Bias Correction for Global Ensemble Forecast

BO CUI

I. M. Systems Group, NOAA/NWS/NCEP/Environmental Modeling Center, Camp Springs, Maryland

ZOLTAN TOTH

NOAA/ESRL/Global Systems Division, Boulder, Colorado

YUEJIAN ZHU AND DINGCHEN HOU

NOAA/NWS/NCEP/Environmental Modeling Center, Camp Springs, Maryland

(Manuscript received 12 January 2011, in final form 8 September 2011)

ABSTRACT

The main task of this study is to introduce a statistical postprocessing algorithm to reduce the bias in the National Centers for Environmental Prediction (NCEP) and Meteorological Service of Canada (MSC) ensemble forecasts before they are merged to form a joint ensemble within the North American Ensemble Forecast System (NAEFS). This statistical postprocessing method applies a Kalman filter type algorithm to accumulate the decaying averaging bias and produces bias-corrected ensembles for 35 variables. NCEP implemented this bias-correction technique in 2006. NAEFS is a joint operational multimodel ensemble forecast system that combines NCEP and MSC ensemble forecasts after bias correction. According to operational statistical verification, both the NCEP and MSC bias-corrected ensemble forecast products are enhanced significantly. In addition to the operational calibration technique, three other experiments were designed to assess and mitigate ensemble biases on the model grid: a decaying averaging bias calibration method with short samples, a climate mean bias calibration method, and a bias calibration method using dependent data. Preliminary results show that the decaying averaging method works well for the first few days. After removing the decaying averaging bias, the calibrated NCEP operational ensemble has improved probabilistic performance for all measures until day 5. The reforecast ensembles from the Earth System Research Laboratory's Physical Sciences Division with and without the climate mean bias correction were also examined. A comparison between the operational and the bias-corrected reforecast ensembles shows that the climate mean bias correction can add value, especially for week-2 probability forecasts.

1. Introduction

Over the last decade, a global forecast model–based global ensemble forecast system [such as the National Centers for Environmental Prediction's (NCEP) Global Ensemble Forecast System (GEFS)] has been found to be useful for medium-range probabilistic forecasting. Ensemble forecasting has been embraced as a practical way of estimating the uncertainty of weather forecasts and of making probabilistic forecasts (Toth and Kalnay 1993, 1997; Molteni et al. 1996; Houtekamer et al. 1996). However, ensemble forecasts still suffer from model and ensemble formation related shortcomings (i.e., imperfect model physics, initial conditions, and boundary conditions for regional ensembles). As Toth et al. (2003) indicated, ensemble forecasts contain systematic errors and these systematic errors remain and cause biases in the first and second moments of the ensemble distribution. To make a skillful medium-range forecast, it is necessary to run postprocessing algorithms to remove these systematic errors before the ensemble forecasts can be used.

A large variety of numerical weather prediction postprocessing methods have been proposed and tested by many investigators (Gel 2007; Hacker and Rife 2007; Yussouf and Stensrud 2006; Cheng and Steenburgh 2007). These techniques are designed for deterministic forecasts and work well for near-surface variables. There

Corresponding author address: Dr. Bo Cui, NOAA/NWS/NCEP/Environmental Modeling Center, 5200 Auth Rd., Camp Springs, MD 20746.
E-mail: bo.cui@noaa.gov

DOI: 10.1175/WAF-D-11-00011.1

have been many attempts to postprocess ensemble forecasts to provide reliable probability forecasts. Recent work includes ensemble model output statistics (Gneiting et al. 2005), gene-expression programming (Bakhshaii and Stull 2009), and the Bayesian model averaging method (Raftery et al. 2005; Krzysztofowicz and Evans 2008). These methods gained broad attention in the ensemble postprocessing community. However, these techniques are still in development. The need remains for ensemble statistical postprocessing. Further research in this direction is desirable.

Postprocessing of ensemble forecasts is a necessary and important step for the daily operational runs at numerical weather prediction centers. Reliability, accuracy, and efficiency are the most important issues for daily operations. In this paper, we first introduce a statistical postprocessing algorithm to adjust the first moment of ensemble forecasts. This statistical postprocessing method applies an adaptive [Kalman filter type (KF)] algorithm to accumulate the decaying averaging bias. In statistics the Kalman filter is a mathematical method named after R. E. Kalman (Kalman 1960). It is mainly used to estimate system states that can only be observed indirectly or inaccurately by the system itself and its process is carried out iteratively. The estimates produced by this method tend to be closer to the true values than the original measurements because the weighted average has a better estimated uncertainty than either of the values that went into the weighted average (Kalman 1960). The basic ideas of KF are straightforward and the KF method turns out to be useful for many applications in science, engineering, and economics. We design a specific algorithm based on the KF concept to estimate ensemble forecast errors and we call the bias estimation and correction process the decaying averaging method.

This method was developed by the National Weather Service (NWS) at NCEP and was implemented operationally in 2006 at NCEP to reduce the bias of the NCEP and Canadian Meteorological Centre (CMC) ensemble forecasts. The calibrated NCEP and CMC global ensembles are then are merged to form a joint ensemble within the North American Ensemble Forecast System (NAEFS; Zhu et al. 2012b). NAEFS is an operationally joined multimodel ensemble forecast system, which combines the NCEP and CMC ensemble forecasts after bias correction (Zhu et al. 2012a; Zhu and Cui 2012).

We also test three different calibration experiments that are designed to assess and mitigate ensemble biases in the first (mean) moment of the ensemble on the model grid with respect to analysis fields: a decaying averaging bias calibration method with a short sample, a climate mean bias calibration method, and a bias calibration method using dependent data. The decaying averaging bias calibration method with a short sample is similar to the technique implemented in NCEP's GEFS, but uses training data for a fixed long period. The second calibration method uses a climate mean bias to do the bias correction, which is supported by a 25-yr ensemble reforecast experiment. This reforecast experiment is another ensemble run operationally at NCEP with a frozen analysis/modeling system developed by scientists at the Earth System Research Laboratory's Physical Sciences Division (ESRL/PSD; formerly the Climate Diagnostics Center, CDC). The ESRL/PSD reforecast is run operationally to produce a dataset of historical weather forecasts generated with a fixed numerical model, the 1998 version of NCEP's Global Forecast System (GFS; http://www.esrl.noaa.gov/psd/forecasts/reforecast/). A reforecast for each day since 1979 has been made with this GFS version, which is composed of a 15-member ensemble forecast run out to 15 days (Hamill et al. 2004, 2006). From the collected 25 yr of reforecast data, climate mean forecast errors are diagnosed and the reforecast data are calibrated by removing these errors to increase the reforecast ensemble skill. The design and usages of the second method aim at taking advantage of week-2 forecasts from the long historical reforecast data.

The purpose of this study is to introduce and compare several statistical postprocessing methods to assess and mitigate ensemble biases in the first (mean) moment of the ensemble. The decaying averaging method NCEP runs operationally is described in section 2. How and why a specific decaying parameter is chosen will be discussed. The three different experiments [i.e., the decaying average method with a short sample, the ESRL/PSD reforecast calibration method, and the bias-correction method using dependent data (optimal calibrated ensemble)] are described in section 3. Some statistical evaluation methods used in this paper are also reviewed in section 3. Section 4 contains the evaluation of the several calibration methods. Results from the calibrated NCEP/GEFS and CMC/GEFS datasets will be compared with the raw NCEP and CMC ensembles to evaluate the performance of the operational bias-correction method. The results from the three calibration experiments (i.e., the decaying averaging bias correction of NCEP/GEFS, the climate mean bias correction of ESRL/PSD reforecast, and the optimal calibrated ensembles) are also compared. The relative merits of using the current best analysis/modeling system with a small sample, versus the merits of an older and frozen analysis/modeling system that has a long forecast sample for the bias correction, will be examined. In general, these two calibration methods for the NCEP/GEFS and ESRL/PSD reforecast ensembles are incompatible because of their different uses of model systems. However, these comparisons are not of the superiority of one method over another but to help us illustrate the possibility of improving the current operational decaying averaging method to improve week-2 forecasts. Finally, the preliminary conclusions of this research and our future plans are summarized in section 5.

2. The design of the NCEP bias-correction method—Decaying average

NCEP implemented a statistical postprocessing algorithm (i.e., the decaying averaging bias-correction method) to calibrate global ensemble forecasts in 2006. The bias-correction method is applied to the NCEP and CMC global ensembles. The operational NCEP/GEFS system configuration is described in Toth et al. (2012) and Wei et al. (2008). The 20-member ensembles are produced at T126L28 horizontal and vertical resolutions. The perturbations of the initial conditions are from the ensemble transform with rescaling (ETR) technique. The operational CMC ensemble is described in Charron et al. (2010). A single dynamical core [i.e., the Global Environmental Multiscale (GEM) model] is used to produce the ensembles. A multiparameterization approach and stochastic perturbations are used in order to sample model error for the 20 members of the ensembles. Due to the different ensemble configurations, the bias estimation and correction processes of NCEP/GEFS and CMC/GEFS will be adjusted to meet their specific characteristics.

The operational environment requires that the ensemble postprocessing algorithms be relatively applicable and flexible for implementation. The decaying averaging method applies an adaptive algorithm, and its application includes two steps. The first step is to estimate the first-moment bias with respect to the analysis field, which is called the decaying averaging mean error. The second step is to remove the error from the ensemble forecasts. Both the bias assessment step and the bias-correction step are carried out separately at each forecast lead time, on each individual grid point and for each initial cycle.

a. Bias estimation

The bias $b_{i,j}(t)$ for each lead-time t (a 6-h interval up to 384 h), and each grid point (i, j), is defined as the difference between the analysis $a_{i,j}(t)$ and forecast $f_{i,j}(t)$ at the same valid time t_0, on the latest available analysis:

$$b_{i,j}(t) = f_{i,j}(t) - a_{i,j}(t). \qquad (1)$$

b. Decaying average

The average bias $B_{i,j}(t)$ will be updated by considering the prior period bias $B_{i,j}(t-1)$ and current bias $b_{i,j}(t)$ by using the decaying average with the weight coefficient w:

$$B_{i,j}(t) = (1 - w)B_{i,j}(t - 1) + wb_{i,j}(t). \qquad (2)$$

This decaying average bias estimation method is a convenient way to consider the most recent behavior of weather systems. Once initialized, the bias estimate can be updated by considering just the current forecast error with regard to the stored bias fields. The weight factor w controls how much influence to give the most recent data. A w equal to 2% is used for the NCEP/GEFS and CMC/GEFS bias accumulations, which include mainly the past 50–60 days of information (Fig. 1). Experiments with a choice of 2% weight and other values (0.25%, 0.5%, 1%, and 10%, respectively) have been conducted. The details will be discussed in section 4.

c. Bias correction

The new bias-corrected forecast $F_{i,j}(t)$ will be generated by applying the decaying average bias $B_{i,j}(t)$ to current forecasts $f_{i,j}(t)$ at each lead time and each grid point:

$$F_{i,j}(t) = f_{i,j}(t) - B_{i,j}(t). \qquad (3)$$

Steps 1–3 allow users to accomplish the bias-correction procedure for both NCEP/GEFS and CMC/GEFS. Note that this procedure contains two options. The first is that the NCEP/GEFS and CMC/GEFS can be grouped together before postprocessing, and then the bias correction is applied to the joint ensemble. The second option is to apply the bias correction to the NCEP and CMC ensembles separately and then the NCEP/GEFS and CMC/GEFS are grouped together after postprocessing. Although the first option is an easy approach, it may not provide the best results since each participating ensemble may have unique biases due to its own ensemble generation configuration. Specific treatments are needed for each participating ensemble. NCEP uses one model and perturbed initial conditions to create its ensemble (Toth et al. 2012; Wei et al. 2008). The model-related systematic errors grow with lead time and it is assumed that the forecast errors obtained from the ensemble mean can stand in for the systematic errors. The NCEP/GEFS biases are estimated from the ensemble mean with respect to the NCEP analysis, and the same bias estimation is applied to each ensemble member during the calibration. On the other hand, the Canadian ensemble includes 20 perturbed forecasts and 1 control forecast. All are performed with the GEM model but use different physics parameterizations, data assimilation cycles, and sets of perturbed observations (http://www.weatheroffice.gc.ca/ensemble/index_e.html). Therefore, the individual bias is estimated and used to correct each individual member

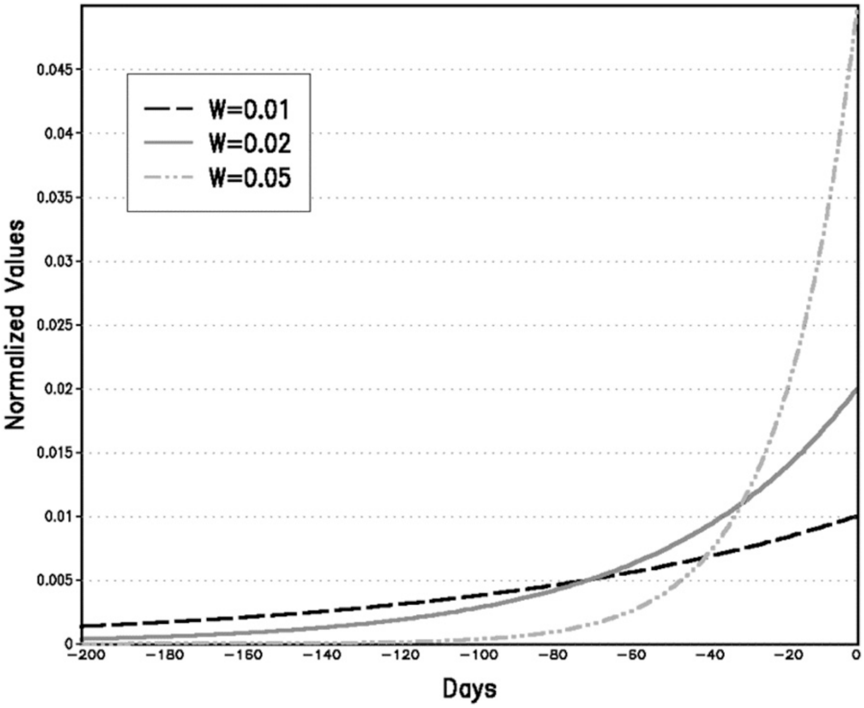

FIG. 1. Historical information (days) used for different decaying average weights (0.01, 0.02, and 0.05). Accumulated area (under the curve) is equal to 1.0.

independently. For ease and efficiency, each participating ensemble calibrates its raw forecast against its own analysis.

The bias-correction procedure generates ensemble output on 1° × 1° latitude–longitude grids for 35 selected variables. Table 1 lists all the variables that are postprocessed for both the NCEP/GEFS and CMC/GEFS. The selection of these variables depends heavily on the assumption that the forecast variable is well represented by the Gaussian distribution. It will not work very well for non-Gaussian-distributed variables, such as precipitation. A new method is required for non-Gaussian-distributed variables.

3. Experiments for comparing the usage of reforecast information

In addition to the decaying averaging bias-correction method used operationally for the NCEP/GEFS and CMC/GEFS, three other postprocessing methods were designed in this study to assess and mitigate ensemble biases in the first moment of the ensemble. The discrepancies among these three methods are a way of estimating bias, which are through using decaying averaging with a short sample, climate mean errors from the ESRL/PSD reforecast, and from dependent data, respectively. Discrepancies contribute to the error estimation and the differences show the amount of improvement.

a. *Bias estimation from decaying averaging with short training data*

The first method applies the Kalman-filter-type algorithm to get bias estimations through the following procedure: First, a prior estimate starts the procedure. At a given day T, we calculate the time mean forecast errors between days $T - 46$ and $T - 17$ to create an initial average. Second, the average is updated by setting it to the weighted average of the new forecast error at day $T - 16$ with a weight of w, and to the previous average with a weight of $1 - w$ $(0 \leq w < 1)$. Third, we repeat the second step every day from day $T - 15$ to $T - 1$; we call this cycling. Experiments with different decay weights w (1%, 2%,

TABLE 1. List of postprocessed variables for the NCEP/GEFS and CMC/GEFS ensembles.

Ensemble	CMC/GEFS (20 members) and NCEP/GEFS (20 members, control, and GFS)
Grid	1° × 1°
Domain	Global
Format	World Meteorological Organization (WMO) gridded binary (GRIB) format
Hours	6 hourly out of 384 h
GZ, TT, *U*, *V*	200, 250, 500, 700, 850, 925, 1000 hPa
MSLP, surface pressure	Mean sea level pressure, surface pressure
TT, Tmax, Tmin, *U*, *V*	2im temp, 2-m max and min temperature, 10-m *U* and *V*

and 10%, respectively) were conducted and a detailed discussion of these results can be found in Section 4.

This method is different from the operational decaying averaging technique as it chooses a fixed-length period of training data. The application of this method is not applicable in operations since it requires a huge amount of disk space to save 46 days of forecasts online. The design and testing of this method was done before the final operational technique was selected. The current operational decaying averaging technique is adapted from it and is much simpler to implement. The bias estimate in operations is carried out iteratively, which takes the use of a long training dataset but does not require this extra dataset to be saved onto disk. The operational method is more flexible in practice and avoids the issue involved with the disk space required to save data for days T–46 through T − 1. However, results from the bias-corrected ensemble using 46 days of training data are still useful. Comparisons between this approach and the other calibration techniques will help to identify their strengths and weaknesses.

b. *Climate mean bias estimation from ESRL/PSD reforecast ensemble*

A second method of assessing the bias is by using the climatological mean forecast error, which is obtained from the ESRL/PSD 25-yr reforecast ensemble (from 1978 to 2003). Hamill et al. (2004) thought that it was not effective to perform a bias correction with only a short set of prior forecasts because systematic errors may not be well established if only a few cases are used in the tests, but that errors may be more obvious with the larger sample afforded by a reforecast. With the model output statistics (MOS) techniques (Glahn and Lowry 1972; Carter et al. 1989; Vislocky and Fritsch 1995) and a frozen forecast model, their results show that dramatic improvements in medium- to extended-range probabilistic forecasts are possible by using retrospective forecasts. Motivated by their success, especially for the week-2 probabilistic forecast, we introduced the climatological mean forecast error into our bias correction and utilized it as the bias estimate. Following the first-moment bias-correction procedure mentioned above, the climatological mean forecast error is removed from the ESRL/PSD reforecast for each forecast lead time and individual grid point. The reforecast ensembles with and without the climatological bias correction are then examined and compared to the NCEP operational ensembles calibrated from decaying averaging with short training data. Please note that it is only for convenience that we classify the two ensemble datasets as the operational and reforecast ensembles, because the reforecast is also being run operationally at ESRL/PSD.

c. *Bias estimate using dependent data*

A third way of estimating the first-moment bias of the ensemble is through the calculation of a 31-day running-mean forecast error centered on day *T*. The implementation of this method is not feasible operationally but is used as an optimal benchmark. The optimal scenarios, therefore, are compared to the raw and calibrated ensembles to show how large the improvement in the ensemble forecast could possibly be when using the first-moment adjustment technique.

The three bias-correction techniques discussed above are applied to the NCEP/GEFS operational and ESRL/PSD reforecast ensembles, respectively. The bias estimation and bias correction are carried out separately at each forecast lead time and for each individual grid point. The bias correction is applied to all ensemble member forecasts. The fields studied include the NCEP/GEFS and ESRL/PSD reforecast ensemble 500-hPa geopotential heights and 850-hPa temperatures for the period 1 March 2004–28 February 2005 for the 0000 UTC initial cycle. Other calibrated fields available from the operational ensemble include the 2-m temperature and 10-m U and V components (not shown in this paper). Each data source creates three different ensembles: the raw, bias-corrected, and optimal ensembles. For the operational NCEP/GEFS ensemble, the three ensembles are named OPR_RAW, OPR_DAV2% (removing the decaying average bias estimate), and OPR_OPT, respectively. For the ESRL/PSD reforecast ensemble, the three ensembles are named RFC_RAW, RFC_COR (removing the climatological mean bias estimate), and RFC_OPT, respectively. All of the ensemble forecasts and analyses are on grid points with a spacing of 2.5° × 2.5° globally. The NCEP operational analysis is used for the bias estimation and verification calculations.

d. *Evaluation methods*

Several probabilistic (for ensemble distribution) and deterministic (for ensemble mean) verification methods are used to evaluate the ensemble forecast performance, such as the ranked probability skill score (RPSS), the relative operating characteristics (ROC) skill score, excessive outlier, the pattern anomaly correlation coefficient (PAC), the root-mean-square error of the ensemble mean (RMS), and the relative economic value.

The RPSS score is one of the most important measures for evaluating the performance of probabilistic forecasts (Toth et al. 2003). The higher the RPSS score (the maximum is 1), the better the probabilistic system is by being both reliable and exhibiting high resolution. The best probabilistic system would be rewarded

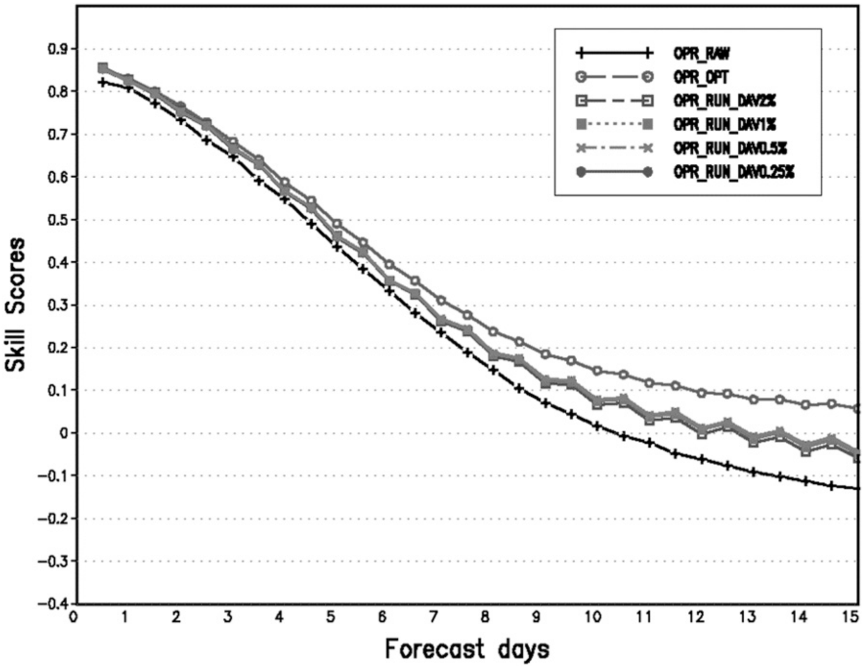

FIG. 2. RPSS of 500-hPa geopotential height averaged from 1 Mar 2004 to 28 Feb 2005 for the Northern Hemisphere with decaying weights of 0.25%, 0.5%, 1%, and 2%.

by a RPSS score of 1. The ROC skill score is a measure of the forecast system resolution (Y. Zhu et al. 1996; unpublished manuscript, 2002). A ROC skill score of 1 corresponds to a perfect forecast while 0 indicates no skill above the sample climatology. The continuous ranked probability skill score (CRPSS) measures the reliability and resolution. For statistics over a long period, CRPSS is very similar to RPSS (Zhu and Toth 2008). Therefore, either one of these two measures are used, whichever is more convenient.

4. Results and discussion

Issues examined in this section include (a) the choice of decaying averaging weight factors, (b) the results of bias-correction techniques applied to the NCEP/GEFS and CMC/GEFS ensembles, and (c) comparisons of the three experimental calibration techniques.

a. Tests of different decaying averaging weights

The first issue when applying the decaying averaging method is the choice of the decaying weight w. Different w's had been chosen and tested. Figure 2 shows some of the decaying weight sensitivity test results. Among the six curves in Fig. 2, the curve OPR_OPT is for the calibrated NCEP/GEFS result using dependent data and the curve OPR_RAW is for the raw NCEP ensemble forecast. The other four curves (OPR_RUN_DAV2%, OPR_RUN_DAV1%, OPR_RUN_DAV0.5%, and OPR_RUN_DAV0.25%) are for the RPSSs of 500-hPa geopotential height that are averaged from 1 March 2004 to 28 February 2005 for the Northern Hemisphere with decaying weights of 2%, 1%, 0.5%, and 0.25%, respectively. All four calibrated ensembles show improvements compared with the raw ensemble OPR_RAW for all lead times. There is little room for further improvement compared with the OPR_OPT test for short lead times until day 4. Though the four curves are close together, for short lead times OPR_DAV2% is better than the other decaying weights. On the other hand, OPR_DAV0.25% is the best for week-2 forecasts. Other statistics are calculated that show the 2% ensemble produces large improvements in ROC and BSS scores over the Northern and Southern Hemispheres. The improvement of these scores in summer is more significant than in spring. A higher decaying weight (10%) is also investigated and compared with 2%. The choice of a 10% weight works better for the tropics compared to 1% or 2%. In general, the 2% weight works better for most regions and seasons (not shown). For an optimal result, the decaying accumulated bias with a 2% weight is used in operations and is updated every day for all 35 selected variables.

For a better understanding and explanation of the above results, another issue related to the postprocessing algorithm design is quickly reviewed here—a comparison

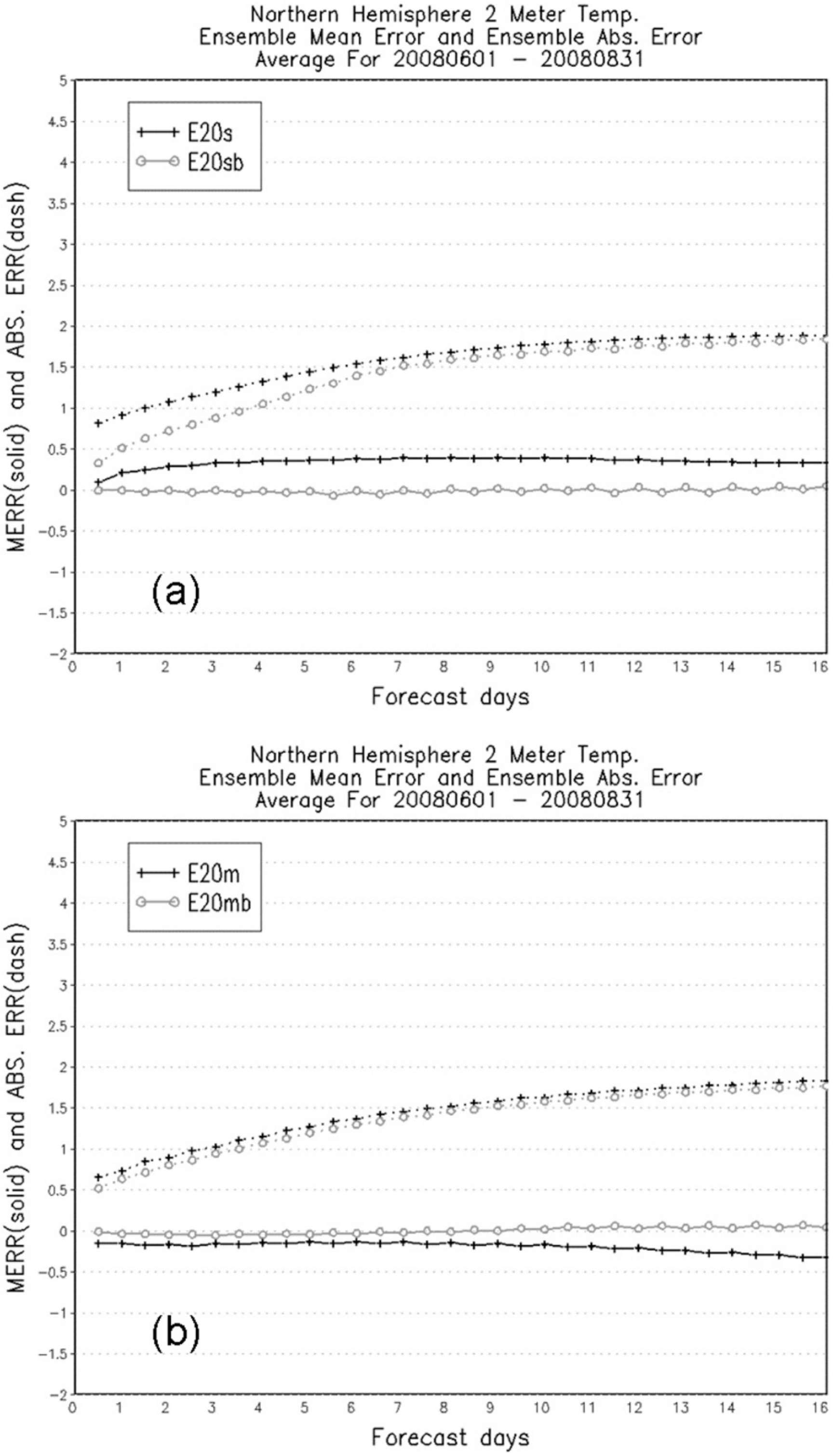

FIG. 3. Mean error (solid) and mean absolute error (dashed) of 2-m temperature averaged from 1 Jun to 31 Aug 2008 for the Northern Hemisphere for the (a) NCEP/GEFS and (b) CMC/GEFS ensembles. E20s/E20m is for the raw ensemble forecast and E20sb/E20mb is for the bias-corrected ensemble forecast.

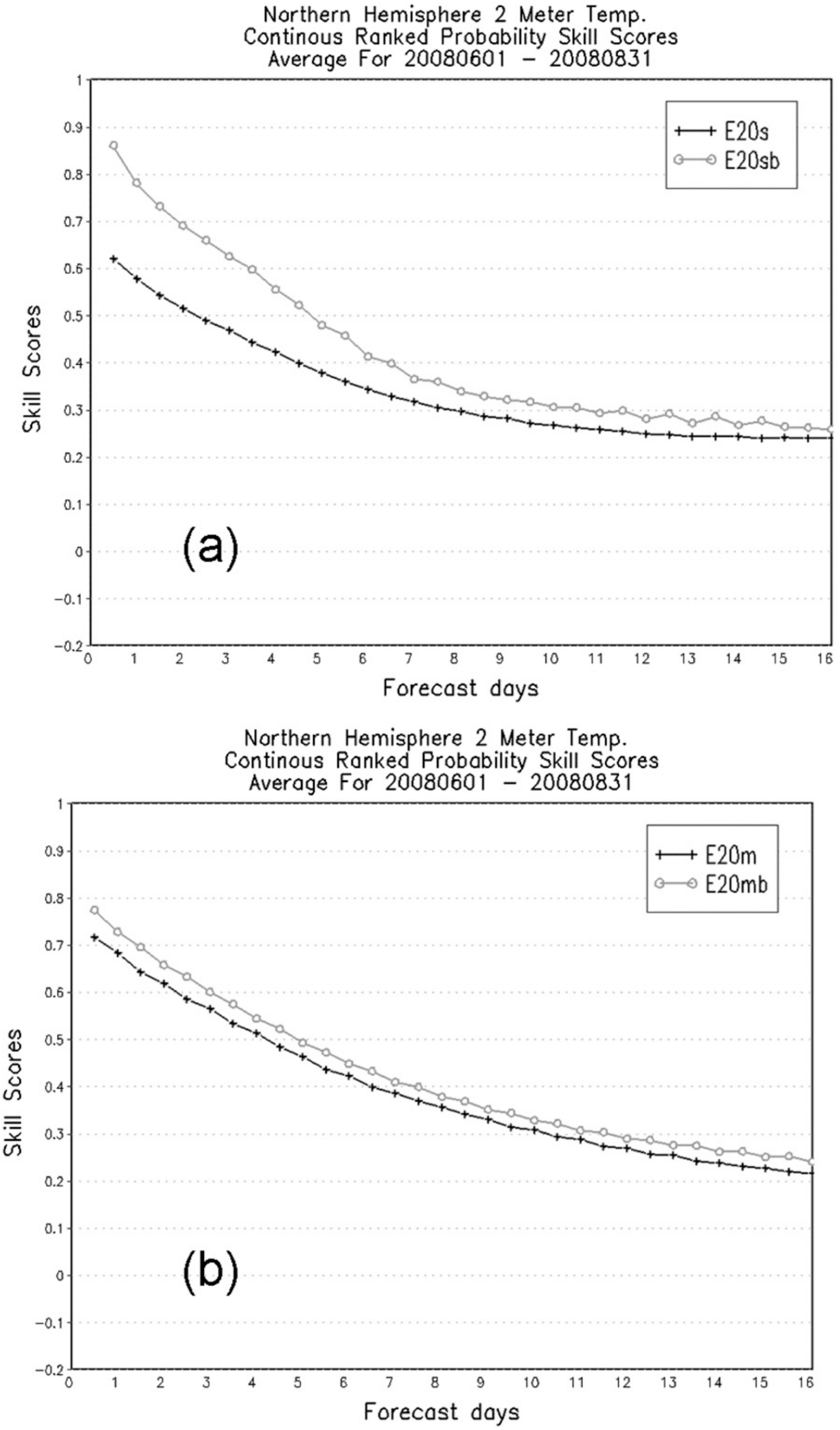

FIG. 4. CRPSS of 2-m temperature averaged from 1 Jun to 31 Aug 2008 for the Northern Hemisphere for (a) NCEP/GEFS and (b) CMC/GEFS. E20s/E20m is for the raw ensemble forecast, and E20sb/E20mb is for the bias-corrected ensemble forecast.

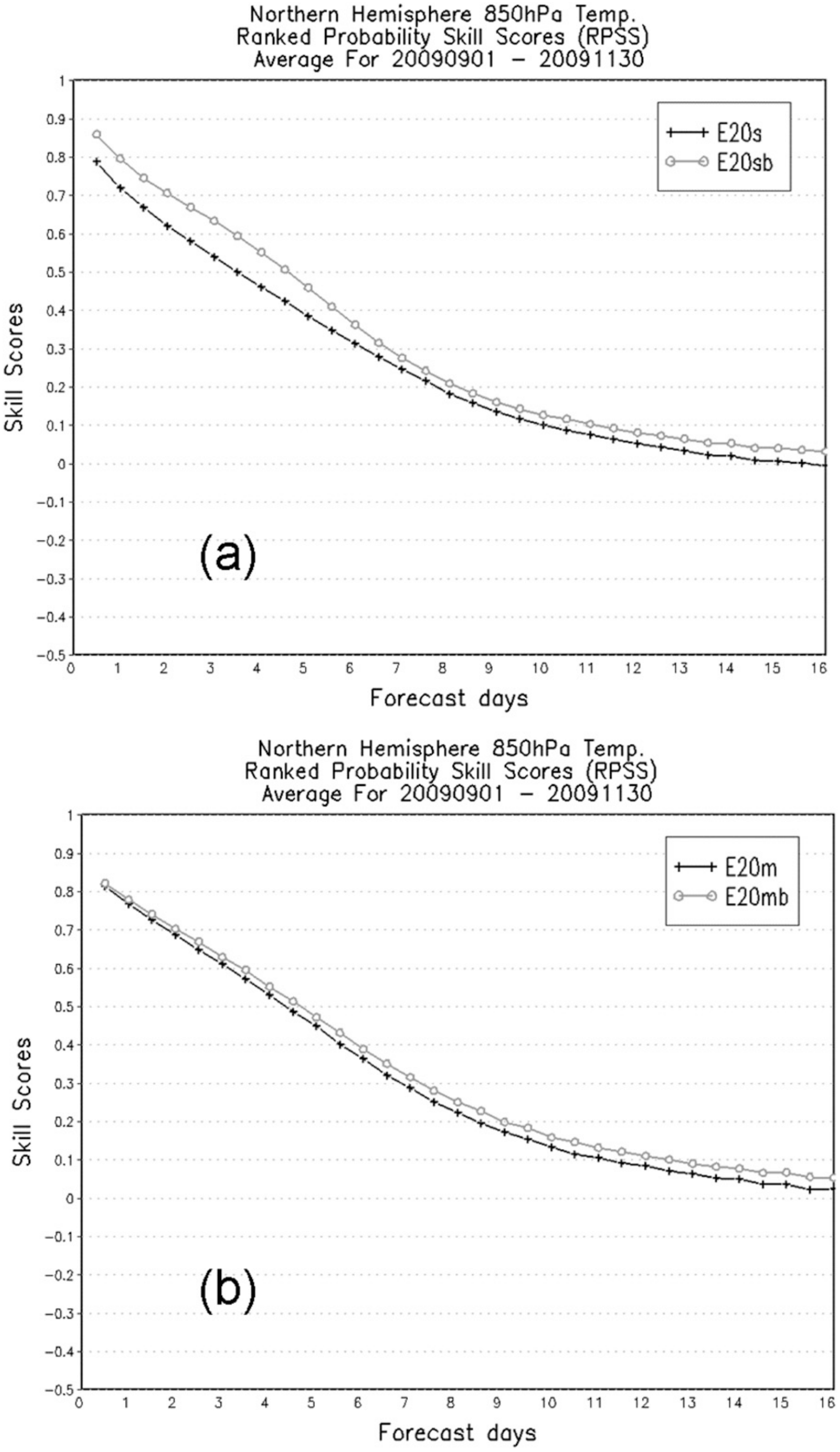

FIG. 5. RPSS of 850-hPa temperature averaged from 1 Sep to 30 Nov 2009 for the Northern Hemisphere for (a) NCEP/GEFS and (b) CMC/GEFS. E20s/E20m is for the raw ensemble forecast and E20sb/E20mb is for the bias-corrected ensemble forecast.

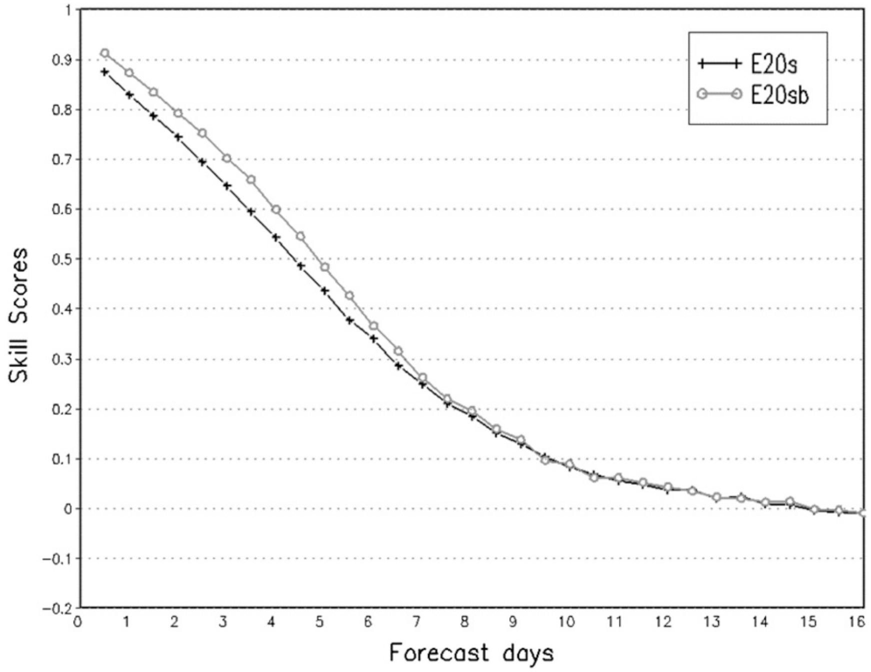

FIG. 6. NCEP/GEFS CRPSS of 1000-hPa height averaged from 1 Dec 2007 to 29 Feb 2008 for the Northern Hemisphere. E20s is for the raw ensemble forecast and E20sb is for the bias-corrected ensemble forecast.

between the decaying average and equal weight bias estimate approaches. The equal weight approach also makes a first-moment bias calculation over some previous days but with equal weighting for each day. We applied the equal weight bias estimate method to two seasons of the 2004 operational NCEP/GEFS ensemble and compared the results with the decaying weight OPR_DAV2%. Results from the equal weight and decaying weight are very similar (less than 2% weight for a longer forecast; not shown). The reasons to choose the decaying weight and not the equal weight include (a) the decay method having a higher weight for the latest information, which is good for a flow-dependent system (short-term forecast), and (b) the application of the decay weight method being operationally cost effective. There is no need to save extra data on the central computer system, and bias estimates can include more historical information through continuous updates once the latest analysis is available. In general, the result from the decaying average will be better than any single average (equal weight) method.

b. NCEP/GEFS and CMC/GEFS bias correction in operation

Figure 3 shows the mean forecast error (solid) and the mean absolute forecast error (dash) of 2-m temperature for the Northern Hemisphere. Curves E20s/E20m are for the raw ensembles and E20sb/E20mb are for the bias-corrected ensembles. All are from 20-member ensembles. The NCEP/GEFS raw ensemble has a warm bias tendency (curve E20s; see Fig. 3a) and CMC/GEFS displays a cold bias tendency (curve E20m; see Fig. 3b) at all lead times. After bias correction, the mean errors of both the NCEP and CMC ensembles are very close to zero. Meanwhile, mean absolute errors of the bias-corrected ensembles are also reduced, indicating that the change of the mean error does not come from the balance of positive and negative values. This suggests that the bias-correction technique can effectively alleviate the ensemble forecast system from overpredicting (too warm) or underpredicting (too cold) temperatures. The calibrated ensemble forecasts become closer to the actual temperatures. However, the extents to which the bias corrections can improve the raw ensembles are different for NCEP/GEFS and CMC/GEFS because of the different characteristics of the NCEP and CMC ensemble systems.

Figure 4 shows CRPSS scores of the 2-m temperature for 2008 summer, verified over the Northern Hemisphere. Values added from the bias correction are noticeable for both NCEP/GEFS and CMC/GEFS. Figure 5 shows the RPSS of 850-hPa temperature for fall 2009. Both the NCEP/GEFS and CMC/GEFS are improved due to calibration. Overall, the bias reductions globally

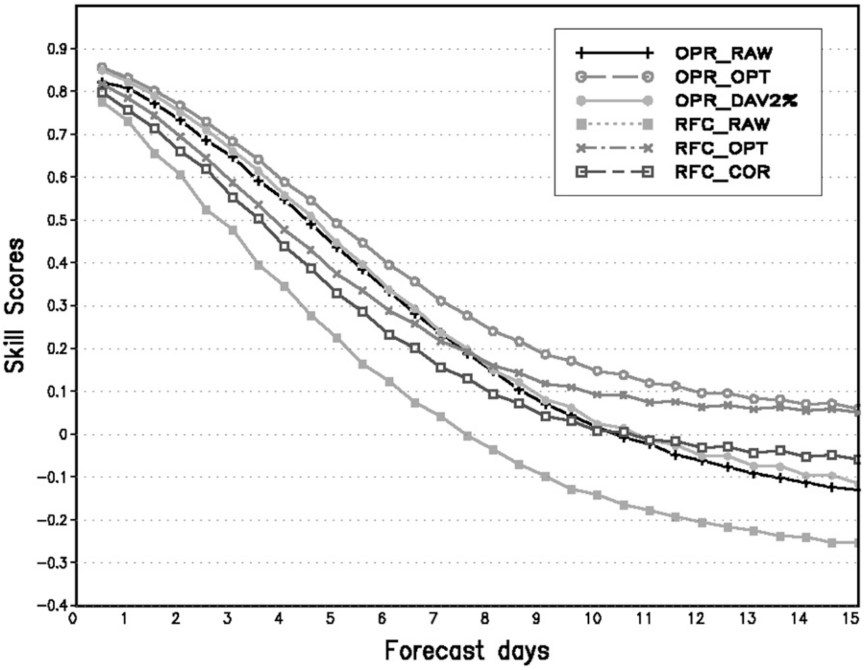

FIG. 7. RPSS of Northern Hemisphere 500-hPa geopotential height from 1 Mar 2004 to 28 Feb 2005 comparing the NCEP operational forecast (OPR) and ESRL reforecast (RFC): OPR_RAW is the NCEP operational raw ensemble forecast, OPR_OPT is the NCEP operational forecast using optimal bias correction, OPR_DAV2% is the NCEP operational forecast using a 2% weight for bias correction, RFC_RAW is the raw reforecast, RFC_OPT is the reforecast after optimal bias correction, and RFC_COR is the reforecast after removing the climatological mean bias.

(absolute value) after calibration can reach up to 75% for days 0–3, 60% for days 2–8, and 45% for days 8–15 for the 500-hPa height field (not shown). More evaluation results such as PAC and relative economic value (Zhu et al. 2002) are also examined. These results are available online (http://www.emc.ncep.noaa.gov/gmb/yzhu/html/opr/naefs.html) and are updated seasonally. In general, there are skill gains for forecast days 1–6 due to bias correction.

However, the calibration technique does not work well in all circumstances. For some seasons and variables there is no skill improvement for the week-2 forecast. Figure 6 shows there is almost no improvement for the 1000-hPa height field RPSS score after day 7. Previous studies have indicated that there remains room for improvement in the week-2 forecasts, as can be seen when comparing the calibrated and the optimal ensembles displayed in Fig. 2. How can we improve the current calibration technique? Do we need a hindcast for the calibration of the week-2 forecast? These questions will be discussed in the next section.

c. Calibration techniques comparison

Figure 7 shows the annual mean RPSS scores of the 500-hPa geopotential heights verified over the Northern Hemisphere. Of the three operational NCEP/GEFS ensembles, the one with optimal bias correction (OPR_OPT) gets the highest RPSS scores among the six curves. The decaying average bias-correction algorithm also works well. The RPSS of the OPR_DAV2% is improved versus the OPR_RAW for all lead times, but especially in the short range, which can be judged from the small distance between the two curves OPR_DAV2% and OPR_OPT.

For the three reforecast ensembles, it is not surprising to note that the optimal bias-corrected ensemble (RFC_OPT) shows the best performance when compared with the raw and climatological bias-corrected ensembles (RFC_RAW and RFC_COR). A comparison between RFC_RAW and RFC_COR shows that RFC_COR has a noticeable RPSS improvement versus RFC_RAW, especially for the week-2 forecasts. Using the climatological mean bias estimate, it is possible to make probabilistic week-2 forecasts more skillful than the raw reforecast. The RFC_COR comes from a reforecast ensemble with an older version of the model, which has initial data of relatively poorer quality compared with the operational ensembles, including OPR_RAW and OPR_DAV2%. However, RFC_COR has an even better level of performance than OPR_RAW

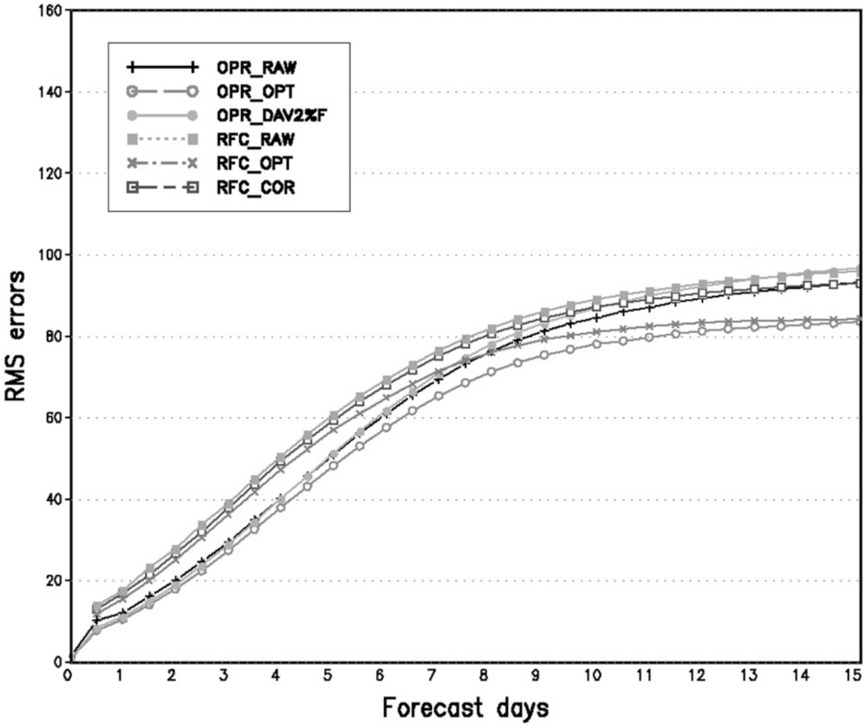

FIG. 8. As in Fig. 7, but for RMS errors from the ensemble mean.

and OPR_DAV2% after day 10 (Fig. 7), indicating the effectiveness of a large data sample in improving week-2 forecasts.

Figure 8 shows the annual mean RMS error of the ensemble mean forecasts for 500-hPa height, verified over the Northern Hemisphere. The six curves coming from the six ensembles for forecast days 1–6 are divided into two clusters that belong to the operational and reforecast ensemble forecast groups, respectively. Of the three operational ensembles, OPR_OPT has the lowest RMS error among the six ensembles. The OPR_DAV2% has reduced RMS errors for the first week compared with OPR_RAW but its RMS becomes larger for week-2 forecasts. However, the two similar curves of OPR_OPT and OPR_DAV2% for the first week suggest that there is only a limited opportunity for future improvement in bias correction for the first few days. The big distance between the OPR_OPT and OPR_DAV2% curves for week 2 indicates that the OPR_DAV2% calibration technique has the potential to improve extended forecasts.

Of the three reforecast ensembles, RFC_COR has smaller RMS values than does RFC_RAW for all lead times, even for week 2. A comparison between the operational and reforecast ensembles shows that the operational ensemble mean (OPR_RAW) has a much lower RMS error than the ESRL/PSD hindcast (RFC_RAW). The RFC_RAW short-range error is around 50% larger than that of OPR_RAW. Though the reforecast runs start from relatively poorer quality initial data than are used in the operational ensemble, RFC_COR works for short-range forecasts and its curve with reduced RMS error comes close to the OPR_RAW curve after day 10. Both Figs. 7 and 8 show that the decaying averaging with a 2% weight and 45 days of training data works very well in the short range. All measures are improved until day 5.

Figures 9 and 10 are the RPSS and ROC skill scores of 850-hPa temperature for the summer of 2004, verified over the Northern Hemisphere. Results are similar to those shown for the 500-hPa height. The OPR_DAV2% has better performance than OPR_RAW, and RFC_COR also shows noticeable improvement versus the RFC_RAW. Notice that there is poor performance in the short range for RFC_COR versus RFC_RAW, starting from the initial time. We think this was caused by the bias between the reanalysis, which was used during the climatological bias estimation, and the operational analysis, which was used for the ensemble evaluation. The climatological bias estimate is calculated from the difference between the daily forecast and verification climatologies as a function of forecast lead time. The climatologies are computed from 31-day running means using data from 1979 to 2003. Therefore, there is an existing bias between the operational analysis and the verification climatologies. Such a bias can be mitigated or removed through bias correction.

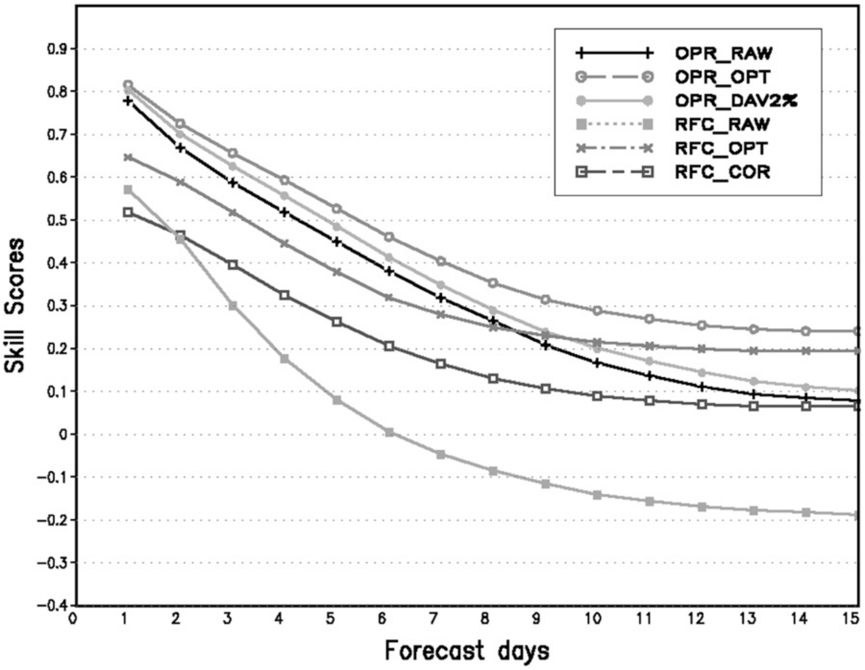

FIG. 9. As in Fig. 6, but for the RPSS of 850-hPa temperature for June–August 2004.

Other bias-corrected variables include 2-m temperature and 10-m *U* and *V* components (not shown). Based on the results from different variables, some tentative conclusions may be drawn.

The decaying averaging (DA) with a 2% weight and 45 days of operational training data works very well over the short range (almost as well as the ‘‘optimal), which makes its application possible for frequent updates of the DA/NWP modeling system. On the other hand, the climatological mean bias correction can add value, especially for week-2 probability forecasts. Since the operational analysis–modeling system that supports the NCEP/GEFS ensemble undergoes frequent (once or twice a year) changes, it would be a very large computing problem if the reforecast method requiring the same model used for operational forecasting was also used for the reforecasting. No such long-term archive based on the most recent analysis–modeling system is available for the reforecast ensemble. The generation of a large hindcast ensemble is expensive but may be helpful. The use of up-to-date data assimilation/NWP techniques is imperative at all ranges.

5. Summary and future plans

A statistical postprocessing algorithm (i.e., the decaying average method) has been applied to the NCEP/GEFS and CMC/GEFS to generate calibrated forecasts. The implementation of this technique is expected to improve NCEP and CMC global ensemble forecasts in order to provide more accurate NAEFS products. Due to the different ensemble configurations, calibration strategies applied to the NCEP and CMC ensembles have been adjusted. The NCEP/GEFS is created by using one model with perturbed initial conditions. We assume the biases from one model have a kind of similarity, and ensemble mean biases are thought to be able to represent these systematic errors. Therefore, NCEP/GEFS uses the ensemble mean bias to calibrate each member. For CMC/GEFS, the bias for each individual ensemble member is calculated and used for that member. Both the NCEP/GEFS and CMC/GEFS benefit from the application of bias correction. Several studies have shown that NAEFS, when compared to the CMC and NCEP ensemble system, shows significant improvements both in terms of reliability and resolution (Zhu and Toth 2008; Candille 2009). Even with the attractive properties of the decaying average method, its limitations and performance for some variables in week 2 forecasts in some seasons represent a major drawback. There is room for future improvement from (a) adjusting weights to allow a longer training time and (b) to take advantage of reforecasts/hindcasts.

To further improve the current operational bias-correction technique, three other experiments were designed and assessed using annual retrospective experiments from 1 March 2004 to 28 February 2005. Results show that the decaying average bias estimation method with a short sample works well for the first few

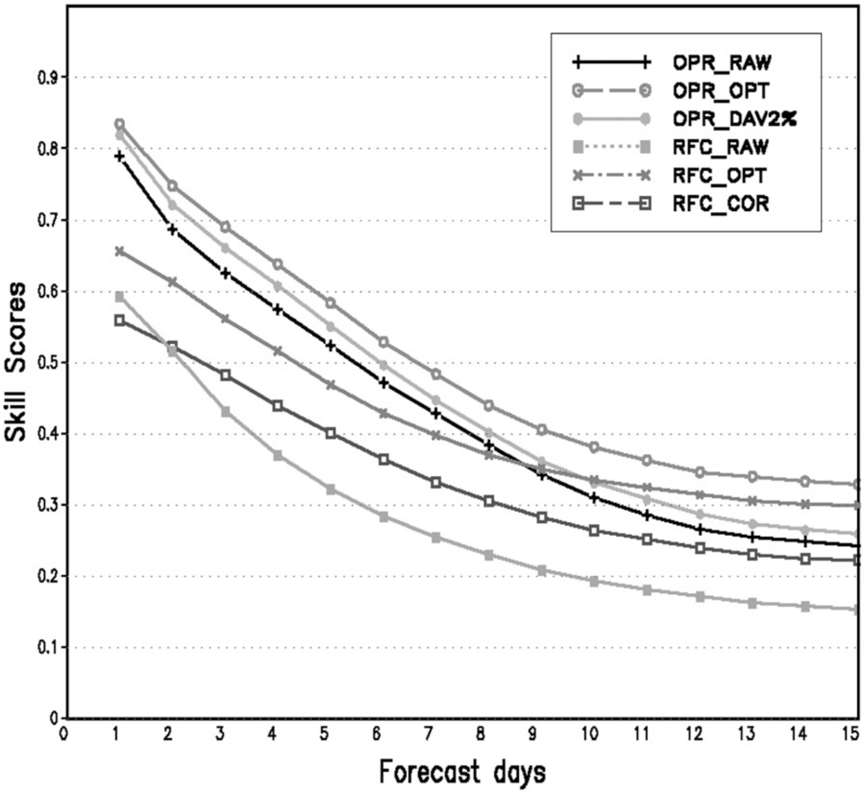

FIG. 10. As in Fig. 6, but for the ROC score of 850-hPa temperature for June–August 2004.

days. The calibrated NCEP/GEFS ensemble, after removing time mean forecast errors for the most recent period, has an improved probabilistic performance for all measures until day 5. The reforecast ensembles from ESRL/PSD with and without a climate mean bias correction are also examined. A comparison between the NCEP/GEFS and ESRL/PSD bias-corrected ensemble forecasts shows that a climate mean bias correction can add value, especially for week-2 probability forecasts. This conclusion is very similar to the studies by Hamill et al. in multiple papers (e.g., Hamill et al. 2004, 2006).

The major drawback of climate mean bias correction is the need for a long training dataset, and since a reforecast works best with a frozen model, the database must be completely rebuilt whenever the model is updated, which requires large computing resources. In this way, routine improvements to the model are incorporated in the reforecast-based products as soon as they are implemented. However, due to the good performance of the climate mean bias correction, the current reforecast ensemble uses an old low-resolution version of the model system and it is worth the effort to generate the reforecast dataset and apply it to the ensemble postprocessing. NCEP has plans with ESRL/PSD to jointly implement a real-time hindcast experiment in the 2011–12 time frame and utilize additional resources to generate a set of historical ensemble reforecasts (20 yr). The operational forecast model will be applied to the reforecast configurations. Our postprocessing study will benefit from this new high-resolution reforecast dataset. Using the reforecast dataset, we will be able to test our postprocessing methodology and compare it with the calibration method developed by ESRL/PSD. New bias-correction methods developed under The Observing System Research and Predictability Experiment (THORPEX) project will also be considered for use in the NAEFS statistical postprocessing system.

Acknowledgments. We gratefully acknowledge the support of Jeff Whitaker (NOAA/Earth System Research Laboratory/Physical Sciences Division), Tom Hamill (NOAA/Earth System Research Laboratory/Physical Sciences Division), Richard Verret (CMC), and Richard Wobus (NCEP). This study would have been much more difficult without their assistance. We also thank Mary Hart (NCEP) for editing our English.

REFERENCES

Bakhshaii, A., and R. Stull, 2009: Deterministic ensemble forecasts using gene-expression programming. *Wea. Forecasting,* **24,** 1431–1451.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.,* **137,** 1655–1665.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting,* **4,** 401–412.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian Ensemble Prediction System. *Mon. Wea. Rev.,* **138,** 1877–1901.

Cheng, W. Y. Y., and W. J. Steenburgh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Wea. Forecasting,* **22,** 1304–1318.

Gel, Y. R., 2007: Comparative analysis of the local observation-based (LOB) method and the nonparametric regression-based method for gridded bias correction in mesoscale weather forecasting. *Wea. Forecasting,* **22,** 1243–1256.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118.

Hacker, J., and D. L. Rife, 2007: A practical approach to sequential estimation of systematic error on near-surface mesoscale grids. *Wea. Forecasting,* **22,** 1257–1273.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

Houtekamer, P. L., L. Lefaivre, and J. Derome, 1996: The RPN ensemble prediction system. *Proc. ECMWF Seminar on Predictability,* Vol. II, Reading, United Kingdom, ECMWF, 121–146. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *Trans. ASME—J. Basic Eng.,* **82,** 35–45.

Krzysztofowicz, R., and W. B. Evans, 2008: Probabilistic forecasts from the National Digital Forecast Database. *Wea. Forecasting,* **23,** 270–289.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 137–163.

——, Y. Zhu, and R. Wobus, cited 2012: March 2004 upgrades of the NCEP Global Ensemble Forecast System. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.]

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP Global Operational Forecast System. *Tellus,* **60A,** 62–79.

Yussouf, N., and D. J. Stensrud, 2006: Prediction of near-surface variables at independent locations from a bias-corrected ensemble forecasting system. *Mon. Wea. Rev.,* **134,** 3415–3424.

Zhu, Y., and Z. Toth, 2008: Ensemble based probabilistic forecast verification. Preprints, *19th Conf. on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc., 2.2. [Available online at http://ams.confex.com/ams/pdfpapers/131645.pdf.]

——, and B. Cui, cited 2012: GFS bias correction. [Available online at http://www.emc.ncep.noaa.gov/gmb/yzhu/imp/i200711/1-GFS_bc.pdf.]

——, Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 73–83.

——, B. Cui, and Z. Toth, cited 2012a: December 2007 upgrade of the NCEP Global Ensemble Forecast System (NAEFS). [Available online at http://wwwt.emc.ncep.noaa.gov/gmb/yzhu/imp/i200711/IMP_PLAN_final_v08_brief.pdf.]

——, Z. Toth, R. Wobus, M. Wei, and B. Cui, cited 2012b: May 2006 upgrade of the GEFS and first implementation of NAEFS systems. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.]