Global Sea Surface Temperature Forecasts Using a Pairwise Dynamic Combination Approach

SHAHADAT CHOWDHURY AND ASHISH SHARMA

School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, Australia

(Manuscript received 22 January 2010, in final form 16 November 2010)

ABSTRACT

This paper dynamically combined three multivariate forecasts where spatially and temporally variant combination weights are estimated using a nearest-neighbor approach. The case study presented combines forecasts from three climate models for the period 1958–2001. The variables of interest here are the monthly global sea surface temperature anomalies (SSTA) at a $5\times5$ latitude–longitude grid, predicted 3 months in advance. The forecast from the static weight combination is used as the base case for comparison. The forecasted sea surface temperature using the dynamic combination algorithm offers consistent improvements over the static combination approach for all seasons. This improved skill is achieved over at least 93% of the global grid cells, in four 10-yr independent validation segments. Dynamically combined forecasts reduce the mean-square error of the SSTA by at least 25% for 72% of the global grid cells when compared against the best-performing single forecast among the three climate models considered.

1. Introduction

There are many modeling alternatives for seasonal hydroclimatic prediction. The use of a single best model over an entire ensemble often leads to poorer predictive performance as a result of the greater associated structural uncertainty. This paper investigates options for combining multisite responses from multiple models with consideration for temporal variability in individual model skills. While linear combinations of multimodel responses have been used to reduce the predictive uncertainty of climatic and hydrological variables (Barnston et al. 2003; Fraedrich and Smith 1989; Peña and Van den Dool 2008; Peng et al. 2002; Raftery et al. 2005; Sanders 1963; Thompson 1977), little has been done to take advantage of increased skills offered by individual models under specific conditions or over specific periods. We present here a basis for a linear model combination that allows explicit consideration of changes in model skill, by allowing the model combination weight (or the proportional contribution each ensemble member makes to the model ensemble) to change with time. Such an approach is also similar in intent to the rationale behind dynamic linear models (Huerta and Sansó 2007; Lundberg et al. 2000; West and Harrison 1997), where the parameters of the model (analogous to the model combination weights here) vary with time.

The effectiveness for model combination weights that vary over time has been investigated by Chowdhury and Sharma (2009a, hereafter CS2009). Such weights are termed dynamic model combination weights [or dynamic weights (DW)]. The improvement resulting from dynamic weight for 3-month-ahead forecasts of the Niño-3.4 index in contrast to temporally invariant weights [or static weights (SW)] has been documented in CS2009. While the CS2009 study was limited to the prediction of a univariate response (Niño-3.4), this paper extends the method for prediction of gridded global sea surface temperature anomalies (SSTAs). The existing developments (which mainly lie outside of climate science discipline) that set the background of our current research are discussed in Chowdhury (2009, p. 108) and in Chowdhury and Sharma (2009b).

This paper is organized as follows. We first describe the multivariate dynamic weight formulation and propose a way for forecasting the dynamic weights forward in time. Then we present the application of the proposed methodology to dynamically combine global sea surface forecasts from three predictive models. Finally, we secuiritize the results with relevant discussion and conclusions.
2. Methodology

The forecasts issued by various models are combined using parameters referred to as combination weights in this paper. Let us introduce the combination of two single-model predictions, \( \hat{u}_{1,t} \) and \( \hat{u}_{2,t} \), at a time \( t \). The dynamic combination logic can be expressed as

\[
y_t = \hat{u}_{1,t}w_t + \hat{u}_{2,t}(1 - w_t) + \tilde{e}_t,
\]

where \( y_t \) is observed predictand (SSTA in this paper) at time \( t \), \( w_t \) is model 1 dynamic weight for time \( t \), and \( \tilde{e}_t \) is the residual of the combined prediction at time \( t \).

The algebraic solution of the weight \( w_t \) in Eq. (1) that leads to \( \tilde{e}_t = 0 \) is

\[
w_t = \frac{e_{2,t}}{e_{2,t} - e_{1,t}},
\]

where \( e_{1,t} = (y_t - \hat{u}_{1,t}) \) and \( e_{2,t} = (y_t - \hat{u}_{2,t}) \) are the respective residuals for models 1 and 2.

Analytically, \( w_t \) may span from \(-\infty\) to \(+\infty\). The methodology sensibly expects that the individual models are unbiased, and hence weights are constrained to be \( \{w_t \in [0, 1]\} \). This constraint allows for combination only when two forecasts are bracketing the true value. Alternatively, when both forecasts exhibit the same directional bias, the more accurate forecast is selected.

The dynamic combination problem can now be considered as a multivariate time series forecasting problem, where the response is the dynamic weight \( \{w_t\} \) conditional to predictors that could be chosen from past lags of \( w \) and selected exogenous variables. CS2009 applied a mixture regression approach to forecast the observed weight time series \( \{w_t; t = 1, 2 \ldots, t_{\text{max}}\} \) forward in time. In contrast to CS2009, this study applies a distribution-free alternative, nearest-neighbor sampling as the basis for forecasting the dynamic weight ahead in time. The nonparametric-weighted nearest-neighbor approach is known as KNNW (Mehrotra and Sharma 2006). The \( K \) of KNNW stands for \( K \) number of neighbors and \( W \) depicts the influence load of each predictor. The KNNW approach aims to ascertain the conditional dependence of predictands \( \{w_t\} \) on a specified set of predictors by identifying weighted nearest neighbors of the predictors in the historical record. The approach adopted to maintain spatial dependence in the predicted field involves transforming the predictor field for a given time step with a spatial smoother, centered at each grid point, and using the resulting spatially smoothened predictors as the basis for deriving the dynamic weight for each grid cell. Readers may refer to Chowdhury (2009, 113–116) for more details.

In summary, the dynamic weight combination method begins with deriving the historical weights; see Eq. (2) from the hindcast series of the paired single models. Then the raw predictors of the derived weights at each grid point are selected from a pool of candidate predictor variables. The selected raw predictors are smoothened across its neighbor. Once the predictors are ascertained, the weights are forecasted using the nonparametric method KNNW approach.

3. Application

a. The case study

The method is applied to improve the 3-month-ahead prediction of global SSTAs at a 5° × 5° grid of the global sea surface between 60°N and 40°S. The base of the anomalies was the climatology derived from the Global Ocean Surface Temperature Atlas (GOSTA) from 1951 to 1980 (Bottomley et al. 1990; Reynolds and Smith 1995). The observed dataset is the extended SSTA set, reconstructed at the National Climatic Data Center (Smith and Reynolds 2003). Three model predictions were combined, two of which were prepared by the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) project (Palmer et al. 2004) of the European Centre for Medium-Range Weather Forecasts (ECMWF). The first model was developed at ECMWF and is referred to as ECM (Wolff et al. 1997); the second model originates from Météo-France (Madec et al. 1997) and is referred to as MetF. The DEMETER models (ECM and MetF) are global coupled ocean–atmosphere models. The third model was developed at the Climate Prediction Center of the National Oceanic and Atmospheric Administration and is referred to as the CPC model (Van den Dool 2000; Van den Dool et al. 2003). The CPC model used a statistical technique known as “constructed analog” to forecast SSTA as linear combinations of all past observations during the same month. All SSTA time series were downloaded from the data library of International Research Institute for Climate and Society. The common period of hindcasts among these three models extends from March 1958 to December 2001. Note that this study accepts the single models as black boxes and ignores any minor biases they may contain. The removal of any apparent bias by looking only at one time window can affect any unique strength of the model when assessed at an alternative time window.

Readers may note that the dynamic weight formulation and forecasting methods developed in section 2 are based on two single models only. The application of this method to a higher number of single models requires a hierarchical pairwise combination tree. The formation of the pairwise combination tree for the three single
models are MetF, ECM, and CPC.

The combination tree of the three single models. The broken line is showing the static weight (SW) combination, and the solid line is the dynamic weight (DW) combination. The single models are MetF, ECM, and CPC.

models is illustrated in Fig. 1. The design of the tree may be influenced by a number of considerations, such as residual covariance and existence of static combination nodes. First, the two global circulation models (ECM and MetF) were combined using static weight, the static combination being indicated by a broken line in Fig. 1. The static combination method used was similar to the method described in the section below. Then, the joined ECM + MetF model was combined with CPC using the dynamic weight logic. The next step requires the selection of predictors to forecast these notional observed weight time series as defined in Eq. (2).

b. Static weight combination

Static model combination weights, in their various forms, embody a major portion of the current state of practice; hence, it is used as a benchmark for our proposed method. Accordingly, we compared the performance of dynamically combined prediction to the predictions obtained by the static model combination. We largely followed the method used by Robertson et al. (2004) when conducting the static combination. Each of the three models (ECM, MetF, and CPC) was first combined against the GOSTA climatology in isolation using the weight that minimized the sum of the squared error. The three sets of weights were normalized at the next step. The spatial noises of the normalized raw weights were condensed by taking an average value of weights of all surrounding grid cell within ±20° distance. These smooth static weights were used to combine the three single models and form the static weight alternative of our proposed method.

c. Predictor selection

The predictors for the forecasting weights are ascertained from lagged representations of the two derived functions of past weights, the mixture ratio \( r_t \) and the residual ratio \( \hat{r}_t \), presented below:

\[
r_t = \frac{|e_{2,t}|}{(|e_{2,t}| + |e_{1,t}|)} \quad \text{and} \quad (3)
\]

\[
\hat{r}_t = \frac{e_{2,t}}{e_{1,t}}. \quad (4)
\]

Here, \( e_{1,t} \) and \( e_{2,t} \) are residuals of the paired single models 1 and 2. The residual ratio \( \hat{r}_t \) is constrained to fall within \( \{-1 \rightarrow 2\} \), so as to avoid numerical instability when \( e_{1,t} \rightarrow 0 \).

A common set of predictors is used for the entire global sea surface for simplicity. This is achieved by first identifying the geographical spread of the concentration of high loadings of the first few principal components of observed \( w \). We narrowed our predictor search by primarily aiming to forecast \( w \) at these identified locations. At each identified location, we select a set of predictors based on the partial autocorrelation to the response, backward stepwise selection using an F test (Chambers 1992; Hastie et al. 2001), and partial mutual information (Sharma 2000). This research uses the same predictor variables for all the seasons. Seasonality is represented by using a set of 12 intercepts that varies gradually from one month to another. Based on the predictor selection procedure outlined above, the optimal predictors were identified as autoregressive lags of the mixture ratio \( \{r_t\} \) of order \( t \leq 3 \), \( t \leq 6 \), and \( t \leq 12 \) months and the residual ratio \( \{\hat{r}_t\} \) of \( t \leq 3 \) and \( t \leq 12 \) months. The approach of using minimal predictors ensures strong parsimony, an important feature of any forecasting model. Note that the predictors are limited to persistence only \( \{r_{t-1}, \ldots, r_{t-6}\}; \) to maintain simplicity in our presentation, we do not attempt to search for any predictors exogenous to the observed weights.

Any excessive variation of the raw predictor values along the neighboring grid is smoothed by taking weighted averaged predictor values among the neighbors. In this study the correlation coefficient depicts “weights” of the weighted average. An exhaustive cross-correlation analysis of raw predictors at each grid point against all other grid locations concludes that the influence (i.e., correlation) beyond ±20° distances reduces significantly (<0.4). Accordingly, each raw predictor vector is smoothed by a weighted linear combination (weights = correlation) of neighboring predictors within a radius of ±20°.

d. Forecasting dynamic weights

The value of \( K \) in KNNW method is estimated based on the square root of calibration data length (Lall and Sharma 1996). The loadings (\( W \) of KNNW) are estimated as scaled squared linear regression coefficients (Mehrotra and Sharma 2006) of the predictand (historical weight) versus the predictor set. Only one set of the influence load (\( W \) of KNNW) is used for the entire globe for simplicity. This spatially uniform influence load is ascertained by first
estimating linear regression coefficients at all grid points. The marginal distribution for each regression coefficient across the entire grid space is determined next, and the value corresponding to the highest probability density (mode) is selected as representative for the entire world.

4. Results

The forecast obtained by the dynamic combination is compared against the static combination, which is in turn compared against the selection of the best-performing single model (MetF in this study). All results presented here refer to forecasts in four sets of 10-yr validation outputs reflecting notional real forecast. Note that for the sake of brevity, instead of drawing four sets of graphs, all the figures are based on 40 years of collective validation results. First, we compared the unconditional probability density of pooled SSTA at all grid locations. The dynamic combination forecast density more closely resembled the observed SSTA density than that of static weight or the MetF forecast. The more detailed analysis is discussed next.

a. MSE

This study used mean-square error (MSE) as a relative measure of skill between different models. Figure 2 illustrates the spatial spread of MSE of the three single forecasts and that of the combined forecast. The spatial extent of improvement is evident. A consistent reduction of MSE was achieved because of the model combination across all four validation blocks. No increase in the MSE after the dynamic combination was noted in any of the validation blocks assessed (aggregated in the figure). We further examined the performance of the combination methods on a month-by-month basis by analyzing (as shown in Table 1) the percentage reduction of MSE in each month with respect to the MSE of the best single model, MetF. The dynamic combination reduced the MSE compared to the static combination, which in turn is better than any single-model prediction. Interestingly in Table 1, unlike in dynamic weight combination, there are months having worse performance using post-static weight combination at specific locations (e.g., Niño-3.4 and North Pacific). The dynamic combination is more resilient against returning worse results.

The MSE of the static combination predictions minus the dynamic combination predictions across the global sea surface grid is presented in Fig. 3. Note that during the four independent validation decades, 93%–95% of grid cells exhibited at least 5% reduction of MSE due to the dynamic combination method compared to the static combination method. The decrease of mean-square error...
is found statistically significant when analyzed by a paired one-tailed \( t \) test (\( p = 2.94 \times 10^{-2} \)). This demonstrates the superior outcome of the dynamic combination to the static combination.

### b. Spatial features

The representation of spatial dependence is analyzed by measuring linear interdependence across the global grid. The correlation of the SSTA time series at a reference grid cell to all other cells in the grid is measured first. The reference grid point \( l \) is then moved across all the available grid cells \( \{l = 1, 2, \ldots, l_{\text{max}}\} \), resulting in \( \frac{1}{2}l_{\text{max}}(l_{\text{max}} - 1) \) estimates of the correlation, where \( l_{\text{max}} \) is the total number of cells in the grid. A random subset (1000 values) of these correlations is compared against their historical values as illustrated in Fig. 4. Few systematic losses (or gains) of spatial correlation were evident, except where the observed correlation is low. The artificial inflation of linear dependence around cells having low correlations can be attributed to the underlying single-model predictions. We have drawn a least squares fitted line to correlations for both the dynamically combined prediction and the best single prediction, MetF. The fitted line corresponding to the dynamically combined prediction is closer to the 1:1 line, indicating a better match to the observed dependence attributes.

Our case study found that the improvements due to the dynamic combination are spatially and temporally consistent. We observed that because of the lower accuracy of the single models over the extratropics (hence, a higher scope of improvement), the reduction of MSE around extratropics is higher than that over the equatorial region. The weaker improvements around the equator concur with our earlier experience of forecasting Niño-3.4 (CS2009). This observation is analogous to past studies that had experienced lesser skill of combined models in the tropics (Colman and Davey 2003).

The spatial consistency of improvement is scrutinized as follows. First, a spatial spread of minimum MSE out of three models (ECM, MetF, and CPC) separately at each grid point is selected. The percent reduction of MSE over this \emph{potential minimum} MSE spread is illustrated in Fig. 5. We found a spatially consistent reduction of MSE as a result of the dynamic combination: about 72% cells exhibited reductions in excess of 25%. The reduction in MSE of SSTA forecast is important, as it further improves the forecast skill of other dependent environmental variables, such as river flow (Chowdhury and Sharma 2009b).

### TABLE 1. Percentage reduction in MSE after combination compared to the MetF model. DW and SW columns list reductions after DW and SW combination methods, respectively. Negative values denote increased MSE. The results of all four 10-yr validation blocks (1962–2001) are aggregated here. North P is the 30°–40°N and 175°–225°E box in the North Pacific Ocean.

<table>
<thead>
<tr>
<th>Months</th>
<th>Global</th>
<th>Niño-3.4</th>
<th>North P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DW SW</td>
<td>DW SW</td>
<td>DW SW</td>
</tr>
<tr>
<td>Jan</td>
<td>16.2 9.4</td>
<td>11 −70.2</td>
<td>46.5 30.3</td>
</tr>
<tr>
<td>Feb</td>
<td>10.8 3.9</td>
<td>19.1 −104.5</td>
<td>48.5 38.4</td>
</tr>
<tr>
<td>Mar</td>
<td>8.7 1.8</td>
<td>29.5 −4.5</td>
<td>32.1 9.0</td>
</tr>
<tr>
<td>Apr</td>
<td>19.1 9.8</td>
<td>53.1 37.2</td>
<td>58.6 −18.8</td>
</tr>
<tr>
<td>May</td>
<td>15.0 2.8</td>
<td>31.7 8.5</td>
<td>38.3 −107.6</td>
</tr>
<tr>
<td>Jun</td>
<td>16.8 7.2</td>
<td>51.0 20.9</td>
<td>50.4 −49.5</td>
</tr>
<tr>
<td>Jul</td>
<td>29.3 19.6</td>
<td>75.9 57.6</td>
<td>63.4 23.3</td>
</tr>
<tr>
<td>Aug</td>
<td>13.1 7.6</td>
<td>82.7 81.1</td>
<td>71.3 61.3</td>
</tr>
<tr>
<td>Sep</td>
<td>8.1 0.1</td>
<td>76.7 75.1</td>
<td>1.3 −54.1</td>
</tr>
<tr>
<td>Oct</td>
<td>16.2 3.7</td>
<td>69.5 68.3</td>
<td>62.1 −50.7</td>
</tr>
<tr>
<td>Nov</td>
<td>13.3 0.2</td>
<td>60.0 57.0</td>
<td>63.3 −41.3</td>
</tr>
<tr>
<td>Dec</td>
<td>11.9 2.5</td>
<td>16.7 17.6</td>
<td>39.5 −52.0</td>
</tr>
</tbody>
</table>

FIG. 3. The percent reduction in prediction MSE due to DW combination compared to SW combination. Positive reduction denotes improvement.
5. Discussion and conclusions

a. Realistic predictive skill

It should be noted that the forecast combination adds parameters supplementary to each single model and hence any combination exercise is prone to overfitting (Peña and Van den Dool 2008). Moreover, the dynamic combination introduces additional parameters and complexity to the overall prediction scheme in comparison to a static combination approach. It is imperative that the performance of any such “complex” model be evaluated in a carefully designed cross-validation setting. We attempt to do so by estimating model prediction error, reflective of the performance of the model in a pure forecast setting. This study validates the results by applying the models in four “10 year” blocks. For example, results from 1992 to 2001 are based on parameters that have been calibrated for the period 1958–91 only. We believe that the 10-yr gap would minimize any likely artificial inflation of skill because of boundary influences. It is pertinent to note the caution (DelSole and Shukla 2009) against inadvertently instilling artificial skill in validation results by biased predictor selection. The possibility of overestimating the validation performance arises if the validation period is not removed prior to predictor selection. In this research the predictors are limited to only three autoregressive terms, the choice of which does not change by expunging the validation data. Besides, as recommended by DelSole and Shukla (2009), the predictors are chosen based on a variety of rigorous statistical analyses rather than simple correlation, which may exhibit spurious signals. Overall, we believe that our validation results reflect the realistic predictive skill of the models.

b. Temporal and spatial variation of weights

The results presented earlier in the section illustrate how the dynamic combination approach exploits the presence of localized persistence in individual model skill
for sustained periods of time. The issues that still need to be addressed are whether this localized temporal skill is uniform over all the models being combined and whether the associated persistence attributes vary from one model to another. Figure 6 presents the stacked CPC and ECM weights (the remainder being the MetF weight) ascertained over a single 10-yr validation period, over two selected regions. It is interesting to note that the skill associated with each model remains relatively invariant over the Niño-3.4 region as compared to the North Pacific region evaluated. In contrast, the significant variation in forecast skill over the North Pacific region illustrated is reflective of the poor skill that can be associated for any of the three models, creating a situation where the dynamic combination offers significant improvements over any single model considered. Some of the errors in the North Pacific SSTA forecast may emanate from errors in the individual model specific to tropical SSTA patterns and their intensity.

c. Effect of any model bias

One of the underlying assumptions behind our weight formulation is that the single forecasts are unbiased; however, in our application above, we did not attempt to remove any single forecast bias prior to combination. We took the view that bias correction of the single model should be incumbent on the model developer and that the postprocessing (bias correction) looking at a certain time window without appreciating the structure may not be robust. Are the improvements presented earlier a result of the reduction in bias after combination? Would the combination of the fully unbiased single forecasts yield a similar extent of improvement?
The proposed linear combination intends to reduce variance and to converge to the weighted mean of the single-model bias. So, the improvements demonstrated earlier in this paper mainly result from merging the variance rather than from combining the associated biases. This can be proven by achieving improvement even after combining purely unbiased forecast. So, this study examined the extent of improvement of the bias-corrected forecast as follows. First, the biases of the three single models are removed in a leave 10-yr cross-validated mode. Then, the unbiased single forecasts are dynamically combined. The MSE of the single model and combined models are tabulated in Table 2, for simplicity the entire globe is pooled together. As expected, there are overall reductions of MSE due to bias correction. The MSE combination after bias correction is about 75% of the best model (MetF) MSE; this factor is similar to the 85% figure of our original case study.

We further analyzed the spatial spread of the percent reduction of MSE over the potential minimum MSE (out of ECM, MetF, and CPC). We found that the dynamic combination after bias correction is also able to achieve a similar extent of improvement as shown earlier in Fig. 5. In this case about 78% of cells showed reductions of at least 25% or more; this compares well with our earlier count of 72% of cells. As evident from a similar improvement after the combination of purely unbiased forecasts, the improvement in the dynamic combination is achieved because of the reduction in variance.

d. Other issues

In this study, the dynamic combination is only applied at the highest pair of the combination tree (Fig. 1), unlike the pairwise dynamic weight combination method of CS2009. This simplification in the combination method architecture did not worsen the MSE results for our case study. Another aspect of the combination tree is the order in which the models should be paired. Generally, two models, whose residuals show low covariance, are paired, as lower covariance indicates a higher chance of improvement after combination (see CS2009 for more on tree architecture). The effect of varying the combination tree architecture and the underlying mathematics is beyond the scope of current research.

This paper demonstrated the use of the KNNW method to forecast weights instead of the mixture regression method used in CS2009. The relative improvements due to the choice of the forecasting method used are evaluated in detail in Chowdhury (2009, 121–131) and are not discussed here for conciseness. In general, we recommend the use of the KNNW as a forecast model because of the need for fewer parameters, assumptions, and computation cost in the case of the combination of multivariate forecasts. There are other issues related to forecasting methods that may be investigated in future studies; for example, should the weights be derived after smoothing the residual errors across the grids, and should the dimension of the weight matrix be reduced by orthogonal analysis prior to grid-by-grid weight forecast? Note that the residual error of forecast within a neighborhood varies gradually and hence prior smoothing is not crucial. We do not suggest any orthogonal transformation, as it complicates the intuitive meaning of weights being a proportional contribution of a single forecast to the final forecast at a point of interest.

This study does not develop any statistical or process-based climate model. The intent of this study is to combine various model responses that are readily available. Models are treated as black box only, and no alteration of any single model is proposed. The success of the method relies on the presence of a skillful single forecast in the mix.

e. Conclusions

This study dynamically combined globally gridded sea surface temperature anomaly forecasts 3 months in advance. The predicted sea surface temperature using the dynamic combination algorithm consistently exhibited better accuracy, both in space and time, to that of the static combination or the best single model. This improvement is possible because of the identification of both the temporal and spatial persistence of a superior single forecast among the three alternatives combined.

Acknowledgments. This research was funded by the Australian Research Council and the Sydney Catchment Authority. We acknowledge the assistance of Upmanu Lall, Andrew Robertson, and Lisa Goddard of Lamont-Doherty Earth Observatory of Columbia University, New York. The computation was performed using the freely available R statistical computing platform.
REFERENCES


