Precipitation Calibration Based on the Frequency-Matching Method

YUEJIAN ZHU

NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland

YAN LUO

NOAA/NWS/NCEP/Environmental Modeling Center, and I. M. Systems Group, Inc., College Park, Maryland

(Manuscript received 26 April 2013, in final form 19 May 2014)

ABSTRACT

A postprocessing technique is employed to correct model bias for precipitation fields in real time based on a comparison of the frequency distributions of observed and forecast precipitation amounts. Essentially, a calibration is made by defining an adjustment to the forecast value in such a way that the adjusted cumulative forecast distribution over a moving time window dynamically matches the corresponding observed distribution accumulated over a domain of interest, for example, the entire conterminous United States (CONUS), or different River Forecast Center (RFC) regions in the cases examined herein. In particular, the Kalman filter method is used to catch the flow dependence and bias information. Calibration is done on a pointwise basis for a specified domain. Using this unique technique, the calibration of precipitation forecasts for the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) was implemented in May 2004. To further satisfy various users, a recent upgrade to the May 2004 implementation has been made for higher resolution with better analyses. From this study, it was found that this method has a positive impact on the intensity-dominated errors but has some common limitations with extreme events and dry bias elimination like other precipitation calibration methods. Overall, the frequency-matching algorithm substantially improves NCEP Global Forecast System (GFS) and GEFS systematic precipitation forecast errors (or biases) over a wide range of forecast amounts and produces more realistic precipitation patterns. Moreover, this approach improves the deterministic forecast skills measured by most verification scores through applying this method to GFS and GEFS ensemble means.

1. Introduction

There are many important applications that require more accurate quantitative precipitation forecasts (QPFs) and probabilistic quantitative precipitation forecasts (PQPFs) (Demargne et al. 2014; Brown et al. 2012; Mascaro et al. 2010; Marty et al. 2013). For instance, water management decisions are crucially dependent on forecast information regarding the possible future evolution of precipitation. The QPF- and ensemble-based PQPF forecast products were implemented into National Centers for Environmental Prediction (NCEP) operations in the late 1990s (Zhu et al. 1998; Zhu 2005), providing crucial guidance for water management decision makers. Furthermore, another important use is for downstream applications. Hydrologic models need accurate precipitation forecasts from numerical weather prediction (NWP) as forcing inputs. Therefore, a realistic representation of the precipitation field in forecasts is very important.

However, many studies have demonstrated systematic errors (or biases) in the model precipitation products due to model deficiencies. It has long been recognized that model precipitation uncertainty affects the accuracy of hydrologic modeling (Demargne et al. 2014; Brown et al. 2012), because the performance of lumped, distributed, physically based hydrologic models depends greatly on the quality of the precipitation input data. The intermittent and space–time-scale-dependent features of precipitation fields make precipitation extremely difficult to predict, imposing a great challenge on precipitation forecasts (Brown et al. 2012; Yuan et al. 2008). For both of these reasons, statistical postprocessing techniques

Denotes Open Access content.

Corresponding author address: Yuejian Zhu, NOAA/NWS/NCEP/Environmental Modeling Center, 5830 University Research Ct., College Park, MD 20740.
E-mail: yuejian.zhu@noaa.gov

DOI: 10.1175/WAF-D-13-00049.1

© 2015 American Meteorological Society

have been developed and applied to reduce these biases in precipitation and better quantify the uncertainty therein. Better-calibrated QPFs and PQPFs could benefit the short- and medium-range forecasts, and extend the forecast predictability (Eckel and Walters 1998; Zhu and Toth 1999). Many studies have demonstrated some success with precipitation forecasts through statistical postprocessing (Hamill et al. 2008; Bentzien and Friederichs 2012). There are several studies on calibration of PQPFs; some focus on the methodology (Krzysztofowicz and Sigrest 1999; Primo et al. 2009) and others use reforecast information (Hamill et al. 2008; Fundel et al. 2010). Yuan et al. (2007) applied an artificial neural network as a postprocessor to calibrate PQPFs from the NCEP Regional Spectral Model (RSM) ensemble forecast system. Voisin et al. (2010) described two bias correction methods with spatial disaggregation (BCSD) and an analog technique for downscaling and calibrating errors from ensemble precipitation forecasts. An analog method has been developed by using large samples of reforecasts (Hamill and Whitaker 2006) and is experimentally run at the Earth System Research Laboratory (ESRL). ESRL's products provide additional guidance for NCEP Weather Prediction Center (WPC) forecasters.

In this study, we have developed a method for precipitation calibration in real time called the frequency-matching method. A similar concept has been investigated for different applications, such as multimodel ensembles (Ebert 2001), and the relationship between radar reflectivity and rain rate (Rosenfeld et al. 1993). Basically, the methodology employed here is a statistical adjustment based on cumulative frequency distributions of forecast and observed precipitation amounts. Two steps are undertaken for calibration with frequency matching. First, it requires an observation dataset at the same spatial and temporal resolutions as the model forecast output and a reasonable number of days of prior forecasts to construct the respective cumulative frequency distributions for forecasts and observations. In addition, it introduces a decaying average for appropriate historical sampling that makes the cumulative frequency distribution of forecasts match that of the observations. The second step in this method makes use of the cumulative frequency distributions of observations and forecasts; in this way a frequency match is performed between prior observations and forecasts using bilinear interpolation. The resulting correction factor is applied to adjust a target forecast value at each grid point and each grid point is treated individually.

In this paper, we first briefly describe the "bias correction" procedure implemented on 4 May 2004 at NCEP. A bias [or frequency bias (FB)] is defined as the ratio of forecast and observation frequency counts at each threshold. A ratio equal to one means a perfect (or unbiased) forecast. The 2004 implementation was first developed as a pioneering version of precipitation calibration with frequency matching for application in precipitation forecasts with 24-h accumulations at 2.5° resolution (Zhu and Toth 2004). As with any other numerical weather prediction model, QPFs from the Global Forecast System (GFS) at NCEP suffer from biases due to model deficiencies. PQPFs based on the Global Ensemble Forecast System (GEFS) at NCEP are biased as well because of imperfections in the model and ensemble formation. Typically, model precipitation bias is dependent on the model version, lead time, and location. In most cases, small amounts of precipitation are overforecast while large amounts are underforecast. By calibrating each member of the ensemble based on verification statistics accumulated over the conterminous United States (CONUS), the bias in QPF (first moment) is practically eliminated, and the PQPF (second moment) is substantially improved. By following the approach of the 2004 implementation with the timely availability of higher-resolution model output and a better analysis, named the climatology-calibrated precipitation analysis (CCPA; Hou et al. 2014), we pursue a similar application at 1° resolution and every 6 h out to 384 h (~16 days) globally.

To provide a better proxy of the truth for the precipitation field over the CONUS at high spatial and temporal resolutions, CCPA has been developed and evaluated at NCEP by Hou et al. (2014). It is a precipitation analysis dataset generated by statistically adjusting the NCEP stage IV analysis to make its long-term average and climate probability distribution closer to that of the Climate Prediction Center's (CPC) Unified Gauge-Based Analysis of Daily Precipitation over CONUS. The dataset takes advantage of the higher climatological reliability of the CPC dataset and the higher temporal and spatial resolutions of the stage IV dataset (Lin and Mitchell 2005). Thus, CCPA is reliable and quality controlled, with high spatial and temporal resolutions. It is available as 6-h accumulations from 2002 onward. The CCPA data are first produced on the 4-km Hydrologic Rainfall Analysis Project (HRAP) grid, the same as in the NCEP stage IV dataset over the CONUS, as a primary product and then interpolated onto 1°, 0.5°, 0.125°, and 5-km National Digital Guidance Database (NDGD) grids by a volume conservation scheme as by-products. The 1° CCPA is applied in this study as it exactly matches the model output grid.

We continue to investigate here the method that applies to the NCEP GFS/GEFS precipitation model output with CCPA. Then, we analyze aspects of the bias correction of ensemble precipitation forecasts, including precipitation forecast skill and reliability. Our objective is to produce bias-corrected precipitation ensemble forecasts through the postprocessing for near-real-time forecast applications.

The remainder of the paper is organized as follows. Section 2 describes the frequency-matching method for precipitation calibration. Section 3 reviews the background of the 2004 implementation. A few cases demonstrating the success of this method are presented. Section 4 applies and evaluates the bias correction approach for higher resolutions using CCPA, and in the last section we present our conclusions with suggestions for future work that will further improve the calibration of precipitation.

2. Methodology

A systematic difference (or error) between forecast and observed precipitation amounts can be progressively removed using information provided by observations. However, it cannot be processed directly from the accumulated difference of precipitation amounts because it is a non-Gaussian distribution. In this study, the systematic difference information can be estimated through comparing forecast and observed precipitation frequency distributions (i.e., FB). A general frequency-matching method proceeds as follows. First, we conduct an FB assessment by constructing a cumulative distribution function (CDF) for the preceding forecast and corresponding observed precipitation amounts. Given a set of precipitation thresholds in ascending order, the CDF is calculated as the number of grid points over a given domain where the forecast or observed precipitation values exceed a threshold. The CDFs for both forecast and observed precipitation amounts are updated with the Kalman filter method, which is similar to the bias correction method in the North American Ensemble Forecast System (NAEFS; Cui et al. 2012; B. Cui et al. 2013, unpublished manuscript), expressed as

$$\mathrm{FB}_{i,j} = \frac{\mathrm{FFC}_{i,j}}{\mathrm{OFC}_{i,j}}, \tag{1}$$

where $\mathrm{FFC}_{i,j}$ stands for the forecast frequency counts at threshold i for day j, while $\mathrm{OFC}_{i,j}$ stands for observation frequency counts, and $\mathrm{FB}_{i,j}$ is a frequency bias. We use $\mathrm{FB}_{i,j} = 1$ for the perfect frequency unbiased forecast. Variable $\overline{\mathrm{CDF}_{i,j}}$ is the decaying averaged CDF at threshold i for day j, while $\overline{\mathrm{CDF}_{i,j-1}}$ is the prior decaying averaged CDF for day $j - 1$ in

$$\overline{\mathrm{CDF}_{i,j}} = (1 - W)\overline{\mathrm{CDF}_{i,j-1}} + W(\mathrm{CDF}_{i,j}). \tag{2}$$

The newly counted CDF at threshold i for day j is $\mathrm{CDF}_{i,j}$, and W is the decaying weight between 0 and 1, defined simply by an approximated time moving window nd (nd cannot equal zero), that is, a number of days for decaying:

$$W = \frac{1}{\mathrm{nd}}. \tag{3}$$

Here, a time moving window nd (or decaying weight 1/nd) is chosen to make a weighted average of these CDF values over the domain depending on how far it is from the target forecast day, which is illustrated in Fig. 1. For day j, the previous day's $j - 1$ $\overline{\mathrm{CDF}_{i,j-1}}$ contributes a weight of $1 - W$ to $\overline{\mathrm{CDF}_{i,j}}$. From an iteration procedure following Eq. (1), the previous day's $j - k$ $\overline{\mathrm{CDF}_{i,j-k}}$ contributes a weight of $(1 - W)^k$ to $\overline{\mathrm{CDF}_{i,j}}$, which becomes smaller and smaller and eventually approaches zero as k goes toward infinity. Therefore, the higher the decaying weight, the faster the decaying speed (which indicates that there is a higher overall weight on the most recent data and less on the oldest data) and vice versa. Our strategy is to specify a number of prior forecast days (an approximated time moving window, or decaying weight) for each grid point and each lead time as a pool for sampling appropriate historical information from forecasts and observations. For instance, a 50-day window ($W = 0.02$) means that training data are accumulated over the most recent 50-day period, with the most weight on the most recent data (see Fig. 1 for $W = 0.02$). Thus, the idea behind the adaptive method is to catch the dynamic flow dependence and statistics of the observations. The time moving window (or decaying weight) can be tuned from short (or large) to long (or small) times (weights) to ensure the best performance of the method. In our adaptation of the frequency-matching method, there are two ways of constructing CDFs for forecasts and observations. We call the CDF based on the whole CONUS domain the CONUS CDF, and the CDFs based on each River Forecast Center (RFC) region (see Fig. 2; Table 1) are RFC CDFs. For each grid point within a specific domain (e.g., CONUS or any RFC) and for each forecast lead time, the observed and forecast CDFs are derived using the same time moving window (or decaying weight). To be useful for applications, this method needs to handle the initial CDFs, which is termed spinup. For example, one can use 1-month or 1-yr averages of CDF pairs as a cold start, update the CDF pairs for a certain period and throw away old results, and select new statistics and calibrated products for evaluation of the method and application of the products.

Second is the forecast adjustment. To keep the spatial and temporal coherence of a forecast as similar as possible to that of the observation, we match the cumulative

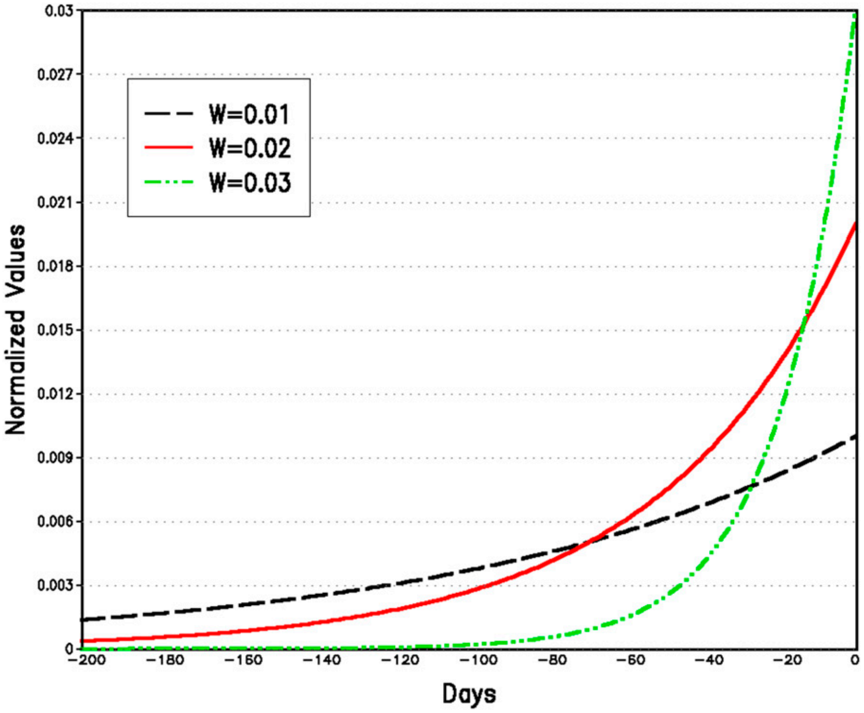

FIG. 1. Decaying averaged weight as a function of preceding days (weighting function for decaying average of preceding days). The dashed curve denotes a weight beginning with a max value of 0.01 at day 0. The solid curve denotes a weight beginning with a max value of 0.02 and the dashed–dotted curve denotes a weight beginning with a max value of 0.03 at day 0. All curves gradually approach zero depending on how far away the preceding days are from day 0. The larger the weight at day 0, the faster the decaying speed, which indicates greater weight on the most recent data and less on the oldest data.

frequency distribution of the forecast to that of the observation using a frequency-matching algorithm. Here, the updated CDF values from Eq. (1) form cumulative frequency distributions. As illustrated in Fig. 3, according to this pair of distributions, for a raw forecast value (“RAW”) we find and assign an observed value that has the same frequency within a given domain as the forecast value of the corresponding calibrated forecast (“CAL”). Consequently, the information related to systematic difference is estimated based on the paired and updated CDFs for the forecast and corresponding observed values. For example, in Fig. 3 in the case of a CDF (forecast) greater than the CDF (observed), model precipitation tends to be overforecast; so to match the frequency, a correction factor of less than one will be expected to reduce a forecast value. In doing this matching process, linear interpolation is applied twice in real calculations to derive a correction factor for each grid point. Mathematically, the linear interpolation applied here is polynomial interpolation using two data points. Given a vector of precipitation thresholds $\mathbf{T}(n)$, $T_1, T_2, \ldots, T_n$ in ascending order as the abscissas; and a vector of observed CDF values $\mathbf{O}(n)$ at $\mathbf{T}(n)$, $O_1, O_2, \ldots, O_n$ as ordinates; a vector of calibrated thresholds $\mathbf{T}^*(n)$, $T_1^*, T_2^*, \ldots, T_n^*$ is derived based on a vector of forecast CDF values $\mathbf{F}(n)$ at $\mathbf{T}(n)$, $F_1, F_2, \ldots, F_n$ through the first linear interpolation. Here, n is the length of the vector. Consequently, n source pairs $(\mathbf{O}, \mathbf{T})$ are linearly interpolated to n targets $(\mathbf{F}, \mathbf{T}^*)$, which implies that the forecast CDF F_i at T_i^* $(i = 1, \ldots, n)$ is equal to the observed CDF O_i at T_i $(i = 1, \ldots, n)$, just as in what we call frequency matching. That is,

$$\begin{aligned} O_1(T_1) &= F_1(T_1^*), \\ O_2(T_2) &= F_2(T_2^*), \\ &\vdots \\ O_n(T_n) &= F_n(T_n^*). \end{aligned}$$

In practice, the selected thresholds use logarithmic transformation in the first interpolation. Next, a correction factor Ri is calculated as the ratio of a calibrated threshold to its related threshold (i.e., Ri = T_i^*/T_i, $i = 1, \ldots, n$). Once again, given a vector of thresholds $\mathbf{T}(n)$, $T_1, T_2, \ldots, T_n$ as the abscissas and a vector of correction factors $\mathbf{R}(n)$, $R_1, R_2, \ldots, R_n$ as ordinates, for a forecast value (RAW) at any grid point a single correction

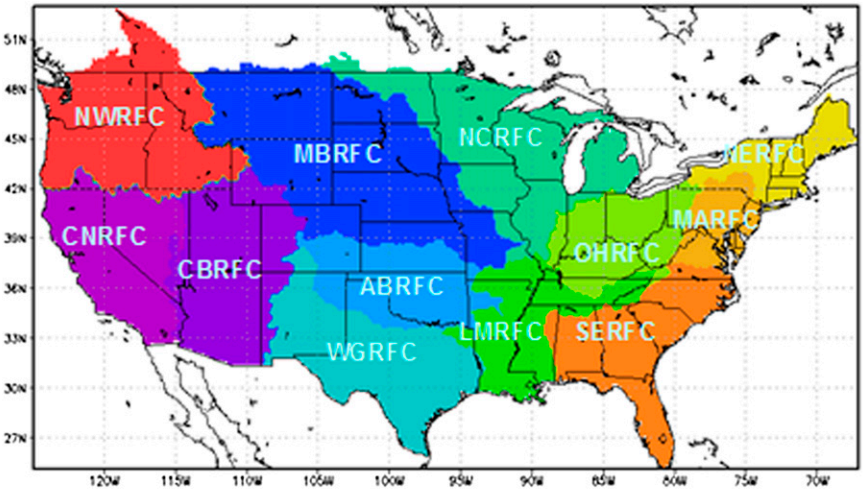

Fig. 2. The domains of the 12 RFCs. Note that the CCPA covers the 12 RFCs (acronym expansions and index numbers associated with the scale are provided in Table 1) across the CONUS.

factor r, the ratio of a calibrated forecast value (CAL) to its corresponding raw forecast value RAW, is derived by linear interpolation from n source pairs ($\mathbf{T}$, $\mathbf{R}$) to one target (RAW; r). Then, the single correction factor is applied to the raw forecast value (RAW) to compute the final calibrated forecast value [CAL = r(RAW)]. This correction is applied to each model grid point, which implies that the correction is a function of the forecast value. No adjustment of a zero precipitation forecast value is made in order to prevent an unrealistic negative precipitation value due to interpolation. In addition, all resulting negative precipitation values from interpolation become zero for the same reason.

This calibration technique with its frequency matching should work with any model output as long as observations are available and are processed to be at model grid points. However, our experience with this technique indicates that some important considerations must be addressed. That is, precautions must be taken regarding the selections of thresholds and the number of decay days, particularly when the CDF is calculated for each RFC rather than the CONUS because there will be a much smaller sample size, as implied from Table 1. For example, an insufficient amount of nonzero sample data is very likely to cause more than two equal values of zero as CDFs for adjacent highest thresholds, though this situation is not allowed in this method as it may lead to a failure in the interpolation. To deal with this problem,

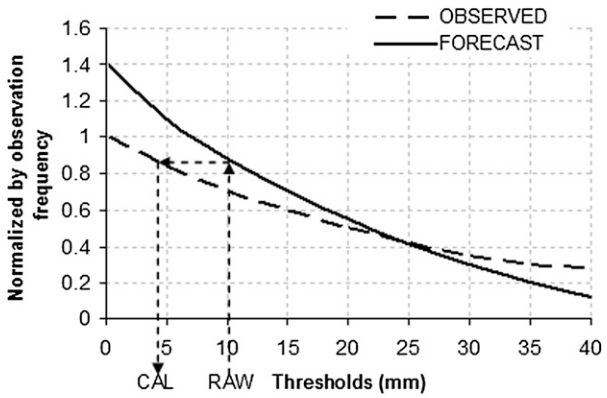

Fig. 3. Schematic of the frequency-matching algorithm demonstrated as precipitation distributions normalized by observation frequency varying with threshold. The dashed line is for observed and the solid line is for forecast precipitation. See text for details.

Table 1. Total gridpoint counts of CONUS and each RFC at 1° spacing.

Index	RFC	Count
1	California–Nevada RFC (CNRFC)	67
2	Colorado basin RFC (CBRFC)	83
3	Missouri basin RFC (MBRFC)	152
4	Arkansas-Red River RFC (ABRFC)	56
5	West Gulf RFC (WGRFC)	75
6	North-central RFC (NCRFC)	105
7	Lower Mississippi RFC (LMRFC)	51
8	Ohio RFC (OHRFC)	47
9	Northeast RFC (NERFC)	31
10	Mid-Atlantic RFC (MARFC)	21
11	Southeast RFC (SERFC)	61
12	Northwest RFC (NWRFC)	95
Total	CONUS	844

selecting a reasonable range of thresholds is necessary to produce nonequal CDF values. From our precipitation calibration practice, it is better to have closed equal distances in a logarithm coordinate and to consider the common thresholds as well. Another solution is found by choosing a proper number of decaying days. If the number of decaying days is too small, it will be problematic since there will not be sufficient sample data, especially when dry-climate regions experience a long-duration drought. Therefore, there is an inevitable trade-off as to the number of decaying days when tuning for optimal calibration performance. It is believed that potential difficulties in CDF construction in dry regions are related to the small number of days with precipitation, imposing a practical challenge to this method. When the above statistical deficiencies and operational limitations are avoided, the method should be computationally realistic and feasible for real-time implementation.

3. Background review of the 2004 frequency bias correction technique

As part of the 2004 implementation (Zhu and Toth 2004), the calibration system was designed to apply all 0000 UTC forecasts (only), including high- and low-resolution control forecasts, and all ensemble member forecasts for 24-h amounts at 2.5° resolution.

A frequency bias assessment is approached separately for the GFS high-resolution and ensemble control (low resolution) forecasts at each lead time because the model behaviors, especially for precipitation forecasts, depend closely on the model resolutions. Data are sampled from prior forecasts and observations with a 30-day average of the whole CONUS domain as the cold-start sampling. Later, the corresponding decaying weight used is 1/30 (or $W = 0.333$). The observations with 24-h accumulations come from the RFC rain gauge network, with about 10 000 observation station reports after regridding to the common 2.5° model grid. A set of thresholds of 0.2, 2.0, 5.0, 10.0, 15.0, 25.0, 35.0, 50.0, and 75.0 mm day^{-1} were selected for the 24-h accumulation amount to ensure the performance of the interpolation in the calibration procedure. The selection of thresholds is based on approximated similar distances of logarithm values, as well as on common thresholds in the daily applications. The values of CDF may be slight different from the selection of thresholds. The frequency bias assessment based on CONUS CDF may be applied to the global domain, when assuming that the frequency bias information over the CONUS is much the same as over other parts of the globe, which may not be an optimum application. This application can be improved when global precipitation observations become

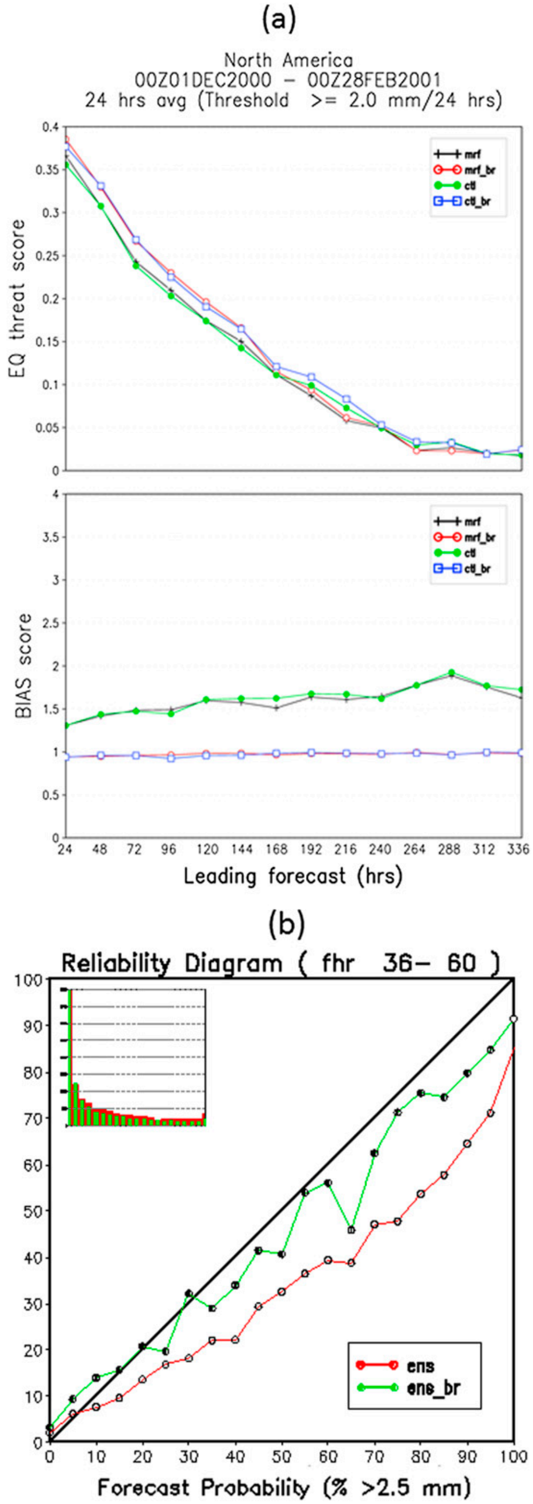

FIG. 4. Examples from the 2004 implementation. Results were selected for the period 1 Dec 2000–28 Feb 2001. (a) Averaged ETS and frequency bias scores of the raw GFS (mrf) and GEFS control (ctl) forecasts and their calibrated forecasts (mrf_br and ctl_br) at a threshold of 2.0 mm day^{-1}. (b) Reliability of the 2.5 mm day^{-1} GEFS raw (ens; red) and calibrated (ens_br; green) forecasts at 36–60-h lead time. The inset histogram denotes the frequency of forecast usage of each probability bin.

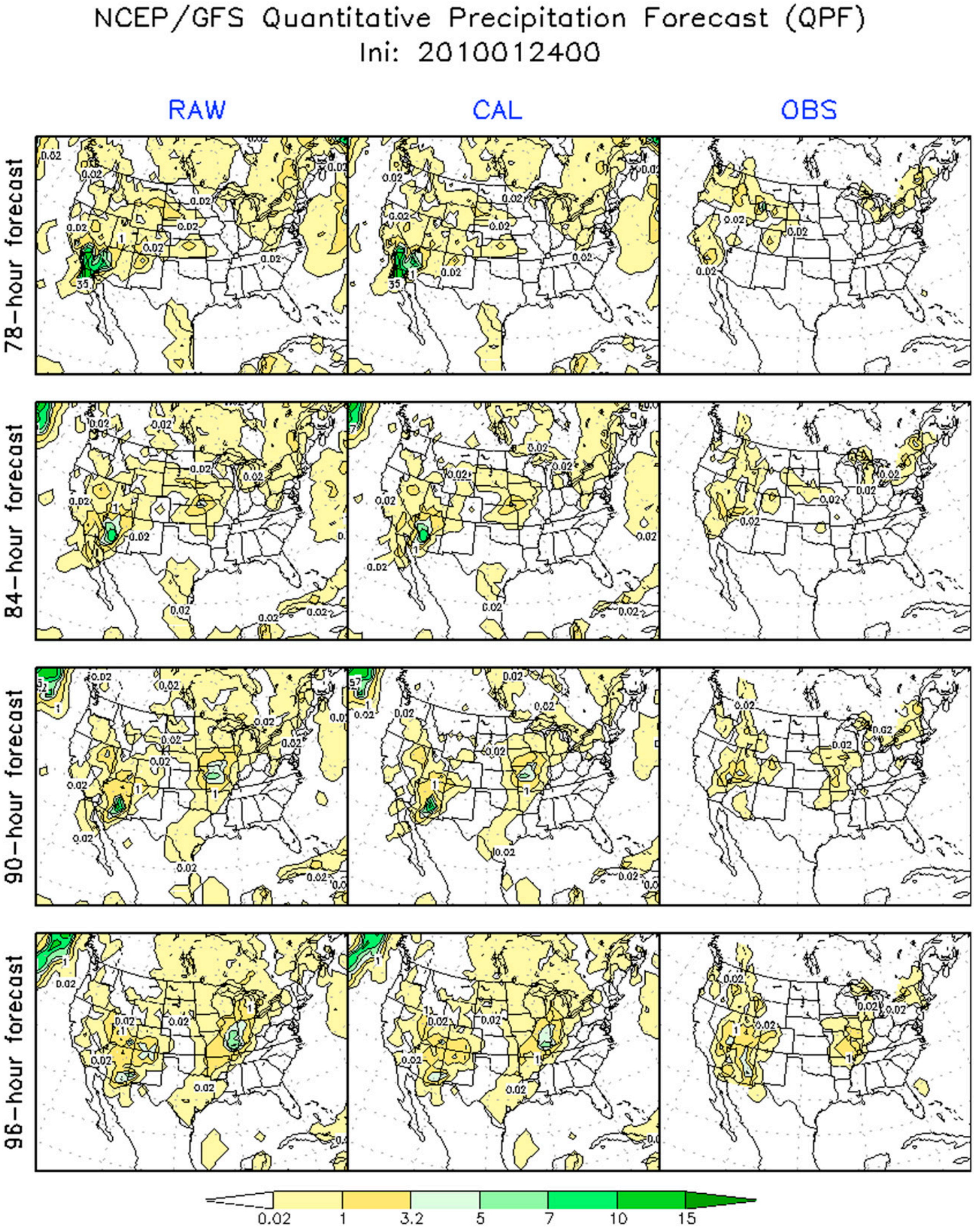

FIG. 5. Comparisons of 6-hourly accumulated precipitation (mm) initialized at 0000 UTC 24 Jan 2010 from the (left) raw GFS and (middle) calibrated forecasts against (right) the CCPA products that are valid at corresponding time periods.

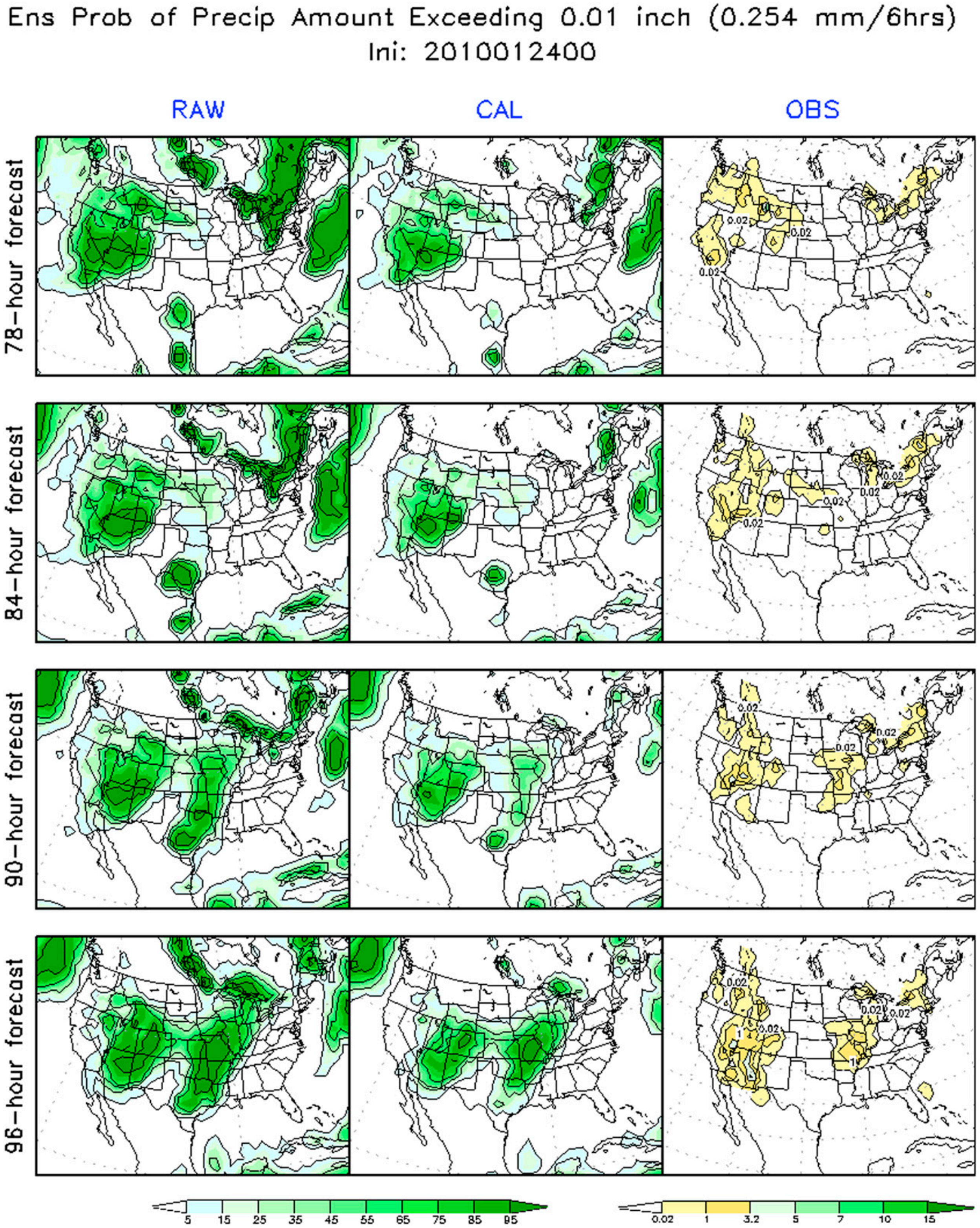

FIG. 6. (left) Raw GEFS probabilities, (middle) calibrated probabilities of the 6-h precipitation amount exceeding 0.01 in. initialized at 0000 UTC 24 Jan 2010, and (right) CCPA precipitation estimates for 6-h precipitation that are valid at the corresponding time periods.

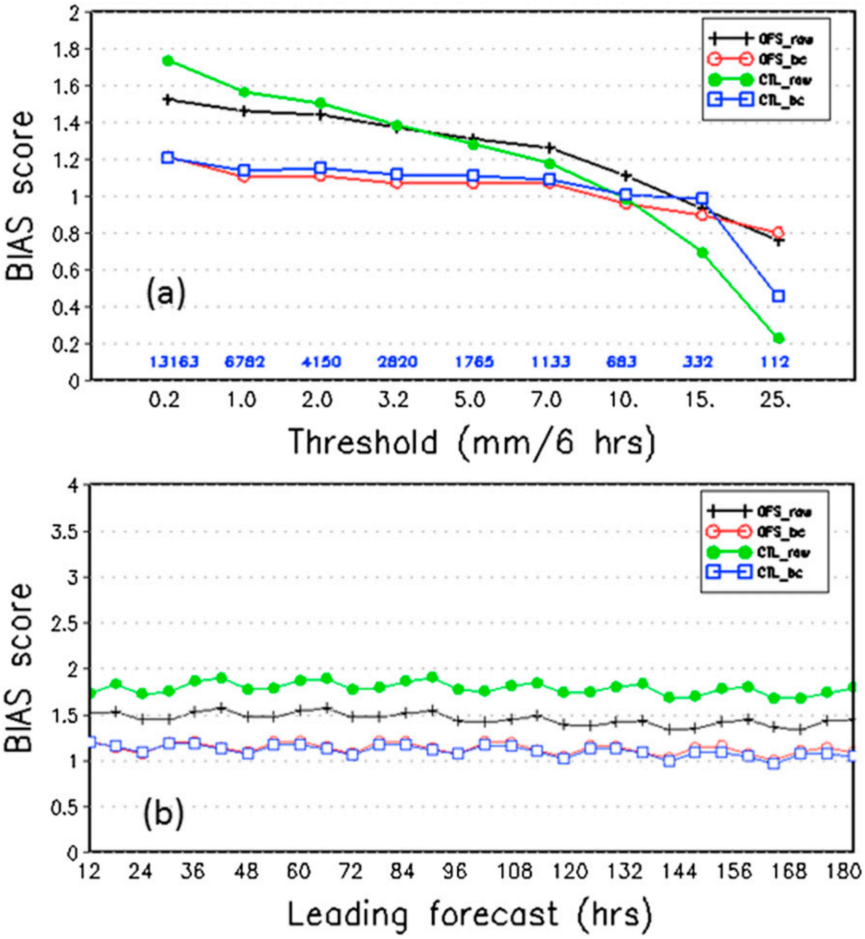

FIG. 7. Frequency bias scores of raw (GFS_raw, black; CTL_raw, green) and calibrated (GFS_bc, red; CTL_bc, blue) forecasts with increasing lead times for 6-h precipitation averaged between 1 Dec 2009 and 28 Feb 2010 (a) as a function of threshold (where the numbers in the plot above the x axis are for the total number of boxes verified) and (b) at a 0.2-mm threshold.

available in real time. The calibration system runs once daily at the 0000 UTC cycle and typically the daily runs are completed within a minute in real time on a supercomputer.

The evaluation period for this implementation was chosen to be from 1 December 2000 to 28 February 2001. Comparisons of the calibrated forecast against the raw forecast in terms of some scores were made and are shown in Fig. 4. Figure 4a presents equitable threat scores (ETSs) and frequency bias scores (Wilks 2006; Jolliffe and Stephenson 2003) at the 2.0-mm threshold for each forecast lead time. Figure 4b provides the 36–60-h reliability diagram at the 2.5-mm threshold for the 20-member ensemble forecast, validated for all grid points in the CONUS. A perfectly reliable forecast is shown by the diagonal line in Fig. 4b. The calibrated forecast shows a remarkably improved frequency bias score over the CONUS at all thresholds. Not only is the frequency bias reduced, but the postprocessing through frequency matching helped to improve the probabilistic forecast, as well (a full discussion of the probabilistic score/verification is presented in section 4). There was a much reduced ensemble mean frequency bias in the calibrated forecasts compared to the raw forecasts, indicating a great improvement in the reliability for higher-probabilistic forecasts (Fig. 4b), but not for lower-probability values, such as 5%. The reason for no improvement in lower-probability values is still unclear, however, this may affect overall reliability score due to high frequency of lower-probability events. For this particular winter period, the overall reliability scores are 0.016 for the raw forecast and 0.003 for the calibrated forecast. Ideally, the reliability curve approaches the diagonal line, which means the forecast is perfect or more reliable; a smaller value of overall reliability is better.

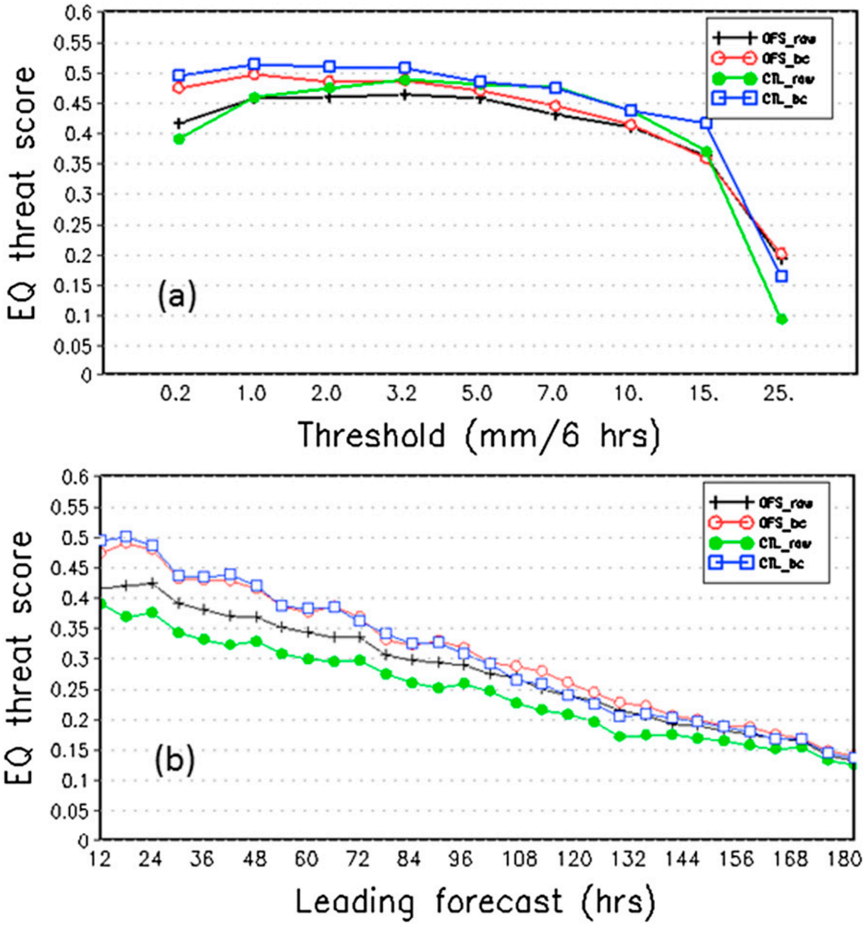

FIG. 8. As in Fig. 7, but for ETS.

4. Applications and evaluations of an improved frequency bias correction technique

In this section, earlier work is expanded to upgrade the calibration system and make it capable of frequency bias correction at higher temporal and spatial resolutions. More specifically, the current application works at 1° resolution with 6-h accumulations. This section describes how the higher-resolution precipitation forecasts are calibrated so that their cumulative frequency distribution matches that of the observations.

The operational NCEP GFS/GEFS forecast system runs 4 times per day (0000, 0600, 1200, and 1800 UTC) and produces 1° global ensemble precipitation forecast products for 6-h accumulations. The system contains 22 ensemble members: a high-resolution GFS run, a low-resolution GEFS control run, and 20 perturbed runs using the bred vector–ensemble transform with rescaling (BV-ETR) method (Wei et al. 2006, 2008). Technical information about NCEP's latest GEFS ensemble forecast system is available online (Zhu et al. 2012). Unlike in the 2004 implementation, here all 6-hourly 1° forecasts for the four cycles are directly calibrated with respect to the gridded precipitation analysis CCPA at the same resolution as the forecasts. To be more realistic and to better sample the statistics, 12 RFC CDFs are derived for each lead time to construct the cumulative frequency distributions. For each category among the nine thresholds [0.2, 1, 2, 3.2, 5, 7, 10, 15, and 25 mm $(6\,\mathrm{h})^{-1}$], a CDF is calculated as the number of grid points over each RFC where the forecasts or observed precipitation amounts are greater than the threshold. Again, to reduce the computational burden, only one set of CDFs is derived from the high-resolution GFS run, and another set of CDFs is developed from the low-resolution GEFS control run. Then, the latter set of CDFs is applied to the 20 ensemble members since all of them are low-resolution forecasts from the same forecast model, resulting in 2 rather than 22 sets of CDFs per lead time per threshold per RFC region. In each calibration run, there is a total of 6912 forecast–observation CDF pairs for 64 forecast lead times for the low-resolution runs and 3240 pairs for 30 forecast lead times for the high-resolution runs, summed

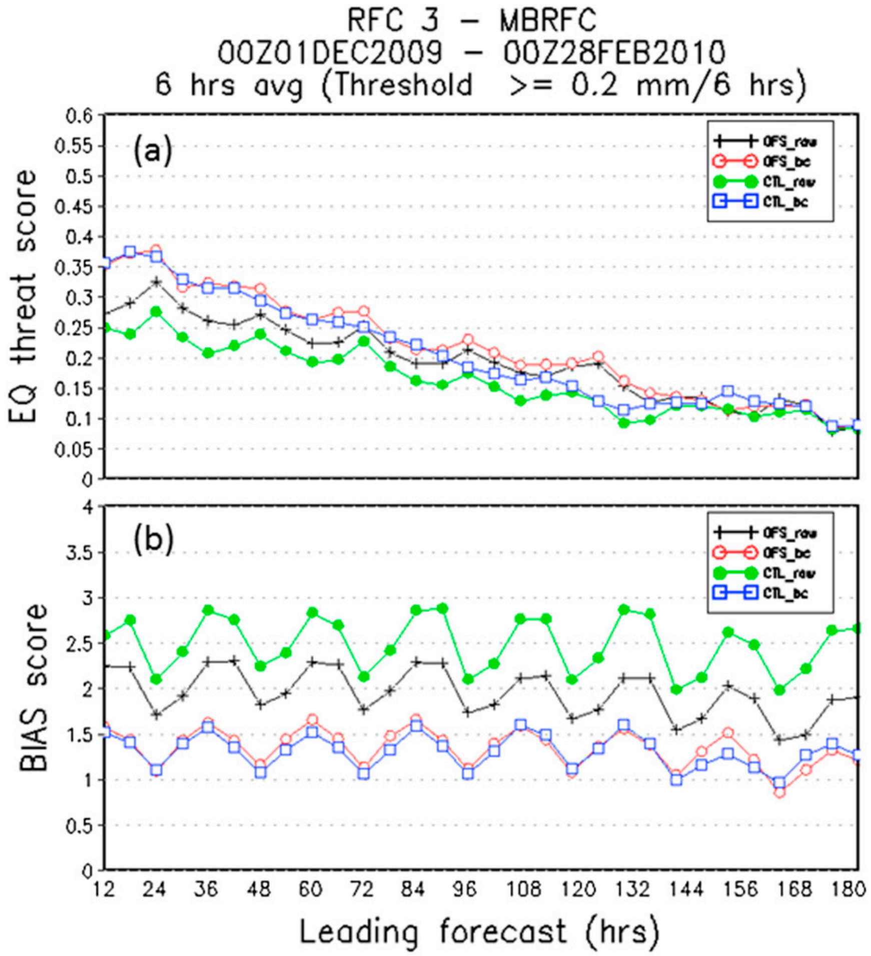

Fig. 9. Comparison of raw (GFS_raw, black; CTL_raw, green) and calibrated (GFS_raw, red; CTL_raw, blue) forecasts with increasing lead times for 6-h precipitation averaged between 1 Dec 2009 and 28 Feb 2010 and analyzed for the MBRFC region for (a) ETS and (b) frequency bias score at a 0.2-mm threshold.

for a total of 9 thresholds and 12 RFC regions. A decaying weight coefficient of 0.02 (or 2%) has been used to accumulate historical samples to build up CDFs of forecast and observation frequency counts (Fig. 1). The calibration process is applied 4 times per day to each 6-hourly forecast at each grid point globally and to each forecast lead time independently.

The operational forecasts initialized daily at 0000 UTC from 1 March 2009 through 28 February 2010 will be assessed. These forecasts produced with the same modeling suite were used to produce the calibrated forecasts. Both sets of forecasts will be examined out to 384 h with precipitation accumulation output available every 6 h. Although the method we developed can apply to global forecasts, in this study our evaluation domain is the CONUS, which allows evaluations of this method using the 1° CCPA dataset. The evaluation focuses on the frequency biases and skill levels of the calibrated ensemble precipitation forecasts with respect to raw forecasts. Several examples are analyzed and presented with some verification statistics. The verification statistics will be stratified by either lead time or threshold.

Figure 5 shows one application of this calibration for the high-resolution GFS forecast for the four selected forecast lead times (78, 84, 90, and 96 h). The comparison is of 6-hourly accumulated precipitation (mm) initialized at 0000 UTC 24 January 2010 for the raw GFS forecast (left), calibrated forecast (middle), and the observations (CCPA; right). Apparently, the GFS overforecast for the CONUS in general, and the calibrated forecast reduced the forecast amount accordingly. Figure 6 shows the ensemble PQPF (same time period) for the 0.254 mm $(6\,\mathrm{h})^{-1}$ threshold with the raw ensemble PQPF (left), the calibrated PQPF (or CPQPF; middle), and the observations (right). The forecast area of the PQPF is reduced; the quantity (value) of the PQPF is diminished

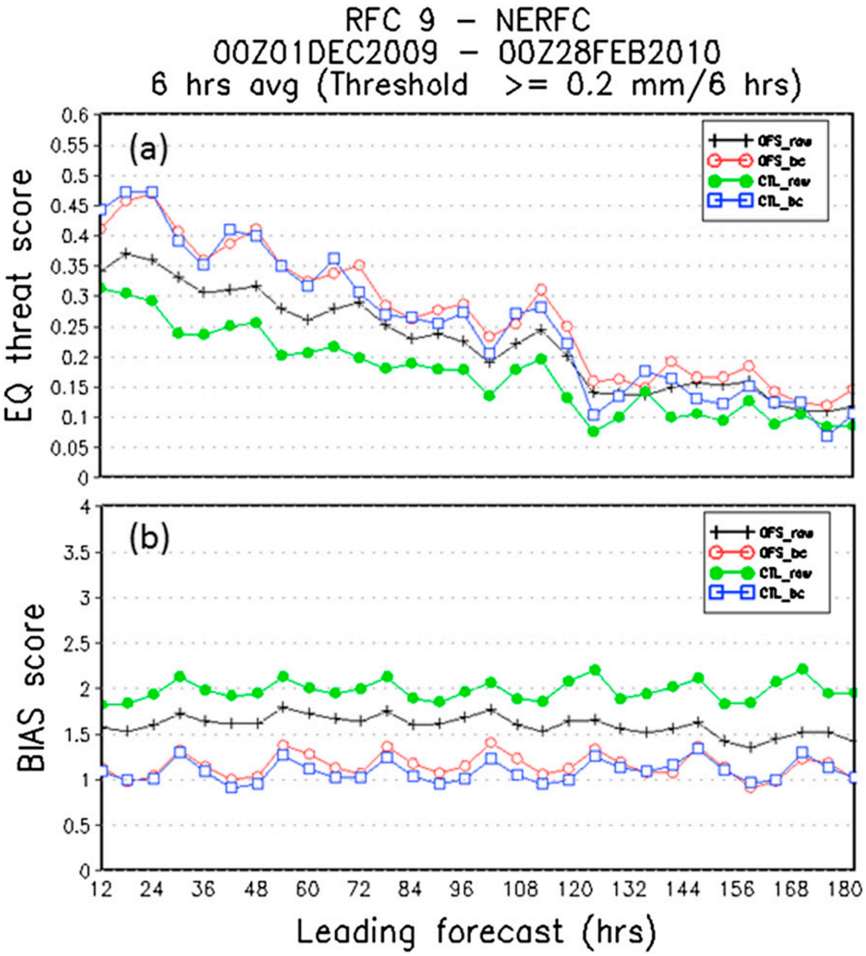

FIG. 10. As in Fig. 9, but for the NERFC region.

in the calibrated PQPF, which matches better with the observations from this case.

To demonstrate the benefits from this calibration, several different scores have been presented for the seasonal and yearly averages. The frequency bias scores and ETSs for the CONUS for the period 1 December 2009–28 February 2010 (winter season) are shown in Figs. 7 and 8. Figure 7a is for the 0–6-h forecast frequency bias of the different thresholds, and Fig. 7b shows forecast lead times out to 180 h for greater than 0.2 mm $(6\,h)^{-1}$. The numbers above the thresholds in Fig. 7a indicate the sample size of the 1° × 1° forecast box we have verified. Overall, the frequency bias is reduced and the ETS is increased in the calibrated forecasts for both the GFS (comparing GFS_raw to GFS_bc) and control (comparing CTL_raw to CTL_bc) for all lead times, and the improvement of ETS tends to be especially more effective for shorter lead times. Similar improvements in frequency bias scores and ETSs are also observed for the RFC regions, such as the Missouri basin RFC (MBRFC) and the Northeast RFC (NERFC) shown in Figs. 9 and 10, respectively, although they exhibit slightly larger diurnal variabilities. In the summer season, the forecast skill is very limited, especially for extended-range forecasts. The calibration of FMM could improve the frequency bias score, but not the ETS (figures not shown). For higher thresholds, the improvements are also limited due to fewer training samples and large seasonal variations (figures not shown).

The root-mean-square error (RMSE) and mean absolute error (MAE) across the CONUS for the period 1 March 2009–28 February 2010 (1 yr) for every 6-h accumulated precipitation forecast are shown in Figs. 11 and 12. Figure 11 is for the GFS forecast (RMSE_raw is for raw forecasts and RMSE_bc is for frequency-bias-corrected forecasts) and Fig. 12 is for the GEFS control forecast. Based on this year's worth of statistics, RMSE is reduced considerably for the GFS, but not for the (lower resolution) GEFS control. This difference might be related to the model resolution and version (the operational GFS model version is slightly different from GEFS for this period due to different implementation times). In

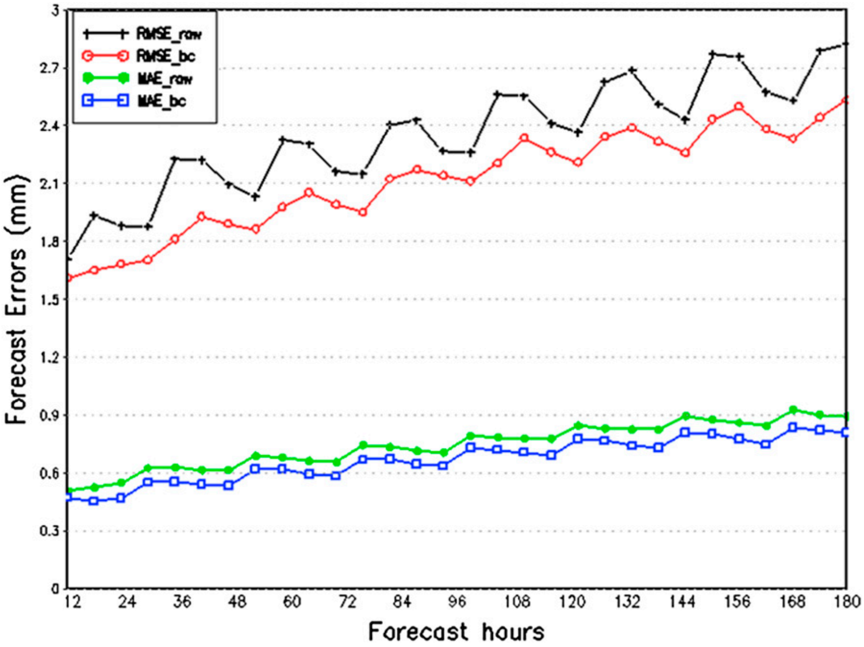

FIG. 11. RMSEs with increasing lead times for 6-h precipitation from the GFS high-resolution raw (RMSE_raw, black) and calibrated (RMSE_bc, red) forecasts, and MAEs with increasing lead times for 6-h precipitation from the GFS raw (MAE_raw, green) and calibrated (MAE_bc, blue) forecasts.

particular, the higher-resolution model produces larger errors compared to the lower-resolution version due to the resolution and forecast sharpness (Figs. 11 and 12). It may be better to separately verify the forecast intensity and pattern (or position). The results could be different if different verification methods are applied, such as the Method for Object-Based Diagnostic Evaluation (MODE; Davis et al. 2006a,b). However, the RMSEs are

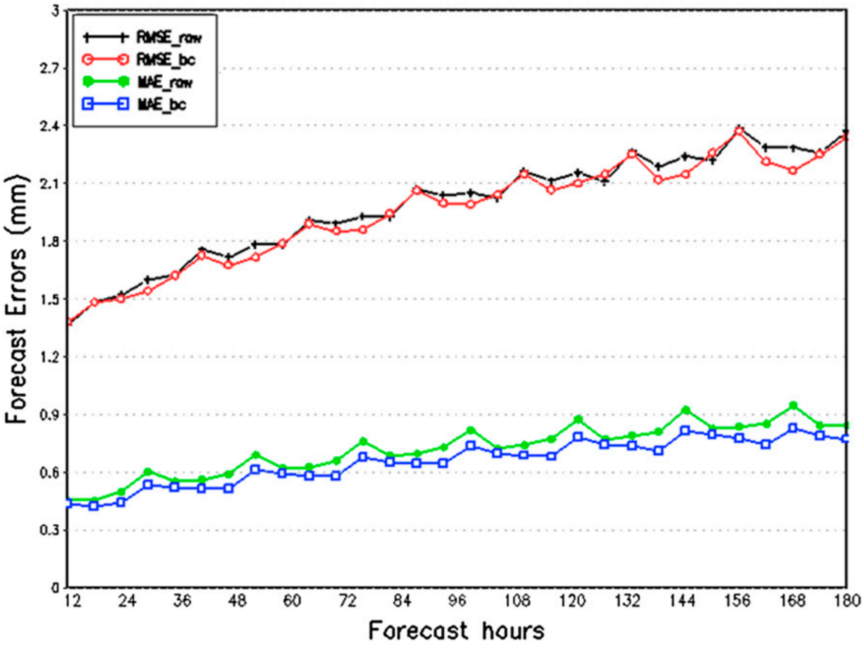

FIG. 12. As in Fig. 11, but for GEFS CTL forecasts.

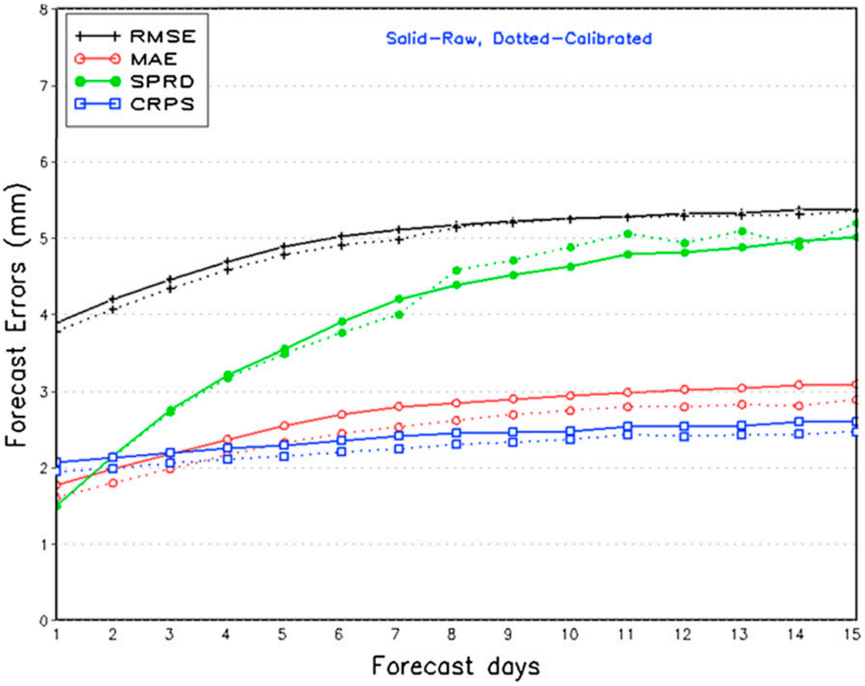

FIG. 13. The RMSE (black), MAE (red), SPRD (green), and CRPS (blue) with increasing lead times for 24-h precipitation from the GEFS ensemble mean (RMSE and MAE) and members (spread and CRPS) for raw (solid lines) and calibrated (dotted lines) forecasts.

very similar after calibration for both the higher- and lower-resolution model forecasts. Meanwhile, for this 1 yr of statistics, MAE is reduced for both the GFS and GEFS control runs at all lead times.

For the ensemble forecast, RMSE and MAE of the ensemble mean, ensemble spread (SPRD; defined as the ensemble standard deviation from its mean), and continuous ranked probability score (CRPS; Hersbach 2000; Zhu and Toth 2008) have been calculated for the period 1 March 2009–28 February 2010 and are displayed in Fig. 13. This is a year's worth of verification against CCPA for every 24-h accumulated precipitation forecast. The results indicate that 1) the RMSE is marginally reduced (similar to the ensemble control in Fig. 11) and the MAE for the ensemble mean is reduced, too; 2) CRPS is improved; and 3) ensemble spread is increased for longer-lead-time forecasts with larger forecast errors, but is the same (or with less change) for short-lead-time forecasts, which may be due to reduced RMSE. The improved spread and CRPS could be explained as a by-product of the frequency-matching method. The algorithm not only matches the precipitation frequency (reducing the frequency bias), but also adjusts the amount of precipitation forecast by each ensemble member (adjusting the distribution). A comparison of the Brier scores (Wilks 2006; Jolliffe and Stephenson 2003) between the raw and calibrated forecasts is also shown in Fig. 14. The Brier score is negatively oriented, which means the smaller the score value the better the results. As expected, the score is reduced after frequency bias correction (dotted curves) for all lead times.

5. Conclusions and future plans

The frequency-matching method is developed and applied to the NCEP QPF and PQPF forecasts for the first precipitation calibration since 2004. The latest version will be implemented in 2013 for finer temporal (every 6 h out to 16 days) and spatial (1° × 1°) resolutions. The prior CDFs of the forecasts and observation can be easily generated from the GFS/GEFS precipitation forecast and CCPA through applying the Kalman filter method (or decaying average). In real-time operations, a postprocessing with the frequency bias correction technique is carried out to produce calibrated ensemble precipitation products soon after raw forecast outputs are produced. In the postprocessing procedure, the frequency bias statistics that include all CDF values both for observations and forecasts are updated in terms of decaying weight on the daily basis to catch as much of the latest frequency bias information as possible.

The performance of this method has been investigated with respect to a year's worth of operational GEFS

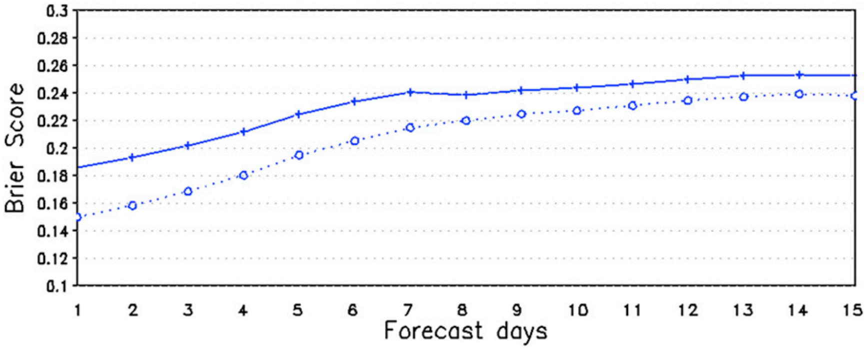

FIG. 14. The Brier score at a 0.2-mm threshold with increasing lead times for 24-h precipitation from the GEFS ensemble mean raw (solid line) and calibrated (dotted line) forecasts.

precipitation products. Results show that model frequency bias has been effectively reduced and some skill (ETS) scores have been improved in the calibrated forecasts. The good performance of the frequency bias correction is obviously due to the fact that it can dynamically catch systematic model biases in most cases. The frequency bias correction works effectively for both the higher- and lower-resolution model forecasts and for wet bias elimination. Another attractive advantage of this method is that it saves a significant amount of both computer and human resources. Unlike other statistical postprocessing methods, it is not heavily reliant on a huge amount of data for model training, so it takes up much less disk space on the computer systems and is able to update the model frequency bias when a model is upgraded as well.

One important issue to consider is that the method has a limitation in its frequency bias correction for extreme events. Such seldom-happening extreme events may not be captured by the current short-term data pooling scheme. The method may not perform well for extreme events, unlike the analog method that makes use of an extensively long training set of reforecast data. Another important issue is the validity of the frequency-matching method. The method used in this study is based on a certain knowledge of model information drawn from past verification statistics. Remember that as mentioned in section 2, this method is not perfect as it is unable to make adjustments to areas that have no precipitation. As a result, there is no adjustment of the forecast probability of precipitation (PoP) although the observed PoP is possibly higher; therefore, this kind of dry bias can never be removed, though this is also the case with other traditional precipitation bias correction methods. Generally, model forecasts include two kinds of errors; intensity and pattern errors. This method appears to have a positive impact on intensity-error-dominated cases. However, it has a neutral or negative impact on pattern-error-dominated cases (Fig. 12), causing a poorer sampling of frequency bias information. In this case, frequency bias is reduced at the expense of an increase in random error. Further investigation is needed to fully understand the performance of this method and to determine where and when it has a significantly positive impact and the usefulness of the calibrated products.

In this study, the decaying average weight is constantly selected as 0.02 (except for the 2000–01 application) for all lead times. Actually, the decaying average weights really depend on the experiment and could range from 0.01 to 0.5. In general, the weight is varied for different forecast lead times; a larger weight is good for short lead times, which can catch up quick moving systems, and a smaller weight is more favorable for long-lead-time forecasts (not shown). Therefore, choosing an optimum weight for each lead time could be a constructive way to improve the calibration system in the future. Meanwhile, the weight is varied for geographical locations and seasons. There are two improvements we are expecting to validate through future study. One is an optimum weight, which will need large samples for experiments. The weights should be a function of lead time, location, and season. The second is a downscaling process to produce a much finer–resolution forecast (5- and 2.5-km resolutions).

Acknowledgments. The authors thank Drs. Bo Cui and Dingchen Hou and the other members of the ensemble and postprocessing team at NCEP/EMC for helpful suggestions during this work. The authors also thank Mr. Binbin Zhou for useful comments. The first author gratefully acknowledges Drs. Zoltan Toth and Stephen Lord for their long-time support of this calibration work. We also thank Ms. Mary Hart (NCEP) and Dr. Glenn White for editing our English and three anonymous reviewers for many valuable comments.

REFERENCES

Bentzien, S., and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting*, **27**, 988–1002, doi:10.1175/WAF-D-11-00101.1.

Brown, J. D., D. J. Seo, and J. Du, 2012: Verification of precipitation forecasts from NCEP's Short Range Ensemble Forecast (SREF) system with reference to ensemble streamflow prediction using lumped hydrologic models. *J. Hydrometeor.*, **13**, 808–836, doi:10.1175/JHM-D-11-036.1.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, doi:10.1175/WAF-D-11-00011.1.

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, doi:10.1175/MWR3145.1.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, doi:10.1175/MWR3146.1.

Demargne, J., and Coauthors, 2014: The science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, doi:10.1175/BAMS-D-12-00081.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, doi:10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132–1147, doi:10.1175/1520-0434(1998)013<1132:CPQPFB>2.0.CO;2.

Fundel, F., A. Walser, M. A. Liniger, C. Frei, and C. Appenzeller, 2010: Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts. *Mon. Wea. Rev.*, **138**, 176–189, doi:10.1175/2009MWR2977.1.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, doi:10.1175/MWR3237.1.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:10.1175/2007MWR2411.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, in press.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 240 pp.

Krzysztofowicz, R., and A. A. Sigrest, 1999: Calibration of probabilistic quantitative precipitation forecasts. *Wea. Forecasting,* **14**, 427–442, doi:10.1175/1520-0434(1999)014<0427:COPQPF>2.0.CO;2.

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at http://ams.confex.com/ams/pdfpapers/83847.pdf.]

Marty, R., I. Zin, and Ch. Obled, 2013: Sensitivity of hydrological ensemble forecasts to different sources and temporal resolutions of probabilistic quantitative precipitation forecasts: Flash flood case studies in the Cévennes–Vivarais region (southern France). *Hydrol. Processes*, **27**, 33–44, doi:10.1002/hyp.9543.

Mascaro, G., E. R. Vivoni, and R. Deidda, 2010: Implications of ensemble quantitative precipitation forecast errors on distributed streamflow forecasting. *J. Hydrometeor.*, **11**, 69–86, doi:10.1175/2009JHM1144.1.

Primo, C., C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, 2009: Calibration of probabilistic forecasts of binary events. *Mon. Wea. Rev.*, **137**, 1142–1149, doi:10.1175/2008MWR2579.1.

Rosenfeld, D., D. B. Wolff, and D. Atlas, 1993: General probability-matched relations between radar reflectivity and rain rate. *J. Appl. Meteor.*, **32**, 50–72, doi:10.1175/1520-0450(1993)032<0050:GPMRBR>2.0.CO;2.

Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 1603–1627, doi:10.1175/2010WAF2222367.1.

Wei, M., Z. Toth, R. Wobus, Y. Zhu, C. H. Bishop, and X. Wang, 2006: Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus*, **58A**, 28–44, doi:10.1111/j.1600-0870.2006.00159.x.

——, ——, ——, and ——, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP Global Operational Forecast System. *Tellus*, **60A**, 62–79, doi:10.1111/j.1600-0870.2007.00273.x.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 648 pp.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network. *Wea. Forecasting*, **22**, 1287–1303, doi:10.1175/2007WAF2006114.1.

——, J. A. McGinley, P. J. Schultz, C. J. Anderson, and C. Lu, 2008: Short-range precipitation forecasts from time-lagged multimodel ensembles during the HMT-West-2006 campaign. *J. Hydrometeor.*, **9**, 477–491, doi:10.1175/2007JHM879.1.

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability. *Adv. Atmos. Sci.*, **22**, 781–788, doi:10.1007/BF02918678.

——, and Z. Toth, 1999: Calibration of probabilistic quantitative precipitation forecasts. Preprints, *17th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 214–215.

——, and ——, 2004: May 2004 implementation of bias-corrected QPF and PQPF forecasts. NOAA/NWS/Environmental Modeling Center. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.]

——, and ——, 2008: Ensemble based probabilistic forecast verification. *19th Conf. on Predictability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 2.2. [Available online at https://ams.confex.com/ams/pdfpapers/131645.pdf.]

——, ——, E. Kalnay, and S. Tracton, 1998: Probabilistic quantitative precipitation forecasts based on the NCEP global ensemble. Preprints, *12th Conf. on Numerical Weather Prediction*, Phoenix, AZ, Amer. Meteor. Soc., 286–289.

——, D. Hou, M. Wei, R. Wobus, J. Ma, B. Cui, and S. Moorthi, cited 2012: GEFS upgrade—AOP plan—major implementation. NOAA/NWS/Environmental Modeling Center. [Available online at http://www.emc.ncep.noaa.gov/gmb/yzhu/html/imp/201109_imp.html.]