coefficients that minimize the error between predictions f and observations y. SVR calculates an ε-insensitive error norm as:

$$\left|y - f(\mathbf{x},\mathbf{w})\right|_{\varepsilon} = \begin{cases} 0, & \text{if } \left|y - f(\mathbf{x},\mathbf{w})\right| < \varepsilon \\ \left|y - f(\mathbf{x},\mathbf{w})\right| - \varepsilon, & \text{otherwise} \end{cases}, \qquad (2)$$

i.e., the error is ignored when smaller than ε, and approximates the mean absolute error when large. For a given training data set with sample size *N*, SVR calculates **w** and *b* coefficients that minimize the following error function *R*,

$$R = \frac{C}{N}\sum_{i=1}^{N}\left|f(\mathbf{x}_i,\mathbf{w}) - y_i\right|_{\varepsilon} + \frac{1}{2}\left\|\mathbf{w}^2\right\| \qquad , \qquad (3)$$

where *C* is a parameter that controls the relative influence of the two terms in the error function. Large values of *C* emphasize the first (ε-insensitive error) term at the expense of the second (regularization) term. Conversely, low values of *C* lead to a more prominent regularization term, penalizing (i.e. suppressing) complex models which tend to need weight vectors **w** with larger magnitude. Both *C* and ε are user-selected parameters, commonly referred to as hyperparameters.

By linearizing a nonlinear regression problem via (1), SVR eliminates the need for nonlinear optimization and problems associated with local minima of a nonlinear error function. In this regard, SVR presents an advantage over the neural network approach to nonlinear regression. The remaining concern is that the mapping function $\mathbf{\Phi}(\mathbf{x})$ may transform the data into a high dimensional space, and thereby become computationally expensive to solve. This problem is addressed through the use of a kernel trick, which replaces the inner product $<\mathbf{\Phi}(\mathbf{x}), \mathbf{\Phi}(\mathbf{x}')>$ with a scalar kernel function $K(\mathbf{x}, \mathbf{x}')$. For the

purposes of this study, a linear and a nonlinear kernel have been used for linear and nonlinear SVR, respectively. The linear kernel is simply the inner product $K(\mathbf{x}, \mathbf{x}_i) = <\mathbf{x}, \mathbf{x}_i>$. While the linear kernel does not allow the SVR to model nonlinear relations, linear SVR is not the same as the classic linear regression, as the ε-insensitive error norm renders it a robust linear regression method. Nonlinear SVR was performed with the following Gaussian kernel:

$$K(\mathbf{x},\mathbf{x}_i) = \exp\left(\frac{-\left\|\mathbf{x}-\mathbf{x}_i\right\|^2}{2\sigma^2}\right). \quad (4)$$

The use of this kernel adds σ as an additional hyperparameter, describing the width of the kernel.

The selection of hyperparameters has a significant impact on the performance of the SVR model. Although objective methods for selecting some hyperparameters have been proposed, in practice model optimization requires additional tuning. Here, the approach of Cherkassky and Ma (2002) has been used to generate initial objective estimates of the hyperparameters, which are subsequently tuned using a fine grid search.

3.b Post-processing Model Testing

Hindcasts of the NAO and PNA indices were post-processed with three regression algorithms; multivariate linear regression (LR), linear support vector regression (LSVR), and nonlinear support vector regression (NLSVR). Post-processing models use five predictors: the teleconnection hindcast from the unprocessed model output, and the time

series of four leading patterns from one of the following decompositions of GCM output: i) principal components (PCs) of the Z500 anomaly field over the northern hemisphere (45-90^{o}N), ii) PCs of tropical Pacific SST anomalies (20^{o}S-20^{o}N, 120-270^{o}E), or iii) a singular value decomposition (SVD) (i.e. maximum covariance analysis; von Storch and Zwiers 2001) of these two fields, giving spatial patterns for both. In the present paper, anomaly refers to deviations from climatology, with monthly mean fields subtracted from the raw data. As HFP2 does not predict the ocean state, SST predictors for HFP2 were derived from the static SST anomaly driving the AGCM3. CHFP1 predictors were derived from the forecast state of both SSTs and Z500. The predictor sets were chosen for their similarity to sets used in previous post-processing studies (e.g. Lin et al. 2008), and for their relative simplicity. The first step in model training selects one of the four PC or SVD time series as the best predictor set, which is then used in subsequent training steps. The result is that individual post-processing models use the unprocessed teleconnection index with additional spatial data related to a single field (SST or Z500), summarized with a single decomposition method (PC or SVD). Although it is possible that combining various predictors from the three GCM decompositions could generate better forecasts, there are concerns that the skill of forecasts produced with screened predictor sets can be overstated (DelSole and Shukla 2009). The approach used here represents a compromise between limiting the complexity of predictor screening, while maintaining some flexibility to account for the possibility that the relative value of the ocean and atmospheric component models may change as a coupled run progresses.

For CHFP1 data, separate regression models were built for each of ten overlapping three-month periods. Hindcasts used begin in December, and run through to the end of the

following November. The run is then split into ten forecast periods; December/January/February (DJF), JFM, etc. through to SON. Predictors are derived from hindcast data valid for a given forecast period; i.e. the DJF post-processing model uses the hindcast output averaged over DJF. DJF forecasts represent prediction at a zero-month lead, JFM at one-month lead, etc. For each lead time, the best predictor set was first identified, using a coarse 5-fold cross-validation linear regression procedure to compare their relative value.

The selected predictor set was then used by the LR, LSVR and NLSVR methods for post-processing. The use of linear regression to select predictor sets may give LR and LSVR a slight advantage over NLSVR, hence any improved skill by NLSVR over the linear methods is likely to be understated. The three methods were each evaluated using a rigorous double cross-validation scheme, illustrated in Figure 3. Thirty-fold cross-validation produced predictions for each year, using a model trained with the majority of the remaining years in the study period (outer round, or CV1 in Figure 3). Years immediately preceding and following the year being predicted were excluded from the training period, in order to reduce information leakage from autocorrelation.

In the case of LSVR and NLSVR, hyperparameters were selected using a second nine-fold cross-validation applied to the training data (inner round, or CV2 in Figure 3). Values identified as optimal for the training data were then used in an SVR model to predict the testing year in CV1. This double cross-validation scheme is employed to avoid validating forecast skills with data already used for model training and hyperparameter selection.

4. Results

4.a Teleconnection Loading Patterns & Relationship to Predictors

It is helpful to begin with an examination of the physical structure of the teleconnections being predicted, and considering the ability of the forecast models to reproduce them. Figures 1a) and 2a) respectively show the Z500 loading patterns of the PNA and NAO used in the current study, for the DJF period. The PNA features two opposing centres of action, located over the North Pacific and Western North America. Similarly, the NAO places opposing centres over the climatological Iceland Low in the North Atlantic and the Azores High in the subtropical Atlantic. The spatial characteristics of these correlation-derived patterns are essentially identical to those used by the NCDC. Time series of the amplitudes of these patterns also agree very closely to the original NAO and PNA time series used to define them, giving correlation scores above 0.9.

Past studies have shown that an SVD of the Z500 and tropical SST fields used to generate predictor sets in the current work produce patterns related to the teleconnections of interest. Specifically, the first Z500 SVD (SVD1) shares spatial characteristics with the PNA, while the second (SVD2) typically resembles the NAO (Lin & Derome, 2005). This emphasizes the influence of coupled atmosphere-ocean processes on these teleconnections, and provides insight into the physical mechanisms behind these phenomena. In the context of the present study, it follows that an SVD analysis of GCM output can provide an indication of whether the model accurately simulates these mechanisms.

Results of SVD analyses of NCEP data (Figures 1b & 2b), HFP2 output (Figures 1c & 2c), and CHFP1 output (Figures 1d & 2d) for DJF are presented alongside the teleconnection loadings. The NCEP SVDs both feature centres of action similar to the associated loading patterns. However, while the teleconnection patterns de-emphasize regions outside of these centres, the NCEP SVDs also place high loadings in other areas of the Northern Hemisphere. The two teleconnections appear to be somewhat convoluted in the SVD analysis, with SVD1 emphasizing the PNA centres but also introducing weak loadings in the vicinity of the Iceland Low and Azores High. SVD2 emphasizes the Iceland Low and Azores High, but shifts their location relative to the teleconnections. It also introduces a dipole between western and eastern North America, resembling a weakened and eastwards shifted PNA pattern.

SVD results for HFP2 output closely resemble NCEP results, particularly for SVD1. There is again a tendency to convolve PNA and NAO-like patterns, although the NAO dipole is less well-defined. However, CHFP1 output gives very different results. CHFP1 SVDs feature weaker centres of action, and less defined features in general than NCEP or HFP2. This could result from greater variability in atmosphere-ocean interactions, perhaps related to the adjustment process of the component models during the early stages of the coupled runs. Despite this, CHFP1 SVD1 closely resembles the PNA loading pattern, with a well-defined Pacific/North America dipole. The NAO is not captured well in either SVD, with SVD2 showing only an Azores High combined with a weaker PNA dipole. It can be inferred that the model captures PNA-like processes well, but may be misrepresenting those related to the NAO.

Table 1 summarizes relationships between the SVDs and the teleconnection indices, showing correlation scores between SVD time series and observed PNA and NAO indices. The strong relationship between NCEP SVD1 and the PNA is confirmed with high values in both DJF and JFM. The PNA-like aspects of NCEP SVD2 also leads to reasonably strong correlations. However, neither SVD is strongly related to the observed NAO. HFP2 SVD1 also correlates reasonably well with the observed PNA, suggesting this is a relatively useful predictor. This is not the case for CHFP1; although the CHFP1 SVD1 spatial pattern resembles the PNA, the temporal amplitude of this pattern in the CHFP1 output does not provide a useful estimate of the teleconnection's state. Surprisingly, SVD2 does show a statistically significant, if weak, relationship with the observed NAO, suggesting there is a signal related to the NAO embedded in a structurally dissimilar pattern.

Table 2 gives the number of times the four predictor sets described in the previous section were identified as the best basis for a post-processing model, using the linear regression-based screening process. Z500 SVDs were selected most often, likely due to their marked relationship to the PNA and NAO.

4.b Post-processing Results

Performance of unprocessed and processed model simulations are primarily evaluated in terms of correlation between observed and projected teleconnection indices, with results presented in Tables 3 (HFP2) and 4 (CHFP1). Post-processing of CHFP1 data are also evaluated using correlation, mean absolute error (MAE), and index of agreement

(Willmott et al. 1985), reported as skill scores (SS) relative to the unprocessed CHFP1 output (Figures 4 & 5), with SS defined by

$$\mathrm{SS} = \frac{A - A_{\mathrm{ref}}}{A_{\mathrm{perfect}} - A_{\mathrm{ref}}}, \qquad (5)$$

where A is a particular accuracy measure, A_{ref} is the value of A from a reference forecast set (here the unprocessed CHFP1 output), and A_{perfect} is the value for a set of perfect forecasts. As presented here, SS illustrates the improvements of post-processing methods relative to raw CHFP1 output.

It is useful to first compare performance of the raw model projections from the two forecasting systems, both to provide a baseline for post-processing assessment and to evaluate the relative strengths of the AOGCM and AGCM approaches. HFP2 performs reasonably well for both teleconnections at all lead times (Table 3). Correlation scores larger than 0.33 indicate statistical significance at the 95% confidence level, and this value is exceeded by all unprocessed HFP2 forecasts. However, in absolute terms these correlations are relatively low, with projections capturing only 9-25% of teleconnection variability. CHFP1 output is considerably less reliable (Table 4). Although capable of predicting the PNA index relatively well at short lead times (>0.5 for JFM & FMA), CHFP1 scores decrease rapidly at longer leads. NAO simulation is uniformly poor, with the model producing no statistically significant positive correlations. These results suggest the CHFP1 offers a negligible improvement over the simpler HFP2 forecasting system, extending useful PNA forecasts by one forecast period but degrading simulation of the NAO. Although not shown, MAE values are comparable for the two forecasting

systems, ranging from 0.5-1 of the observed teleconnection's standard deviation for all forecast periods.

The post-processing methods used here vary considerably in their effectiveness when applied to HFP2 output (Table 3). Post-processing reduces correlation scores to statistically insignificant values for the NAO, but all methods tend to improve PNA forecasts. This may be due to the fact that the additional predictors (the SVD and PC time series from Z500 and tropical Pacific SST) are less directly related to NAO than to PNA, and extra irrelevant predictors tend to decrease skills. The two linear methods give comparable results and outperform NLSVR, which fails to improve the DJF PNA prediction. The post-processing of CHFP1 output is considerably more effective (Table 4; Figures 4 & 5). Although the methods used here have difficulty improving the few statistically significant correlations found in the raw forecasts (e.g. PNA at JFM & FMA), they frequently improve on the poorer CHFP1 results. In the case of the NAO, the only significant results are generated through post-processing which yields relatively high correlations even at a nine-month lead (SON). Similarly, relatively high PNA correlations are generated for most lead times by one or more of the methods. Although no post-processing method consistently generates the highest correlations, NLSVR has the greatest success rate, giving the best scores for five three-month periods for the NAO and four for the PNA, accounting for almost half of the twenty post-processing periods. By contrast, linear regression gives the best correlation in four instances (FMA, MJJ, JJA, and ASO for the NAO), and LSVR performs best in another four cases (NAO at DJF, and PNA at MAM, AMJ & SON). Unprocessed CHFP1 output gives the best results in only three instances (PNA at JFM, FMA and MJJ). It should be noted that the statistically

significant post-processed correlations still capture only 10-30% of teleconnection variability. With the 30 year sample available, it is difficult to assess the source of this skill. However, the ability of all methods to generate significant correlations at extended leads suggests a useful signal is embedded in the model output.

Figure 4 also presents mean absolute error (MAE) skill, relative to unprocessed CHFP1 forecasts. As with correlation, MAE results indicate post-processing generally improves CHFP1 forecasts, but the relative advantage of any single method is less clear. LR marginally outperforms NLSVR, giving the lowest MAE for eight prediction periods to NLSVR's six. LSVR and the unprocessed model both give the best results for only three periods.

To better illustrate the relative advantages of various post-processing methods, Willmott's index of agreement (IA) has also been used to assess the skill of CHFP1-derived forecasts (Willmott et al. 1985), with

$$\mathrm{IA} = 1 - \left[\frac{\sum_{i=1}^{N} (f_i - y_i)^2}{\sum_{i=1}^{N} (|f_i'| + |y_i'|)^2} \right], \tag{6}$$

where

$$\begin{aligned} f' &= f - \bar{y} \\ y' &= y - \bar{y} \end{aligned}, \tag{7}$$

with f the predicted value, and y the observed. IA incorporates elements of MSE and correlation into a single value between zero and one, with one being a perfect score.

Results are given in Figure 5, again presented as skill scores relative to the raw model forecast. This measure emphasizes the value of nonlinear post-processing, with NLSVR forecasts improving CHFP1 output in all but five cases, and giving the highest positive skill in nine. By contrast, LR and LSVR reduce raw CHFP1 skill in a substantial number of cases (seven each), and combined provide the highest positive skill in only seven instances.

5. Summary & Discussion

Results suggest regression-based post-processing is most effective when applied to coupled model output, improving CHFP1 output considerably but only improving PNA forecasts when applied to HFP2. The inference is that CHFP1 output contains a useful signal that is obscured by biases in the model's integration from a prescribed initial condition. Although CHFP1 may not faithfully reproduce loading patterns associated with the NAO and PNA, it does produce a predictable evolution related to the observed teleconnections. This relationship appears to be nonlinear, allowing NLSVR to outperform linear post-processing methods in the majority of cases. The lower performance of LSVR relative to NLSVR confirms it is the nonlinear capabilities that give this method an advantage, rather than the SVR methodology itself.

Results indicate that HFP2 captures the structure of the NAO and PNA better than CHFP1, and recreates teleconnection time series relatively well prior to post-processing. This is related to the use of an atmosphere-only GCM, which removes the influence of ocean model and coupled atmosphere/ocean error present in the CHFP1 simulations.

However, regression-based post-processing improved only PNA forecasts from HFP2 output, a result not seen in past studies using similar methods (Lin et al. 2005). Reduced performance in the current study may be related to differences in predictor selection, the time period examined, and the double cross-validation methodology employed here. It is also possible that the HFP2 predictors supplied in this study are not sufficiently related to the NAO; with the addition of irrelevant predictors, it is expected that forecast skills will drop. However, similar predictor sets have produced better skill in previous studies (e.g. Lin et al. 2005). Additional sensitivity experiments indicate results are most sensitive to variations in the cross-validation approach, such as changes in the number of cross-validation steps and relaxing the auto-correlation buffer in the outer CV round (Figure 3). In contrast, using predictors and periods identical to those in past studies did not change results appreciably. As the LR-processed skill is lower in the present study than previously reported suggests the approach used here gives a conservative assessment of performance. It follows that performance of processed CHFP1 output and SVR-processed results may also be underestimated, adding weight to the argument that nonlinear post-processing offers valuable improvements to this data set.

Much of the skill in current North American climate forecasts has been attributed to the El Nino/Southern Oscillation (Livezey and Timofeyeva 2008), which produces a persistent, predictable response in large-scale climate. ENSO also exerts a strong influence on the PNA, and has been tentatively associated with the NAO (Hoerling et al. 2001). In order to assess the influence of ENSO on post-processed forecasts, correlation scores for the CHFP1 were recalculated without years initialized with a strong ENSO signal, defined as a NINO3 index of magnitude 1.5 or higher; results are presented in

Figure 5. Removing strong ENSO events reduces raw CHFP1 skill somewhat, leaving only a single significant correlation (PNA at FMA). Post-processing scores are also generally reduced, particularly the LR and LSVR results. However, NLSVR continues to generate statistically significant correlations for almost half of the forecasts. A substantial portion of the skill generated by nonlinear post-processing is apparently independent of the persistent, predictable influence of strong ENSO events. This approach therefore appears to generate the most versatile and robust forecasts, while the raw model, LSVR, and LR forecasts are less suitable for non-ENSO years.

Although the present study demonstrates the value of SVR as post-processing tool in climate forecasting, significant obstacles remain if it is prove useful operationally. Of particular concern is the method's greater complexity relative to traditional linear regression, and a lack of familiarity with this method in the meteorological community. For SVR post-processing to be adopted by forecasting centers, a clear advantage over linear regression must be displayed. The results presented here suggest the additional complexity yields valuable improvements, even when tested on a relatively short set of hindcast experiments. Although PNA and NAO forecasts are of little operational value themselves, these teleconnections are strongly connected with surface temperature and precipitation over North America. The implication is that SVR post-processing should also be capable of improving climate forecasts of these more valuable fields. A cell-by-cell post-processing approach, similar to that presented by Lin et al. (2008), could be used to test this hypothesis. However, such an exercise with SVR is computationally intensive, and beyond the scope of the current paper. Still, the current work suggests the method has some value and could be used to improve seasonal forecasts considerably.

Acknowledgements

J. Finnis and W. Hsieh are grateful for the support from the Canadian Foundation for Climate and Atmospheric Sciences via project GOAPP. W. Hsieh is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

		DJF		JFM	
		SVD1	SVD2	SVD1	SVD2
	Obs.	**0.65**	**-0.48**	**0.88**	**-0.40**
PNA	HFP2	**0.41**	-0.27	**0.57**	-0.16
	CHFP1	0.24	-0.31	0.29	-0.19
	Obs.	0.21	0.32	0.28	0.06
NAO	HFP2	-0.23	-0.13	0.32	-0.17
	CHFP1	0.26	**0.33**	0.16	0.02

Table 1: Pearson correlation scores between teleconnection time series (PNA and NAO), from observed, HFP2 and CHFP1 data, and the first two Z500 SVD time series from observed (NCEP) data. Correlations significant at the 95% are in bold type.

	Decomposition Method	Times Selected	
		NAO	PNA
SST	PC	4	1
	SVD	1	1
Z500	PC	0	2
	SVD	5	6

Table 2: Frequency with which the four predictor sets were chosen for model training, as a best set is chosen for each of the ten lead times and the two teleconnections.

1

	NAO Projections			
	HFP2	LR	LSVR	NLSVR
DJF	**0.361**	0.112	0.185	0.063
JFM	**0.361**	0.184	0.033	-0.093

	PNA Projections			
	HFP2	LR	LSVR	NLSVR
DJF	**0.396**	**0.440**	**0.473**	**0.354**
JFM	**0.487**	**0.621**	**0.607**	**0.577**

2

3 Table 3: Pearson correlation scores between the observed NAO and PNA time series and
4 projections from unprocessed and processed HFP2 data. All three month periods covered
5 by the HFP2 data set are shown. Correlations significant at the 95% level are in bold
6 type.

	NAO Projections			
	CHFP	LR	LSVR	NLSVR
DJF	0.234	0.313	**0.402**	0.326
JFM	0.191	0.056	0.179	**0.566**
FMA	0.150	0.202	0.084	0.013
MAM	0.076	0.200	0.313	**0.466**
AMJ	0.169	0.298	0.243	**0.477**
MJJ	-0.230	0.241	0.130	-0.163
JJA	-0.214	**0.417**	0.163	0.267
JAS	-0.260	**0.337**	0.183	**0.409**
ASO	-0.371	**0.393**	0.292	**0.358**
SON	-0.178	0.140	0.266	**0.332**
Mean	-0.043	0.260	0.225	0.305

	PNA Projections			
	CHFP	LR	LSVR	NLSVR
DJF	0.242	0.279	0.309	**0.340**
JFM	**0.522**	**0.390**	0.277	**0.344**
FMA	**0.573**	**0.520**	**0.529**	**0.541**
MAM	**0.374**	**0.388**	**0.403**	0.277
AMJ	0.084	0.287	**0.442**	**0.376**
MJJ	0.126	-0.123	-0.214	-0.105
JJA	0.320	0.298	0.257	**0.364**
JAS	0.264	0.326	**0.381**	**0.402**
ASO	-0.274	-0.197	-0.291	0.236
SON	0.023	-0.060	0.149	0.118
Mean	0.225	0.211	0.224	0.289

Table 4: Pearson correlation scores between the observed NAO and PNA time series and projections from unprocessed and processed CHFP1 data. All three month periods covered by the CHFP1 data set are shown. Correlations significant at the 95% level are in bold type.

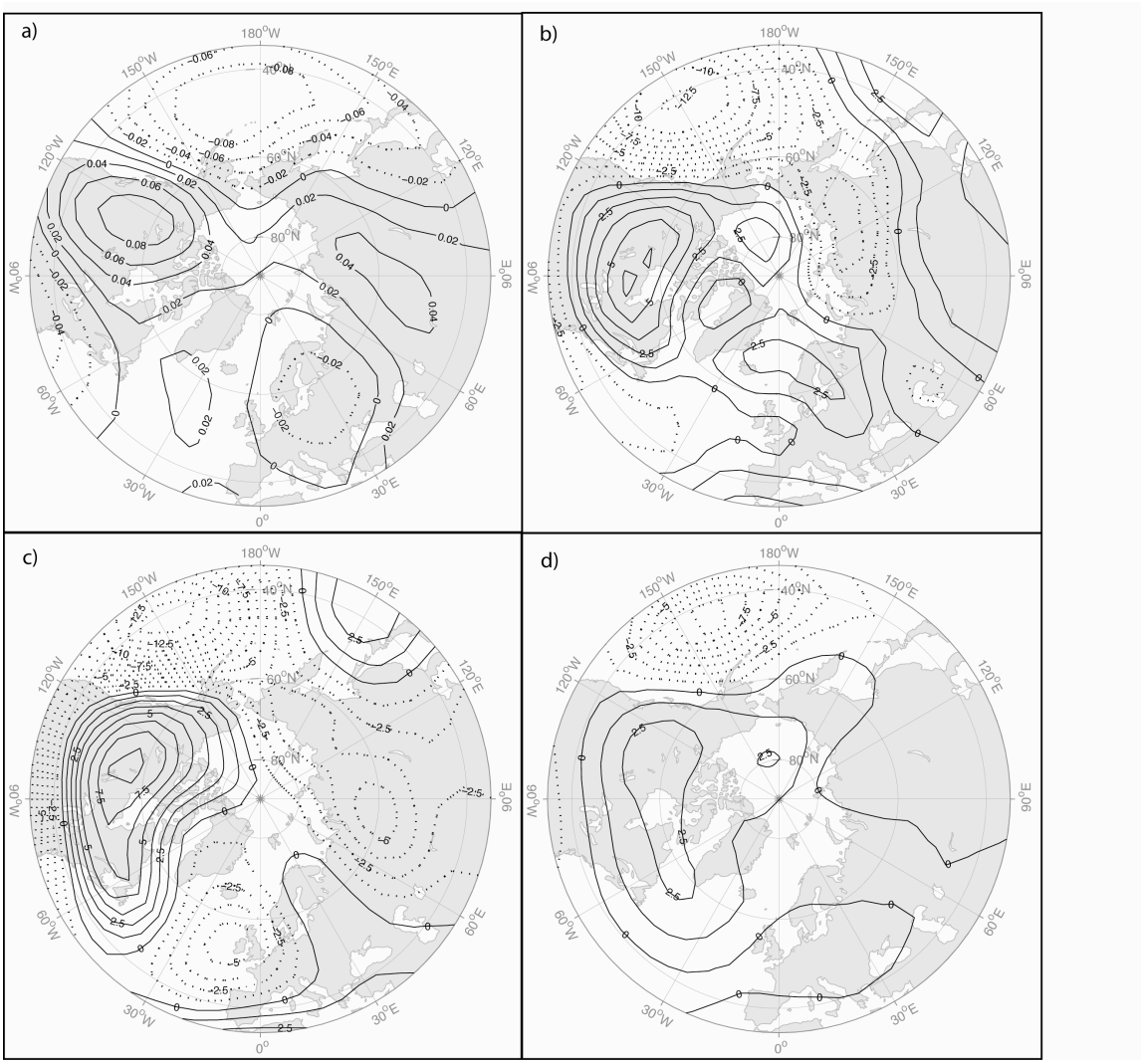
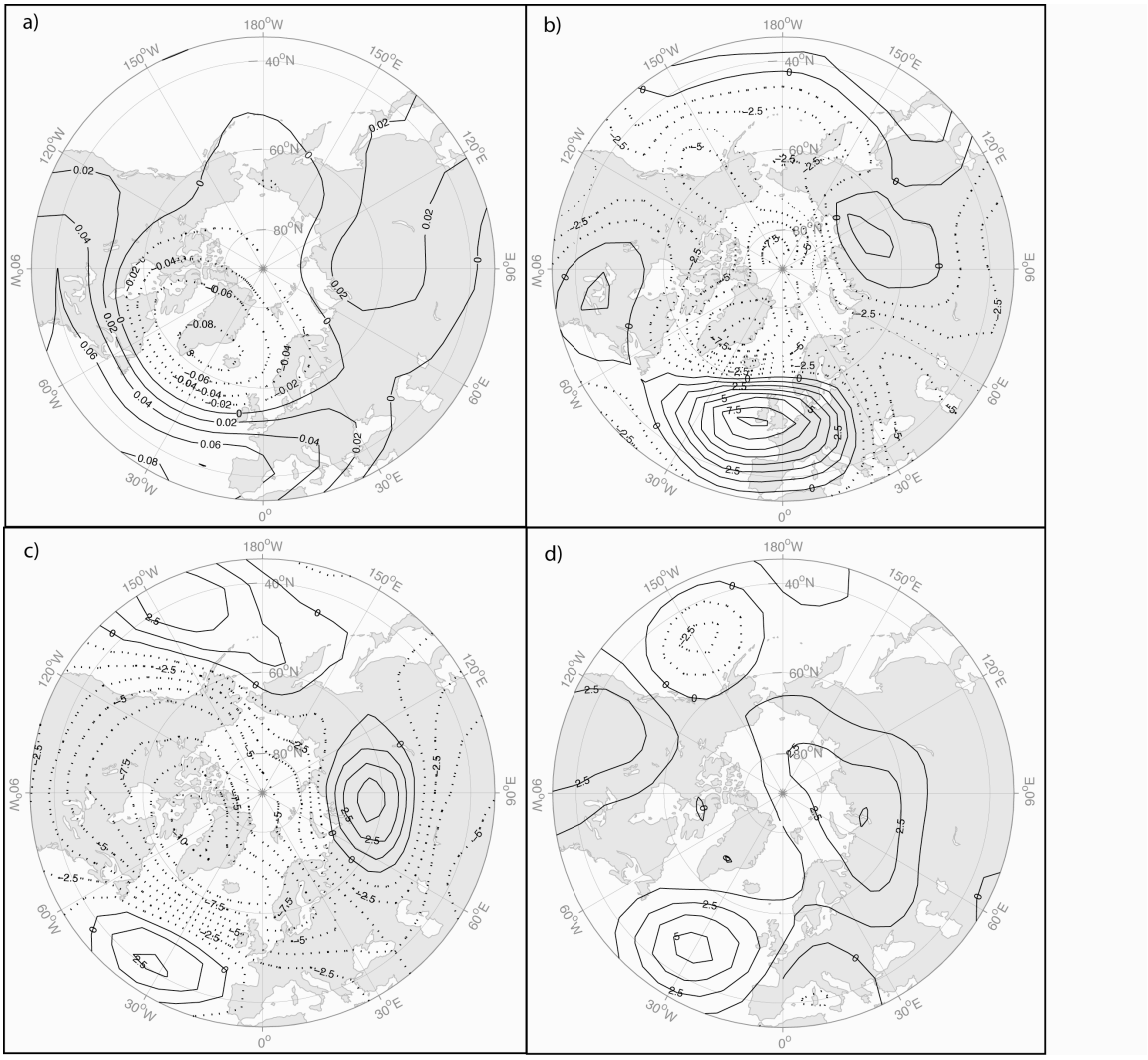


Figure 1: (a) The DJF Z500 PNA loading pattern (unitless), and the first Z500 SVD pattern from (b) NCEP, (c) HFP2, and (d) CHFP1, in units of meters.

1



2

3 Figure 2: (a) The DJF Z500 NAO loading pattern (unitless), with the second Z500 SVD
4 pattern from (b) NCEP, (c) HFP2 and (d) CHFP1, in units of meters.

5

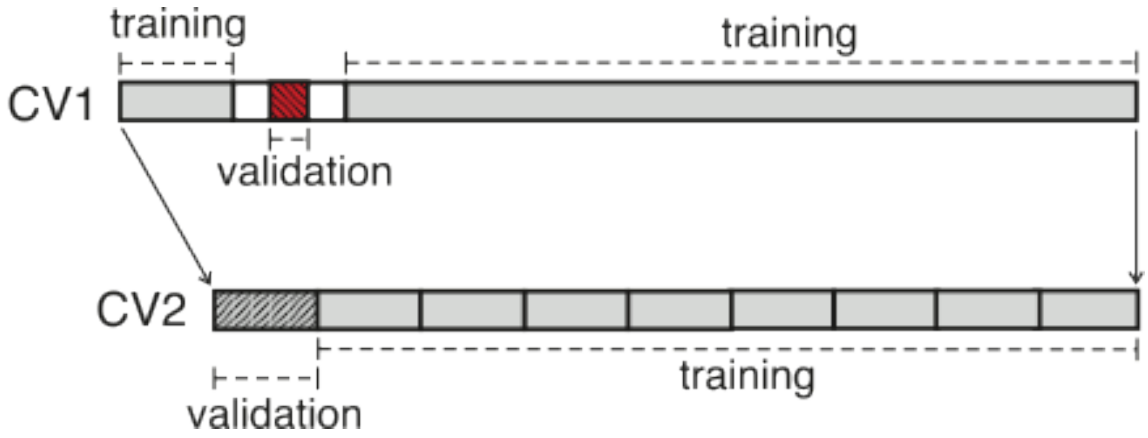


Figure 3: Illustration of the double cross-validation approach to model training and testing. An outer round of validation (CV1) is used to generate hindcast time series, using models with hyperparameters selected through an inner round of cross-validation (CV2). In CV1, the training data are shown in grey and the 1-year validation data shaded. The 1-year data segments (shown in white) bridging the training data and the validation data are not used, to avoid autocorrelation leaking information from the training data to the adjacent validation data. The validation data segment is moved repeatedly in 1-year increments from the start of the data record to the end in this cross-validation loop, so forecast performance is validated over the whole record. Meanwhile in CV2, the training data from CV1 are assembled and divided into 9 segments, with 8 used for training and one (shaded) for validation. Again the training and validation segments are rotated in the loop so all segments are eventually used for validation to determine the optimal model values/hyperparameters. The optimal model determined from CV2 is then used to forecast over the 1-year validation segment in CV1.

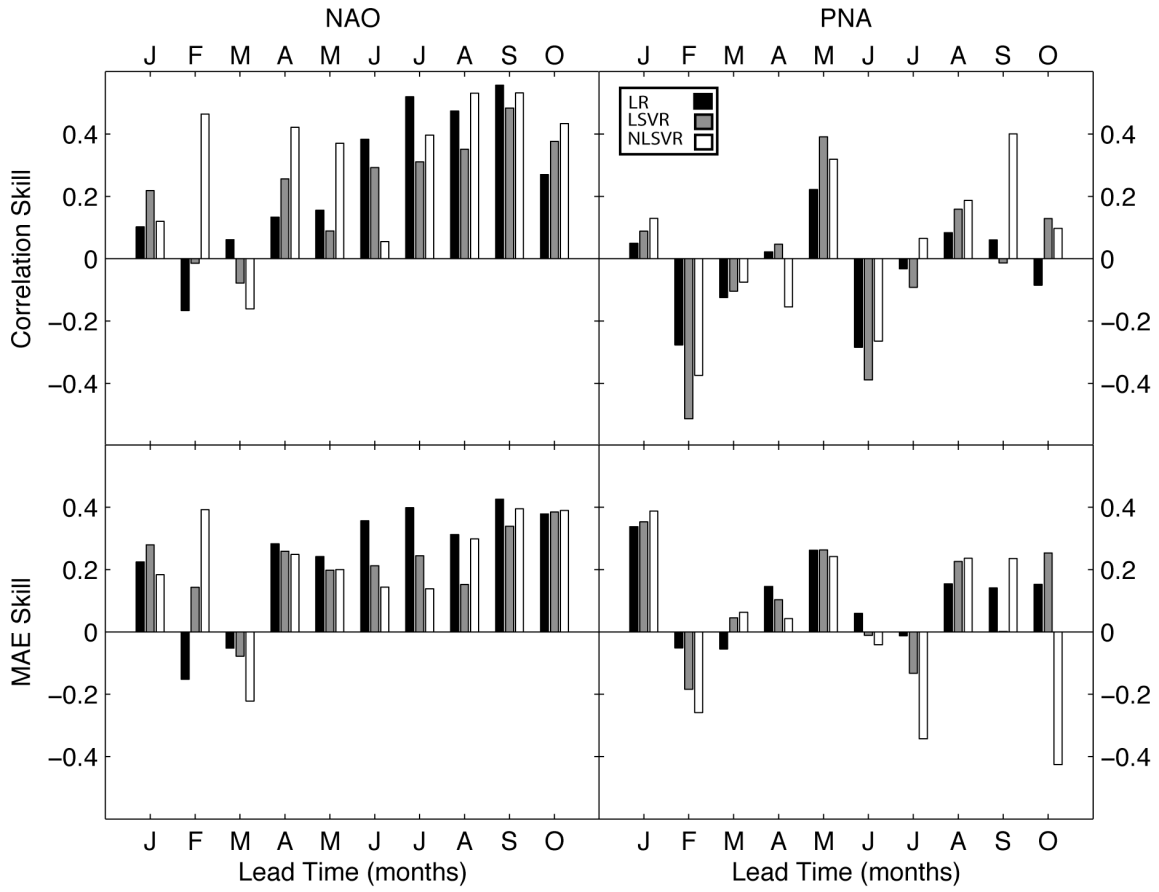
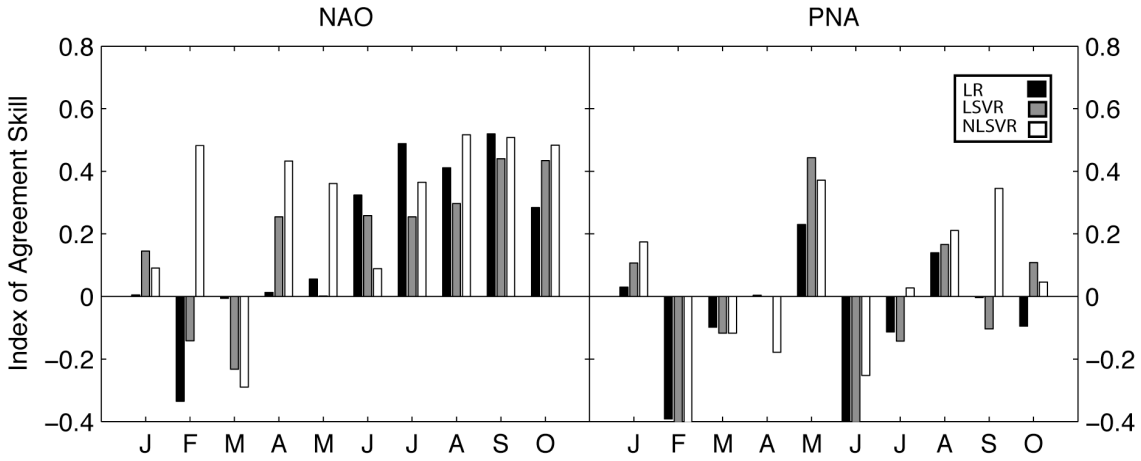


Figure 4: Correlation and mean absolute error skill scores of post-processed CHFP1 output, with the unprocessed CHFP1 output used as the reference prediction. Prediction periods are indicated by the first letter of the central month in each three-month period.

1



2

3 Figure 5: Same as Figure 4, but showing index of agreement skill scores.

4

5

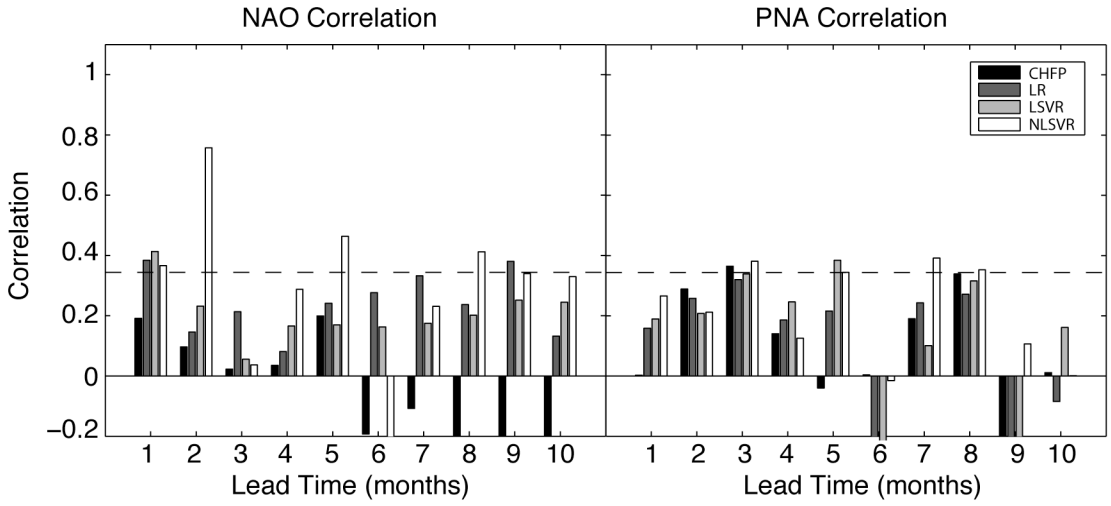


Figure 6: Correlation scores of raw and post-processed CHFP1 predictions, calculated without years initialized during moderate-to-strong ENSO events.

References

Cherkassky, V., and Y. Q. Ma, 2002: Selection of meta-parameters for support vector regression. *Lect Notes Comput Sc*, **2415,** 687-693.

DelSole, T., and J. Shukla, 2009: Artificial Skill due to Predictor Screening. *J Climate*, **22,** 331-345, doi: 10.1175/2008jcli2414.1.

Derome, J., and Coauthors, 2001: Seasonal predictions based on two dynamical models. *Atmos Ocean*, **39,** 485-501.

Gent, P. R., F. O. Bryan, G. Danabasoglu, S. C. Doney, W. R. Holland, W. G. Large, and J. C. McWilliams, 1998: The NCAR Climate System Model global ocean component. *J Climate*, **11,** 1287-1306.

Hoerling, M. P., J. W. Hurrell, and T. Y. Xu, 2001: Tropical origins for recent North Atlantic climate change. *Science*, **292,** 90-92.

Hsieh, W., 2009: *Machine Learning Methods in the Environmental Sciences.* Cambridge University Press.

Hurrell, J. W., and H. van Loon, 1997: Decadal variations in climate associated with the north Atlantic oscillation. *Climatic Change*, **36,** 301-326.

Jia, X. J., H. Lin, and J. Derome, 2010: Improving Seasonal Forecast Skill of North American Surface Air Temperature in Fall Using a Postprocessing Method. *Mon Weather Rev*, **138,** 1843-1857, doi: 10.1175/2009mwr3154.1.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *B Am Meteorol Soc*, **77,** 437-471.

Kharin, V. V., Q. B. Teng, F. W. Zwiers, G. J. Boer, J. Derome, and J. S. Fontecilla, 2009: Skill Assessment of Seasonal Hindcasts from the Canadian Historical Forecast Project. *Atmos Ocean*, **47,** 204-223, doi: 10.3137/Ao1101.2009.

Leathers, D. J., B. Yarnal, and M. A. Palecki, 1991: The Pacific North-American Teleconnection Pattern and United-States Climate .1. Regional Temperature and Precipitation Associations. *J Climate*, **4,** 517-528.

Lin, H., J. Derome, and G. Brunet, 2005: Correction of atmospheric dynamical seasonal forecasts using the leading ocean-forced spatial patterns. *Geophys Res Lett*, **32,** -, doi: 10.1029/2005gl023060.

Lin, H., G. Brunet, and J. Derome, 2008: Seasonal forecasts of Canadian winter precipitation by postprocessing GCM integrations. *Mon Weather Rev*, **136,** 769-783, doi: 10.1175/2007mwr2232.1.

Livezey, R. E., and M. M. Timofeyeva, 2008: The first decade of long-lead US seasonal forecasts - Insights from a skill analysis. *B Am Meteorol Soc*, **89,** 843-854, doi: 10.1175/2008bams2488.1.

Merryfield, W. J., W. S. Lee, G. J. Boer, V. V. Kharin, B. Pal, J. F. Scinocca, and G. M. Flato, 2010: The First Coupled Historical Forecasting Project (CHFP1). *Atmos Ocean*, **48,** 263-283, doi: 10.3137/Ao1008.2010.

Rowell, D. P., 1998: Assessing Potential Seasonal Predictability with an Ensemble of Multidecadal GCM Simulations. *J Climate*, **11,** 109-120, doi: 10.1175/1520-0442(1998)011<0109:APSPWA>2.0.CO;2.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J Climate*, **19,** 3483-3517, doi: 10.1175/JCLI3812.1.

Scinocca, J. F., N. A. McFarlane, M. Lazare, J. Li, and D. Plummer, 2008: Technical Note: The CCCma third generation AGCM and its extension into the middle atmosphere. *Atmos Chem Phys*, **8,** 7055-7074.

Shukla, J., and Coauthors, 2000: Dynamical seasonal prediction. *B Am Meteorol Soc*, **81,** 2593-2606.

Vapnik, V., Golowich, S., Smola, A., 1997: Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, M. Mozer, Jordan, M., Petsche, T., Ed., MIT Press, 281-287.

von Storch, H., and F. W. Zwiers, 2001: *Statistical analysis in climate research.* 1st pbk. ed. Cambridge University Press, x, 484 p. pp.

Willmott, C. J., and Coauthors, 1985: Statistics for the Evaluation and Comparison of Models. *J Geophys Res-Oceans*, **90,** 8995-9005.

Wu, A. M., W. W. Hsieh, and A. Shabbar, 2005: The nonlinear patterns of North American winter temperature and precipitation associated with ENSO. *J Climate*, **18,** 1736-1752.

Wu, A. M., W. W. Hsieh, and B. Y. Tang, 2006: Neural network forecasts of the tropical Pacific sea surface temperatures. *Neural Networks*, **19,** 145-154, doi: 10.1016/J.Neunet.2006.01.004.

Zeng, Z., W. W. Hsieh, A. Shabbar, and W. R. Burrows, 2011: Seasonal prediction of winter extreme precipitation over Canada by support vector regression. *Hydrol Earth Syst Sc*, **15,** 65-74, doi: 10.5194/Hess-15-65-011.