Combination of Multimodel Probabilistic Forecasts Using an Optimal Weighting System

LI-CHUAN GWEN CHEN

Earth System Science Interdisciplinary Center/Cooperative Institute for Climate and Satellites, University of Maryland, College Park, and NOAA/NWS/NCEP Climate Prediction Center, College Park, Maryland

HUUG VAN DEN DOOL

NOAA/NWS/NCEP Climate Prediction Center, College Park, and Innovim, LLC, Greenbelt, Maryland

(Manuscript received 13 June 2017, in final form 5 September 2017)

ABSTRACT

In this study, an optimal weighting system is developed that combines multiple seasonal probabilistic forecasts in the North American Multimodel Ensemble (NMME). The system is applied to predict temperature and precipitation over the North American continent, and the analysis is conducted using the 1982–2010 hindcasts from eight NMME models, including the CFSv2, CanCM3, CanCM4, GFDL CM2.1, Forecast-Oriented Low Ocean Resolution (FLOR), GEOS5, CCSM4, and CESM models, with weights determined by minimizing the Brier score using ridge regression. Strategies to improve the performance of ridge regression are explored, such as eliminating a priori models with negative skill and increasing the effective sample size by pooling information from neighboring grids. A set of constraints is put in place to confine the weights within a reasonable range or restrict the weights from departing wildly from equal weights. So when the predictor–predictand relationship is weak, the multimodel ensemble forecast returns to an equal-weight combination. The new weighting system improves the predictive skill from the baseline, equally weighted forecasts. All models contribute to the weighted forecasts differently based upon location and forecast start and lead times. The amount of improvement varies across space and corresponds to the average model elimination percentage. The areas with higher elimination rates tend to show larger improvement in cross-validated verification scores. Some local improvements can be as large as 0.6 in temporal probability anomaly correlation (TPAC). On average, the results are about 0.02–0.05 in TPAC for temperature probabilistic forecasts and 0.03–0.05 for precipitation probabilistic forecasts over North America. The skill improvement is generally greater for precipitation probabilistic forecasts than for temperature probabilistic forecasts.

1. Introduction

Multimodel ensembles have been proven to provide better climate prediction skill (on average) than any constituent single models (Hagedorn et al. 2005; Weisheimer et al. 2009; Kirtman et al. 2014; Becker et al. 2014). A topic remaining in debate is whether a combination of multiple model predictions can significantly improve forecast skill over a simple multimodel average. Strategies to combine forecasts from multiple models have been studied by many (e.g., Krishnamurti et al. 1999, 2000; Kharin and Zwiers 2002; Peng et al. 2002; DelSole 2007; Weigel et al. 2008; Peña and Van den Dool 2008; Wanders and Wood 2016) with diverse results. Some research has found that the predictive skill of weighted multimodel ensembles is significantly higher than that of individual models and the ensemble mean (e.g., Krishnamurti et al. 1999, 2000; Wanders and Wood 2016). Other studies concluded that skill scores based on optimal combinations of multimodel predictions are only marginally better than those from equally weighted forecasts (e.g., Kharin and Zwiers 2002; Peng et al. 2002; DelSole et al. 2013). Several factors contribute to the discrepancies, including different methods, datasets, and lengths of data used in these studies.

Among the various methods, regression-based weighting schemes are the most common. Regression-based analyses suffer from two main obstacles: collinearity and overfitting. Collinearity occurs when two or more model forecasts are highly correlated, meaning that one can be predicted from the others with a degree of accuracy. Collinearity is often encountered in multimodel combination

Corresponding author: L. Gwen Chen, lichuan.chen@noaa.gov

DOI: 10.1175/WAF-D-17-0074.1

studies and exacerbated when more models are added into the forecasts. One approach to reducing collinearity is through regularization methods, particularly ridging (Phillips 1962; Tikhonov 1963; Tikhonov and Arsenin 1977), as seen in the work by Peng et al. (2002) and Peña and Van den Dool (2008). Another way is to decrease the dimension of the covariance matrix of forecasts (e.g., Yun et al. 2003). Several studies (e.g., Doblas-Reyes et al. 2005; DelSole et al. 2013) suggested that removing models with no skill or negative weights prior to combining them can effectively reduce the collinearity and increase the skill scores. In this study, we propose an iterative elimination approach in conjunction with ridge regression to obtain a subset of models (i.e., varying number of models) with optimal weights for consolidation forecasts.

Another major challenge often encountered in optimization procedures is overfitting. Overfitting occurs when the length of the training dataset is too short compared to the number of input models (Peña and Van den Dool 2008). To combat this issue, Van den Dool and Rukhovets (1994) and Peng et al. (2002) used information from all grid points in the region of analysis to estimate regression parameters, thereby effectively increasing the training dataset. Wanders and Wood (2016) showed significant improvements in the subseasonal precipitation P and temperature T forecast skill of the weighted multimodel ensemble when the forecasts are spatially upscaled to a large region. However, the 22 regions (Giorgi and Francisco 2000) used in their analysis are so large they would only provide a small number of forecasts over the United States. This feature is not desirable in regional operational forecasts that need a depiction of the forecast variation across space. In this study, we focus on gridpoint-wise analysis to meet this need. To reduce overfitting, DelSole (2007) and Peña and Van den Dool (2008) suggested that borrowing strength from training data at neighboring grids can stabilize parameter estimates and improve the performance of regression schemes relative to the multimodel mean. In this study, we propose to accumulate statistics from the nearest eight grid points surrounding the target grid point, namely from a 3×3 spatial mosaic. By doing so, we not only increase the sample size for estimating the regression coefficients, but also smooth the sample statistics as the spatial mosaic moves across space.

The aim of this research is to develop a new weighting system with conservative strategies to address the above two difficulties in regression-based, gridpoint-wise analyses. Specific objectives include 1) the development of an automatic system that optimally selects and combines a number of probabilistic forecasts from a fixed set of available models, 2) incorporation of new strategies and constraints to objectively determine weights for consolidation forecasts, and 3) evaluation of the performance of the new system and weighted multimodel ensemble with a strict three-year-out strategy to ensure a trustworthy skill assessment. A novelty of this research is that we directly work with forecasts expressed as probabilities in a three-class system (above, near, and below normal), in contrast to the ensemble mean forecasts in physical units used in most studies. The investigation is conducted using the 1982–2010 hindcasts from eight models in the North American Multimodel Ensemble (NMME), and the study area is focused on North America.

In the following, section 2 introduces the NMME seasonal forecast data and verification fields. Section 3 describes the weighting methods and system design for deriving the consolidation forecasts. Section 4 explains the cross-validation procedure and performance metrics. Section 5 presents the system evaluation and discusses the results. Section 6 gives forecast examples from the new system. Section 7 summarizes the major findings and conclusions from the investigation.

2. Data

a. NMME seasonal forecast data

NMME is a multimodel forecasting system consisting of coupled climate models from U.S. modeling centers (including NCEP, GFDL, NASA, and NCAR) and the Canadian Meteorological Centre (CMC), aimed at improving subseasonal-to-seasonal prediction capability (Kirtman et al. 2014). The NMME seasonal forecast system has a number of models providing various periods of hindcasts from 1980 to 2012 with monthly initialization. In this study, we select eight models in NMME: CFSv2 (Saha et al. 2006, 2014), CanCM3 (Merryfield et al. 2013), CanCM4 (Merryfield et al. 2013), CM2.1 (Delworth et al. 2006; Gnanadesikan et al. 2006), the Forecast-Oriented Low Ocean Resolution (FLOR) model (Vecchi et al. 2014; Jia et al. 2015), GEOS5 (Vernieres et al. 2012), CCSM4 (Gent et al. 2011; Danabasoglu et al. 2012), and CESM (Hurrell et al. 2013). These models have a common period of hindcasts from 1982 to 2010 and are mapped onto a common grid system of $1° \times 1°$ resolution covering the globe. The number of ensemble members for each model ranges from 10 to 24. Detailed information about the NMME project and its dataset is available on NOAA's Climate Test Bed website (http://www.nws.noaa.gov/ost/CTB/nmme.htm).

b. Precipitation verification data

The verifying precipitation observations used in this study are taken from the precipitation reconstruction (PREC) global land analysis (Chen et al. 2002) from

January 1982 to June 2011. PREC is a gridded monthly precipitation product that interpolated gauge observations from over 17 000 stations collected in the Global Historical Climatology Network (GHCN) and the Climate Anomaly Monitoring System (CAMS). The PREC product is remapped onto the 1° × 1° NMME grid system from its original 0.5° × 0.5° resolution using bilinear interpolation.

c. Temperature verification data

The GHCN_CAMS gridded 2-m temperature data from January 1982 to June 2011 are used to validate temperature forecasts up to 6-month lead time. This dataset combines station observations from the GHCN and CAMS and employs the anomaly interpolation approach with spatially and temporally varying temperature lapse rates derived from the reanalysis for topographic adjustment (Fan and Van den Dool 2008). Similar to the PREC dataset, the GHCN_CAMS data are also remapped onto the 1° × 1° NMME grid system, in order to be consistent with NMME seasonal forecast data.

3. Weighting methods and system design

a. Baseline, equally weighted probabilistic forecasts

Seasonal T and P probabilistic forecasts are computed using the simple count method for a three-class forecast system (Van den Dool 2007; Becker and Van den Dool 2016). For each model, seasonal T and P forecasts at a given start and lead time and location/grid point are classified into three categories (above, near, and below normal) based on the terciles derived from the hindcasts of all members, excluding the target forecast year plus two randomly selected years. Here, we implement a more stringent cross-validation approach than the traditional leave-one-out procedure to reduce the effect of degeneracy in regression-based forecasts when predictor–predictand relationships are weak (Barnston and Van den Dool 1993). For T forecasts, the tercile thresholds are set as mean ± 0.431 × standard deviation by assuming a Gaussian distribution. For P forecasts, the tercile thresholds are the 33th and 67th percentiles determined by fitting a gamma distribution to the hindcasts. The classification applies to each individual member forecast, and the numbers of ensemble members that fell into the three categories are counted for target forecast years from 1982 to 2010. The probability for each forecast year and category is then calculated by dividing the number of counts in the category by the total number of ensemble members for each model.

The equally weighted NMME T or P forecast for a given start and lead time, location, and category is computed by

$$Y(t) = \sum_{i=1}^{m} \alpha_i X_i(t), \tag{1}$$

where $Y(t)$ is the weighted NMME forecast for target year t, α_i is the weight for model i, $X_i(t)$ is the forecast probability for model i at time/forecast year t, and m is the total number of models ($m = 8$ in this study). For equally weighted forecasts, $\alpha_i = 1/m = 1/8$. Please note that this definition of equally weighted NMME probabilistic forecasts (i.e., one model one vote) is different from the one in Becker and Van den Dool (2016) used for real-time NMME forecasts, in which each member forecast is weighted equally (i.e., one member one vote).

b. Ridge regression

When α_i is not assumed to be a constant across models, Eq. (1) can be stated as a consolidation forecast and generalized as a linear combination of m participating predictions, each multiplied by a correspondent weight. With a set of verification data, the best forecast minimizes the average of an error metric (e.g., mean squared error) over a series of training data. In probabilistic forecasts, the quantity to be minimized is the mean squared error in probability terms, known as the Brier score (BS; Brier 1950):

$$\mathrm{BS} = \frac{1}{n}\sum_{t=1}^{n}[Y(t) - O(t)]^2 = \frac{1}{n}\sum_{t=1}^{n}\left\{\left[\sum_{i=1}^{m}\alpha_i X_i(t)\right] - O(t)\right\}^2, \tag{2}$$

where $O(t)$ is the observed probability (either 0 or 1) for target year t at a given start and lead time, location, and category, and n is the total number of target forecast years ($n = 29$ in this study).

Such a problem, minimizing Eq. (2), can be solved classically as

$$\mathbf{A}\boldsymbol{\alpha} = \mathbf{b}, \tag{3}$$

where $\mathbf{A}$ is the $m \times m$ covariance matrix with elements $a_{ij} = \sum_{t=1}^{n} X_i(t)X_j(t)$, $\mathbf{b}$ is the covariance vector with m elements $b_i = \sum_{t=1}^{n} X_i(t)O(t)$, and vector $\boldsymbol{\alpha}$ has m elements $(\alpha_1, \alpha_2, \ldots, \alpha_m)$. Both $\mathbf{A}$ and $\mathbf{b}$ can be computed from hindcasts and thus $\boldsymbol{\alpha}$ can be solved for, in principle.

When $\mathbf{A}$ in Eq. (3) is ill-conditioned or nearly so, the solution of weights becomes unstable, and the weights can be unrealistically large, both positive and negative (Peña and Van den Dool 2008). Weights could also vary greatly from grid point to grid point. One way to reduce

this problem is through ridging (Phillips 1962; Tikhonov 1963; Tikhonov and Arsenin 1977), by repeatedly adding a small positive constant to the main diagonal of **A** until a constraint is satisfied, namely that $\sum_{i=1}^{m}\alpha_i^2$ is "small" (Peña and Van den Dool 2008). In this study, we employ the penalty term proposed by DelSole (2007) to constrain the weights deviating far from $1/m$ within a Bayesian framework. So when the ridging amount increases, model weights gradually converge to equal weights. Instead of the sum of the squared weights being small, we impose a new constraint to restrict the sum of the weights within the range of 0.9–1.05. So the weighted probabilistic forecast derived from ridge regression is compatible with the equally weighted and skill-based weighted forecasts (described in the next section), both with the sum of the weights equal to one.

c. Skill-based weights

In areas where T or P forecast skill is low and collinearity is high, ridge regression does not always yield improvements in weighted forecasts. An alternative solution of weights can be calculated from Eq. (3) based solely on the diagonal information:

$$\alpha_i = \frac{b_i}{fa_{ii}}, \tag{4}$$

where $f = \sum_{i=1}^{m} b_i/a_{ii}$ is a factor that makes $\sum_{i=1}^{m}\alpha_i = 1$. Here, α_i is proportional to the probability anomaly correlation (PAC; Van den Dool et al. 2017) and reflects model forecast skill in probability terms. This set of weights can be viewed as if each model is regressed individually and independently against the observations (Peña and Van den Dool 2008) and hereafter is referred to as the skill-based weights.

d. System design

To further reduce the collinearity and overfitting that often hinder regression-based forecast consolidations, we develop a new weighting system to incorporate the proposed new strategies and constraints. Figure 1 shows the schematic of the system design. Without change of notation in the equations, we directly work with the probability anomaly in the new system. The probability anomaly is defined as the difference between the model (or observed) probability and the climatology value (i.e., 0.333), and it is calculated for NMME probabilistic forecasts and verification fields before being inputted into the system. For each target forecast year at a given start and lead time, location, and forecast category, we first compute the covariance matrix **A** and vector **b** from all eight participating models using training data (see the cross-validation strategy in section 4a) with a 3×3 spatial mosaic centered at the target grid point. If b_i is less than 0.01, indicating that the associated input model has no skill, model i is removed from Eqs. (1)–(3). The remaining models are used to calculate the skill-based weights (all positive) and solve for the system of linear equations by ridge regression. If the ridging-based weight of a model α_i is negative, suggesting that the associated model does not have a "positive" contribution to the weighted forecast (Peña and Van den Dool 2008), that model is further removed from Eqs. (1)–(3), and ridge regression is repeated until all ridging-based weights are positive and have a sum close to one. This step may take several iterations. Because seasonal T and P forecast skill is low in some areas, there are places where all eight models are eliminated. When that happens, there is no solution for the skill-based and/or ridging-based weights, and the system automatically falls back to equal weights.

After the weights are determined from the training data, we compute the target year's weighted forecasts from the available sets of weights (equal, skill-based, and/or ridging-based weights). The above process is repeated for all target forecast years and then the cross-validated BS is calculated for each available weighting method. Although the ridging-based weights are obtained through an optimization method over a series of training data, they are not guaranteed to work the best on independent data compared to other weighting methods. Therefore, the weighting method that gives the smallest cross-validated BS is chosen for that grid point, as intended by Eq. (2), and its correspondent weighted forecasts are denoted as the "consolidated" forecasts at that grid point. This step ensures that the selected set of weights generates the best independent forecasts among the available weighting methods. If the regression-based (skill based and ridging based) weights do not present an advantage over equal weights, the system falls back to the baseline, and hence there is no loss in skill with respect to equally weighted forecasts.

4. Performance assessment

a. Cross validation

To assess the system performance on independent data, we use a leave-three-out cross-validation procedure to reduce the effect of degeneracy and avoid artificial skill (Barnston and Van den Dool 1993). Of the three years excluded, one is the target forecast year and two are randomly chosen among 1982–2010 without repetition. This procedure is applied to the computation of both the tercile thresholds of the count method and the covariance matrix **A** and vector **b** in Eq. (3) for determining regression-based weights, so the target forecast year is not

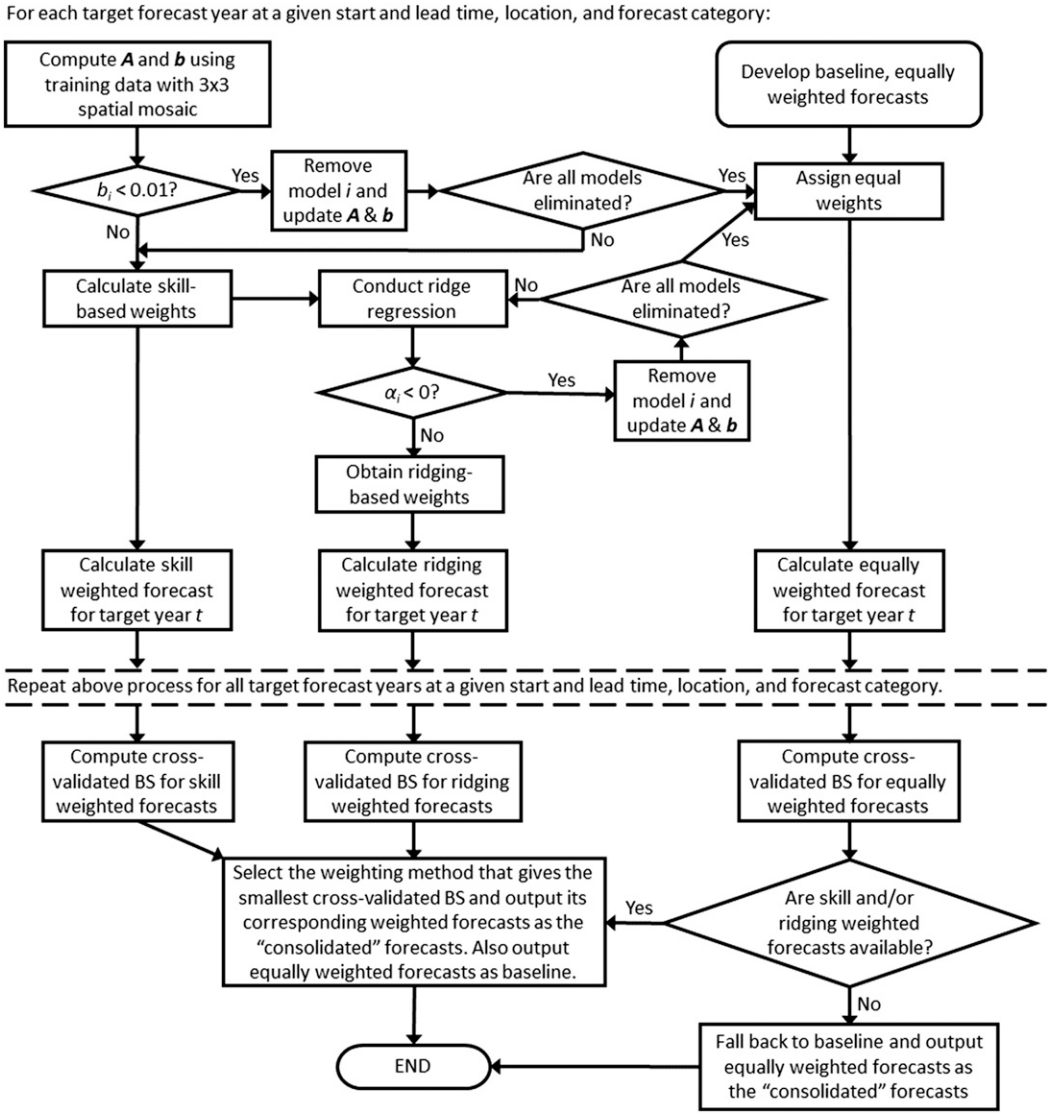

FIG. 1. Schematic of the system design.

involved in any of the training analyses. More specifically, cross validation only applies to the definition of tercile boundaries for the equal-weight average. For skill-based weights and full ridge regression, cross validation also affects **A** and **b**.

b. Probability anomaly correlation

In addition to BS, we employ the PAC as a performance metric to evaluate forecast skill. This metric was first introduced in Chen et al. (2017) for probability composite validation and further developed by Van den Dool et al. (2017) for probabilistic forecast verification and calibration. The PAC, analogous to the anomaly correlation, quantifies the strength of the linear association between the model probability anomaly and the observed probability anomaly. In this study, we use both temporal and spatial PACs for skill assessment. The cross-validated temporal PAC (TPAC) at a given start and lead time, location, and forecast category is defined as

$$\mathrm{TPAC} = \frac{\sum_{t=1}^{n} Y'(t) \times O'(t)}{\sqrt{\sum_{t=1}^{n} Y'(t)^2 \times \sum_{t=1}^{n} O'(t)^2}}, \tag{5}$$

where $Y'(t)$ is the consolidated probability anomaly for target year t and $O'(t)$ is the observed probability anomaly for target year t.

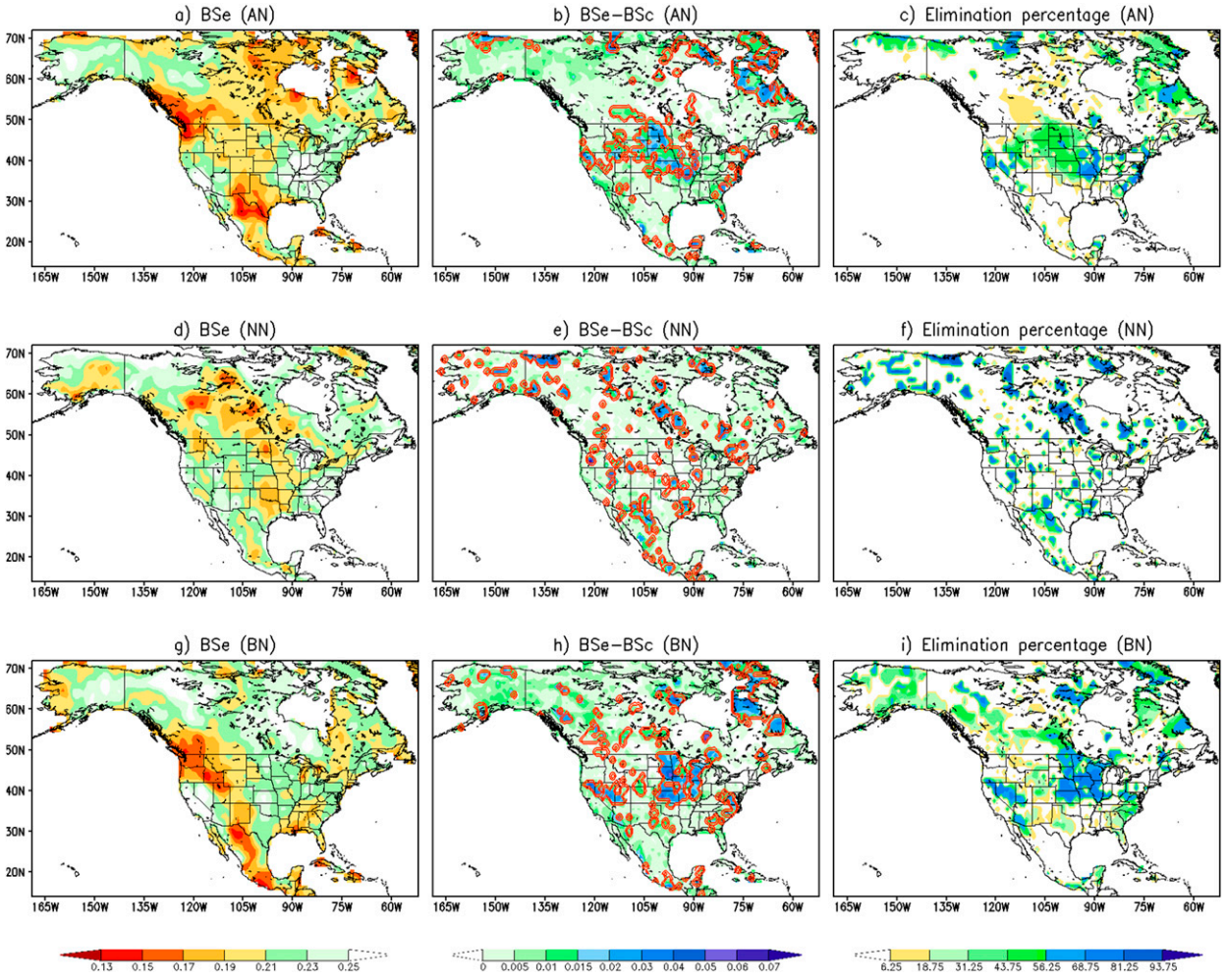

FIG. 2. For FMA T probabilistic forecasts initialized on 1 Jan: (left) cross-validated BS of equally weighted forecasts, (center) difference in BS between the consolidated and equally weighted forecasts, and (right) average model elimination percentage of consolidated forecasts for the (a)–(c) AN, (d)–(f) NN, and (g)–(i) BN categories. In (b), (e), and (h), the areas within the red contours are tested to have statistically significant differences at the 5% level.

Similar to the pattern correlation often used for comparing forecast and observation maps in deterministic forecasts, we can calculate a spatial PAC (SPAC) between the consolidated and observed probability anomaly maps for a target forecast year at a given start and lead time and forecast category by

$$\mathrm{SPAC} = \frac{\sum_{s=1}^{q} w(s)[Y'(s) \times O'(s)]}{\sqrt{\sum_{s=1}^{q} w(s)Y'(s)^2 \times \sum_{s=1}^{q} w(s)O'(s)^2}}, \quad (6)$$

where $Y'(s)$ is the consolidated probability anomaly at grid s, $O'(s)$ is the observed probability anomaly at grid s, q is the total number of land grid points within the North American domain, and $w(s)$ is the areal weighting coefficient based on the latitude L of grid s:

$$w(s) = \cos[L(s)]. \quad (7)$$

5. Results and discussion

We start the assessment by examining the predictive skill of the baseline forecasts. Figure 2 shows an example of the cross-validated BS of equally weighted forecasts (left column) for February–April (FMA) T probabilistic forecasts initialized on 1 January. The top, middle, and bottom rows in Fig. 2 are for the above-, near-, and below-normal (AN, NN, and BN) categories, respectively. The spatial variations of the predictive skill can be clearly seen in the plots. For the AN category,

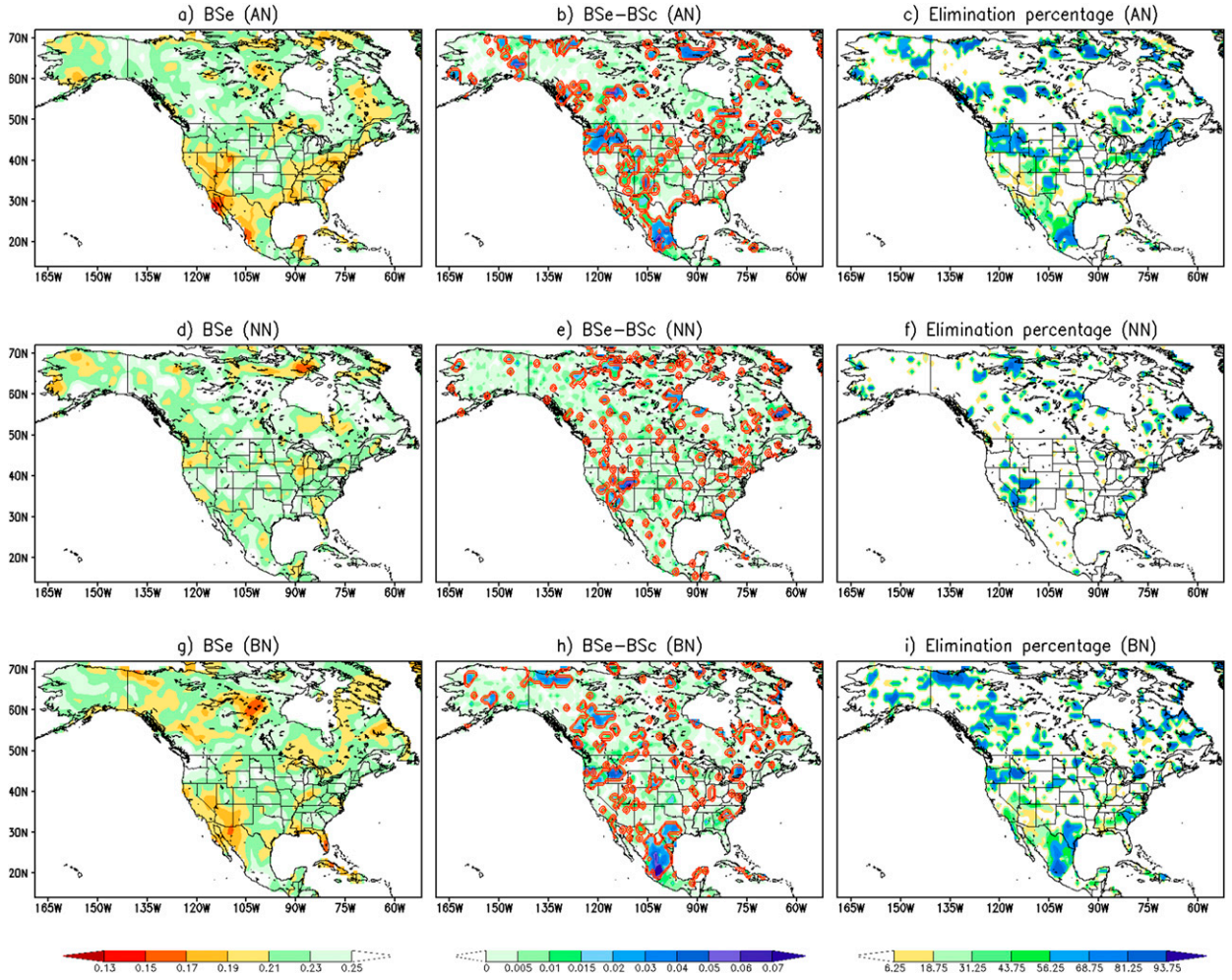

FIG. 3. As in Fig. 2, but for *P* probabilistic forecasts.

equally weighted *T* forecasts are more accurate (low BS) over the Pacific Northwest and along the Texas–Mexico boundary. Forecast accuracy is also high over the Pacific Northwest and northern Mexico for the BN category. Compared to the AN and BN categories, the forecast fidelity of the NN category is much lower. This is consistent with the findings of many studies (e.g., Van den Dool and Toth 1991) that the skill of categorical forecasts has a tendency to be low, if not absent or negative, in the NN category. The center column in Fig. 2 presents the differences in BS between the consolidated and equally weighted forecasts for the three categories. Colors (greens and blues) indicate improvements in *T* forecast skill with weighting schemes. Because of the model elimination and fall back, there is no loss (with respect to equal weights) when the model skill is poor. The areas within the red contours are tested to have statistically significant differences at the 5% level using the method described by DelSole et al. (2013). For the AN and BN categories, there are large improvements in skill with the consolidated *T* forecasts in the central United States. These areas coincide with areas that have large average model elimination percentages (shown in the right column in Fig. 2), suggesting that the strategy of removing a priori models with negative skill/weights is the dominant factor contributing to the improved skill of the consolidated forecasts. The variation of the weights among models that passed is a lesser factor. Note that because of the elimination and fall back, the number of models used in the consolidated forecast of a given category is different for each forecast year and location, varying from one to eight.

Figure 3 shows the same plots but for *P* probabilistic forecasts. Compared to the equally weighted *T* forecasts in Figs. 2a, 2d, and 2g, we can see that all three forecast categories have lower accuracy when making *P* probabilistic forecasts. Different from the *T* forecasts, equally weighted *P* forecasts perform better over the southwestern United States, and there are large

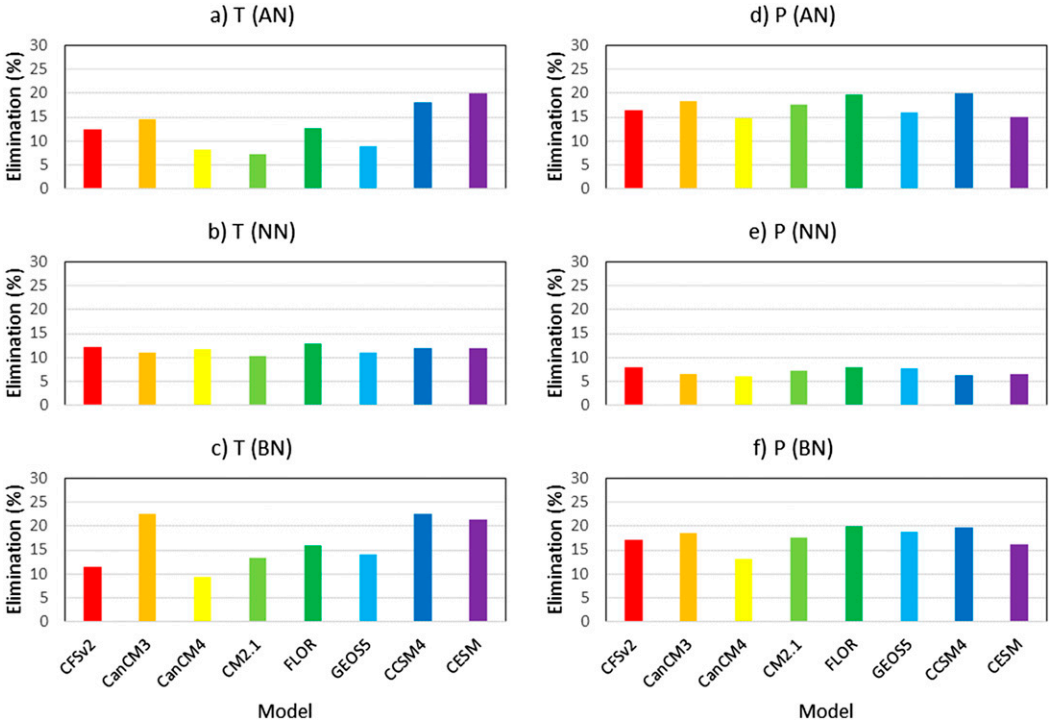

FIG. 4. Elimination percentages of a specific model for FMA probabilistic forecasts initialized on 1 Jan for the (left) temperature and (right) precipitation forecasts for the (a),(d) AN, (b),(e) NN, and (c),(f) BN categories.

improvements seen in skill over the Pacific Northwest and central Mexico with the consolidated forecasts. Because rainfall variability is high across space, areas of improvement (tested to be statistically significant) are more scattered throughout the North American continent. Still, there is strong correspondence between the areas with large skill improvements (Fig. 3, center) and with large model elimination rates (Fig. 3, right), reinforcing that model elimination is a key factor for skill improvements regardless of forecast variable.

In Figs. 2 and 3, we calculate the average model elimination percentage at a given location by dividing the number of eliminated models by the total number of models of all 29 forecast years. This percentage gives a sense of the average number of models that were removed in the consolidated forecasts without specifying the model names. Figure 4 provides the elimination percentage of a specific model for the same FMA *T* and *P* forecasts initialized on 1 January. For a given model in a given forecast category, this elimination percentage reflects how often the specific model was removed from the consolidated forecasts by counting from all land grids within the North American continent and all 29 forecast years. For both *T* (Figs. 4a–c) and *P* (Figs. 4d–f), all models have been removed from the consolidated forecasts some time at some points, suggesting that all models contribute to the consolidated forecasts differently based upon location. This feature can be clearly seen by inspecting the weight maps of all models for all forecast years (not shown due to the large number of figures). For the AN and BN categories, most models were removed from the consolidated forecasts more often for *P* probabilistic forecasts than for *T* probabilistic forecasts, because the predictive skill of *P* probabilistic forecasts generally is lower than that of *T* forecasts. In contrast, the elimination percentages of all models in the NN category are smaller for *P* probabilistic forecasts than for *T* probabilistic forecasts. This is caused by the predictive skill of *P* forecasts in the NN category being so poor, and a large portion of them fall back to equal weights. For a given variable and forecast category, the elimination rate of a specific model somewhat reflects inversely the model's predictive skill. For example, for *T* forecasts in the AN category (Fig. 4a), CCSM4 and CESM generally have lower skill when predicting FMA temperature from a January start compared to other models, and thus they have higher elimination rates.

We further examine the choice of the weighting method at each forecast location to identify any spatial coherence or variation of the preferred weighting method across the North American continent. We use the same FMA forecast example to illustrate the results

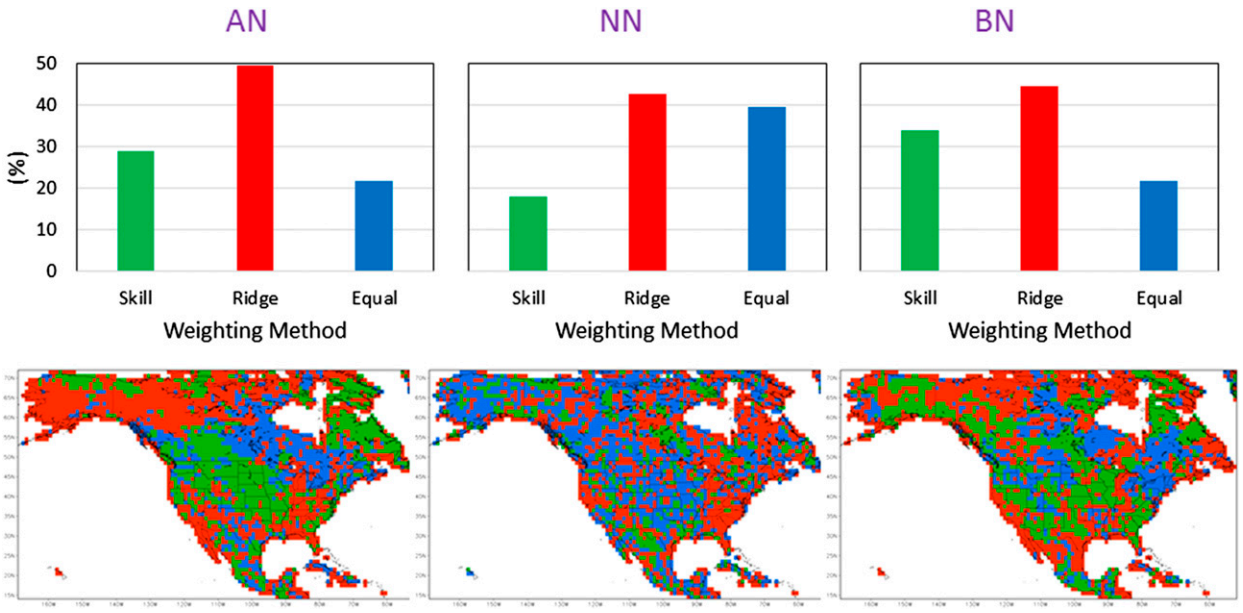

FIG. 5. Choice of weighting method for FMA *T* probabilistic forecasts initialized on 1 Jan: (left) AN, (center) NN, and (right) BN. (top) The fraction of the North American area selected for each weighting method, and (bottom) the selected weighting method at each location using the same colors.

in Figs. 5 (for *T*) and 6 (for *P*). The top row in Fig. 5 presents the fraction of North American land grids where each weighting method is favored, and the bottom row displays the map of the favored weighting methods. Here, favored means producing the lowest cross-validated BS during the test years. Left, center, and right columns are for the AN, NN, and BN categories, respectively. For all three forecast categories, ridge regression is the dominant weighting method over the North American continent. For the AN and BN categories, skill-based weighting has an advantage over equal weights. The maps of the weighting methods for all three forecast categories clearly show clusters of weighting method results in

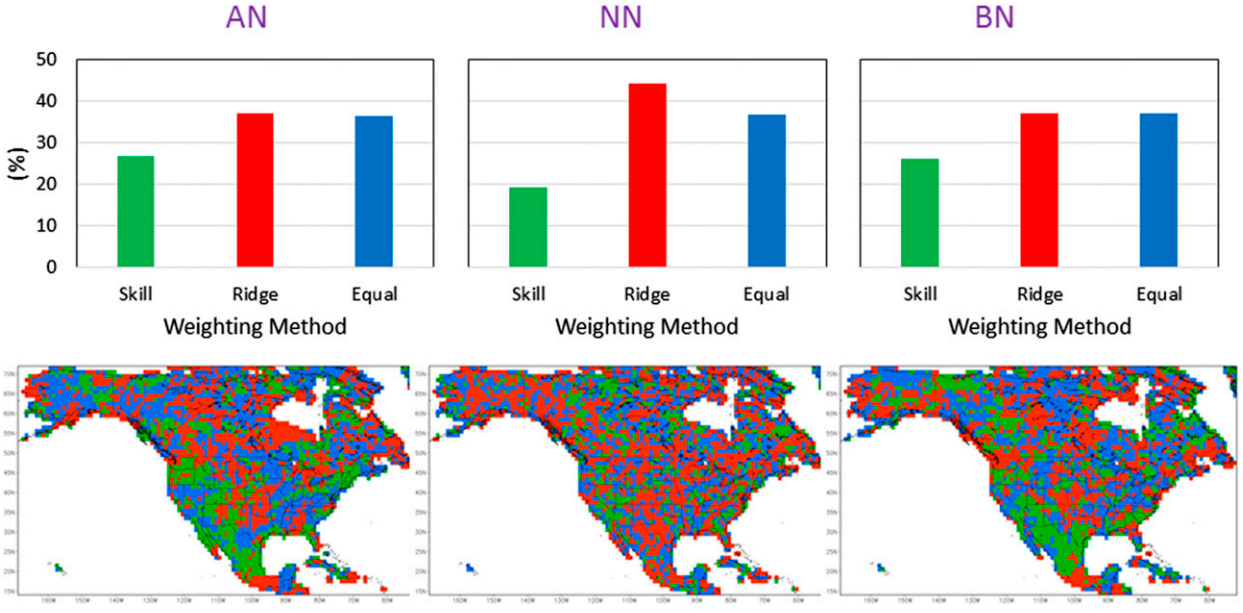

FIG. 6. As in Fig. 5, but for *P* probabilistic forecasts.

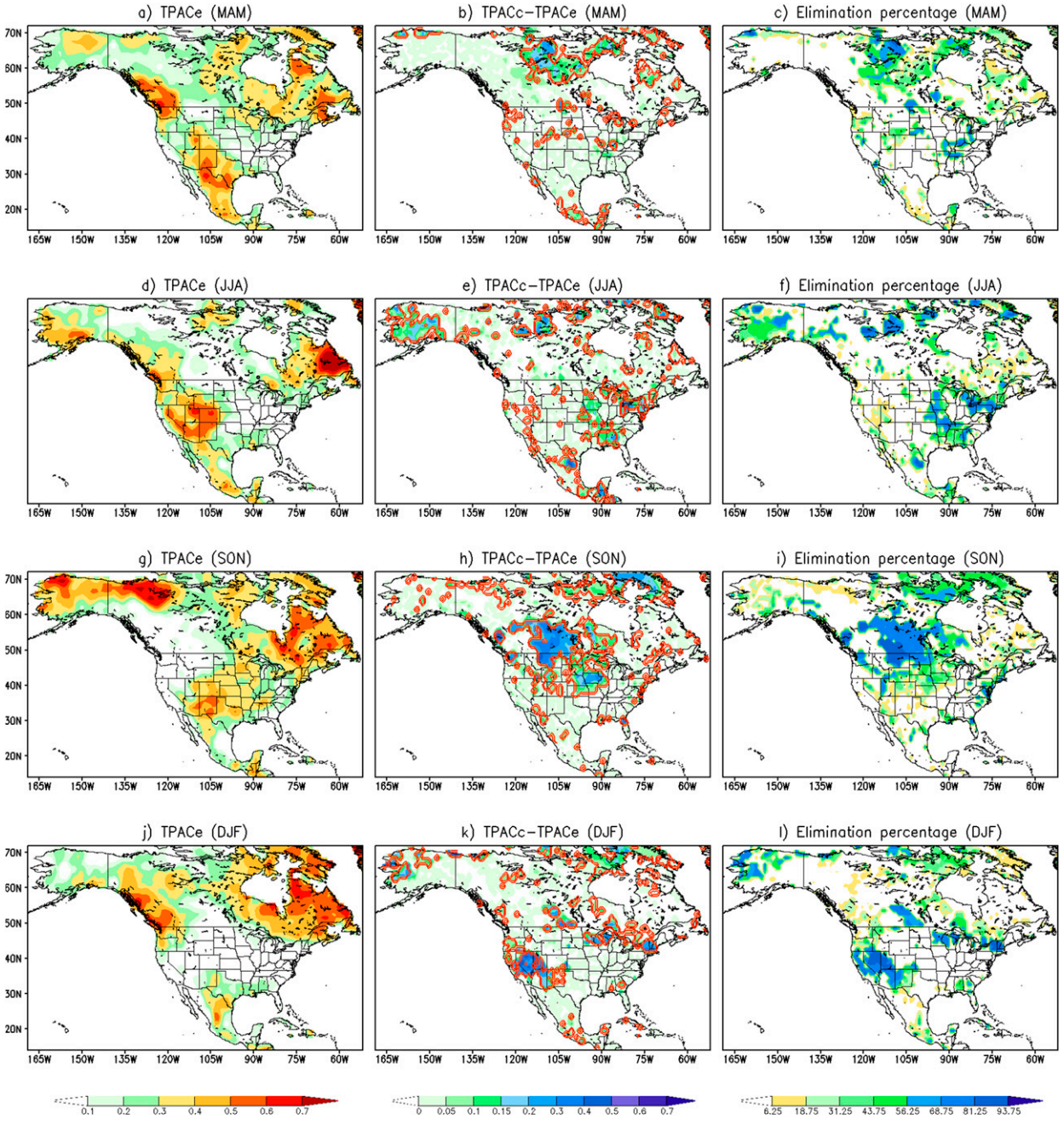

FIG. 7. Lead-1 *T* probabilistic forecasts for the AN category: (left) cross-validated TPAC of equally weighted forecasts, (center) difference in TPAC between the consolidated and equally weighted forecasts, and (right) average elimination percentage of consolidated forecasts for (a)–(c) MAM, (d)–(f) JJA, (g)–(i) SON, and (j)–(l) DJF. In (b), (e), (h), and (k), the areas within the red contours are tested to have statistically significant differences at the 5% level.

space, suggesting that there might be an association between the preferred weighting method and the model forecast skill. For the AN and BN categories, ridge regression is apparent along the West Coast and the south-central United States, and northern Canada, where skill-based weighting is visible across the central United States, southern Mexico, and northeastern Canada.

For the FMA *P* probabilistic forecasts initialized on 1 January (Fig. 6), ridge regression is still the dominant weighting method. In contrast with the *T* forecasts, more grid points in the North American continent fall back to equal weights. In fact, equal weighting is competitive with ridge regression in the AN and BN categories. Although only about 26% of the grid points chose skill-based

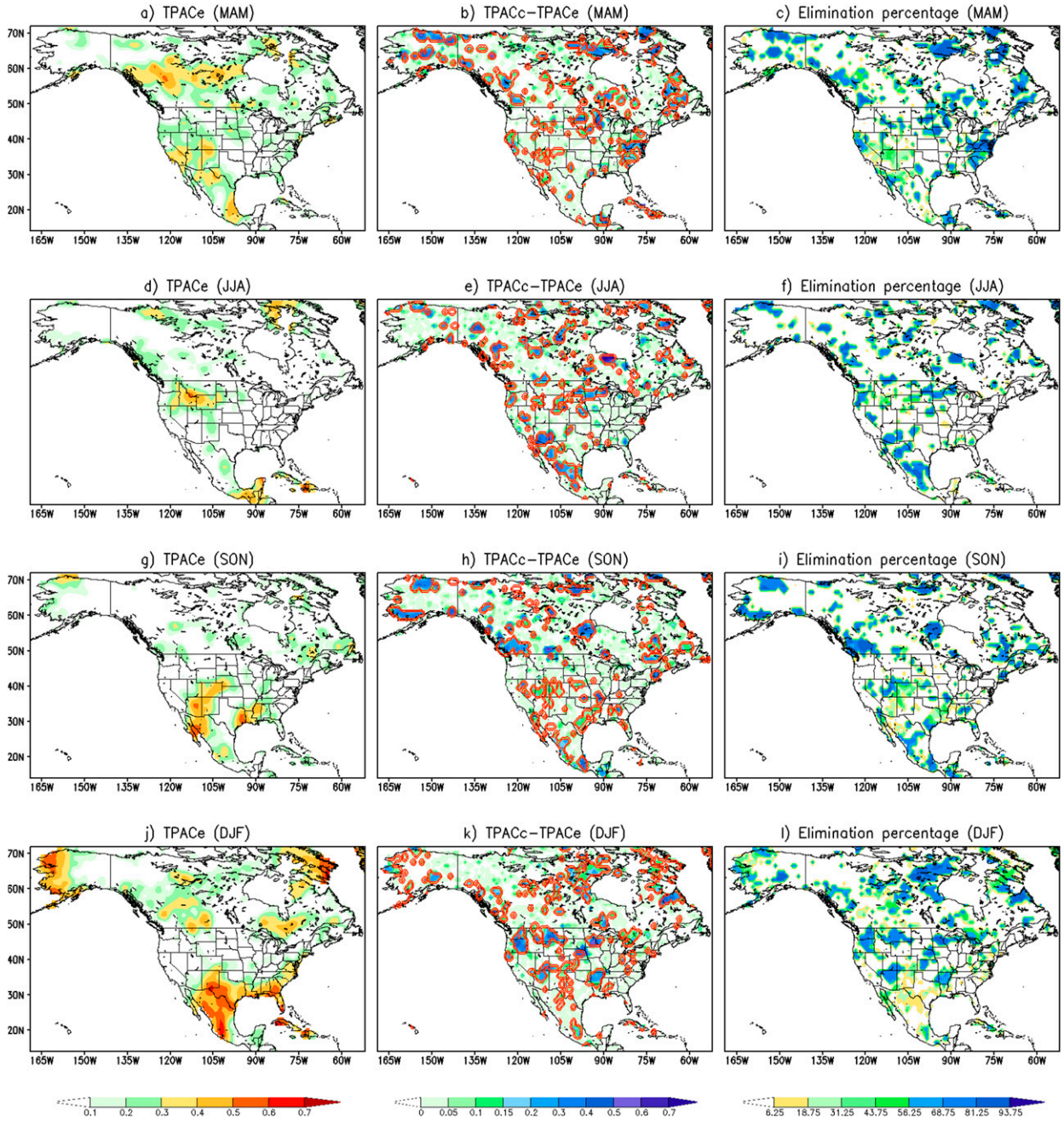

FIG. 8. As in Fig. 7, but for *P* probabilistic forecasts of the BN category.

weighting, most of these grid points are located over the western United States and northern Mexico. Ridge regression is more noticeably favored in the central United States and Canada. For the NN category, there is no clear pattern, and the weighting methods are randomly distributed throughout the North American continent.

To analyze seasonality, Fig. 7 shows the lead-1 *T* probabilistic forecasts in the AN category for four seasons: March–May (MAM; top row), June–August (JJA; second row), September–November (SON; third row), and December–February (DJF; bottom row). Unlike Fig. 2, the left column in Fig. 7 displays the TPAC of the equally weighted *T* forecasts. It can be seen that the predictive skill of the equally weighted forecasts varies not only across space but also with time. For the MAM season, the skill is higher over the Pacific Northwest and southwest monsoon region. For the JJA season, equally weighted forecasts perform better over the western United States.

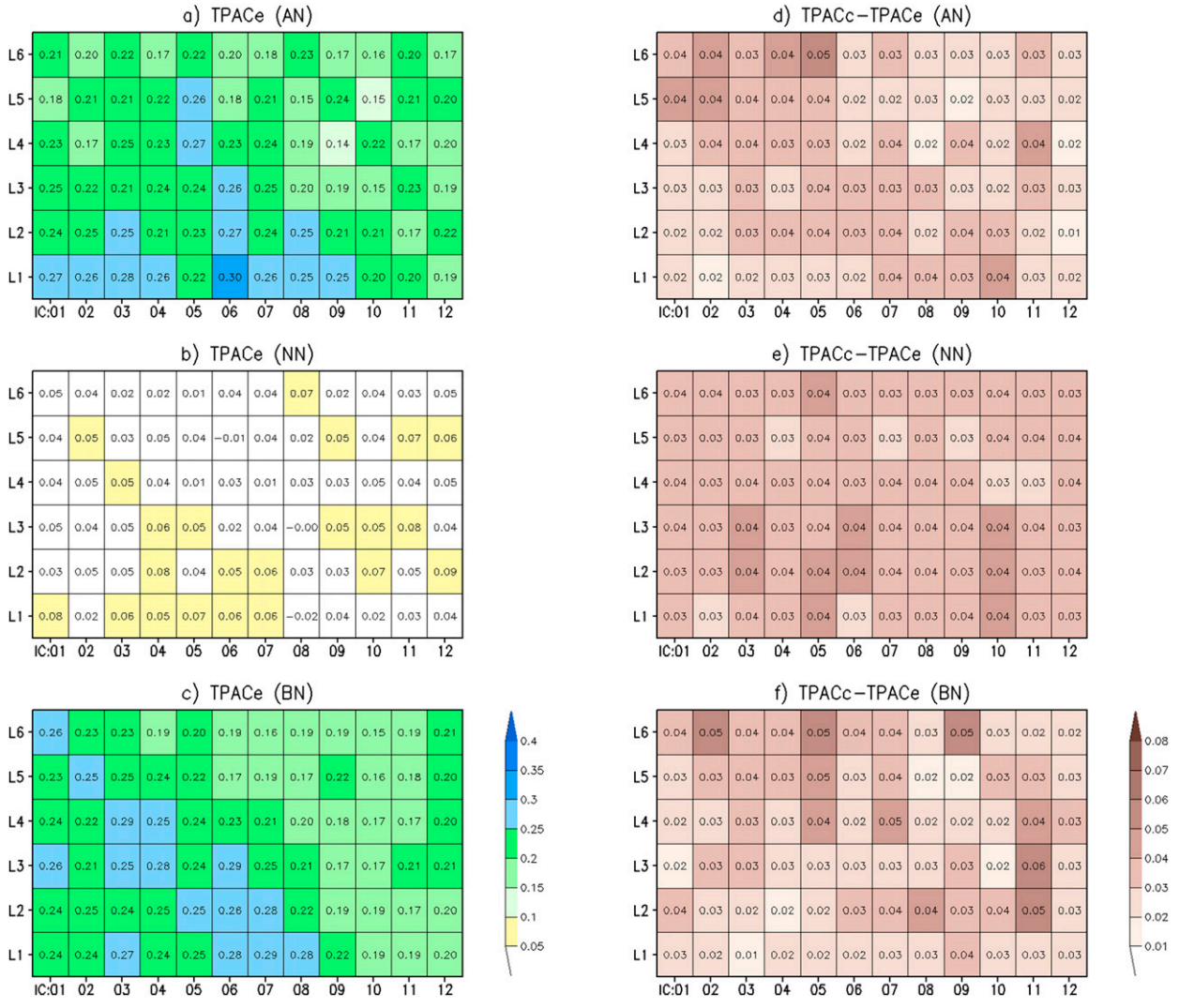

FIG. 9. Matrix charts of cross-validated TPAC averaged over North America for all start and lead months. The (left) equally weighted *T* probabilistic forecasts and (right) differences between the consolidated and equally weighted *T* probabilistic forecasts of the (a),(d) AN, (b),(e) NN, and (c),(f) BN categories.

In the fall (SON season), TPAC is above 0.3 for most areas in the central United States. During wintertime (DJF), equally weighted forecasts are not skillful for the majority of the United States; however, they have skill in predicting above-normal temperatures over eastern Canada, and their skill is high in this region for all seasons. The differences in TPAC between the consolidated and equally weighted forecasts are presented in the center column in Fig. 7. Skill improvements also vary with both space and time. Among the four seasons, SON benefits the most from the weighting schemes, with large areas of improvement seen over the upper Midwest and central Canada. The winter season (DJF) also has a large gain in skill over the western United States. Similar to the findings from Fig. 2, the areas of large skill improvement coincide with the areas of large average model elimination percentage (right column). This finding holds true regardless of the forecast season, category, or lead time.

Figure 8 shows the same plots as those in Fig. 7 but for *P* forecasts in the BN category. Similar to the observations from Fig. 3, the predictive skill of *P* probabilistic forecasts generally is low. Still, it depends on the season and location. For the MAM season, equally weighted *P* forecasts have some skill over the southwest United States. In summertime (JJA), there is not much skill in predicting rainfall over North America except for areas across Idaho, Wyoming, and southern Mexico. In the fall, some skill is observed over the southwest monsoon region and the Gulf Coast states. Rainfall prediction is most skillful in the wintertime

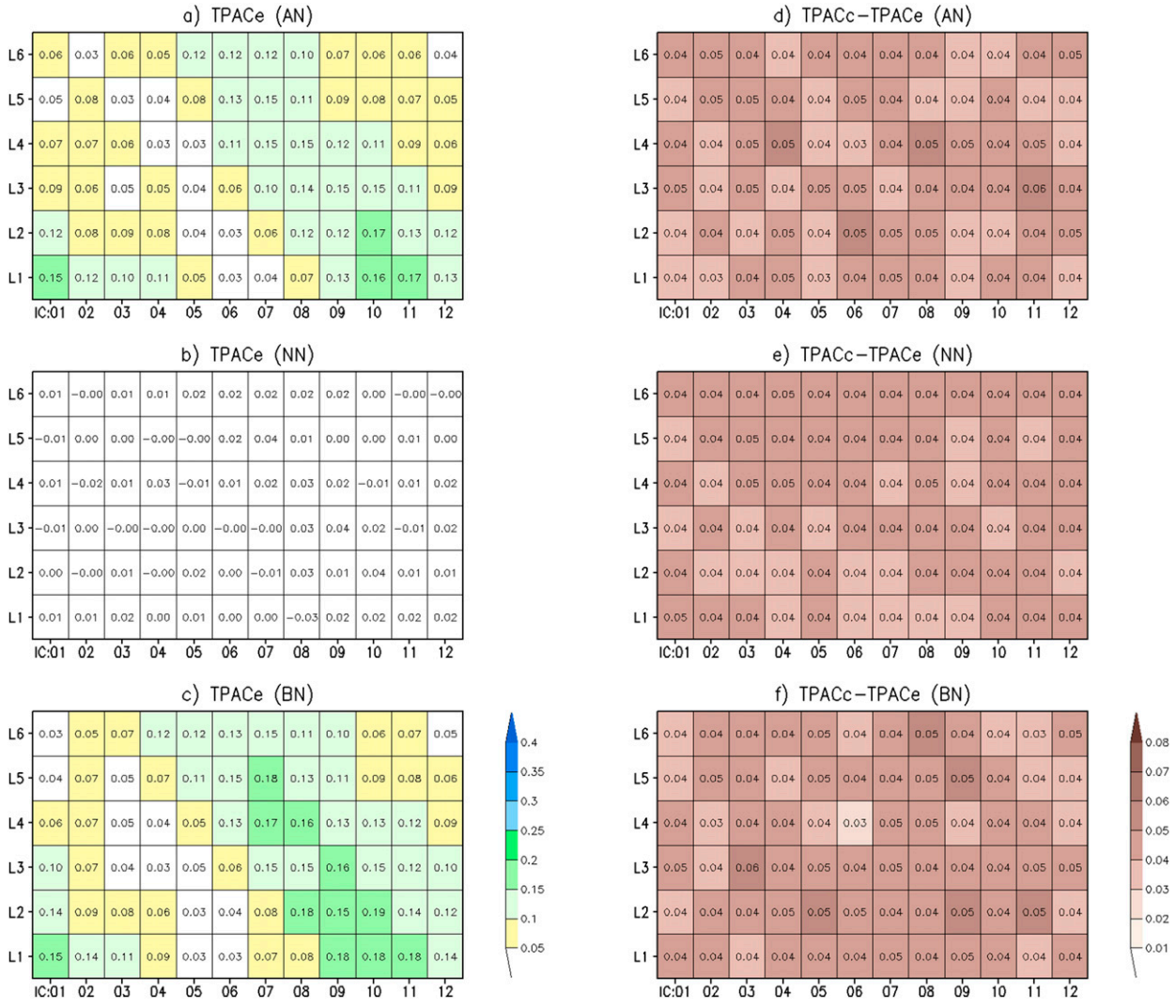

FIG. 10. As in Fig. 9, but for *P* probabilistic forecasts.

(DJF) across the southern United States and northern Mexico, likely as a result of the strong El Niño–Southern Oscillation (ENSO) influence in this season (Chen et al. 2017). The improvements in skill (Fig. 8, center column) are more local and scattered compared with the *T* forecasts. In the spring (MAM), large improvements can be seen for the East Coast, upper Midwest, and northern California. In the summertime (JJA), the upper Midwest, Pacific Northwest, and northern and central Mexico benefit from the weighting schemes. During the SON season, some improvements appear over southwestern Alaska, the central Rocky Mountains, and the Gulf states. For the DJF season, when *P* probabilistic forecasts are most skillful, enhancements are observed in many places over North America, except for some areas along the U. S. East Coast, northern Mexico, and northwestern Canada. The areas of large skill improvement again coincide with the areas of large model elimination rate.

To offer a complete picture of how the skill and its improvements relate to the forecast start and lead times, we compute the cross-validated TPAC averaged over North America for all 12 of the initial months and up to a 6-month lead. Figure 9 presents the results for *T* probabilistic forecasts, and Fig. 10 is for *P* probabilistic forecasts. In both figures, the left column is the average TPAC of equally weighted forecasts, and the right column is the difference between the consolidated and equally weighted forecasts, indicating skill improvements due to weighting schemes. The top, middle, and bottom rows are for the AN, NN, and BN categories, respectively. For *T* probabilistic forecasts, the results of the AN and BN categories are consistent and predictive skill usually is higher at short leads and for all months

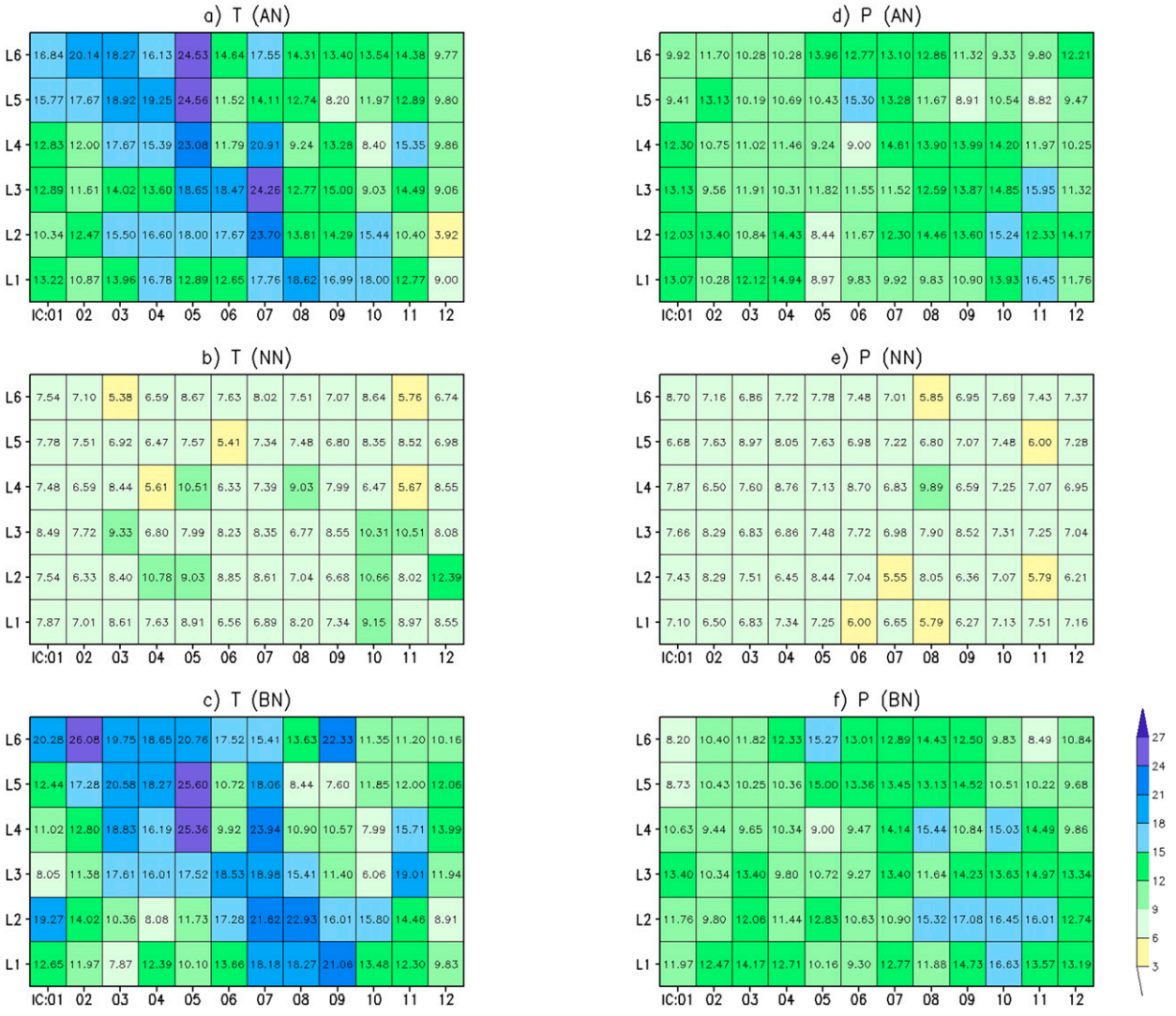

FIG. 11. Matrix charts of the fraction of the North American area with differences between the consolidated and equally weighted forecasts tested to be statistically significant at the 5% level for all start and lead months. The (left) *T* and (right) *P* probabilistic forecasts of the (a),(d) AN, (b),(e) NN, and (c),(f) BN categories.

except late fall. Skill is roughly a function of target forecast month. Predictive skill is very poor for the NN category, as discussed by Van den Dool and Toth (1991). There is no clear pattern in skill improvement associated with the forecast start and lead times. The average increase in TPAC ranges from 0.02 to 0.05 for all three forecast categories.

Compared to the *T* probabilistic forecasts, the predictive skill of *P* probabilistic forecasts evidently is less. However, the increases in TPAC are greater for *P* probabilistic forecasts spanning within 0.03–0.05 and in deeper brown colors. This is caused by the fact that the intermodel variability of the *P* forecast skill is larger than that of the *T* forecast skill for NMME models, so there are more gains by optimally combining models based on hindcast performance. Different from the *T* forecasts, the predictive skill of equally weighted *P* forecasts is higher from late fall to early winter. Similar skill patterns are observed from both the AN and BN categories.

Figure 11 displays the matrix charts of the fraction of the North American land area tested to be statistically significant at the 5% level for all start and lead times. The left column in Fig. 11 shows the results for *T* probabilistic forecasts, and the right column is for *P* probabilistic forecasts. The top, middle, and bottom rows are for the AN, NN, and BN categories, respectively. The fractions of North American land area for which the equal-weighting hypothesis is rejected are about 4%–26% and 6%–17% at the 5% significance

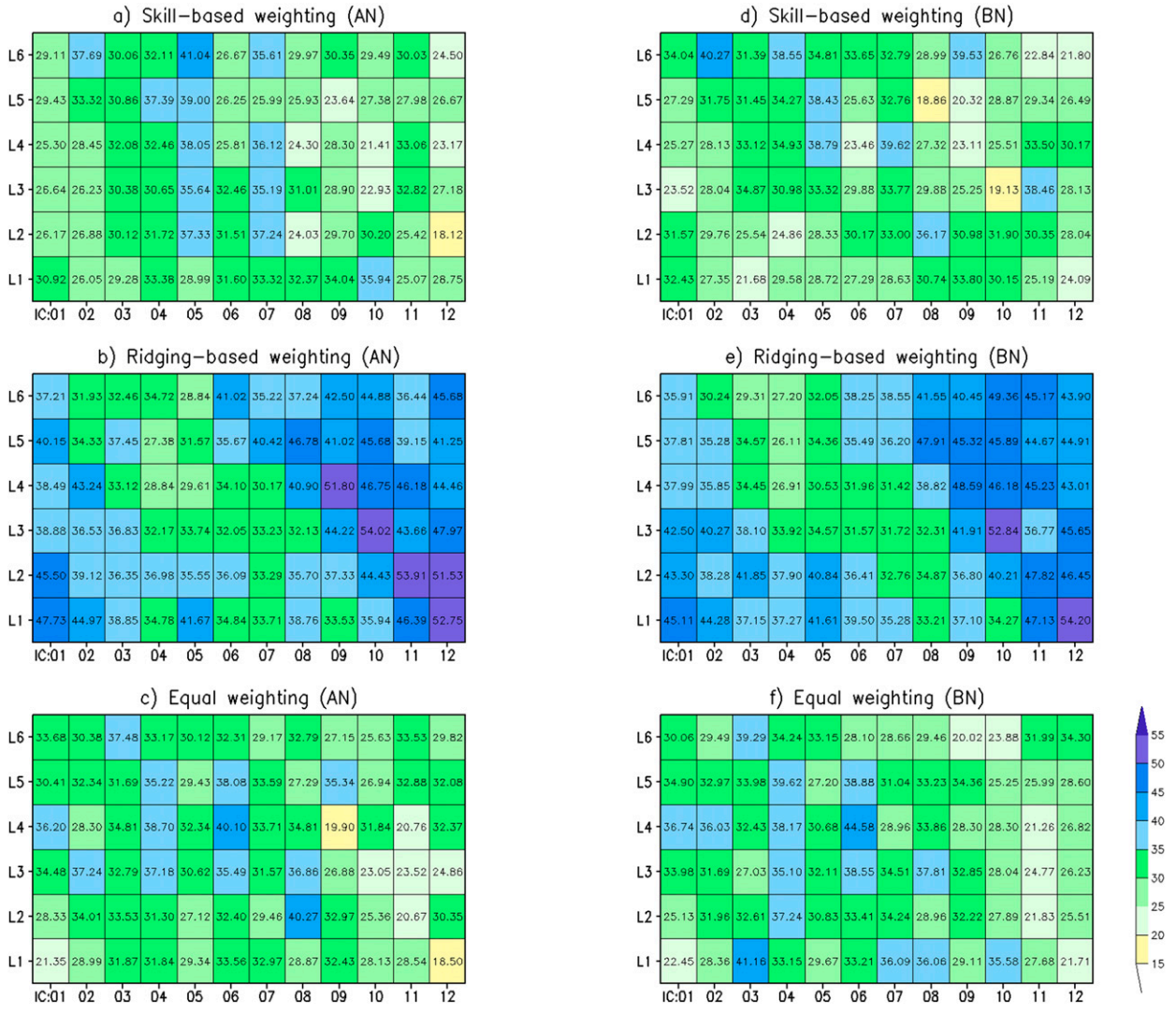

FIG. 12. Matrix charts of the fraction of the North American area selected for (a) skill-based, (b) ridging-based, and (c) equal weighting methods for *T* probabilistic forecasts of the AN category. (d)–(f) As in (a)–(c), but for the BN category.

level for *T* and *P* probabilistic forecasts, respectively, which are both higher than reported in the study by DelSole et al. (2013). Generally, more grid points are tested to have statistically significant differences for *T* probabilistic forecasts than for *P* forecasts. For *T* forecasts, the summer season has greater percentages than other seasons. For *P* forecasts, large fractions appear from late fall to early winter. These seasonal patterns are consistent with the forecast skill patterns observed in Figs. 9 and 10 for the equally weighted *T* and *P* forecasts, respectively, indicating that there is a strong relationship between the area fraction and model forecast skill. When model forecasts are skillful, more grid points are tested to have statistically significant improvement coming from the weighting schemes. For the NN category, in which no skill was found for either *T* or *P* probabilistic forecasts, only a small portion of North American land grids can reject the equal-weighting hypothesis.

Last, we calculate the fraction of North American land grids selected for each weighting method. The results are shown in Fig. 12 for *T* probabilistic forecasts and in Fig. 13 for *P* probabilistic forecasts. Because forecast skill is poor for the NN category and its choice of weighting method does not have a clear spatial pattern (as seen in Fig. 6), we only present the matrix charts for the AN (left column) and BN (right column) categories. The top, middle, and bottom rows are for the skill-based, ridging-based, and equal weighting methods, respectively. For both *T* and *P* probabilistic forecasts, ridge regression is the dominant weighting method among the three options studied here. It is most visible during the winter season for *T*

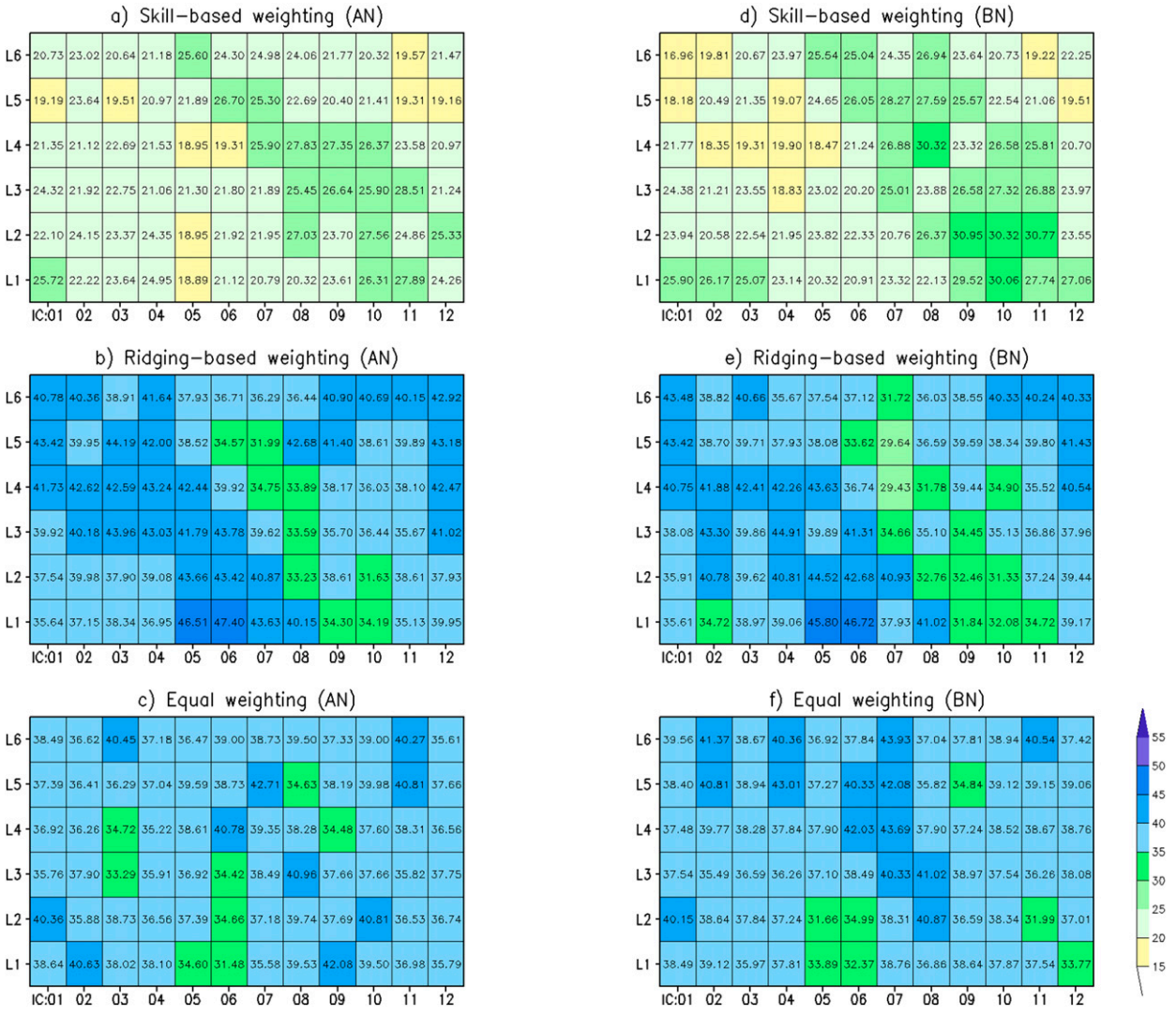

FIG. 13. As in Fig. 12, but for *P* probabilistic forecasts.

probabilistic forecasts (both AN and BN categories) and from late spring to summer for *P* probabilistic forecasts (both AN and BN categories). When model forecasts are skillful (i.e., summer season for *T* forecasts and from late fall to early winter for *P* forecasts), skill-based weighting is more apparent and complementary to the ridge regression method. In fact, the patterns of the matrix charts in Figs. 12 and 13 for the skill-based and ridging-based weighting methods correspond to the patterns of the matrix charts of cross-validated TPAC results shown in Figs. 9 and 10 for both the AN and BN categories. However, because the *P* forecast skill generally is lower than the *T* forecast skill, there are fewer grid points choosing the skill-based weighting and more grid points falling back to equal weights. All these features, in addition to previous observations from Figs. 5 and 6, suggest that the choice of weighting method strongly depends on the model predictive skill.

6. Forecast examples

To inspect the potential impacts of the new weighting system on forecast operations, we generate the consolidated and equally weighted forecast maps for all target forecast years and start and lead months, as well as the observed tercile category maps for comparison. Because the weighting algorithm is applied to the three forecast categories separately, the consolidation forecasts are optimized independently, and the sum of the probability of the three categories is not forced to be one. Most of the time, these violations are within ±5%. Occasionally, discrepancies greater than 20% occur. A similar problem was

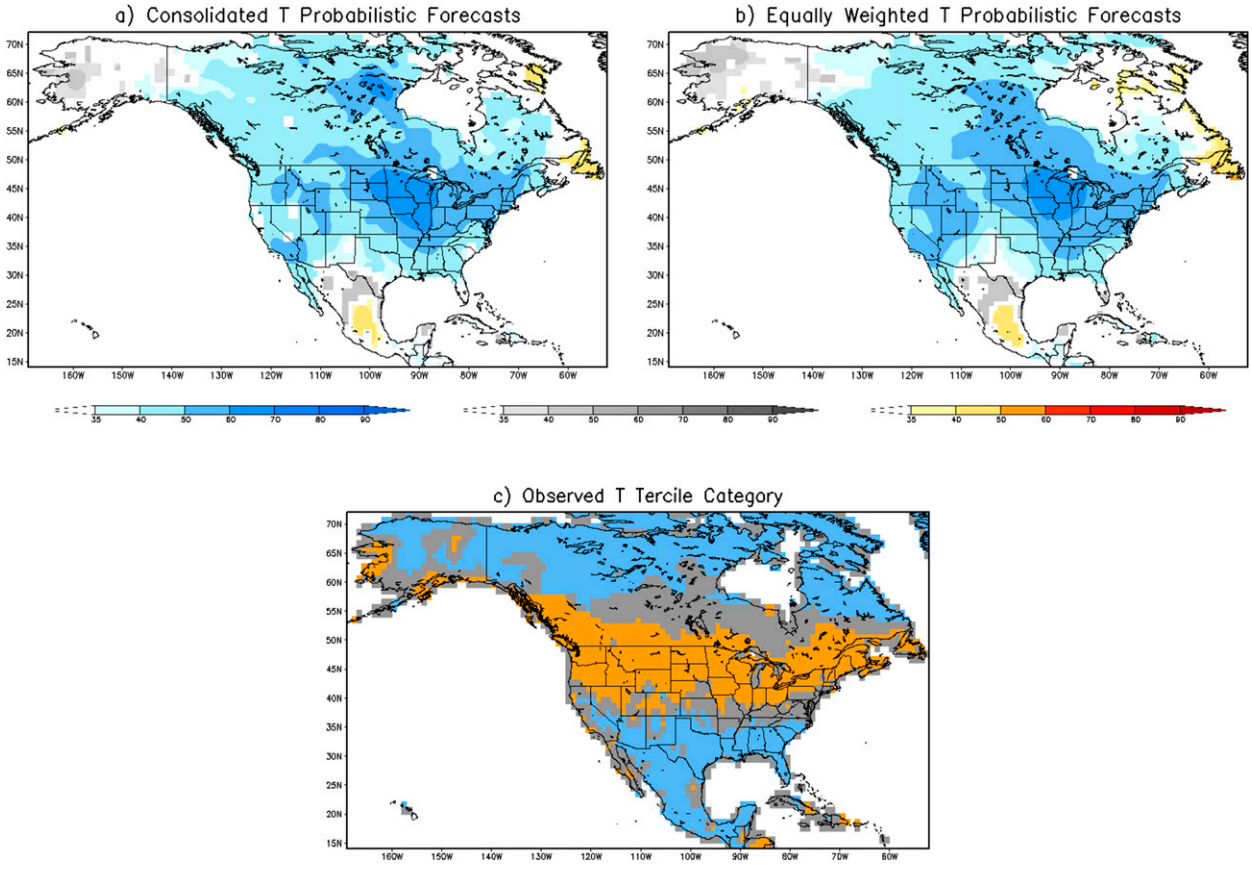

FIG. 14. (a) Consolidated and (b) equally weighted *T* probabilistic forecasts for FMA 1982 initialized on 1 Jan 1982. Blue colors indicate the probability of the BN category, gray colors indicate the probability of the NN category, and yellow-to-red colors indicate the probability of the AN category. (c) Observed tercile categories for the FMA 1982 temperature. Solid gold, gray, and blue shadings are for the AN, NN, and BN categories, respectively.

encountered with PAC-calibrated forecasts, as seen in Van den Dool et al. (2017). We employ the iterative procedure described in their appendix to adjust the probability of the three terciles, so the basic definition of probability can be met.

Figure 14 shows an example of the consolidated (after probability adjustments) and equally weighted *T* probabilistic forecasts of the FMA 1982 season initialized on 1 January 1982. In the forecast maps, the above-normal shading (from yellow to red) at a grid point is shown only when its probability is greater than 35% and the probability of conditions being below normal at the same location is lower than 33%. In contrast, below-normal shading (blue) is shown when its probability is greater than 35% and the probability of above normal at the same location is lower than 33%. Near-normal conditions are shown when the probability of the neutral tercile is more than 35%, and the probabilities of above and below normal are both less than 33%. When no class is dominant (either all categories are under 35% or both above and below normal are over 33%), no shading is shown. Although the probabilities in the consolidated and equally weighted forecast maps are different, they present similar spatial patterns with colder than normal temperature covering most of the North American continent and some warming areas appearing over the southern Mexico and eastern Canada. None of the probabilities captures the above-normal temperatures along the U.S. and Canada boundary as seen in the observations. The SPAC for the equally weighted forecasts of the AN, NN, and BN categories are 0.115, 0.279, and 0.429, respectively. The SPAC for the consolidated forecasts of the AN, NN, and BN categories are 0.138, 0.287, and 0.474, respectively, which are slightly increased from the equally weighted forecasts.

Figure 15 shows the same maps but for the *P* probabilistic forecasts and observations. The overall patterns between the maps of consolidated and equally weighted forecasts are still similar, but their discrepancies are greater when compared to the *T* probabilistic forecasts. Major differences are seen over the Pacific Northwest and central Mexico. For the Pacific Northwest, the rainfall

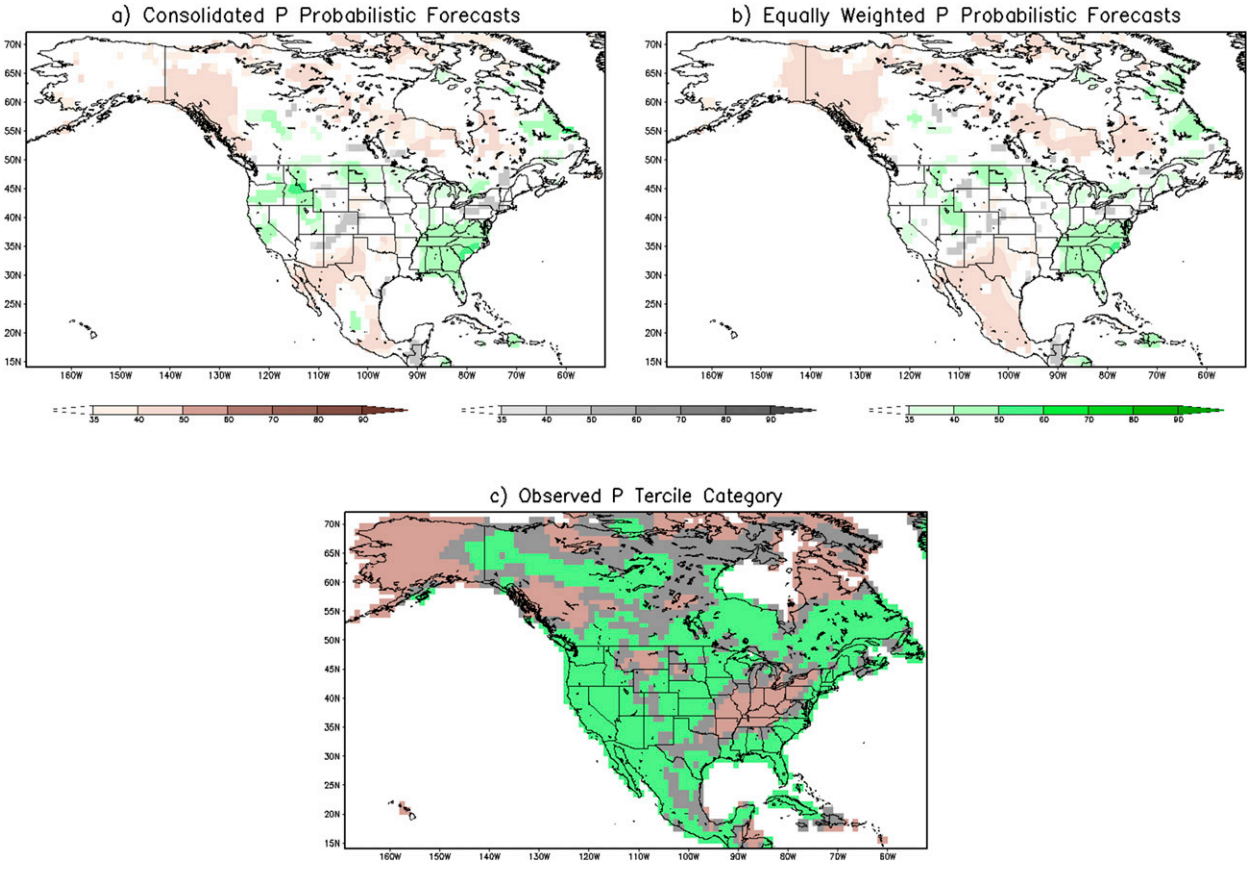

FIG. 15. (a) Consolidated and (b) equally weighted *P* probabilistic forecasts for FMA 1982 initialized on 1 Jan 1982. Brown colors indicate the probability of the BN category, gray colors indicate the probability of the NN category, and green colors indicate the probability of the AN category. (c) Observed tercile categories for the FMA 1982 precipitation. Solid green, gray, and brown shadings are for the AN, NN, and BN categories, respectively.

category changes from near normal to above normal, which is consistent with the observations. For central Mexico, forecasts flip from drier to wetter than normal conditions, verifying well with the above-normal rainfall observations. Compared to the elimination percentage maps of the AN and BN categories in Fig. 3, these two areas coincide with the areas that have high elimination rates and large increases in cross-validated BS, demonstrating the new weighting system's ability to identify skillful models and improve forecast skill. The SPAC for the equally weighted forecasts of the AN, NN, and BN categories are 0.108, 0.049, and 0.107, respectively. The SPAC for the consolidated forecasts of the AN, NN, and BN categories are 0.198, 0.126, and 0.173, respectively. The improvements in SPAC for *P* probabilistic forecasts are greater than those for *T* probabilistic forecasts in this case.

Although Fig. 15 illustrates a successful example of the new weighting system and its potential use for operational forecasts, cases like this (with forecast category changes) are infrequent. Figure 16 displays the time series of SPAC for all target years of FMA probabilistic forecasts initialized on 1 January. For the AN (top row) and BN (bottom row) categories of *T* probabilistic forecasts (left column), only a few years present large increases in SPAC. There are more gains in SPAC for *P* probabilistic forecasts (right column), in both the AN and BN categories. Yet, the increases are small for most years. If we take into account the target years of all forecast start and lead months, about 10% of the time, on average, we see a greater than 0.1 increase in SPAC for an individual forecast (either *T* or *P*). In most cases, the consolidated forecast maps are similar to the equally weighted forecast maps as seen in Fig. 14.

Despite the limited (mainly local) improvements from the new weighting system, several analyses presented in section 5 (e.g., Figs. 11–13) point to a strong linkage between model forecast skill and weighting improvements. When model forecasts are skillful, there are more gains with weighting schemes. When model predictive skill is poor, equal weighting is a competitive method for combining multiple model predictions. This feature may be attributed to the conclusion by some studies in the

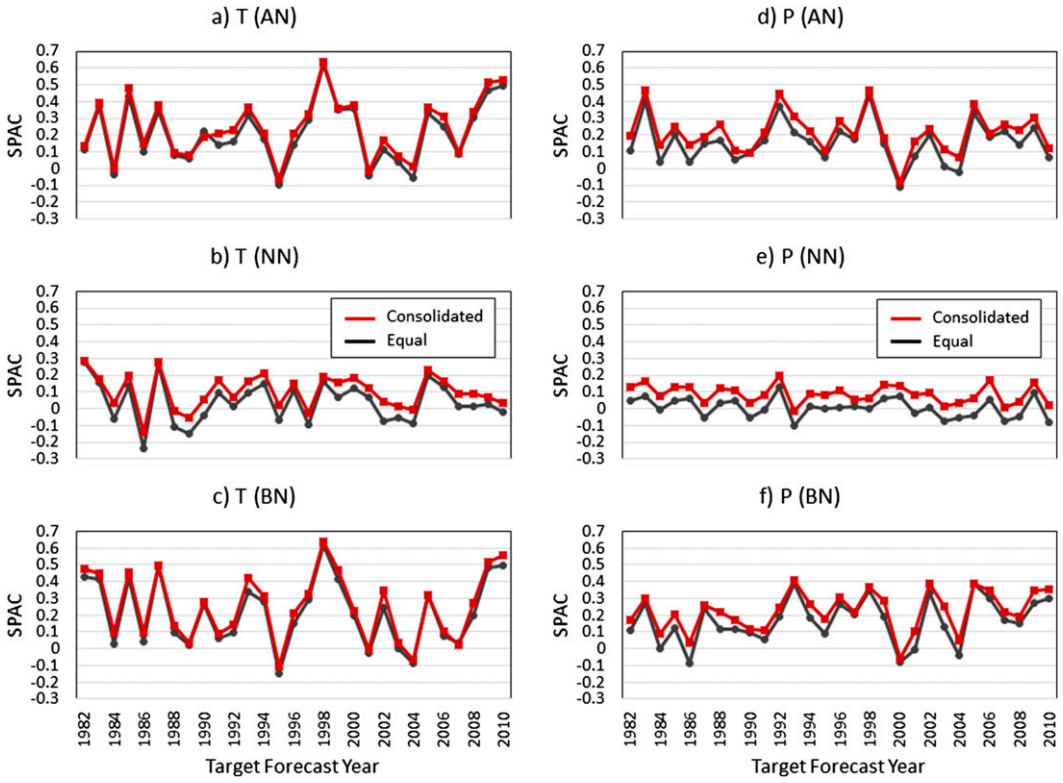

FIG. 16. Time series of SPAC for FMA 1982–2010 probabilistic forecasts initialized on 1 Jan for (left) temperature and (right) precipitation forecasts of the (a),(d) AN, (b),(e) NN, and (c),(f) BN categories.

past decades that weighted forecasts were only marginally better than equally weighted forecasts, since many models in the DEMETER era (Doblas-Reyes et al. 2005; Peña and Van den Dool 2008) have less skill compared to NMME models. As climate models continue to advance, more benefits are expected from weighting schemes with the next-generation models.

7. Summary and conclusions

We have developed an objective weighting system to combine multiple seasonal probabilistic forecasts in the North American Multimodel Ensemble (NMME), in contrast to conventional work on ensemble means. The system is applied to predict precipitation P and temperature T over the North American continent, and the analysis is conducted using the 1982–2010 hindcasts from eight NMME models, including the CFSv2, CanCM3, CanCM4, CM2.1, FLOR, GEOS5, CCSM4, and CESM models, with weights determined by minimizing the Brier score using ridge regression. We exploited two conservative strategies to improve the performance of ridge regression: eliminating a priori models with negative skill (and/or negative weight) and increasing the effective sample size by pooling information from neighboring grids. A set of new constraints is put in place to confine the weights within a reasonable range or restrict the weights from departing wildly from equal weights, which is the fallback. So when the predictor–predictand (model forecast–verifying observation) relationship is weak, the multimodel ensemble forecast returns to an equal-weight combination. The system performance is assessed using a leave-three-out procedure to reduce the effect of degeneracy in cross-validated skill in regression-based forecasts. Our major findings are summarized below.

- The new system is able to optimally select a varying number of skillful models based on hindcast performance and assign weights accordingly. All models contribute to the consolidated forecasts differently based upon location and forecast start and lead times.
- The system shows improved skill from the baseline, equally weighted forecasts. Because of the elimination and fall back, there is no loss (with respect to equal weights) when model skill is poor.
- For a given forecast start and lead time, the amount of improvement over equal weights varies across space and corresponds to the average model elimination percentage. Areas with higher elimination

rates tend to have larger improvements in cross-validated verification scores.
- Some local improvements can be as large as 0.6 in cross-validated temporal probability anomaly correlation (TPAC). On average, improvements are about 0.02–0.05 in TPAC for *T* probabilistic forecasts and 0.03–0.05 for *P* probabilistic forecasts over the North American continent.
- The skill improvement is generally greater for *P* probabilistic forecasts than for *T* probabilistic forecasts.
- The choice of weighting method within the designed system and the significance testing results strongly depend on the model predictive skill.

For years, scientists have debated about whether weighting methods can significantly improve the forecast skill of multimodel combinations from simple multimodel means. We have demonstrated that with proper strategies to reduce collinearity and increase the sample size of the training dataset, it is beneficial to combine multiple model predictions with regression-based weighting schemes, especially when model forecasts are skillful. The strong linkage between model forecast skill and the choice of weighting method suggests that the benefits with regression-based weighting schemes would be even greater if the model skill is higher. When model predictive skill is low, equally weighted forecasts have an advantage. This may have led to earlier conclusions by some studies based on using climate models from a previous era. As climate models continue to advance and their skill continues to improve, we believe that weighting methods will achieve more gains in skill with the next-generation models.

In this study, we have designed a system to fall back to the baseline when all models are eliminated or no skill is enhanced by the weighted forecasts. We use equally weighted forecasts as the baseline for the evaluation to be in alignment with the literature. In operation, forecasts are damped to climatology (equal chance in probabilistic forecasts) when unskillful (i.e., calibrated), and it is desirable to have the system fall back to calibrated forecasts when predictive skill is poor. How to combine multimodel forecast calibration and weighting into a single optimal system with effective strategies (as proposed in this research) is a challenging topic and requires further endeavors to meet this goal.

Acknowledgments. This research was supported by the North American Multimodel Ensemble (NMME) project, a multiagency and multi-institutional research effort led by NOAA's National Weather Service (NWS) Climate Test Bed (CTB) and Climate Program Office (CPO) Modeling, Analysis, Predictions, and Projections (MAPP) program in partnership with DOE, NSF, and NASA, under NOAA Grant NA14OAR4310188 to Huug van den Dool and Grant NA14NES4320003 to the Cooperative Institute for Climate and Satellites at the University of Maryland (CICS-MD). We greatly appreciate the editor and two anonymous reviewers for their positive comments and suggestions to help improve the manuscript.

REFERENCES

Barnston, A. G., and H. M. Van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate*, **6**, 963–977, doi:10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2.

Becker, E., and H. Van den Dool, 2016: Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *J. Climate*, **29**, 3015–3026, doi:10.1175/JCLI-D-14-00862.1.

——, ——, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891–5906, doi:10.1175/JCLI-D-13-00597.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Chen, L.-C., H. Van den Dool, E. Becker, and Q. Zhang, 2017: ENSO precipitation and temperature forecasts in the North American Multimodel Ensemble: Composite analysis and validation. *J. Climate*, **30**, 1103–1125, doi:10.1175/JCLI-D-15-0903.1.

Chen, M., P. Xie, J. E. Janowiak, and P. A. Arkin, 2002: Global land precipitation: A 50-yr monthly analysis based on gauge observations. *J. Hydrometeor.*, **3**, 249–266, doi:10.1175/1525-7541(2002)003<0249:GLPAYM>2.0.CO;2.

Danabasoglu, G., S. C. Bates, B. P. Briegleb, S. R. Jayne, M. Jochum, W. G. Large, S. Peacock, and S. G. Yeager, 2012: The CCSM4 ocean component. *J. Climate*, **25**, 1361–1389, doi:10.1175/JCLI-D-11-00091.1.

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate*, **20**, 2810–2826, doi:10.1175/JCLI4179.1.

——, X. Yang, and M. K. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.*, **139**, 176–183, doi:10.1002/qj.1961.

Delworth, T. L., and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate*, **19**, 643–674, doi:10.1175/JCLI3629.1.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensemble in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

Fan, Y., and H. M. Van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, doi:10.1029/2007JD008470.

Gent, P. R., and Coauthors, 2011: The Community Climate System Model version 4. *J. Climate*, **24**, 4973–4991, doi:10.1175/2011JCLI4083.1.

Giorgi, F., and R. Francisco, 2000: Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dyn.*, **16**, 169–182, doi:10.1007/PL00013733.

Gnanadesikan, A., and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part II: The baseline ocean simulation. *J. Climate*, **19**, 675–697, doi:10.1175/JCLI3630.1.

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*, **57A**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hurrell, J. W., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, doi:10.1175/BAMS-D-12-00121.1.

Jia, L., and Coauthors, 2015: Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. *J. Climate*, **28**, 2044–2064, doi:10.1175/JCLI-D-14-00112.1.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15**, 793–799, doi:10.1175/1520-0442(2002)015<0793:CPWME>2.0.CO;2.

Kirtman, B. P., and Coauthors, 2014: The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction; phase-2 toward developing intra-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

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 multimodel superensemble. *Science*, **285**, 1548–1550, doi:10.1126/science.285.5433.1548.

——, ——, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, doi:10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Merryfield, W. J., and Coauthors, 2013: The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, doi:10.1175/MWR-D-12-00216.1.

Peña, M., and H. M. Van den Dool, 2008: Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *J. Climate*, **21**, 6521–6538, doi:10.1175/2008JCLI2226.1.

Peng, P., A. Kumar, H. M. van den Dool, and A. G. Barnston, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.*, **107**, 4710, doi:10.1029/2002JD002712.

Phillips, D. L., 1962: A technique for the numerical solution of certain integral equations of the first kind. *J. Assoc. Comput. Mach.*, **9**, 84–97, doi:10.1145/321105.321114.

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

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Tikhonov, A., 1963: Solution of incorrectly formulated problems and the regularization method. *Sov. Math. Dokl.*, **4**, 651–667.

——, and V. Y. Arsenin, 1977: *Solutions of Ill-Posed Problems.* Winston, 258 pp.

Van den Dool, H. M., 2007: *Empirical Methods in Short-Term Climate Prediction.* Oxford University Press, 215 pp.

——, and Z. Toth, 1991: Why do forecasts for ''near normal'' often fail? *Wea. Forecasting*, **6**, 76–85, doi:10.1175/1520-0434(1991)006<0076:WDFFNO>2.0.CO;2.

——, and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6–10-day forecast. *Wea. Forecasting*, **9**, 457–465, doi:10.1175/1520-0434(1994)009<0457:OTWFAE>2.0.CO;2.

——, E. Becker, L.-C. Chen, and Q. Zhang, 2017: The probability anomaly correlation and calibration of probabilistic forecasts. *Wea. Forecasting*, **32**, 199–206, doi:10.1175/WAF-D-16-0115.1.

Vecchi, G. A., and Coauthors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27**, 7994–8016, doi:10.1175/JCLI-D-14-00158.1.

Vernieres, G., C. Keppenne, M. M. Rienecker, J. Jacob, and R. Kovach, 2012: The GEOS-ODAS, description and evaluation. NASA Tech. Rep. NASA/TM-2012-104606, NASA Tech. Rep. Series on Global Modeling and Data Assimilation, Vol. 30, 61 pp., https://gmao.gsfc.nasa.gov/pubs/docs/Vernieres589.pdf.

Wanders, N., and E. F. Wood, 2016: Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. *Environ. Res. Lett.*, **11**, 094007, doi:10.1088/1748-9326/11/9/094007.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.*, **134**, 241–260, doi:10.1002/qj.210.

Weisheimer, A., and Coauthors, 2009: ENSEMBLES: A new multi-model 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.

Yun, W., L. Stefanova, and T. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate*, **16**, 3834–3840, doi:10.1175/1520-0442(2003)016<3834:IOTMST>2.0.CO;2.