Improving the Prediction of Winter Precipitation and Temperature over the Continental United States: Role of the ENSO State in Developing Multimodel Combinations

NARESH DEVINENI AND A. SANKARASUBRAMANIAN

Department of Civil, Construction and Environmental Engineering, North Carolina State University, Raleigh, North Carolina

(Manuscript received 24 June 2009, in final form 5 November 2009)

ABSTRACT

Recent research into seasonal climate prediction has focused on combining multiple atmospheric general circulation models (GCMs) to develop multimodel ensembles. A new approach to combining multiple GCMs is proposed by analyzing the skill levels of candidate models contingent on the relevant predictor(s) state. To demonstrate this approach, historical simulations of winter (December–February, DJF) precipitation and temperature from seven GCMs were combined by evaluating their skill—represented by mean square error (MSE)—over similar predictor (DJF Niño-3.4) conditions. The MSE estimates are converted into weights for each GCM for developing multimodel tercile probabilities. A total of six multimodel schemes are considered that include combinations based on pooling of ensembles as well as on the long-term skill of the models. To ensure the improved skill exhibited by the multimodel scheme is statistically significant, rigorous hypothesis tests were performed comparing the skill of multimodels with each individual model's skill. The multimodel combination contingent on Niño-3.4 shows improved skill particularly for regions whose winter precipitation and temperature exhibit significant correlation with Niño-3.4. Analyses of these weights also show that the proposed multimodel combination methodology assigns higher weights for GCMs and lesser weights for climatology during El Niño and La Niña conditions. On the other hand, because of the limited skill of GCMs during neutral Niño-3.4 conditions, the methodology assigns higher weights for climatology resulting in improved skill from the multimodel combinations. Thus, analyzing GCMs' skill contingent on the relevant predictor state provides an alternate approach for multimodel combinations such that years with limited skill could be replaced with climatology.

1. Introduction

Planning and management of water and energy systems are usually carried out based upon the seasonal climate (precipitation and temperature) forecasts over a particular region. Several national and international agencies routinely issue climate forecasts using coupled general circulation models (GCMs; e.g., Saha et al. 2006) as well as using atmospheric GCMs (AGCMs; e.g., Goddard et al. 2003). Forecasts from AGCMs are typically developed in a two-tiered process with sea surface temperatures (SSTs) being predicted first from an ocean–atmosphere model and then the predicted SSTs are forced as boundary conditions into the AGCMs. This two-tiered approach primarily emphasizes that much of the predictability at seasonal time scales primarily stems from the oceanic conditions with the ensembles representing the atmospheric internal variability. However, the skill of the climate forecasts could vary substantially depending on the location, time, and the GCMs itself (Doblas-Reyes et al. 2000; Robertson et al. 2004).

Reducing model uncertainties through the conventional approach of the refinement of parameterizations and improved process representation is time consuming, which led recent efforts to focus on the combination of AGCMs for improving seasonal climatic prediction (Krishnamurti et al. 1999; Doblas-Reyes et al. 2000; Rajagopalan et al. 2002). Research studies from the Prediction of Climate Variations on Seasonal to Interannual Time-scales (PROVOST) experiment and from International Research Institute for Climate and Society (IRI) show that multimodel combination of AGCMs provide better calibration (i.e., improved reliability and resolution) than individual model predictions (Doblas-Reyes et al. 2000; Barnston et al. 2003).

Corresponding author address: A. Sankarasubramanian, Dept. of Civil, Construction and Environmental Engineering, North Carolina State University, Raleigh, NC 27695-7908.
E-mail: sankar_arumugam@ncsu.edu

DOI: 10.1175/2009MWR3112.1

Studies from the Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER) experiments show that multimodel combination of coupled GCMs (CGCMs) also improve the reliability and skill in predicting summer precipitation in the tropics and winter precipitation in the northern extratropics (Palmer et al. 2004). Hagedorn et al. (2005) demonstrate that the superiority of multimodels primarily arises from error cancellation, not increased ensemble members (i.e., pooling of ensemble from single models), resulting in improved reliability and consistency. Reviewing the skill of various forecast products (medium range, monthly and seasonal) over the Europe, Rodwell and Doblas-Reyes (2006) show that multimodel ensembles of CGCMs exhibit higher skill in predicting precipitation and temperature during both winter and summer seasons. Recently, multimodel ensembles developed using 46-yr hindcasts from five CGCMs run from the European Union's ENSEMBLES project have better skill in predicting tropical SSTs than the multimodel ensembles developed using reforecasts from the DEMETER project (Weisheimer et al. 2009).

The simplest approach to developing multimodel combination is to pool the ensembles from all the models by giving equal weights for all the models (Palmer et al. 2000). A different approach to developing multimodel ensembles is to optimally combine multiple GCMs so that the resulting multimodel forecast has better skill than individual models' forecasts (Rajagopalan et al. 2002; Robertson et al. 2004; DelSole 2007). Under the optimal combination approach, weights are obtained for each GCM as a fraction such that the chosen skill metric of the multimodel ensembles is maximized (Rajagopalan et al. 2002; Robertson et al. 2004). Doblas-Reyes et al. (2005) compare the performance of two multimodel combination techniques—equal weighting of all models and optimal combination using multiple linear regression—and show that, except in tropical regions, it is difficult to improve upon the performance of the optimal combination due to small sample size. Studies have also employed simple statistical techniques such as linear regression (Krishnamurti et al. 1999) to advanced statistical techniques such as the canonical variate method (Mason and Mimmack 2002) and Bayesian techniques (Hoeting et al. 1999; Stephenson et al. 2005; Luo et al. 2007) for developing multimodel combinations. DelSole (2007) proposed a Bayesian multimodel regression framework that incorporates prior beliefs about the model weights for estimating regression parameters.

It is well known that anomalous conditions in the tropical Pacific influence the skill of GCMs in predicting precipitation and temperature over North America (Shukla et al. 2000; Quan et al. 2006). Upon diagnosing the sources of seasonal prediction skill over the United States, Quan et al. (2006) showed that the entire skill level of the AGCMs could be explained by El Niño–Southern Oscillation (ENSO) alone. For this analysis, Quan et al. (2006) considered the multimodel mean obtained by averaging all the ensembles (total of 48 simulations) from four AGCMs. Recently, Devineni et al. (2008) proposed a new approach to developing multimodel ensembles of streamflow that combines forecasts from individual models by evaluating their skill contingent on the predictor state.

The main intent of this study is to explore strategies for improving the skill in predicting winter precipitation and temperature over the continental United States by optimally combining multiple GCMs. Given that predicting winter precipitation and temperature over the United States primarily depends on SST conditions over the tropical Pacific (Quan et al. 2006), we combine multiple GCMs by evaluating the skill score of seven AGCMs conditioned on the ENSO state based on the algorithm outlined in Devineni et al. (2008). The skill score of the GCMs contingent on the ENSO state are estimated by averaging the mean square error (MSE) in predictions under similar tropical SST conditions. For this purpose, we consider simulated precipitation and temperature results (i.e., forced with observed SSTs) from seven different AGCMs for developing multimodel combinations. The performance of the developed multimodel tercile probabilities of winter precipitation and temperature are compared with the performance of individual models' as well as with two of the commonly employed techniques for multimodel combination.

For better readability from here onward in this manuscript, we often refer to both individual and multimodel simulations of precipitation and temperature as ''predictions/forecasts'' with an understanding that simulated GCM variables overestimate the potential forecasting skill. Section 2 describes the data and the GCMs used for the study along with a description of the multimodel combination methodology. Section 3 presents the results and analysis by comparing the skill of individual GCMs and multimodels in predicting the observed winter precipitation and temperature. Finally, in section 4, we summarize the findings and conclusions from the study.

2. Multimodel combination contingent on the predictor state: Basis and methodology

a. Data

Seven AGCMs that are commonly employed by various research institutes and agencies are considered for developing multimodel winter (December–February, DJF)

TABLE 1. Details of atmospheric GCMs considered for the study. All models span from 25° to 45°N and from 123.75° to 66.25°W, resulting in a total of 192 grid points. Historical simulations of winter (DJF) precipitation and temperature from the seven GCMs are considered for multimodel combination.

Historical simulations			
Model	Source	Ensemble size	Reference
ECHAM4.5	Max Planck Institute	85	Roeckner et al. (1996)
CCM3v6	National Center for Atmospheric Research (NCAR)	24	Kiehl et al. (1998)
COLA	Center for Ocean–Land–Atmosphere Studies	10	Schneider (2002)
GFDL, AM2p12b	Princeton University	10	GFDL Global Atmospheric Model Development Team (2005)
ECPC	Scripps Institution of Oceanography, University of California, San Diego	10	Kanamitsu et al. (2003)
NCEP	National Oceanic and Atmospheric Administration (NOAA)	10	Saha et al. (2006)
NSIPP-1	National Aeronautics and Space Administration (NASA) GSFC	9	Bacmeister et al. (2000)

precipitation and temperature forecasts over the continental United States. Table 1 gives the details on each model along with the number of ensembles available for predicting precipitation and temperature. Historical monthly simulations of winter precipitation and temperature, which are developed by forcing the AGCMs with observed SSTs, were obtained from the IRI data library (information available online at http://iridl.ldeo.columbia.edu/SOURCES/.IRI/.FD/). Figure 1a shows the grid points (a total of 192) that are considered for developing multimodel predictions. Observed monthly precipitation and temperature data at 0.5° × 0.5°, available from the University of East Anglia's (UEA) Climate Research Unit (CRU; New et al. 2000), are used to assess the skill of each model. Monthly climate anomalies, relative to the 1961–90 mean (New et al. 1999), were interpolated from the station data to develop monthly terrestrial surface climate grids for the period 1901–96. Recent studies of multimodel combination have used the observed precipitation and temperature for the UEA database to show the improvements resulting from multimodel combination (Rajagopalan et al. 2002; Robertson et al. 2004; DelSole 2007). Grid points (0.5° × 0.5°) of monthly precipitation and temperature from UEA were spatially averaged to map the grid points of the GCMs.

We consider Niño-3.4, the index commonly used to denote the ENSO state, as the primary predictor influencing the winter precipitation and temperature over the United States. Niño-3.4 denotes the anomalous SST conditions in the tropical Pacific, which are obtained by averaging the SSTs over 5°S–5°N and 170°–120°W. The average DJF Niño-3.4, which is computed using Kaplan's SST database (Kaplan et al. 1998), is obtained from the IRI data library for the 46 yr (DJF 1951–1996; information available online at http://iridl.ldeo.columbia.edu/SOURCES/.KAPLAN/.Indices/.NINO34/) considered for verification. El Niño (Niño-3.4 > 0.5), La Niña (Niño-3.4 < –0.5), and neutral conditions (|Niño 3.4| ≤ 0.5) are identified resulting in a total of 14 yr of El Niño, 12 yr of La Niña, and 20 yr of neutral conditions from the DJF Niño-3.4 time series.

b. *Basis behind combining individual GCMs contingent on the predictor state*

The multimodel combination methodology proposed in this study is motivated by the premise that model uncertainties could be better reduced by combining the GCMs based on their ability to predict under a given predictor state. Recent studies on seasonal to interannual climate prediction over North America clearly show that the skill of GCMs is enhanced during ENSO years (Branković and Palmer 2000; Shukla et al. 2000; Quan et al. 2006). To understand this further, Fig. 1a shows the correlation between the observed precipitation and ensemble mean of the GCM-predicted precipitation without consideration of the ENSO state (i.e., over the entire period of record), whereas Figs. 1b–d show the skill (correlation) of two GCMs, ECHAM4.5 and that of the Experimental Climate Prediction Center (ECPC), in simulating winter precipitation under El Niño, La Niña, and neutral conditions, respectively. The correlations ($1.96/\sqrt{n_s - 3}$ where n_s denotes the number of samples under each category) that are statistically significant at the 95% confidence interval under El Niño ($n_s = 14$), La Niña ($n_s = 12$), neutral conditions ($n_s = 20$), and over the entire record ($n_s = 46$) are 0.59, 0.65, 0.48, and 0.30, respectively.

Though Fig. 1a shows significant correlation at many grid points (>0.30) for both models, the performances of the models under those grid points are not consistent under three different ENSO conditions. Further, the skill levels of both GCMs are not significant/negative (<0.50)

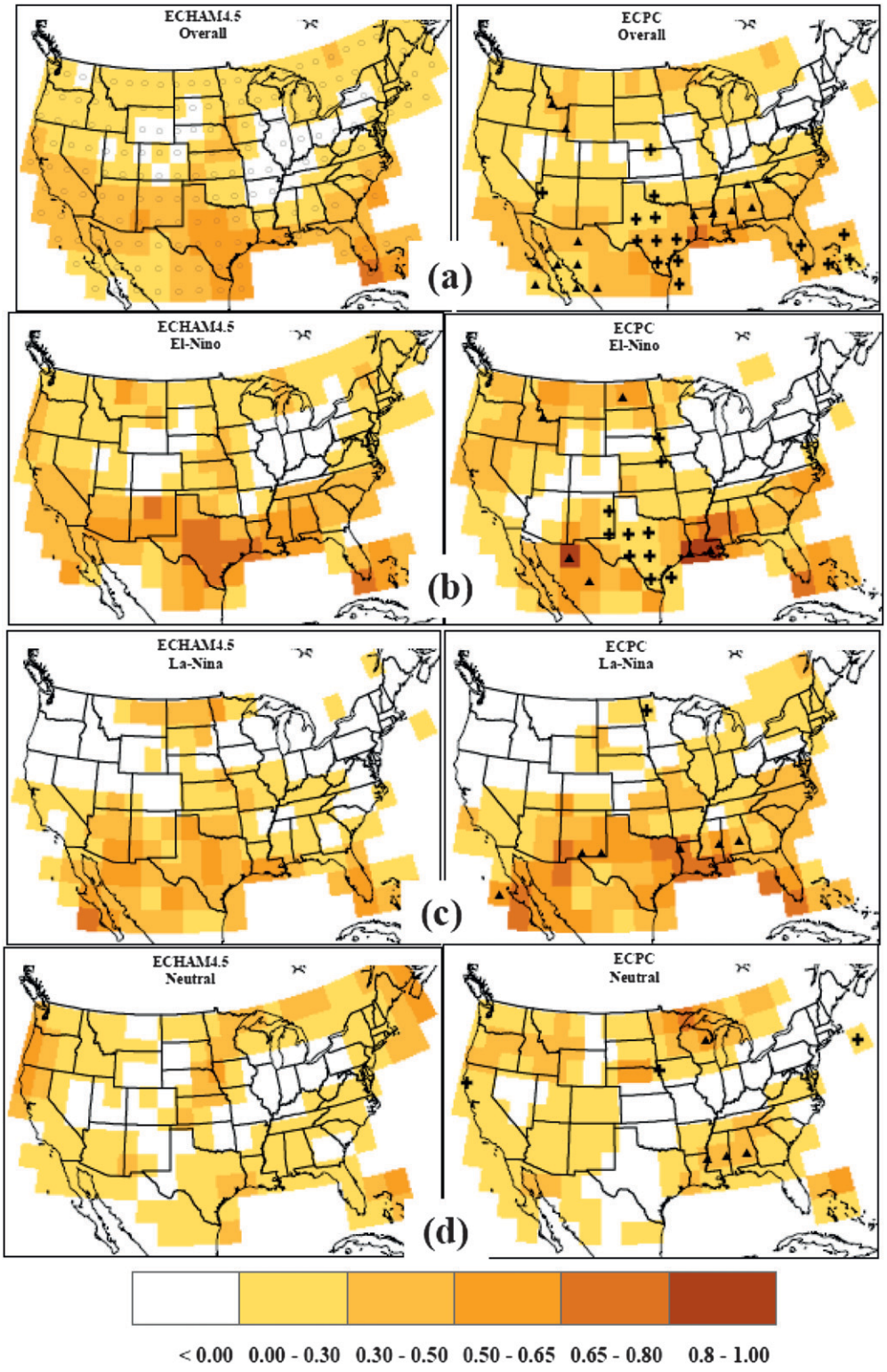

FIG. 1. Skill, expressed as a correlation between the ensemble mean of the GCM and observed precipitation, in simulating the DJF winter precipitation by two GCMs, (left) ECHAM 4.5 and (right) ECPC, (a) over the entire period of record, and (b) under El Niño, (c) La Niña, and (d) neutral conditions. The figure on the right under each category shows a plus (triangle) sign, which indicates that the correlation between DJF precipitation and the ensemble mean of ECHAM4.5 (ECPC) is statistically higher than the correlation between the DJF precipitation and the ensemble mean of ECPC (ECHAM4.5) for that category.

for most grid points under neutral conditions with the skill being mostly significant only under El Niño (>0.59) and La Niña conditions (>0.65). We can also infer from Fig. 1 that grid points exhibiting significant skill by both GCMs also vary spatially. Thus, anyone combining the GCMs purely based on the overall skill would end up giving higher weights to the best-performing GCM at that grid point, which would naturally result in poor predictions during neutral conditions.

We also compare whether the differences in positive correlations exhibited by these two models are statistically significant by using the Hotelling–Williams test (Bobko 1995). The Hotelling–Williams test statistic, $(r_{12} - r_{13})\{(N-1)(1+r_{23})[2(N-1)/(N-3)|R| + \bar{r}^2(1-r_{23})^2]^{-1}\}^{-1/2}$, follows a t distribution with $(N-3)$ degrees of freedom, where r_{12} (r_{13}) denotes the correlation between the observed precipitation and ensemble mean from ECHAM4.5 (ECPC), r_{23} denotes the correlation between the ensemble means of ECHAM4.5 and ECPC, and N denotes the total number of years of observation with $\bar{r} = (r_{12} + r_{23})/2$ and $R = (1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23})$. A plus (triangle) sign on a grid point in Fig. 1b shows that the difference between r_{12} (r_{13}) and r_{13} (r_{12}) is greater than zero, indicating the better performance of ECHAM4.5 (ECPC). We can see from Fig. 1 that, under El Niño conditions, there are 6 (10) grid points with r_{12} (r_{13}) being significant over r_{13} (r_{12}). This primarily shows that the skill levels exhibited by these models under different ENSO states are statistically significant and that they also could be completely different from the overall skill of the model. Hence, we propose a methodology that evaluates the performance of GCMs contingent on the dominant predictor state(s) and assigns higher weights for the best-performing model under those predictor conditions. We also consider climatology as one of the candidate models for developing multimodel ensembles. By including climatology, we believe that if the skill levels of all of the models are poor under neutral ENSO conditions, then the algorithm could give higher weights for climatology in developing multimodel combinations. In the next section, we describe the modified version of multimodel combination algorithm presented in Devineni et al. (2008).

c. *Description of the multimodel combination algorithm*

Figure 2 provides a flow chart of the multimodel combination algorithm that combines tercile predictions/forecasts from multiple GCMs. Historical simulations of winter precipitation and temperature available for each GCM (1951–96) are converted into tercile categories, $Q_{i,t}^m$, where $m = 1, 2, \ldots, M$ ($M = 8$) denotes the model index including climatology, with $i = 1, 2 \ldots, N$ ($N = 3$) representing the categories in year t, which specifies the time index with $t = 1, 2, \ldots, T$ ($T = 46$ yr; for additional details, see Fig. 2). Tercile categories, $Q_{i,t}^m$, are computed from the ensembles of GCMs after the removal of the systematic bias (i.e., each ensemble is represented as an anomaly from the model's seasonal climatology). The squared error (SE_t^m) in predicting the observed precipitation/temperature is computed from the ensemble mean of the simulated precipitation/temperature for each year at 192 grid points over the United States. Based on Quan et al. (2006), we consider the ENSO state indicated by Niño-3.4 as the dominant predictor in influencing the winter precipitation and temperature over the United States.

The objective is to combine individual model simulations by evaluating their skill—represented by the MSE ($\lambda_{t,K}^m$ in Fig. 2)—over K neighboring predictor conditions. Devineni et al. (2008) considered various skill metrics for evaluating the performance of candidate models over similar predictor conditions and found that the mean squared error and average rank probability score perform well in improving the skill of multimodel combinations. Further, the MSE, which is obtained based on the average error in the conditional mean of the forecast over similar predictor conditions, is also a proper skill score (Bröcker and Smith 2007). A skill score is proper if it maximizes the expected score for an observation drawn from a particular distribution only if the issued probabilistic forecast is of the same distribution (Bröcker and Smith 2007). Given that we have candidate GCMs with different ensemble sizes, we did not consider other strictly proper scores such as the ranked probability score (RPS) since their sampling variability heavily depend on the number of ensembles (Weigel et al. 2007). We identify K similar ENSO conditions by calculating the Euclidean distance between DJF Niño-3.4 in the conditioning year t and the rest of DJF Niño-3.4 results observed during 1951–96. *It is important to note that in computing the MSE from K similar climatic conditions, we leave out the skill of the model* (SE_t^m) in that conditioning year.

Based on the MSE computed over K neighbors, weights ($w_{t,K}^m$ in Fig. 2) for each model are computed for each year using which the multimodel tercile categories are computed. Basically, the weighting scheme should be inversely proportional to a chosen increasing function (e.g., linearly or logarithmic) of the prediction error metric. The idea behind the proposed approach in Fig. 2 is that the weights for each model vary according to the ENSO conditions. Thus, if a particular GCM performs well under the El Niño conditions at a given grid point, then higher weights ($w_{t,K}^m$ in Fig. 2) will be given to the tercile probabilities from that GCM in developing multimodel combinations. Using the algorithm in Fig. 2, we

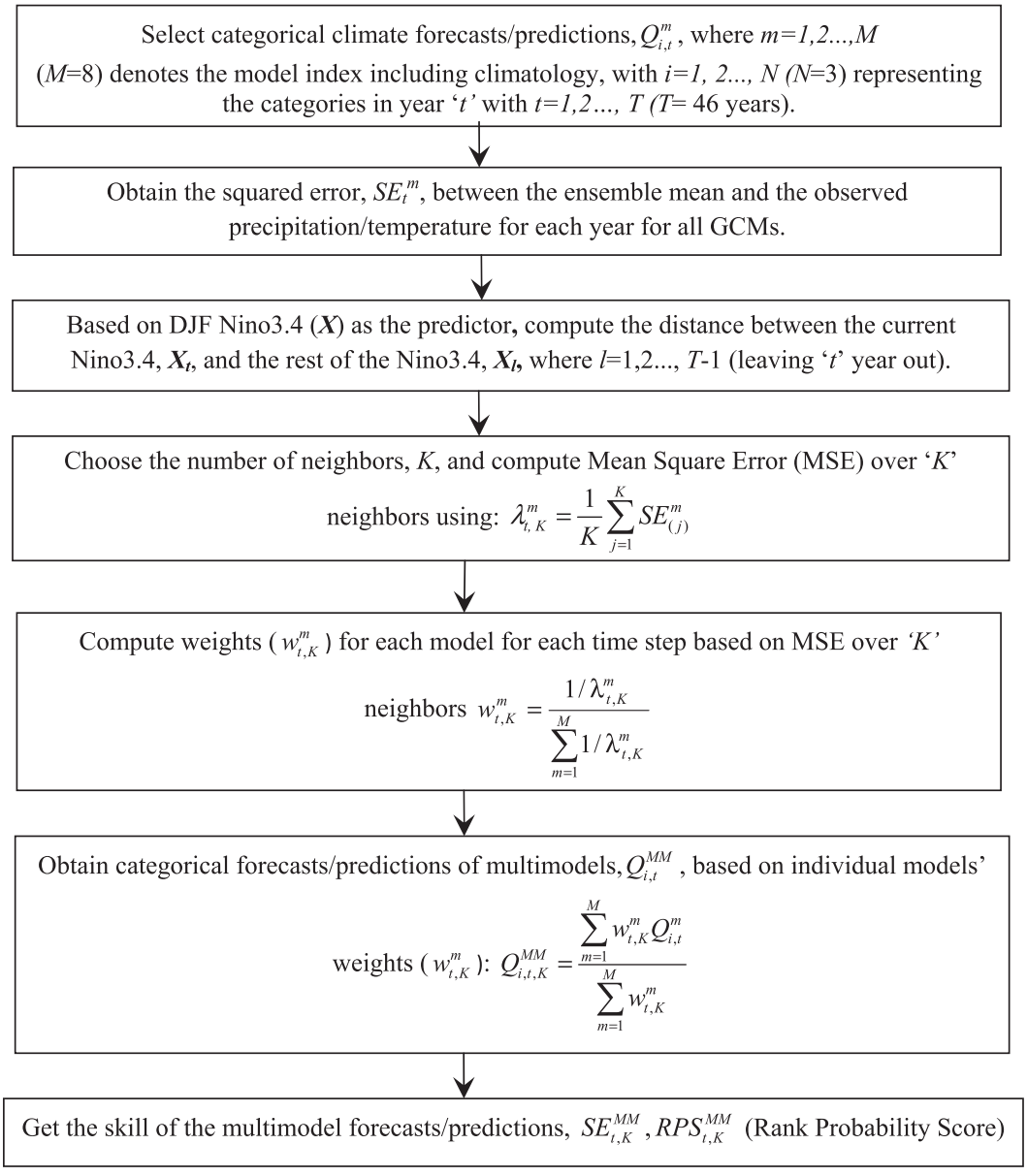

FIG. 2. Flow chart of the multimodel combination algorithm employed in this study (modified from Devineni et al. 2008).

develop six multimodel combinations of winter precipitation (shown in Table 2) and temperature for 192 grid points over the continental United States. We discuss the issue of the selection of the optimal number of neighbors K and various multimodel combination schemes (in Table 2) in the next section.

d. Multimodel schemes and performance analysis

Table 2 provides a brief description on the different multimodel schemes considered in this study. Four multimodel (MM) schemes (MM-1, MM-2, MM-3, and MM-4) are developed using the algorithm in Fig. 2. MM-1 and MM-2 employ fixed neighbors K to obtain the MSE. Under MM-1 and MM-2, if the Niño-3.4 in the conditioning year is under El Niño, La Niña, and neutral states, then we estimate the MSE ($\lambda_{t,K}^{m}$) of the GCM during the observed El Niño (K =13 yr, leaving the conditioning year out), La Niña (K = 11 yr), and neutral (K = 19 yr) years, respectively. Thus, for MM-1 and MM-2, we evaluate the skill of the model only under similar ENSO conditions.

With multimodel schemes (MM-3 and MM-4), we obtain the neighbors, K_t, by performing two-deep cross validation (Stone 1974). The two-deep cross validation is a

TABLE 2. List of multimodel combinations considered for the study.

Multimodel indices–schemes	Brief description
MM-1	Individual models + climatology in one step, with a fixed number of neighbors contingent on the ENSO state
MM-2	Individual models + climatology combined in the first step and the resulting model outputs combined at the second step, with a fixed number of neighbors contingent on the ENSO state
MM-3	Individual models + climatology in one step, using optimized neighbors obtained by two-deep cross validation
MM-4	Individual Model + Climatology combination in the first step and the resulting model outputs combined at the second step using optimized neighbors obtained by two-deep cross-validation
MM-P	Multimodel combination using pooled ensembles
MM-OS	Multimodel combination using weights based on the overall skill in the calibration period.

rigorous model validation technique, which is generally employed to choose optimum model parameters as well as to reduce the overfitting that typically results from multimodel combination (DelSole 2007). The two-deep cross-validation technique obtains model predictions recursively in two stages. In the outer loop, we leave out the predictor (Niño-3.4) and the predictand (DJF precipitation–temperature) in year t and use the remaining $T - 1$ years ($T = 46$) to estimate the optimum K_t. For the samples (which constitute $T - 1$ yr of GCM predictions and Niño-3.4) in the inner loop, we obtain K_t that minimizes the MSE of the multimodel predictions over $T - 1$ yr through a leave-one-out cross validation (i.e., model fitting is done with $T - 2$ years and validated for the left-out year from the $T - 1$ sample). We employ this K_t from the inner loop to develop the multimodel predictions for year t in the outer loop. This procedure is repeated for all possible left-out samples in the outer loop to obtain multimodel predictions for each year. Thus, under MM-3 and MM-4, the number of neighbors (K_t) varies from year to year.

We also employ different strategies for combining individual model simulations with climatological ensembles. MM-1 and MM-3 utilize seven different GCMs (Table 1) along with climatological ensembles to develop multimodel ensembles. MM-2 and MM-4 combine each model with climatology in the first step and then combine the resulting seven models in the second step. Recent studies have shown that a two-step procedure of combining each of the individual model forecasts separately with climatology and then combining the resulting M combinations in the second step improves the skill of the multimodel ensembles (Robertson et al. 2004). For climatology, we simply consider the 45 yr (leaving the conditioning year out) of the observed precipitation and temperature results at each grid point from the UEA dataset. To compute the squared error for each year at the second step of the combination of MM-2 and MM-4, we assume the conditional distribution obtained from the first step to be normal. MM-P is the multimodel combination scheme that is obtained by pooling all the ensembles from seven individual models and climatology. The reason we consider climatology under MM-P is the desire to be consistent for comparisons with other multimodel schemes (MM-1–MM-4). Hence, in the MM-P scheme, we have an increased number of ensembles (203) since we are now pooling ensembles from all the models.

MM-OS combines individual models based on their overall skill (without consideration of the ENSO state), which is specified based on the MSE for the period 1951–96 in predicting the winter precipitation–temperature at a given grid point. Thus, under MM-OS, the weight $\{(\overline{MSE}^m)^{-1}[\sum_{m=1}^{M}(\overline{MSE}^m)^{-1}]^{-1}\}$ for a given model m is obtained based on the inverse of the MSE of model m to the sum of the inverse of MSE of all the models. MM-P and MM-OS provide the baseline comparison with some of the commonly employed techniques in developing multimodel combinations (Palmer et al. 2000; Rajagopalan et al. 2002; Robertson et al. 2004). The performance of multimodel predictions is compared with individual models' skill using standard verification measures such as the average ranked probability score ($\overline{\text{RPS}}$), the average ranked probability skill score ($\overline{\text{RPSS}}$), reliability diagrams, and average Brier scores. Expressions related to these metrics may be found in Wilks (1995) and hence they are not provided here. The next section discusses the performance of multimodels in predicting winter precipitation and temperature over the continental United States.

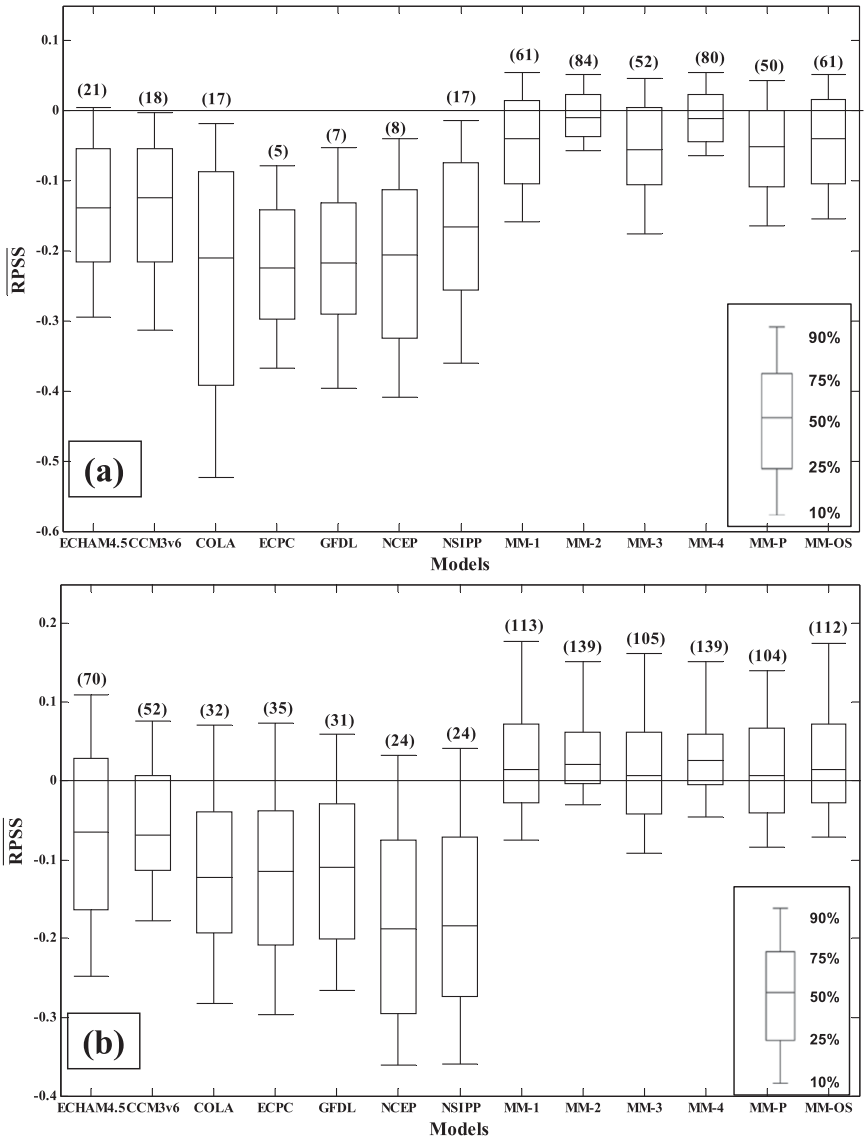

FIG. 3. Box plots of $\overline{\text{RPSS}}$ for predicting winter (a) precipitation and (b) temperature for individual GCMs and various multimodel schemes given in Table 2. Numbers in parenthesis above each box plot indicate the number of grid points having $\overline{\text{RPSS}}$ greater than zero.

3. Results and analysis

Six multimodel predictions (in Table 2) of winter precipitation and temperature are developed by combining seven AGCMs with climatology based on the algorithm shown in Fig. 2. The developed multimodel predictions are represented as tercile probabilities in 192 grid points over the continental United States for the period 1951–96.

a. Baseline comparison between multimodels and individual models

Figure 3 shows the box plot of $\overline{\text{RPSS}}$ for the seven individual models and for six multimodels over the entire United States. $\overline{\text{RPSS}}$ computes the cumulative squared error between the categorical forecast probabilities and the observed category in relevance to a reference forecast (Wilks 1995). The reference forecast is usually composed of climatological ensembles that have equal probabilities of occurrence under each category. A positive score for $\overline{\text{RPSS}}$ indicates that the forecast skill exceeds that of the climatological probabilities. Alternately, if the $\overline{\text{RPSS}}$ is negative, it indicates that the forecast skill is less than that of climatology. Since it evaluates the performance of entire conditional distributions, $\overline{\text{RPSS}}$ is a rigorous metric for evaluating categorical forecasts. Using the multimodels' and individual

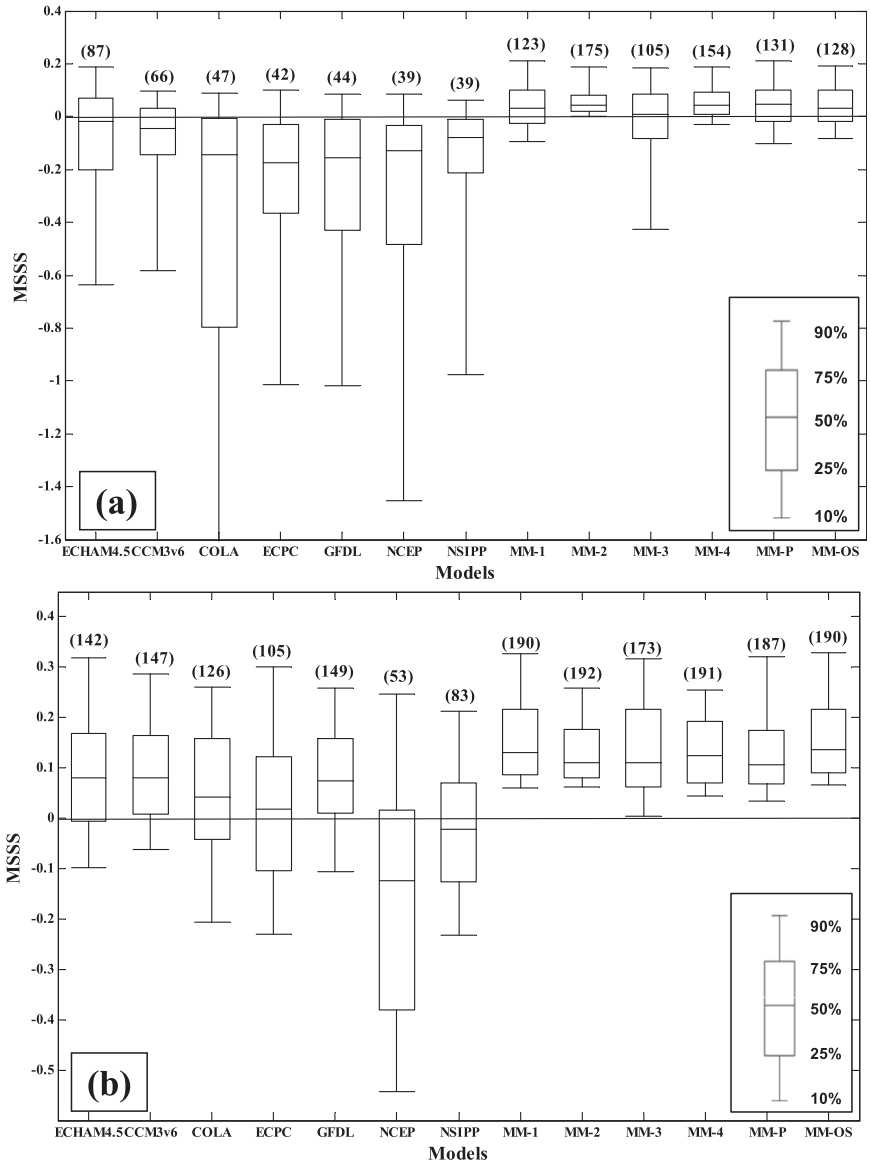

FIG. 4. Box plots of MSSS for predicting winter (a) precipitation and (b) temperature for individual GCMs and various multimodel schemes given in Table 2. Numbers in parenthesis above each box plot indicate the number of grid points having MSSS greater than zero.

models' tercile probabilities, we compute the $\overline{\text{RPSS}}$ for the period 1951–96.

Figures 3a and 3b show the box plots of $\overline{\text{RPSS}}$ for predicting winter precipitation and temperature, respectively, over the United States for the period 1951–96. Figure 3 also shows the number of grid points that have $\overline{\text{RPSS}}$ values greater than zero. Similarly, Figs. 4a and 4b also show the box plots of mean squared error–based skill scores (MSSSs) in predicting winter precipitation and temperature, respectively. For computing MSSS, we assume the conditional distribution resulting from the multimodel combination as being normal. From both Figs. 3 and 4, we can infer that the individual models' $\overline{\text{RPSS}}$ and MSSS values are less than zero in most of the grid points, which implies that the skill of the AGCMs is poorer than climatology. Among the individual models, ECHAM4.5 and version 6 of the Community Climate Model (CCM3v6) perform better than other GCMs in predicting winter precipitation and temperature. Further, we can also see that all six multimodels (in Table 2) perform better than the GCMs, with more grid points having positive $\overline{\text{RPSS}}$ and MSSS trends in predicting winter precipitation and temperature.

Comparing the performance of the multimodels in predicting precipitation (Figs. 3a and 4a) and temperature (Figs. 3b and 4b), we infer that the two-step multimodel combination schemes (MM-2 and MM-4) perform better than the currently employed techniques (MM-P and MM-OS) with a greater number of grid points having positive $\overline{\text{RPSS}}$ and MSSS. MM-2, which uses fixed neighbors contingent on the ENSO conditions in evaluating the skill of the models, also performs better than MM-P or MM-OS in predicting winter precipitation and temperature. Furthermore, we notice from both Figs. 3 and 4 that both individual models and multimodels have better skill in predicting winter temperature in comparison to their skill in predicting winter precipitation. Among the individual models, we see that ECHAM4.5 and CCM3v6 are the best individual models in predicting winter precipitation and temperature. So, all further analyses in quantifying the improvements resulting from multimodels will focus only on comparing with the performance of ECHAM4.5 and CCM3v6.

b. *Statistical significance of multimodel predictions: Hypothesis testing*

To ensure that the improved $\overline{\text{RPSS}}$ exhibited by the multimodel schemes (MM-1-MM-4) is statistically significant compared to the skill of ECHAM4.5, we perform detailed nonparametric hypothesis tests (Hamill 1999) by testing the null hypothesis that $\overline{\text{RPS}}$ of a multimodel scheme is equal to $\overline{\text{RPS}}$ of ECHAM4.5 in predicting the precipitation–temperature at each grid point. With model A denoting ECHAM4.5 and model B denoting one of the multimodel schemes (in Table 2), the null hypothesis for testing $\overline{\text{RPS}}$ could be written as

$$H_O: \overline{\text{RPS}}^{\text{A}} - \overline{\text{RPS}}^{\text{B}} = 0 \quad \text{or} \tag{1}$$

$$H_{\text{A}}: \overline{\text{RPS}}^{\text{A}} - \overline{\text{RPS}}^{\text{B}} \neq 0. \tag{2}$$

The distribution of the null hypothesis, $\overline{\text{RPS}}^{1,*} - \overline{\text{RPS}}^{2,*}$, is constructed by resampling equally likely from the RPS_t^{A} of model A (i.e., ECHAM4.5) and the RPS_t^{B} of model B (MM-1 to MM-4, MM-P, and MM-OS each year). In other words, $\overline{\text{RPS}}^{1,*}$ and $\overline{\text{RPS}}^{2,*}$, the average rank probability score ($\overline{\text{RPS}}$) estimated from 46 yr to construct the null distribution, are resampled equally likely to incorporate RPS_t^{A} and RPS_t^{B}. A total of 10 000 estimates of $\overline{\text{RPS}}^{1,*} - \overline{\text{RPS}}^{2,*}$ are obtained to develop the null distribution for each grid point over the United States. The percentiles at which the observed test statistic at each grid point, $\overline{\text{RPS}}^{\text{A}} - \overline{\text{RPS}}^{\text{B}}$, has fallen in the constructed null distribution are computed. Results from the hypothesis tests, the percentiles of the observed test statistic $\overline{\text{RPS}}^{\text{A}} - \overline{\text{RPS}}^{\text{B}}$ on the constructed null distribution, are plotted on the U.S. map to identify grid points showing significant improvement from multimodel combinations (Figs. 5 and 6). For a significance level of 10%, if the percentile of the observed test statistic is between 0.9 and 1 (0 and 0.1) at a given grid point, then the model B (model A) $\overline{\text{RPS}}$ is statistically lower than the model A (model B) $\overline{\text{RPS}}$. For additional details on the performed nonparametric hypothesis test, see Hamill (1999).

Tables 3 (precipitation) and 4 (temperature) summarize the results from hypothesis tests across the six multimodels. Entries in the upper triangles in Tables 3 and 4 provide the numbers of grid points having the percentiles of observed test statistic between 0.9 and 1 on the constructed null distribution, which implies that the $\overline{\text{RPS}}$ of model B—represented by the columns—is statistically higher than the $\overline{\text{RPS}}$ of model A, which is represented as row entries. For instance, from the upper triangle in Table 3, from the hypothesis tests between MM-1 (Model A) and MM-P (Model B), we find that MM-P's $\overline{\text{RPS}}$ is statistically smaller than the $\overline{\text{RPS}}$ of MM-1 at 24 grid points (with the percentiles of the observed test statistic between 0.9 and 1 in the null distribution). Similarly, results from the same hypothesis tests are also summarized in the lower triangle between the two models, indicating the number of grid points over which the percentiles of the observed test statistic fell between 0 and 0.1 on the constructed null distribution, which implies the MM-1's (model A) $\overline{\text{RPS}}$ is statistically smaller than the $\overline{\text{RPS}}$ of MM-P (model B) at 20 grid points. For both Tables 3 and 4, the best-performing model in terms of the increased number of significant grid points is italicized by its column entry. Thus, between MM-1 and MM-P, we infer that MM-P (underlined by the column) performs better in more grid points in comparison to MM-1 in predicting precipitation.

Figure 5 shows the relative performance of six multimodel combination schemes over the best individual model (ECHAM4.5) in predicting winter precipitation over the entire United States. From Fig. 5a and Table 3, 39 (5) grid points have the percentiles of the test statistic falling between 0.9 and 1(0 and 0.1), which indicates that MM-1(ECHAM4.5) performs better than ECHAM4.5 (MM-1) in those grid points by rejecting the null hypothesis that the difference in $\overline{\text{RPS}}$ between the ECHAM4.5 and MM-1 is zero for a significance level of 10%. We can also see that many grid points fell between 0.1 and 0.9, indicating the difference in skill is not statistically significant at 10%. However, a plus sign is used in Fig. 5 to indicate that the $\overline{\text{RPS}}$ of the corresponding multimodel at that grid point is lesser than the $\overline{\text{RPS}}$ of ECHAM4.5. Even though the difference in $\overline{\text{RPS}}$ is statistically not significant at 10%, we observed that the percentage reduction in using multimodel combinations is around 5%–15% for

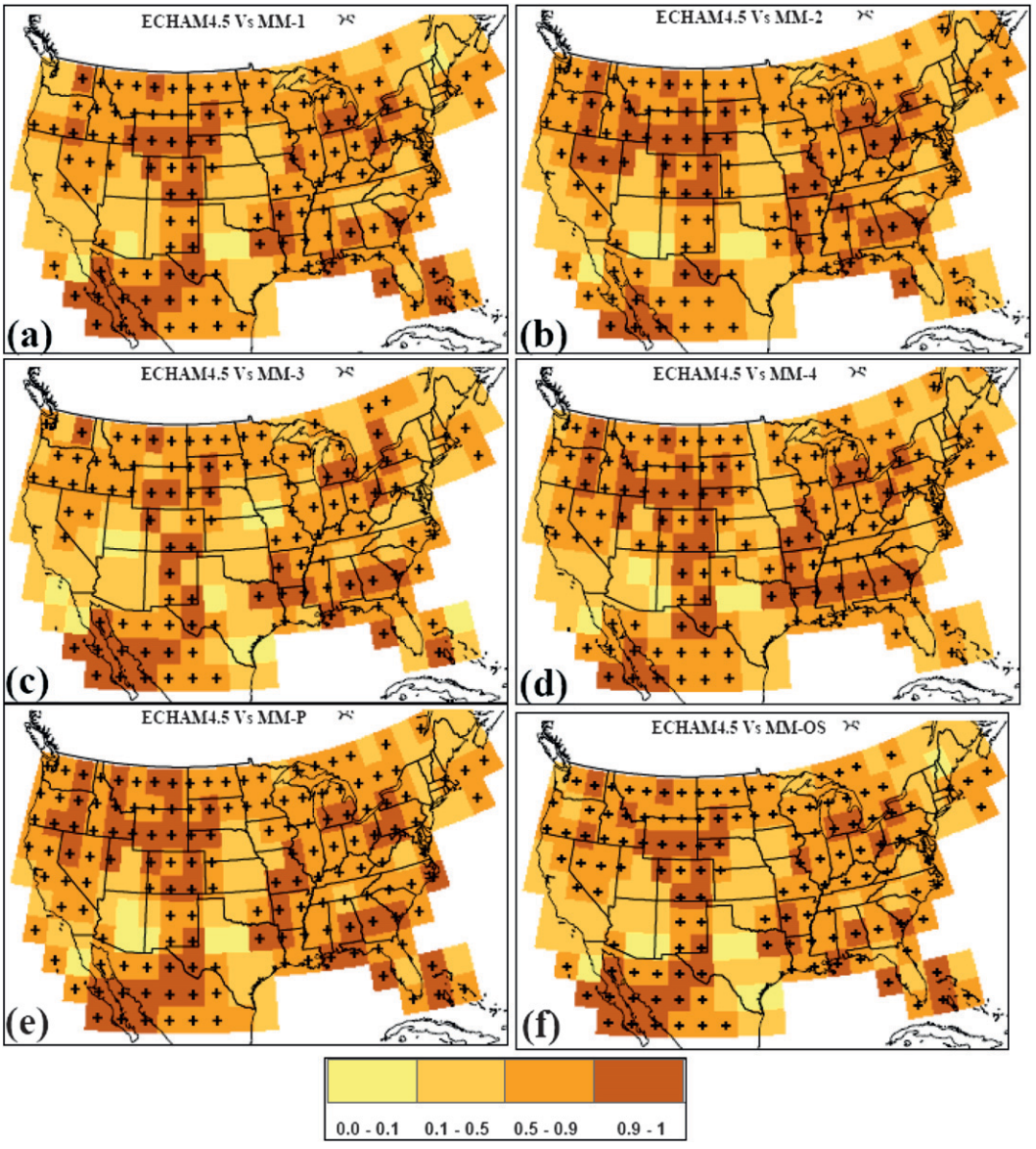

FIG. 5. Performance comparison of multimodels with the best individual model, ECHAM4.5, for predicting U.S. winter precipitation. The background color indicates the percentile of the test statistic ($\overline{\text{RPS}}^{\text{ECHAM4.5}} - \overline{\text{RPS}}^{\text{MM}}$) obtained from the resampled null distribution that represents $\overline{\text{RPS}}^{\text{ECHAM4.5}} - \overline{\text{RPS}}^{\text{MM}}$. A lower (higher) value of the percentiles from the test statistic indicates ECHAM4.5 (multimodel) performs better than multimodel (ECHAM4.5). A plus (blank) sign indicates that the $\overline{\text{RPS}}$ from the multimodel (ECHAM4.5) is lesser than the $\overline{\text{RPS}}$ of ECHAM4.5 (multimodel).

grid points with the percentiles of the observed test statistic, $\overline{\text{RPS}}^{\text{A}} - \overline{\text{RPS}}^{\text{B}}$, from 0.5 to 0.9.

Among the multimodels, MM-2 and MM-4 perform better than the rest of the multimodels, which is indicated by the greater number of grid points (Fig. 3 and Table 3) having statistically significant $\overline{\text{RPS}}$ than the $\overline{\text{RPS}}$ of the rest of the multimodels and ECHAM4.5. From Table 3, we clearly understand that the multimodel scheme proposed in this study (MM-2 and MM-4) perform better than the existing techniques on multimodel combinations (MM-P and MM-OS). It is important to note that both MM-2 and MM-4 employ two-step combinations to develop multimodel predictions. Comparing between MM-2 and MM-4, we infer that in 52 (25) grid points MM-2's (MM-4) $\overline{\text{RPS}}$ is statistically more significant than the $\overline{\text{RPS}}$ of MM-4 (MM-2) with the observed test statistic between the two models falling between 0.9 and 1 (0 and 0.1) on the constructed null distribution. This indicates

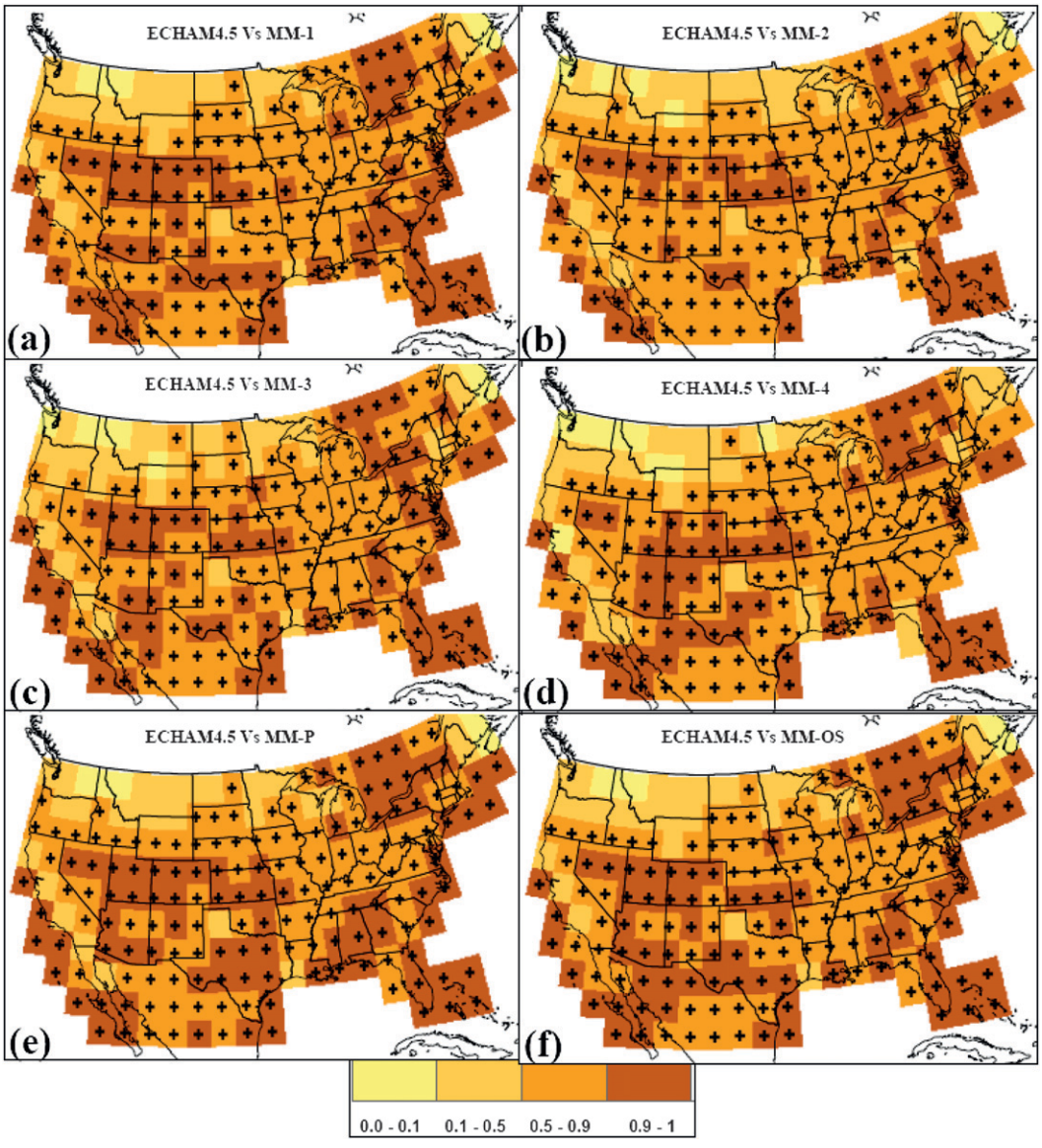

FIG. 6. Performance comparison of multimodels with the best individual model, ECHAM4.5, for predicting U.S. winter temperature. The background color indicates the percentile of the test statistic ($\overline{\text{RPS}}^{\text{ECHAM4.5}} - \overline{\text{RPS}}^{\text{MM}}$) obtained from the resampled null distribution that represents $\overline{\text{RPS}}^{\text{ECHAM4.5}} = \overline{\text{RPS}}^{\text{MM}}$. A lower (higher) value of the percentiles from the test statistic indicates ECHAM4.5 (multimodel) performs better than the multimodel (ECHAM4.5). A plus (blank) sign indicates that the $\overline{\text{RPSS}}$ from the multimodel (ECHAM4.5) is lesser than the $\overline{\text{RPS}}$ of ECHAM4.5 (multimodel).

that the two-step combination seems to be more effective in reducing the $\overline{\text{RPS}}$ of multimodels in predicting precipitation. Recently, Chowdhury and Sharma (2009) showed that combining multimodels that have their least covariance at the first step seems to be more effective for developing better multimodel predictions. Given this, it seems obvious that climatology will have the smallest covariance with individual model predictions, thereby indicating that two-step combinations are very effective in reducing the $\overline{\text{RPS}}$ of multimodels.

Figure 6 and Table 4, which are similar to Fig. 5 and Table 3, summarize the multimodel combination results for temperature. From Fig. 6, it is very clear that all the multimodels perform better than ECHAM4.5 in predicting temperature. Among the multimodels, the MM-1 proposed in the study performs better than the rest of the multimodels. From Table 4, we also infer that the performance of MM-OS is equally good at predicting the winter temperature. Comparing the performance of MM-1 and MM-2, we infer that in 48 (24) grid points

TABLE 3. Number of grid points showing a significant difference in $\overline{\text{RPS}}$ when predicting precipitation based on the hypothesis testing between ECHAM4.5 and the various multimodel schemes given in Table 2. Entries in the top (bottom) triangle of the table summarize the number of grid points having percentile values of the test statistic $\overline{\text{RPS}}^{\text{A}} - \overline{\text{RPS}}^{\text{B}}$ between 0.9 and 1 (0 and 0.1) in the resampled null distribution for hypothesis testing between model A and model B. For values in the top (bottom) triangle, model A (model B) is indicated by its row entry and model B (model A) is indicated by its column entry. The best-performing models in terms of the increased number of significant grid points are italicized.

Models	ECHAM4.5	MM-1	MM-2	MM-3	MM-4	MM-P	MM-OS
ECHAM4.5		*39*	*42*	*36*	*41*	*53*	*40*
MM-1	5		*50*	22	*51*	*24*	*59*
MM-2	5	6		7	25	11	9
MM-3	10	*65*	*57*		*64*	*43*	*81*
MM-4	5	16	*52*	6		11	16
MM-P	6	20	*33*	19	*37*		*20*
MM-OS	8	17	*35*	16	*43*	17	

MM-OS's (MM-1) $\overline{\text{RPS}}$ is statistically more significant than the $\overline{\text{RPS}}$ of MM-1(MM-OS), indicating that combining models purely based on their long-term skill seems to be a good strategy in multimodel combinations. However, among these grid points, if we drop grid points with the $\overline{\text{RPSS}}$ of both models being negative, then we end up with 27 (19) grid points that show the $\overline{\text{RPS}}$ of MM-OS (MM-1) being statistically more significant than the $\overline{\text{RPS}}$ of MM-1 (MM-OS). This indicates MM-1's better performance is more pronounced in grid points exhibiting positive $\overline{\text{RPSS}}$. From Tables 3 and 4, we also understand that the improvements in predicting winter temperature from multimodel combinations is more similar to the improvements in predicting winter precipitation. In section 3e, we discuss in detail improvements resulting from multimodel combinations from a regional perspective over the continental United States, particularly for grid points that exhibit positive $\overline{\text{RPSS}}$.

It is important to note that Figs. 5 and 6 show spatial correlation in the reported percentiles of the test statistic. This is because we resample the RPS_t from models A and B available at each grid point to construct the null distribution. Performing hypothesis tests with a spatially correlated forecast error metric would reduce the effective number of independent samples (Wilks 1997; Hamill 1999). One way to overcome the spatially correlated prediction error metric is to spatially average the verification measure over a region and perform the hypothesis tests over the spatially averaged verification measure. However, we felt that such an approach would first require the identification of homogenous regions for spatial averaging of the skill metric, so this approach is not pursued here.

c. *Comparison of forecast reliability between multimodels and individual models*

Ranked probability scores only quantify the squared errors in forecasted cumulative probabilities for categorical forecasts. In addition, they do not provide information on how the forecasted probabilities for a particular category correspond to their observed frequencies. For this purpose, this section compares the reliability, resolution of multimodel predictions with the reliability, and resolution of individual model predictions. Reliability diagrams provide information on the correspondence between the forecasted probabilities for a particular

TABLE 4. Number of grid points showing significant differences in $\overline{\text{RPS}}$ when predicting temperature based on the hypothesis testing between ECHAM4.5 and the various multimodel schemes given in Table 2. Entries in the top (bottom) triangle of the table summarize the number of grid points having percentile values of the test statistic $\overline{\text{RPS}}^{\text{A}} - \overline{\text{RPS}}^{\text{B}}$ between 0.9 and 1 (0 and 0.1) in the resampled null distribution for hypothesis testing between model A and model B. For values in the top (bottom) triangle, model A (model B) is indicated by its row entry and model B (model A) is indicated by its column entry. The best-performing models in terms of an increased number of significant grid points are italicized.

Models	ECHAM4.5	MM-1	MM-2	MM-3	MM-4	MM-P	MM-OS
ECHAM4.5		*70*	*48*	*65*	*59*	*81*	*77*
MM-1	5		6	27	17	3	*48*
MM-2	6	*39*		*28*	33	*19*	*41*
MM-3	7	*51*	9		23	8	*58*
MM-4	9	*44*	*49*	*36*		24	*46*
MM-P	5	*42*	18	*34*	*31*		*52*
MM-OS	5	24	4	17	9	3	

category (e.g., above-normal, normal, and below-normal categories) and how frequently that category is observed under the issued forecasted probability (Wilks 1995). For instance, if we forecast the probability of the below-normal category as 0.9 over n1 yr, then we expect the actual outcome to fall under the below-normal category for $0.9 \times$ n1 times over the entire forecast verification period.

Figures 7a and 7b compare the reliabilities of three multimodels (MM-2, MM-4, and MM-OS) with the reliabilities of ECHAM4.5 and CCM3v6 in predicting precipitation for below- and above-normal categories, respectively. Similarly, Figs. 8a and 8b compare the reliabilities of MM-1, MM-3, and MM-OS with the reliabilities of ECHAM4.5 and CCM3v6 in predicting the temperature for the below- and above-normal categories, respectively. We did not consider MM-P since it did not reduce the $\overline{\text{RPS}}$ over many grid points in comparison to the rest of the multimodels in predicting precipitation and temperature (Tables 3 and 4).

For developing reliability plots, the tercile probabilities for 46 yr under each category are grouped at an interval of 0.1 over all grid points ($46 \times 192 = 8832$ forecasts for a tercile category for each model). The observed category is also recorded; using this, the observed relative frequency under each forecasted probability is calculated for each tercile category. The inset in each of the reliability plots shows the attribute diagram indicating the logarithm of the number of forecasts that fell under each forecast probability bin for a given model. Figures 7 and 8 also show the perfect (diagonal) reliability line with one to one correspondence between the forecasted probability and its observed relative frequency.

From Figs. 7 and 8, we observe that the selected multimodels improve the reliability of forecasts showing better correspondence between forecasted probabilities and their observed relative frequencies. The basic reason multimodel predictions result in better reliability is by reducing the overconfidence of individual model predictions. This could be understood from the attribute diagram, which clearly shows a reduction in the number of predictions with high forecast probabilities (0.8–1) under individual models (ECHAM4.5 and CCM3v6). These findings are in line with earlier studies (Weigel et al. 2008). On the other hand, multimodels show increases in the number of predictions under moderate forecast probabilities (0.4–0.7), thereby resulting in the reduction of false alarms. Similarly, under low forecast probabilities, individual models seem to be less reliable, indicating a higher frequency of occurrence, whereas multimodels have better reliability resulting in a reduction in the number of missed targets. To better quantify the information in Figs. 7 and 8, we summarize the ability of a model to predict a particular tercile category using the average Brier score ($\overline{\text{BS}}$; Wilks 1995).

The Brier score, which summarizes the squared error in the categorical forecast probabilities, can be decomposed into reliability, resolution, and uncertainty (Wilks 1995). For $\overline{\text{BS}}$ to be close to zero, it is important that the reliability term should be close to zero and the resolution term should be large. Figures 9a (9c) and 9b (9d) provide the reliability, resolution, and $\overline{\text{BS}}$ for ECHAM4.5, CCM3v6, and all six of the multimodels in predicting the below- and above-normal categories of precipitation (temperature). From Figs. 9a and 9b, we infer that all multimodels have smaller reliability scores in comparison to the reliability scores of individual models under both tercile categories, thereby contributing to the reduction in the $\overline{\text{BS}}$. Among the multimodels, MM-2 has the smallest reliability score when compared to the remaining five multimodels in predicting precipitation. In terms of resolution, ECHAM4.5 has a larger resolution score than the resolution scores of CCM3v6 and the other multimodels in predicting precipitation. Among the multimodels, we clearly see that MM-2 has the largest resolution score, which leads to MM-2 and MM-4 having the lowest $\overline{\text{BS}}$ in predicting precipitation. Similarly, from Figs. 9c and 9d, we infer that MM-1 and MM-OS have the smallest $\overline{\text{BS}}$, which results primarily from smaller reliability scores and larger resolution scores in predicting the temperature in the below- and above-normal categories.

To summarize the multimodel schemes, MM2 and MM4 (MM1 and MM-OS), perform better than individual models as well as yielding better results than the remainder of the multimodels in predicting winter precipitation (temperature) over the continental United States (Figs. 9a–d). The proposed multimodel combination schemes in this study (MM-2 and MM-4) have the lowest $\overline{\text{BS}}$ among all the models in predicting precipitation, whereas MM-1 also performs equally well (in comparison to the best multimodel, MM-OS) in predicting winter temperature. To understand why the multimodel combination schemes proposed in Fig. 2 result in improved predictions, we plot the weights (w_k^m) obtained for each of the multimodel schemes in the next section and analyze how they vary conditioned on the ENSO state.

d. Analysis of weights

Figure 10 shows box plots of the ratio of the weights ($w_{t,k}^m|$MM-1) for each model under the MM-1 scheme to the weights ($w^m|$MM-OS) obtained for each model based on the MM-OS scheme in predicting temperature. The weight ratios plotted in Fig. 10 are grouped into two categories, namely grid points exhibiting significant skill

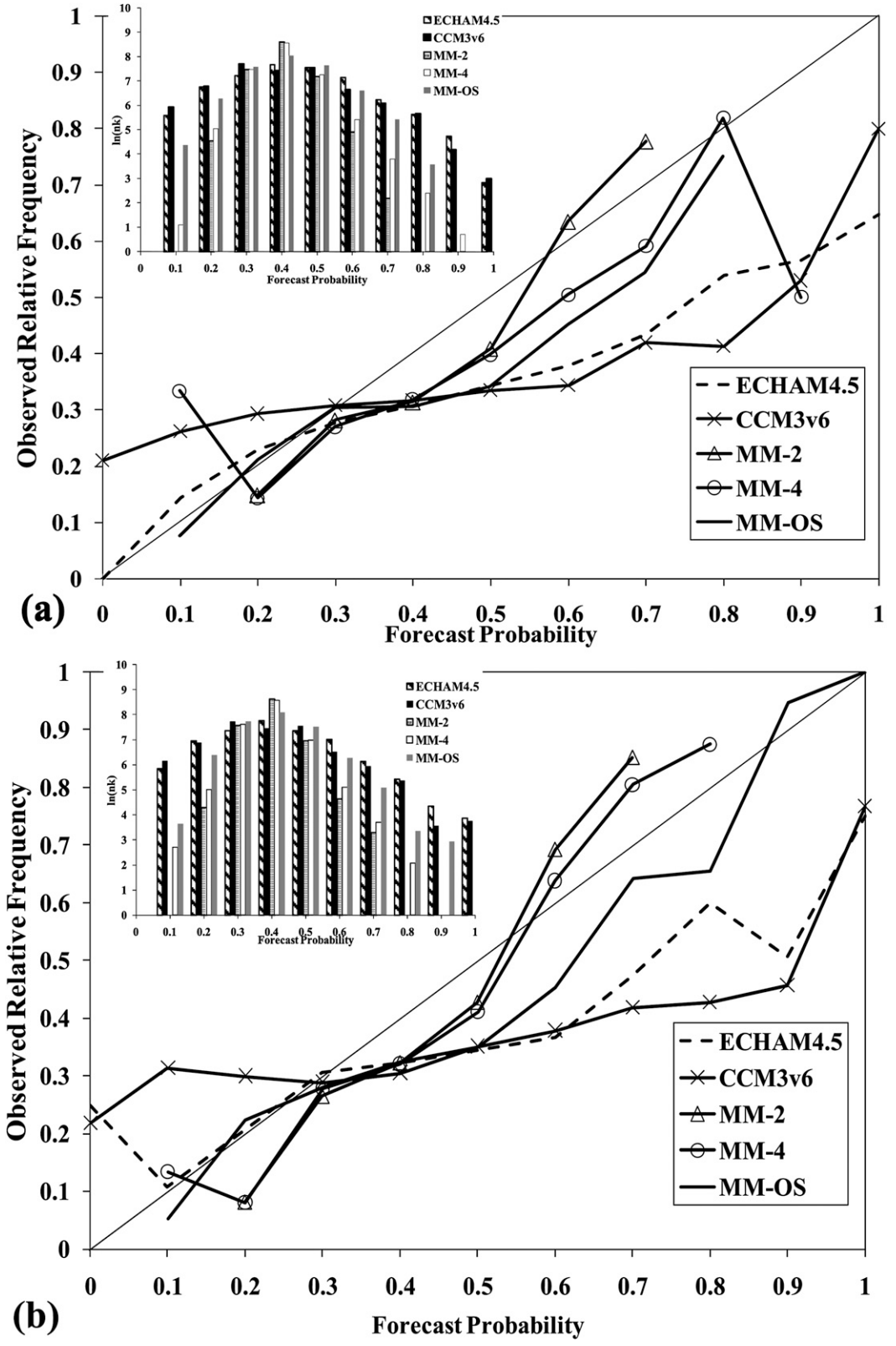

FIG. 7. Reliability diagrams for individual models, ECHAM4.5 and CCM3v6, and for various multimodel combination schemes in predicting (a) below- and (b) above-normal categories of precipitation.

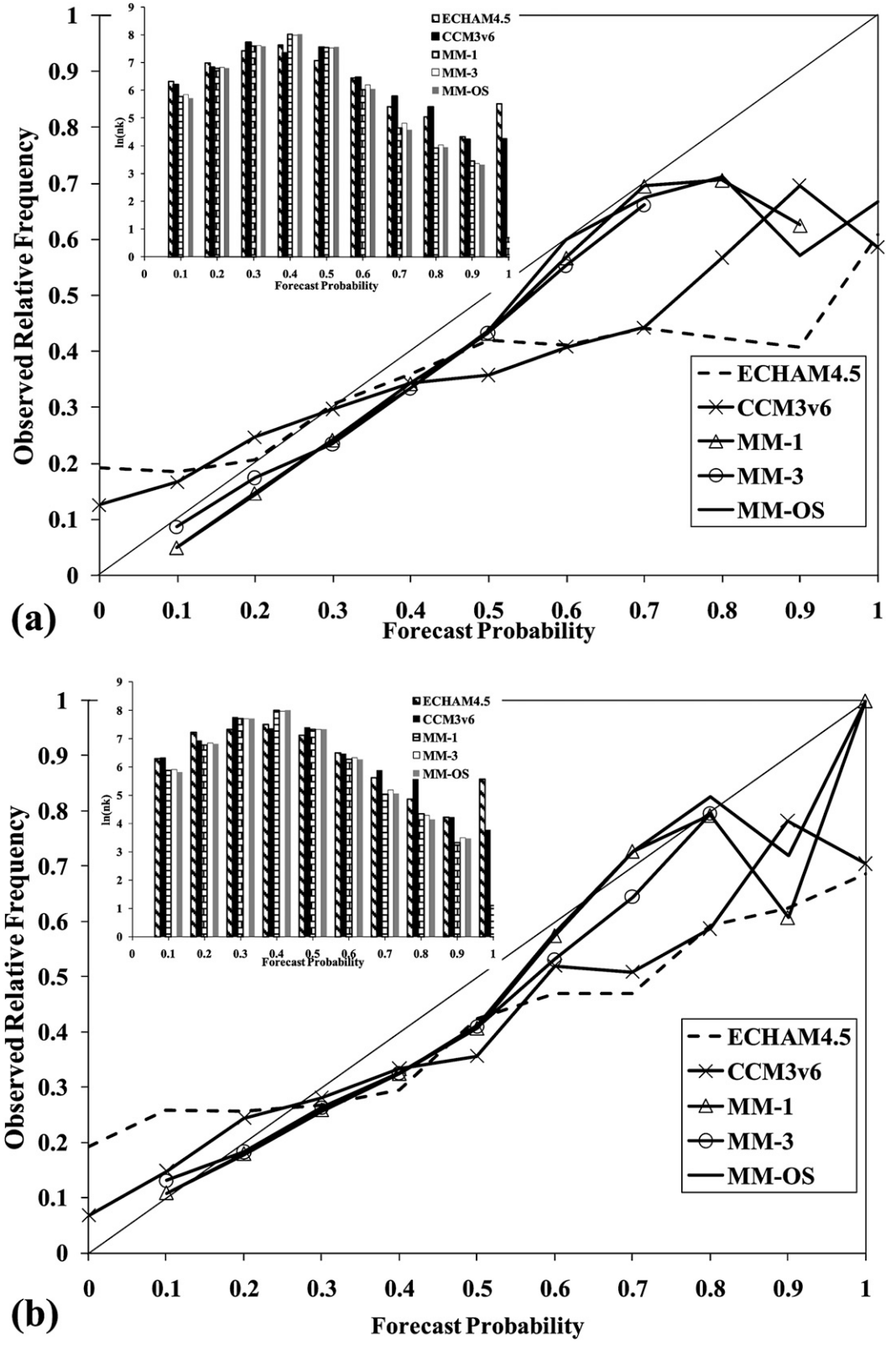

FIG. 8. Reliability diagrams for individual models, ECHAM4.5 and CCM3v6, and for various multimodel combination schemes in predicting (a) below- and (b) above-normal categories of temperature.

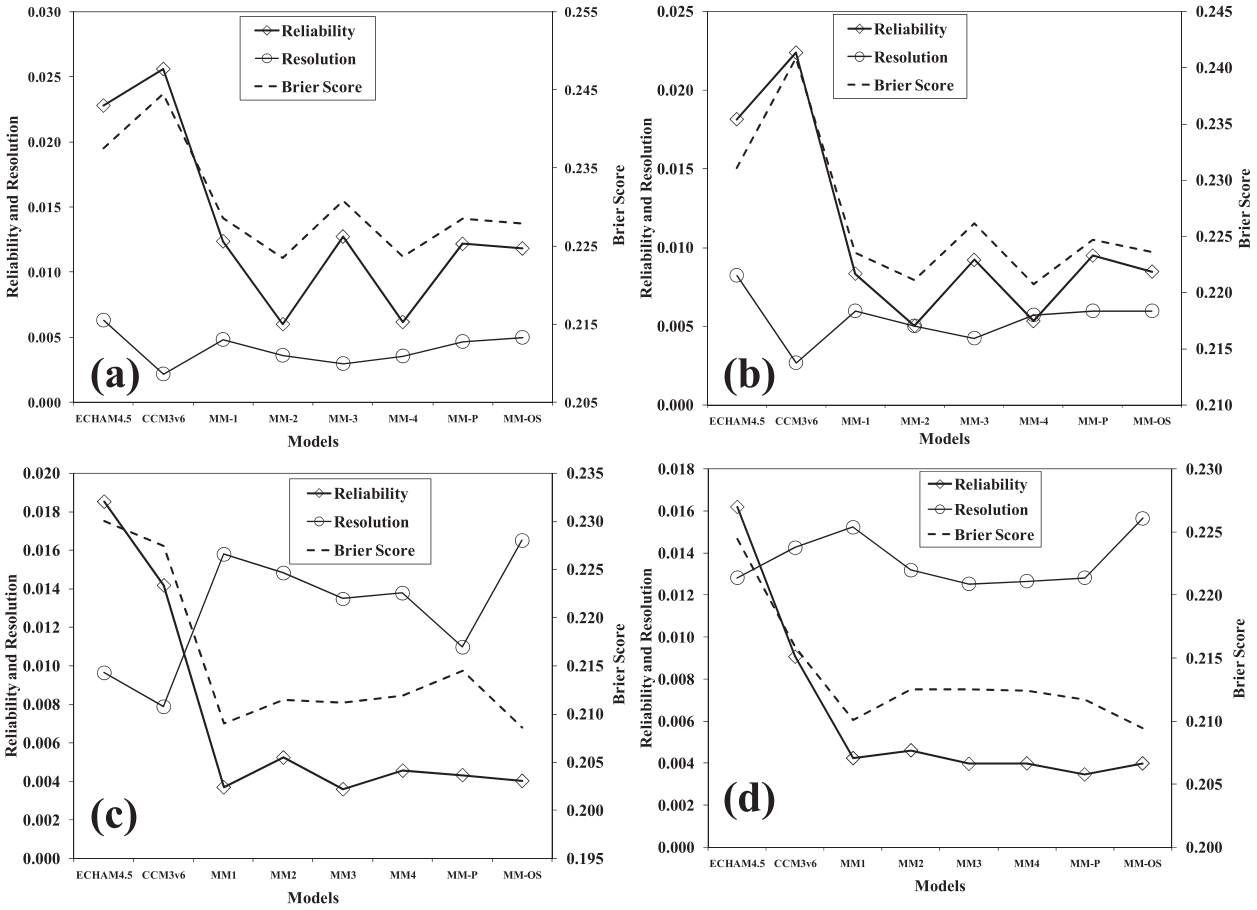

FIG. 9. Performance comparison of individual models, ECHAM4.5 and CCM3v6, with various multimodels based on the Brier score and its components (reliability and resolution) in predicting below normal [(a), precipitation; (c), temperature] and above-normal [(b), precipitation; (d), temperature] events.

under El Niño years (Fig. 10a) and La Niña years (Fig. 10b) by ECHAM4.5. This resulted in a total of 20 and 43 grid points exhibiting significant skill (based on the correlation between the observed temperature and the ensemble mean) under El Niño and La Niña years, respectively. The weights ($w^m_{t,k}|$MM-1, $w^m|$MM-OS) are pooled over 46 yr and the computed weight ratios for these grid points are shown as separate box plots (as columns) conditioned on the ENSO state (El Niño, La Niña, and neutral). Weight ratios above 1 indicate that the MM-1 weights for a given model are higher than the weights assigned by the MM-OS scheme.

From Fig. 10a, which shows the weight ratios for grid points showing significant skill during El Niño years, we can clearly see that weight ratios are greater than 1 for ECHAM4.5 around 25% of the time and lesser than 1 for climatology around 85% of the time during El Niño conditions (first column in Fig. 10a). However, the weight ratios for the Geophysical Fluid Dynamics Laboratory (GFDL) model are higher than 1 (around 75% of the time), indicating GFDL's better performance during El Niño years for the grid points considered in Fig. 10a. This implies that if a particular GCM performs well during El Niño years, then higher weights are assigned for that GCM during those conditions in comparison to the weights based on the long-term skill of the model (MM-OS). Further, the weights assigned for climatology under the MM-1 scheme are less since all GCMs have good skill in predicting temperature during El Niño conditions.

On the other hand, during neutral conditions (last column in Fig. 10a), the weight ratios are substantially less than 1 for both ECHAM4.5 and GFDL, whereas the weight ratios are greater than 1 for climatology (around 90% of the time). Under La Niña conditions for grid points exhibiting significant skill during El Niño years (Fig. 10a, middle column), we can clearly infer that the weights for ECHAM4.5 from the MM-1 schemes are higher in comparison to the weights for ECHAM4.5 received from the MM-OS scheme during La Niña years. This analysis again

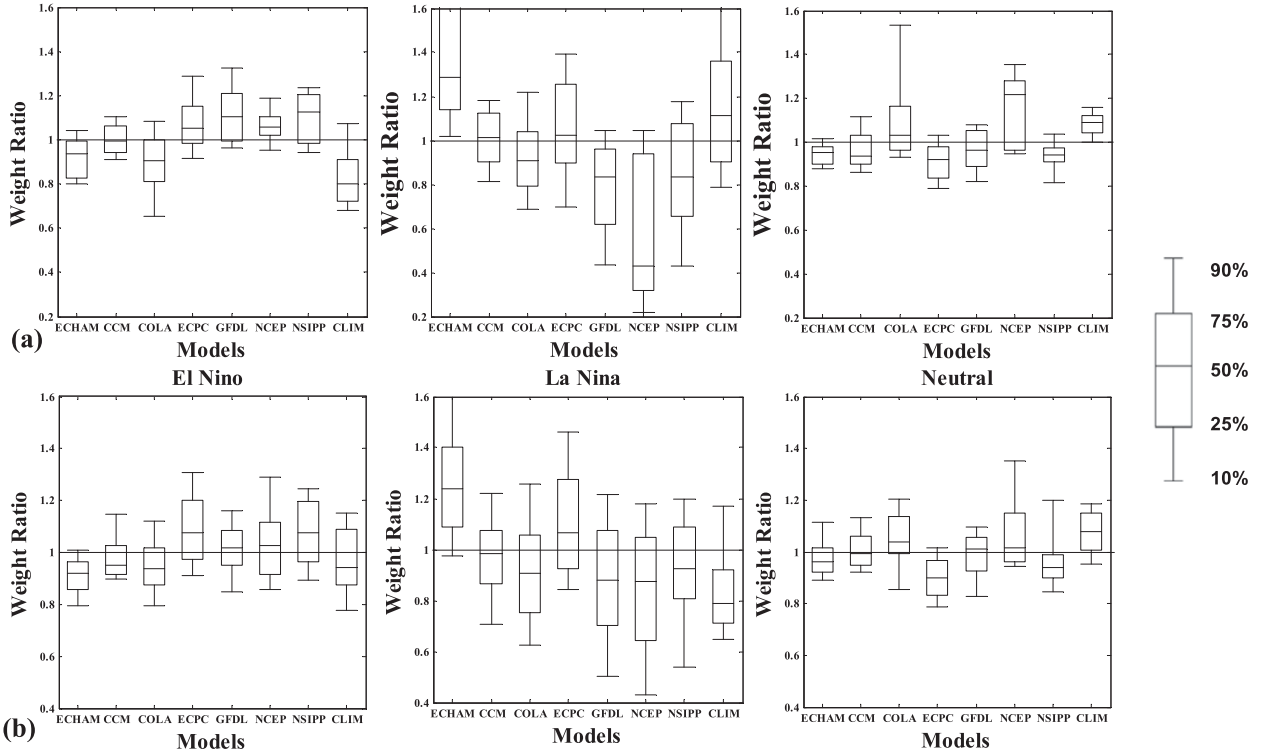

FIG. 10. Box plots of the ratio of the weights ($w^m_{t,K}$|MM-1) for each model using the MM-1 scheme to the weights for each model (w^m|MM-OS) using the MM-OS scheme in predicting temperature. The weights ratio, pooled (as columns) under three ENSO conditions, for grid points having significant correlations with the observed precipitation and the ensemble mean from ECHAM4.5 during (a) El Niño and (b) La Niña years. A weight ratio above 1, for a given GCM scheme, indicates that the weight under MM-1 is higher than the weight assigned based on the MM-OS scheme.

confirms our argument that if a GCM's skill is poor during certain predictor conditions, then it is better to consider climatology as the best information available. Our multimodel combination algorithm shown in Fig. 2 basically implements this by evaluating the models' skill contingent on the dominant predictor state and assigns a higher weight for the model that exhibits better skill during those predictor conditions.

Figure 10b shows similar results for grid points exhibiting significant skill in predicting DJF temperature by ECHAM4.5 during La Niña years. From Fig. 10b, we can clearly see that the weight ratios for ECHAM4.5 are mostly less than 1, with ECPC's ratio being higher than 1 during El Niño conditions (Fig. 10b, leftmost column). On the other hand, during La Niña conditions, the ECHAM4.5 weight ratios are greater than 1 (around 85% of the time), which forces the weight ratios for climatology to be substantially less than 1 (around 60% of the time; Fig. 10b, middle column). Under neutral conditions (Fig. 10b, rightmost column), with none of the models exhibiting significant skill, the weights assigned by the MM-1 scheme for climatology are higher than the weights assigned by the MM-OS scheme (around 90% of the neutral conditions in those grid points). Similar analysis on weights under MM-2 in predicting precipitation revealed the same pattern (figure not shown). Thus, our study clearly shows that combining multiple climate models based on their ability to predict under a given predictor (or predictors) condition is a potential strategy for improving the skill of multimodel climate forecasts.

4. Discussion

The main advantage of the proposed multimodel combination technique is in its ability to evaluate GCM skill contingent on the predictor state and to assign higher weights for GCMs that perform better during those (predictor) conditions. Similarly, the methodology could also assign higher weights to climatology if all the GCMs have limited skill under that conditioning state. On the other hand, pursuing multimodel combinations purely based on the overall skill (MM-OS) could result in higher weights for GCMs under conditions during which the model might exhibit poor/limited skill. Further, the proposed approach combines both models' skill (as quantified

by MSE) and optimization (choosing the number of neighbors, K_t, in MM-3 and MM-4 under two-deep cross validation) to estimate weights as opposed to obtaining model weights based purely on optimization (e.g., Rajagopalan et al. 2002), which could lead to choosing one or two models alone in a multimodel combination. To overcome these difficulties, Robertson et al. (2004) proposed a two-step combination. Analysis of the weights (Fig. 10) shows clearly that model weights are linked to their skill, with GCMs weights being higher during ENSO conditions and climatology receiving higher weights during neutral conditions.

Figure 11 shows the skill, expressed as $\overline{\text{RPSS}}$, in predicting DJF precipitation (panel a) and temperature (panel b) with each grid point's $\overline{\text{RPSS}}$ being indicated by the best-performing individual model or the multimodel. Table 5 shows the number of grid points with each individual model and requirements of having the highest $\overline{\text{RPSS}}$ over 192 grid points shown in Fig. 11. Figure 11 and Table 5 summarize the performance results of the models (individual model and multimodels) only if the $\overline{\text{RPSS}}$ of at least one model is greater than zero at a given grid point. Thus, if the $\overline{\text{RPSS}}$ of all individual models and multimodels are lesser than zero, then climatology provides the best information for those grid points. From Fig. 11 and Table 5, we can clearly see that multimodels proposed in this study (MM-1, MM-2, MM-3, and MM-4) perform better than the individual models and better than the existing multimodel combination techniques (MM-P and MM-OS). Among the multimodels, MM-4 seems to be the best-performing multimodel in predicting precipitation and temperature, whereas the Center for Land–Ocean–Atmosphere (COLA) and ECHAM4.5 models seem to be the best-performing individual models in predicting precipitation and temperature, respectively. Comparing Figs. 11a and 11b, we infer that the prediction of temperature seems to benefit more from multimodel combinations in comparison to the improvements resulting for precipitation.

From Fig. 11a, we understand that the improvements resulting from multimodel combinations in predicting DJF precipitation predominantly lies over the southern United States as well as over certain grid points in the Midwest and Northeast. In the case of temperature (Fig. 11b), with the exception of the Midwest, we infer that $\overline{\text{RPSS}}$ is greater than zero for most of the regions, indicating better skill (in comparison to climatology), which is demonstrated by both the individual models and the multimodels. From Fig. 11, we can see that there is a significant improvement in $\overline{\text{RPSS}}$ for the multimodel schemes proposed in the study (shown as open circles) over the southeast and southwest regions of the United States and over northwest Mexico. Figures 5 and 6 also show similar spatial structures with the $\overline{\text{RPS}}$ of multimodels being statistically more significant than the $\overline{\text{RPS}}$ of ECHAM4.5. The reason for this improved performance over these regions is primarily due to the strong correlation between the ensemble mean of the individual models with the observed precipitation/temperature under ENSO conditions.

It is important to note that this study has employed historical simulations of precipitation and temperature from AGCMs to demonstrate the utility of the multimodel combination algorithm presented in Fig. 2. Historical simulations from AGCMs that employ observed SSTs as boundary conditions typically overestimate the real predictive skill (Goddard et al. 2002; Sankarasubramanian et al. 2008). Further, to apply the proposed methodology within a forecasting context, one may have to use the forecasted Niño-3.4 from multiple CGCMs as the conditioning variable (Tippett and Barnston 2008). Given that the peak ENSO activity typically coincides during the DJF season, 3-month multimodel forecasts of ENSO indices issued in December exhibit very high skill with correlations of above 0.8 and root-mean-square errors of around 0.2°–0.4°C (Jin et al. 2008; Weisheimer et al. 2009). Due to these results, the identified similar DJF ENSO conditions using the forecasted multimodel mean of the DJF Niño-3.4 could slightly differ from the identified conditions using the observed DJF Niño-3.4. More importantly, employing retrospective forecasts from AGCMs forced with forecasted SSTs could result in reduced skill from the proposed multimodel scheme if the skill levels of the retrospective forecasts from AGCMs are better than that of climatology under the conditioned state. But, if the skill levels of the retrospective forecasts from the AGCMs are poorer than that of climatology (which is highly likely based on Figs. 3 and 4), then we expect that the proposed multimodel scheme is likely to be more beneficial in replacing AGCMs forecasts with climatology. Our future investigation will evaluate the utility of the proposed methodology in combining real-time precipitation and temperature forecasts from CGCMs contingent on the forecasted DJF Niño-3.4 state.

5. Summary and conclusions

A methodology for combining multiple GCMs is proposed and evaluated for predicting winter precipitation and temperature over the United States. The methodology assigns weights for each GCM by evaluating their skill, quantified by mean square error, over similar predictor conditions. Considering Niño-3.4 as the primary predictor influencing the winter precipitation and temperature (Quan et al. 2006), the study combines seven GCMs with climatological ensembles to develop multimodel predictions over the continental United States. In

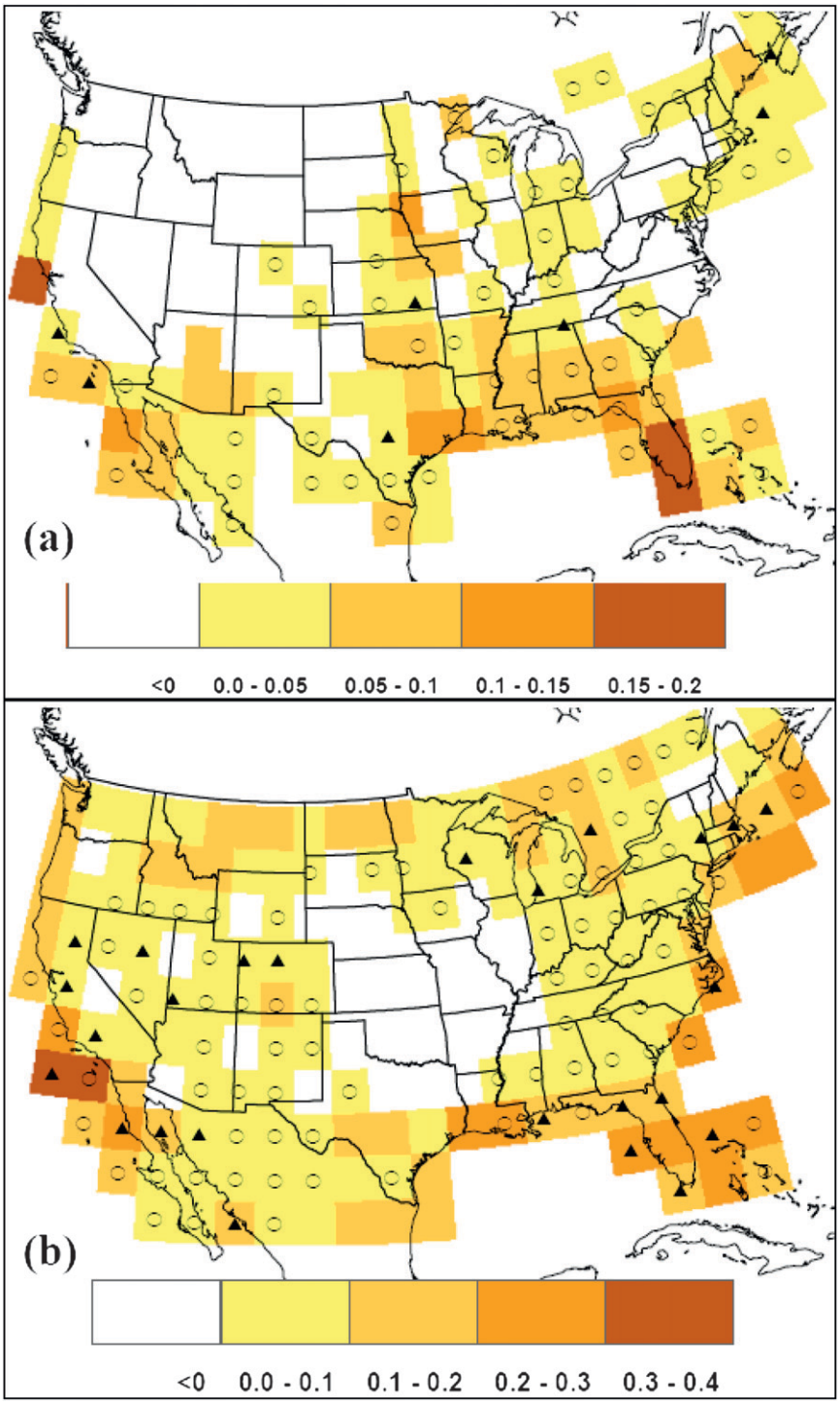

FIG. 11. Performance of multimodels and individual models, expressed as $\overline{\text{RPSS}}$, in predicting DJF winter (a) precipitation and (b) temperature. Grid points having $\overline{\text{RPSS}}$ lesser than zero are shown in white. Grid points with no markers, open circles, and triangles indicate the best-performing model (having the highest $\overline{\text{RPSS}}$) at that grid point being individual GCMs (all seven models in Table 1), multimodels proposed in this study (MM-1, MM-2, MM-3, and MM4), and existing multimodel techniques (MM-P and MM-OS), respectively.

TABLE 5. Numbers of grid points with each individual GCM and multimodels having the highest $\overline{\text{RPSS}}$ in predicting winter precipitation and temperature.

Model	Precipitation	Temp
ECHAM4.5	8	32
CCM3v6	9	6
COLA	11	8
ECPC	0	2
GFDL	1	3
NCEP	3	3
NSIPP-1	6	0
MM-1	4	8
MM-2	12	20
MM-3	9	18
MM-4	27	31
MM-P	3	10
MM-OS	4	15

total, six different multimodel schemes are considered with their performance being compared with individual models based on various verification measures such as ranked probability skill score, reliability and resolution scores, and Brier score. The improvements resulting from (reduction in $\overline{\text{RPS}}$) from multimodel combinations over individual models are also tested through a rigorous nonparametric hypothesis based on resampling.

The study clearly shows that the proposed multimodel combination algorithm perform better, in terms of improving the $\overline{\text{RPSS}}$, than individual models and also multimodel combinations based on pooling and long-term skill. Further, the proposed multimodel combination methodology also improves the reliability and resolution of tercile probabilities resulting in reduced Brier scores. The improved reliability of multimodel predictions primarily arises from reducing the overconfidence of individual model predictions, which in turn results from the reduced number of false alarms and missed targets in the categorical forecasts. Analysis of estimated weights also shows that the proposed methodology assigns higher (lower) weights for GCMs and lower (higher) weights for climatology during anomalous (neutral) ENSO conditions at the grid points. These analyses show that combining multiple models contingent on the dominant predictor state is an attractive strategy for improving the skill of multimodel forecasts.

Acknowledgments. This study was supported by the North Carolina Water Resources Research Institute. The writers also would like to thank the three anonymous reviewers whose valuable comments led to significant improvements in the manuscript. Useful discussions with Dr. Lisa Goddard of IRI were very helpful in improving the analysis presented in the manuscript.

REFERENCES

Bacmeister, J., P. J. Pegion, S. D. Schubert, and M. J. Suarez, 2000: Atlas of seasonal means simulated by the NSIPP 1 atmospheric GCM. NASA/TM-2000-104505, Vol. 17, 194 pp.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. Dewitt, and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.,* **84,** 1783–1796.

Bobko, P., 1995: *Correlation and Regression: Principles and Applications for Industrial/Organizational Psychology and Management*. McGraw-Hill, 283 pp.

Branković, Č., and T. N. Palmer, 2000: Seasonal skill and predictability of ECMWF PROVOST ensembles. *Quart. J. Roy. Meteor. Soc.,* **126,** 2035–2067.

Bröcker, J., and L. A. Smith, 2007: Scoring probabilistic forecasts: On the importance of being proper. *Wea. Forecasting,* **22,** 382–388.

Chowdhury, S., and A. Sharma, 2009: Long-range Niño-3.4 predictions using pairwise dynamic combinations of multiple models. *J. Climate,* **22,** 793–805.

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate,* **20,** 2810–2826.

Devineni, N., A. Sankarasubramanian, and S. Ghosh, 2008: Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations. *Water Resour. Res.,* **44,** W09404, doi:10.1029/2006WR005855.

Doblas-Reyes, F. J., M. Deque, and J. P. Piedelievre, 2000: Multimodel spread and probabilistic seasonal forecasts in PROVOST. *Quart. J. Roy. Meteor. Soc.,* **126,** 2069–2087.

——, R. Hagedorn, and R. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus,* **57A,** 234–252.

GFDL Global Atmospheric Model Development Team, 2005: The new GFDL global atmospehre and land model AM2–LM2: Evaluation with prescribed SST simulations. *J. Climate,* **17,** 4641–4673.

Goddard, L., and S. J. Mason, 2002: Sensitivity of seasonal climate forecasts to persisted SST anomalies. *Climate Dyn.,* **19,** 619–631.

——, A. G. Barnston, and S. J. Mason, 2003: Evaluation of the IRI's ''net assessment'' seasonal climate forecasts: 1997–2001. *Bull. Amer. Meteor. Soc.,* **84,** 1761–1781.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A,* **57,** 219–233, doi:10.1111/j.1600–0870.2005.00103.x.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.,* **14,** 382–401.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dyn.,* **31,** 647–664.

Kanamitsu, M., and K. C. Mo, 2003: Dynamical effect of land surface processes on summer precipitation over the southwestern United States. *J. Climate,* **16,** 496–509.

Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.,* **103,** 18 567–18 589.

Kiehl, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate,* **11,** 1131–1149.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999:

Improved weather and seasonal climate forecasts from a multimodel superensemble. *Science,* **286,** 1548–1550.

Luo, L., E. F. Wood, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.,* **112,** D10102, doi:10.1029/2006JD007655.

Mason, S. J., and G. M. Mimmack, 2002: Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate,* **15,** 8–29.

New, M., M. Hulme, and P. D. Jones, 1999: Representing twentieth-century space–time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *J. Climate,* **12,** 829–856.

——, ——, and ——, 2000: Representing twentieth-century space–time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate. *J. Climate,* **13,** 2217–2238.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multimodel ensemble integrations. *Quart. J. Roy. Meteor. Soc.,* **126,** 2013–2033.

——, and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Quan, X., M. Hoerling, J. Whitaker, G. Bates, and T. Xu, 2006: Diagnosing sources of U.S. seasonal forecast skill. *J. Climate,* **19,** 3279–3293.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.,* **130,** 1792–1811.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.,* **132,** 2732–2744.

Rodwell, M. J., and F. J. Doblas-Reyes, 2006: Medium-range, monthly, and seasonal prediction for Europe and the use of forecast information. *J. Climate,* **19,** 6025–6046.

Roeckner, E., and Coauthors, 1996: The atmospheric general circulation model ECHAM4: Model description and simulation of present-day climate. Max-Planck-Institut für Meteorologie Rep. 218, Hamburg, Germany, 90 pp.

Saha, S., S. Nadiga, C. Thiaw, and J. Wang, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–3517.

Sankarasubramanian, A., U. Lall, and S. Espuneva, 2008: Role of retrospective forecasts of GCM forced with persisted SST anomalies in operational streamflow forecasts development. *J. Hydrometeor.,* **9,** 212–227.

Schneider, E. K., 2002: Understanding differences between the equatorial Pacific as simulated by two coupled GCMs. *J. Climate,* **15,** 449–469.

Shukla, J., and Coauthors, 2000: Dynamical seasonal prediction. *Bull. Amer. Meteor. Soc.,* **81,** 2593–2606.

Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Malmaseda, 2005: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus A,* **57,** 253–264.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Stat. Soc.,* **36A,** 111–147.

Tippett, M. K., and A. G. Barnston, 2008: Skill of multimodel ENSO probability forecasts. *Mon. Wea. Rev.,* **136,** 3933–3946.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **135,** 118–124.

——, ——, and ——, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.,* **134,** 241–260.

Weisheimer, A., and Coauthors, 2009: ENSEMBLES: A new multimodel ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.,* **36,** L21711, doi:10.1029/2009GL040896.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Science*. Academic Press, 467 pp.

——, 1997: Resamping hypothesis tests for auto-correlated fields. *J. Climate,* **10,** 66–82.