Improvement of Statistical Postprocessing Using GEFS Reforecast Information

HONG GUAN

NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland, and System Research Group, Inc., Colorado Springs, Colorado

BO CUI

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

YUEJIAN ZHU

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

(Manuscript received 9 October 2014, in final form 18 February 2015)

ABSTRACT

The National Oceanic and Atmospheric Administration/Earth System Research Laboratory (NOAA/ESRL) generated a multidecadal (from 1985 to present) ensemble reforecast database for the 2012 version of the Global Ensemble Forecast System (GEFS). This dataset includes 11-member reforecasts initialized once per day at 0000 UTC. This GEFS version has a strong cold bias for winter and warm bias for summer in the Northern Hemisphere. Although the operational decaying average bias-correction approach performs well in winter and summer, it sometimes fails during the spring and fall transition seasons at long lead times (>~5 days). In this paper, 24- (1985–2008) and 25-yr (1985–2009) reforecast biases are used to calibrate 2-m temperature forecasts in 2009 and 2010, respectively. The reforecast-calibrated forecasts for both years are more accurate than those adjusted by the decaying average method during transition seasons. A long training period (>5 yr) is necessary to help avoid a large impact on bias correction from an extreme year case and keep a broader diversity of weather scenarios. The improvement from using the full 25-yr, 31-day window, weekly training dataset is almost equivalent to that from using daily training samples. This provides an option to reduce computational expenses while maintaining a desired accuracy. To provide the potential to improve forecast accuracy for transition seasons, reforecast information is added into the current operational bias-correction method. The relative contribution of the two methods is determined by the correlation between the ensemble mean and analysis. This method improves the forecast accuracy for most of the year with a maximum benefit during April–June.

1. Introduction

Several weather centers worldwide routinely produce skillful weather predictions using an ensemble forecast system (Toth and Kalnay 1993, 1997; Wilks and Hamill 2007). The North American Ensemble Forecast System (NAEFS), officially launched in November 2004, is a successful example of applying a multicenter, multimodel ensemble forecast system to estimate the uncertainty of weather forecasts and to make high quality probability forecasts. The NAEFS combines two ensemble forecast systems: the Global Ensemble Forecast System (GEFS) of the National Weather Service (NWS) and the Canadian Meteorological Centre Ensemble (CMCE) of the Meteorological Service of Canada (MSC), which produces a more reliable forecast than either of the forecast systems when used alone (Candille 2009).

Ensemble forecasts are contaminated by system bias and random errors (Toth et al. 2003; Wilks and Hamill 2007). In the last decade, various statistical postprocessing methods have been developed and applied to reduce the bias of the ensemble forecast system and improve the skill of probability forecasts. These methods include logistic regression (Wilks and Hamill 2007), Bayesian model averaging (BMA; Raftery et al.

Corresponding author address: Dr. Hong Guan, NOAA/NWS/NCEP/Environmental Modeling Center, 5830 University Research Ct., College Park, MD 20740.
E-mail: hong.guan@noaa.gov

DOI: 10.1175/WAF-D-14-00126.1

2005; Wilson et al. 2007), nonhomogeneous Gaussian regression (Gneiting et al. 2005), Gaussian ensemble dressing (Roulston and Smith 2003; Wang and Bishop 2005; Bishop and Shanley 2008), artificial neural networks (Yuan et al. 2007), ensemble MOS (Wagner and Glahn 2010), analog techniques (Hamill et al. 2004, 2006, 2013; Hamill and Whitaker 2007; Monache et al. 2011, 2013; Hagedorn et al. 2012), Kalman filtering (Cheng and Steenburgh 2007), and decaying averaging (Cui et al. 2012; Glahn 2014).

Reducing the systematic bias, for both the first and second moments in the ensemble forecast, is a major goal for the NAEFS Statistical Post Processing (SPP). The current SPP includes bias correction and downscaling (B. Cui et al. 2012, unpublished manuscript). The bias correction is mainly a first-moment adjustment by applying the decaying average method to estimate the bias, giving more weight to recent samples than to older ones (B. Cui et al. 2012, unpublished manuscript). The algorithm was developed by the NWS at the National Centers for Environmental Prediction (NCEP) and was implemented operationally in 2006 to reduce the bias of the NAEFS forecasts. This method is fast and does not need to store a large amount of sample data once initialized, which makes the application straightforward in a daily operational system. Operational statistical verification since 2006 reveals that the NAEFS product is significantly enhanced by the decaying bias-correction method. However, the method sometimes fails to improve the forecast skill during the spring and fall transition seasons for long-lead-time forecasts.

It should also be mentioned that in the past, several works (Hamill and Whitaker 2007; Monache et al. 2011, 2013; Hagedorn et al. 2012) have demonstrated that their techniques improve raw forecasts over a simple bias correction like decaying average and running mean bias-correction methods. The analog techniques in Hamill and Whitaker (2007) and Monache et al. (2011, 2013) generally require storing large amounts of data from a past sample dataset to find the past forecasts that are similar to the current forecast. This is its disadvantage relative to the decaying method. In the future, we will compare the analogy and decaying techniques to see if we can find an optimal analog method that fits the computing resources of daily operations while performing better than the decaying technique. Hagedorn et al. (2012) show that the nonhomogeneous Gaussian regression method improves raw forecasts over a simple bias-correction procedure. However, the work in Glahn (2014) demonstrates that the decaying method wins when compared with the regression method. Obviously, future study will be needed to address the inconsistencies of the above studies.

Recently, the NOAA/Earth System Research Laboratory (ESRL) generated an ensemble reforecast dataset using the 2012 version of GEFS. This multidecadal dataset has been applied to precipitation calibration, diagnosis of the ability of GEFS to forecast uncommon phenomena, and the initialization of regional reforecasts (Hamill et al. 2013). In this study, we use a 26-yr reforecast dataset to improve the current operational NAEFS bias-correction process. The decaying average and reforecast bias-correction methods are described in section 2. The GEFS model and reforecast dataset are introduced in section 3. The evaluation of the two calibration methods and the sensitivity of the reforecast calibration to sample size are discussed in section 4. The improvement from combining the reforecast method with the decaying method is highlighted in section 5. Conclusions are given in section 6.

2. Bias-correction methods

a. Bias estimation

In this study, the bias b for each lead time t (6-h intervals up to 384 h for the operational product) and each grid point (i, j) is defined as the difference of the best analysis $a_{i,j}(t_0)$ and forecast $f_{i,j}(t)$ at the same valid time t_0:

$$b_{i,j}(t) = f_{i,j}(t) - a_{i,j}(t_0). \qquad (1)$$

b. Decaying average method

The details of the decaying average method can be found in Cui et al. (2012). Here, we introduce its basic equation. The decaying average bias $B^p_{i,j}(t)$ is updated by combining the bias from the previous forecast with the current bias by using a weighting coefficient w. Experiments using different weights (0.01, 0.02, 0.05, 0.1, and 0.2) show that a weight of 0.02 gives the best overall verification score. Recently, Glahn (2014) applied the decaying average method to the bias correction of station-based forecasts. Sensitivity tests with four weights (0.025, 0.05, 0.075, and 0.1) reveal that only the smaller weights (0.025 and 0.05) improve the bias and mean absolute errors of MOS forecasts for the CONUS region. The value of 0.025 is similar to the optimal weight (0.02) used in the NCEP bias correction:

$$B^p_{i,j}(t) = (1 - w)[B^p_{i,j}(t - 1)] + w[b_{i,j}(t)]. \qquad (2)$$

c. Bias correction using reforecasts

The basic idea for this method is to use knowledge about the forecast errors of the same model during a similar period in previous years to calibrate the current forecast. The average reforecast bias $B^h_{i,j}(t)$ is the climatological mean forecast error, obtained from the multiyear N reforecast ensemble:

$$B^h_{i,j} = \frac{\sum_{k=1}^{N} b_{i,j,k}(t)}{N}. \qquad (3)$$

d. Bias correction

To remove the lead-time-dependent bias from a model grid, a new (or bias corrected) forecast F is generated by applying the decaying average bias and the reforecast bias to the raw forecast at each grid point, for each lead time, and each parameter:

$$F_{i,j} = f_{i,j} - r^2[B^p_{i,j}(t)] - (1 - r^2)[B^h_{i,j}(t)], \qquad (4)$$

where r is the correlation coefficient estimated by linear regression from the most recent joint samples (ensemble mean and analysis). To avoid storing a large dataset, the mean values used in computing r were generated from a decaying average with a weight of 0.10. The calculation mainly uses the most recent 10–15 days of bias information. The data between the most recent 15 and 40 days is a minor contributor. The relative contribution of the reforecast and decaying average bias was quantified by r^2. The high correlation indicates that the model can capture the temporal variation in the 2-m temperature analysis during the most recent period well. This implies that bias features are likely dominated by systematic error during the training period, which is highly predictable. Here, we use r^2 as an approximate measure of the validity of the decaying average bias. For the two special cases of $r = 0$ and $r = 1$, the equations represent the reforecast bias correction and decaying bias correction, respectively.

e. Methodology of verification

The calibration of the ensemble forecast system is evaluated via the mean forecast error (Wilks 2006), mean absolute forecast error (Wilks 2006), root-mean-square error (RMSE; Zhu and Toth 2008), and continuous rank probability score (CRPS; Zhu and Toth 2008). The CRPS is frequently used for evaluating the performance of probabilistic forecasts (Zhu and Toth 2008; Glahn et al. 2009; Friederichs and Thorarinsdottir 2012). The score represents the difference of the cumulative distribution functions of the ensemble forecast and observation. The lower the CRPS, the better the probabilistic system performs.

3. Model and reforecast data

The current operational GEFS version (10) was implemented on 14 February 2012 at the NCEP. It consists of 21 members (1 control member and 20 perturbation members) and is run four times daily (at 0000, 0600, 1200, and 1800 UTC). All members use an identical set of physical parameterizations (Zhu et al. 2007). The model is run at a horizontal resolution of T254 (~55 km) for the first 8 days and T190 (~70 km) for the last 8 days, with 42 hybrid levels. The Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) is used to initialize the simulation. The perturbed initial conditions use the ensemble transform with rescaling (ETR) technique (Wei et al. 2008). The model uncertainty is estimated using the stochastic total tendency perturbation (STTP) method (Hou et al. 2008).

The reforecast data were generated from the above GEFS version but including only 11 members (1 control member and 10 perturbation members). The model was only run at the 0000 UTC cycle for the 10 members. The control member was run at both 0000 and 1200 UTC. The dataset used here was bilinearly interpolated onto 1° × 1° latitude and longitude grids from the native resolution. These data are available starting in 1985 and forward (+29 yr). We use a subset of the data from 1985 to 2010 (26 yr), obtained from NOAA/ESRL. A more detailed description of the model and dataset can be found in Hamill et al. (2013).

The time series of 2-m temperature errors over the Northern Hemisphere (NH) for 120- and 240-h forecasts are displayed in Fig. 1. It is evident that there is a warm bias for April–August (warm season), while a cold bias is prevalent for the rest of the year (cold season). The sharpest error change occurs between March and May with change rates of ~0.5° and 0.6°C month^{-1} for the 120- and 240-h forecasts, respectively. The large change in bias during the spring season can make it difficult to do bias correction with the current decaying average postprocessing algorithms, because the forecast errors in recent periods will not be fully representative of the current forecast error. The difference in errors among different 5-yr periods is relatively small. We did not find a significant improvement in forecast skill from the late 1980s to the most recent year. The bias curve for the last 5-yr period (2005–09) shifts only slightly in the direction of positive bias from the first 5-yr period (1985–89). This suggests that the selection of sample periods may not be a big issue in calibrating 2-m temperature forecasts.

Figure 2 depicts the distribution of global 2-m temperature errors for the cold and warm seasons. Large bias occurs mainly over or near the continents, most likely because of the complex topography and deficient physical parameterizations over land. The semiannual change in bias over the continents of the Northern

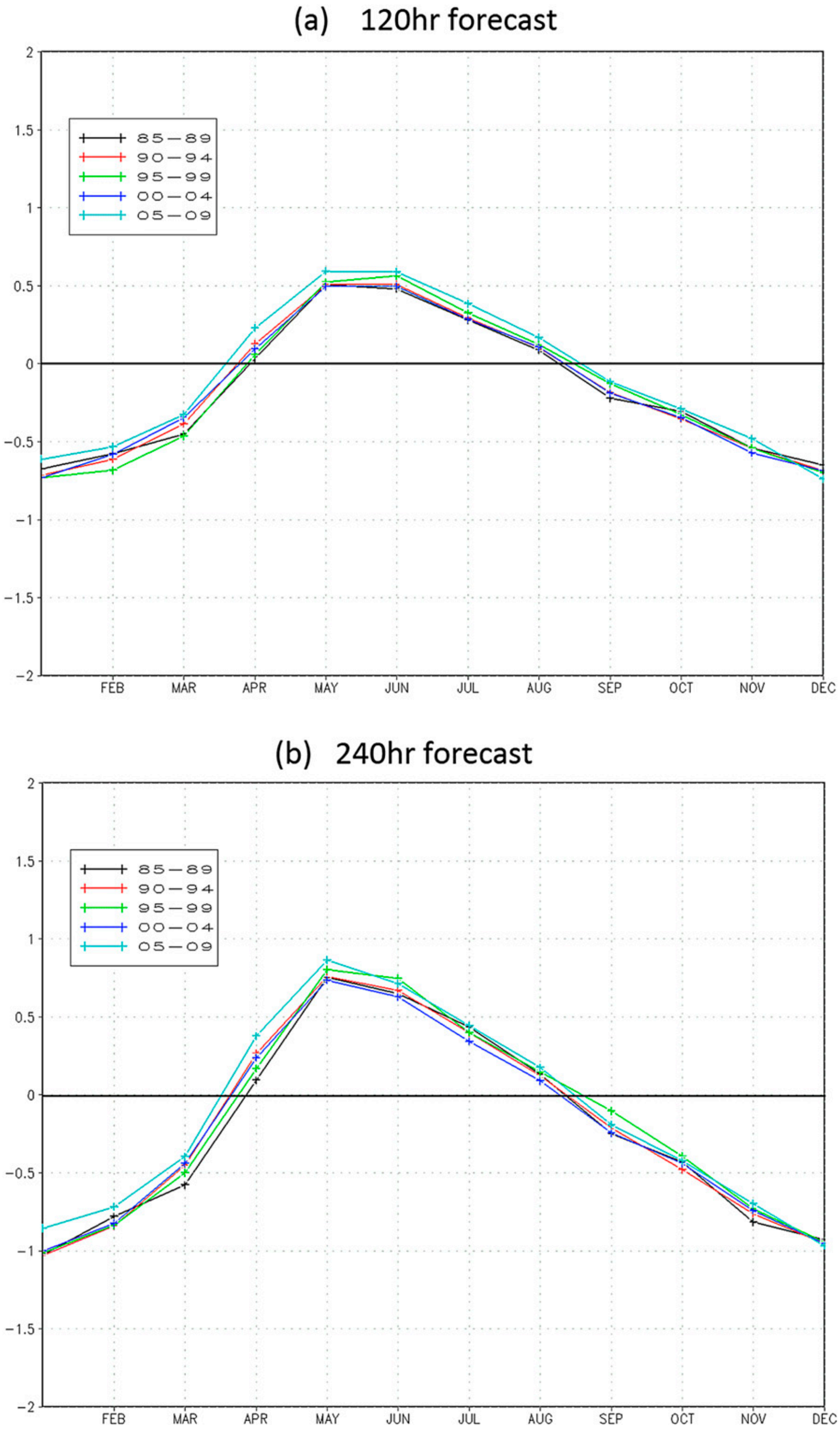

FIG. 1. Errors in valid 2-m temperature forecasts averaged over 5-yr periods for the NH during the reforecast period for (a) 120- and (b) 240-h projections. Black lines indicate the errors for 1985–89, red lines indicate the errors for 1990–94, green lines indicate the errors for 1995–99, blue lines indicate the errors for 2000–04, and light blue lines indicate the errors for 2005–09. Thick black lines indicate bias = 0.

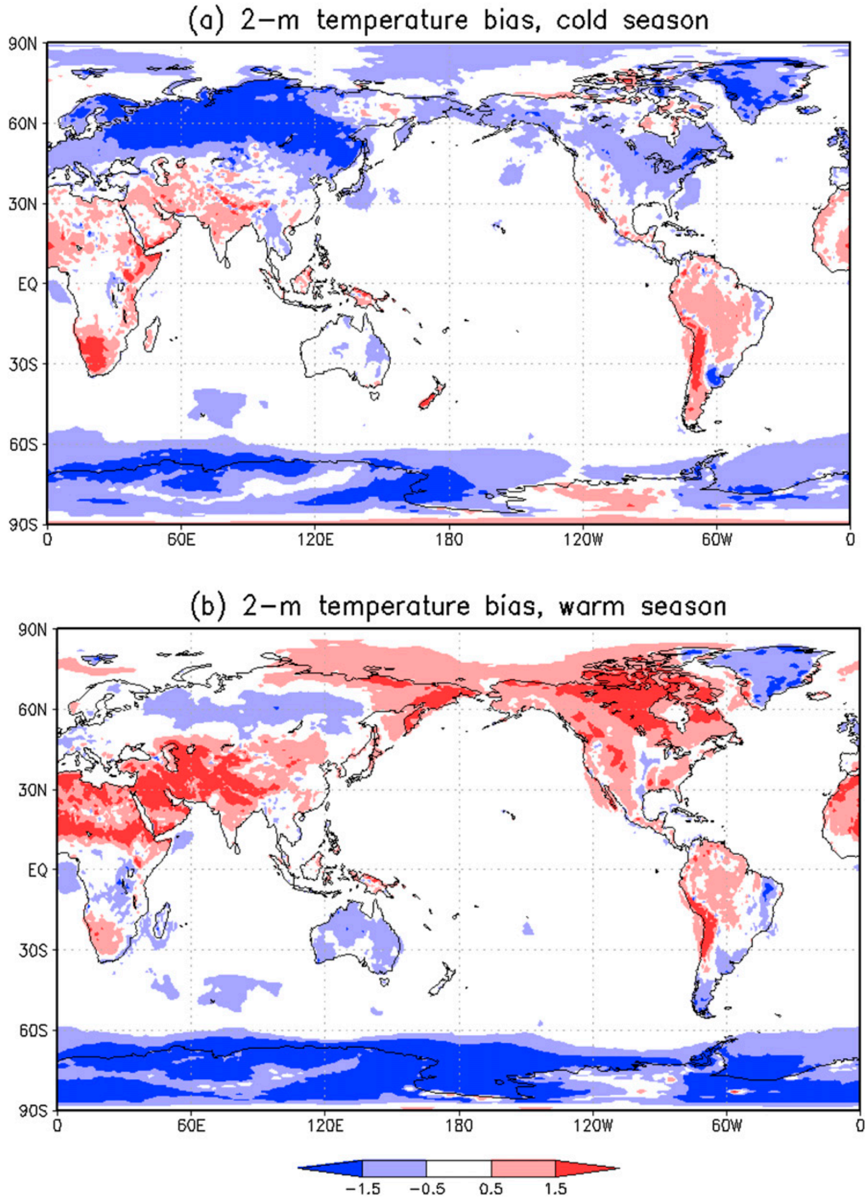

FIG. 2. Global 2-m temperature error averaged over the 25-yr reforecast period for 120-h forecasts during the (a) cold season (from 1 Jan to 15 Mar and from 16 Aug to 31 Dec) and (b) warm season (from 16 Mar to 15 Aug).

Hemisphere is more dramatic than that of the Southern Hemisphere (SH), suggesting that the decaying method faces a challenge mainly in the continents of the NH. For example, in the warm season, the positive bias (Fig. 2b) is dominant over North America with a considerable area having a bias exceeding 1.5 K. During the cold season (Fig. 2a), the maximum negative bias also exceeds 1.5 K. In contrast, the change in bias is much smaller in the continents of the SH. This is possibly due to the fact that most of the landmass in the SH is in the tropics and subtropics, while the NH has much more landmass at higher latitudes.

4. Experiments and results

We calibrate 2-m temperature for 2009 and 2010 using the prior 24- (1985–2008) and 25-yr biases (1985–2009), respectively. We also calibrate the 500-hPa height for 2009 using the 24-yr bias, but a preliminary check shows that it is very hard to improve the forecast skill of this

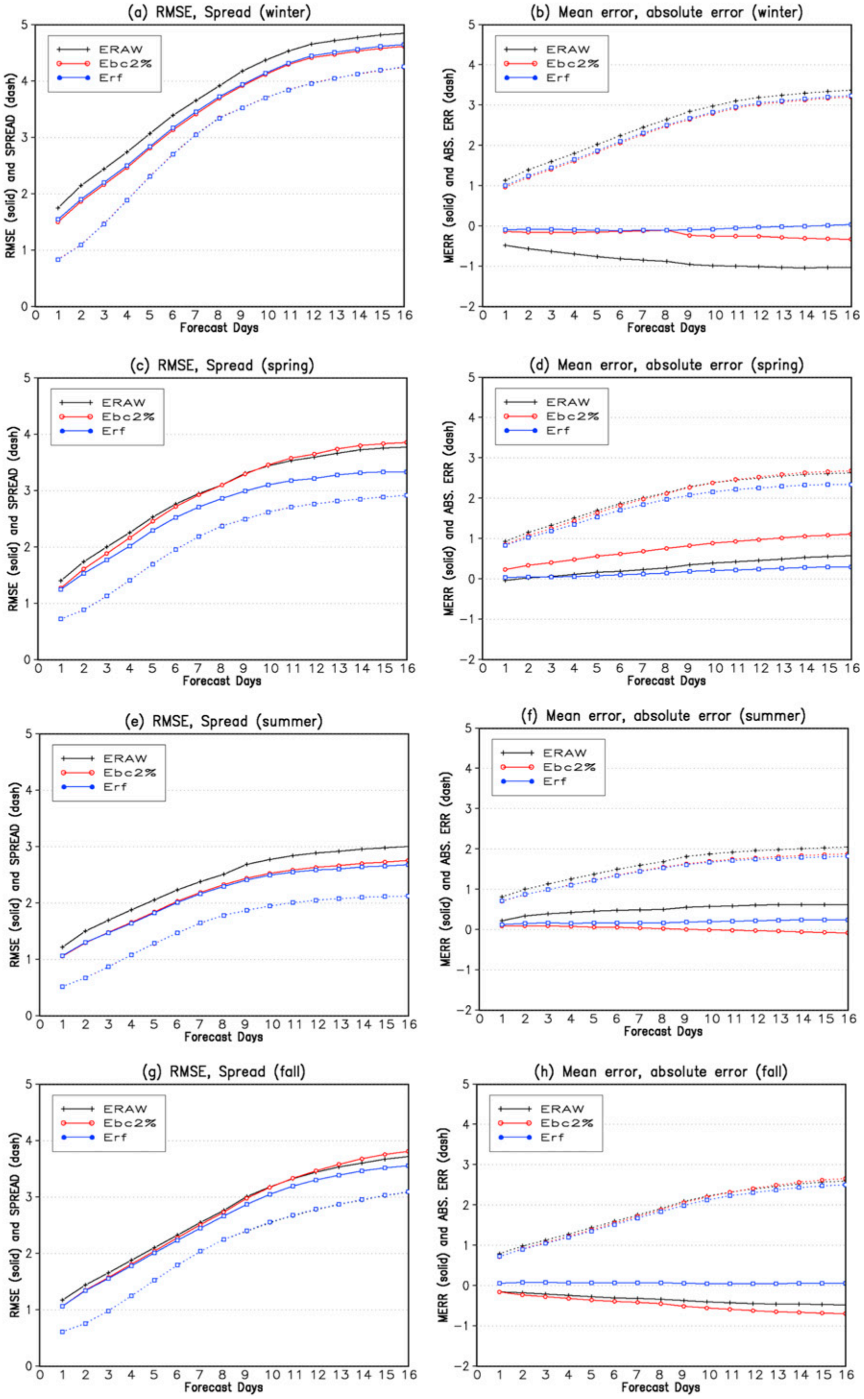

FIG. 3. For 2-m temperature averaged over the NH during the four seasons in 2010, the (left) ensemble mean RMSE (solid lines) and spread (dashed lines), and (right) ensemble mean error (solid lines) and absolute error (dashed lines). In the legends, ERAW, Ebc2%, and Erf represent the raw (black lines), decaying-bias-corrected ensemble (red lines), and reforecast-bias-corrected ensemble (blue lines) forecasts, respectively.

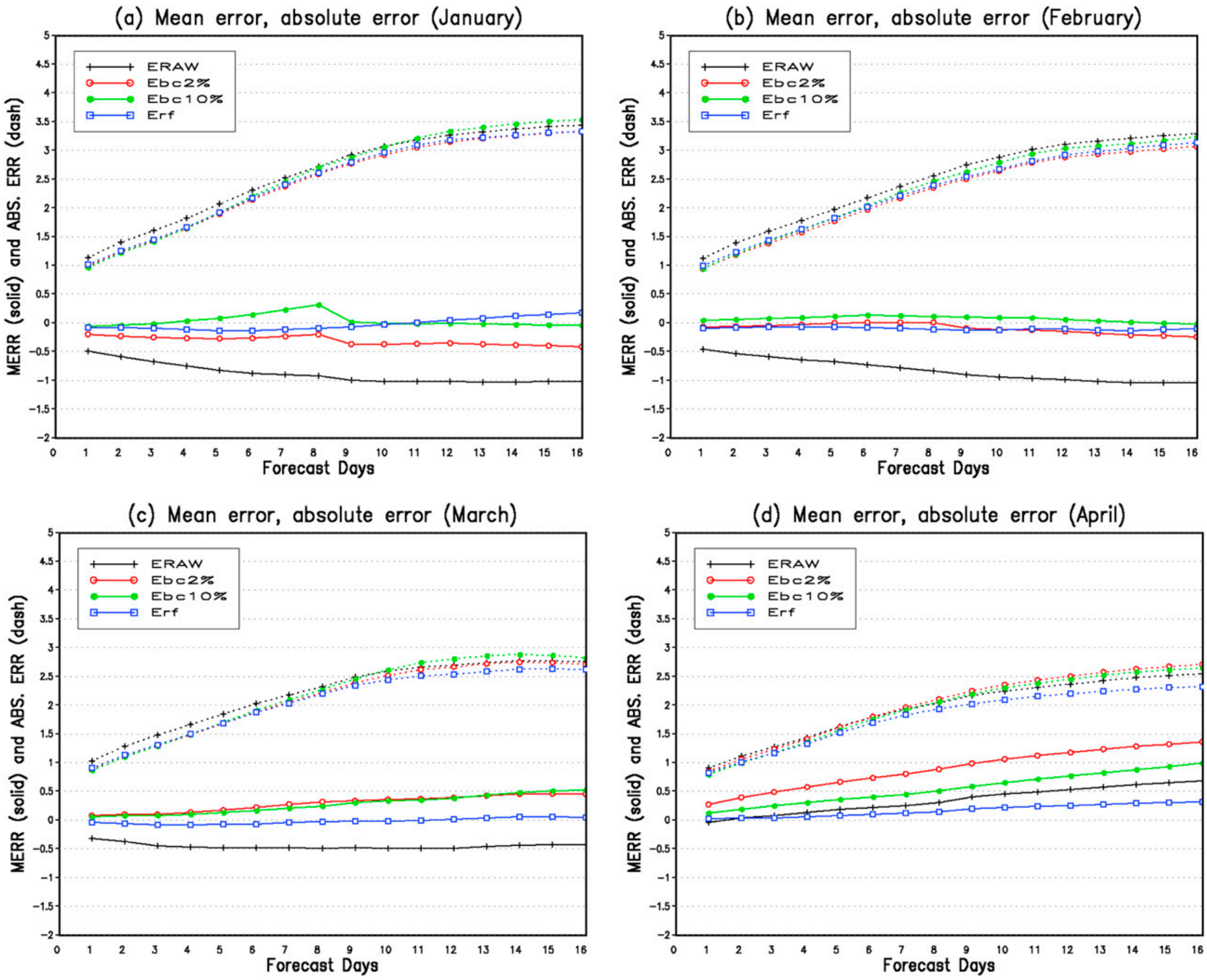

FIG. 4. Mean error (solid lines) and mean absolute error (dashed lines) of 2-m temperature averaged over the NH for (a) January, (b) February, (c) March, and (d) April 2010. In the legends, ERAW, Ebc2%, Ebc10%, and Erf represent the raw (black lines), decaying-bias-corrected method using a 2% weight (red lines), decaying-bias-corrected method using a 10% weight (green lines), and reforecast-bias-corrected ensemble (blue lines) forecasts, respectively.

variable, possibly because of its relatively small bias or sensitivity. Thus, our focus will be on the calibration of 2-m temperature. We explore the sensitivity of the calibration to the number of training years by using the bias from the most recent 2 (2008–09), 5 (2005–09), 10 (2000–09), and 25 (1985–2009) yr of training data, and evaluate the last year (2010) of independent forecasts. We compare the calibrations using three different training-data windows (1, 31, and 61 days) centered on the corresponding forecast date in each of the training years (25 yr). The impact of the sampling interval on the calibration is estimated by comparing verification scores with a sampling interval of 1 day (daily) and 7 days (weekly) for the 31-day window. We first calculated the 2-m temperature error for each day. From this dataset, we created weekly sampling data by using the error every seventh day from the starting date (1 January 1985) of the reforecast. For each year, the daily and weekly sampling creates 31 and 4 or 5 datasets, respectively. Finally, we apply reforecast information to the NCEP operational GEFS product.

a. *Calibrating the 2010 forecasts using the 25-yr training dataset*

Figure 3 shows the verification for 2-m temperature over the Northern Hemisphere for the four seasons. For the 2009/10 winter season, only January and February are included in the verification in order to keep the same training sample size or the same training years. We present a comparison of the results of the raw ensemble forecast (ERAW) and two calibrated forecasts (Ebc2% and Erf). Here, Ebc2% and Erf denote the decaying

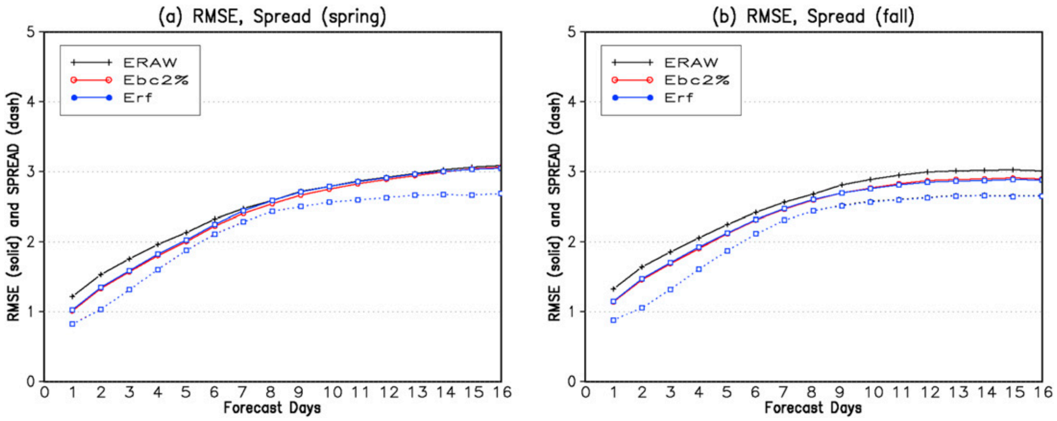

FIG. 5. Ensemble mean RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the SH for (a) spring and (b) autumn 2010. In the legends, ERAW, Ebc2%, and Erf represent the raw (black lines), decaying-bias-corrected (red lines), and reforecast-bias-corrected ensemble (blue lines) forecasts, respectively.

average and reforecast bias-correction methods, respectively. A weight of 2% is used for the decaying method.

The GEFS model is underdispersed for all seasons and lead times (Figs. 3a,c,e,g). Our focus here is on the first-moment adjustment. Improvement for the second-moment adjustment will be addressed in a future spread adjustment paper.

The raw ensemble forecast (black lines) has a cold bias during the winter (Fig. 3b) and autumn (Fig. 3h). Conversely, a warm bias is prevalent during the spring (Fig. 3d) and summer (Fig. 3f). These biases are almost completely corrected by the Erf method (blue lines). The corrected bias is closer to zero for all forecast lead times and the corresponding absolute error and RMSE are also smaller than for the raw ensembles, hinting at the effectiveness of the calibration methods in reducing the systematic error of the ensemble forecast. The Ebc2% also does a good job in the nontransitional seasons (winter and summer), and even performed slightly better than the Erf method in winter. However, this technique does not work well in all circumstances, as pointed out in Cui et al. (2012). Figures 3d and 3h reveal that applying the decaying method leads to a degradation of forecast accuracy during transition seasons throughout almost all lead times. The maximum degradation occurs in spring. As indicated in Figs. 3a, 3c, 3e, and 3g, the simple bias-correction methods do not change the ensemble spread since the bias of the ensemble mean is applied to each ensemble member.

The mean errors in the Ebc2% method are larger than those of the ERAW forecast in the spring and autumn (Figs. 3d,h). To determine the underlying reason, we display the month-to-month evolutions of mean error and mean absolute error of 2-m temperature for the three experiments over the Northern Hemisphere in Fig. 4. In addition to the above three experiments, the results from the decaying method with a weight of 10% are also added to the comparison. We note a persistent cold bias of the raw forecast in the winter (January and February). In the beginning of spring (March), the cold bias becomes smaller and eventually turns into a warm bias in April for almost all lead times. For example, the bias is about −1°C for the day-8 forecasts in January (Fig. 4a) and February (Fig. 4b).The corresponding values are −0.5°C in March (Fig. 4c) and 0.3°C in April (Fig. 4d), respectively. In the two winter months, the performance by the Ebc2% method is very similar to that for Erf, yielding a more accurate forecast than the raw ensembles. This is due to the ensemble forecast error being relatively consistent during the nontransitional months. The 2% and 10% decaying averages incorporate the most recent 50–60 and 10–15 days of bias information (Cui et al. 2012) with the highest weight for the latest information. The Ebc2% fails to improve the forecasts in March and April, when error characteristics change dramatically within a period of ~50–60 days. In April, Ebc2% uses a cold bias, accumulated from winter and early spring, to calibrate a warm bias in spring. This outdated information degrades the forecast (i.e., increases the warm bias), which is most pronounced for longer forecast lead times. This is likely due to a larger separation of training data from the actual forecast day of interest. In other words, the longer-lead-time forecasts are being trained on forecasts made further back because the more recent forecasts were not used to compute the error as their valid date has not passed yet. During the transition seasons, Erf has an obvious advantage over Ebc2% and Ebc10%, particularly for the long-lead forecasts. The Ebc10% method is slightly

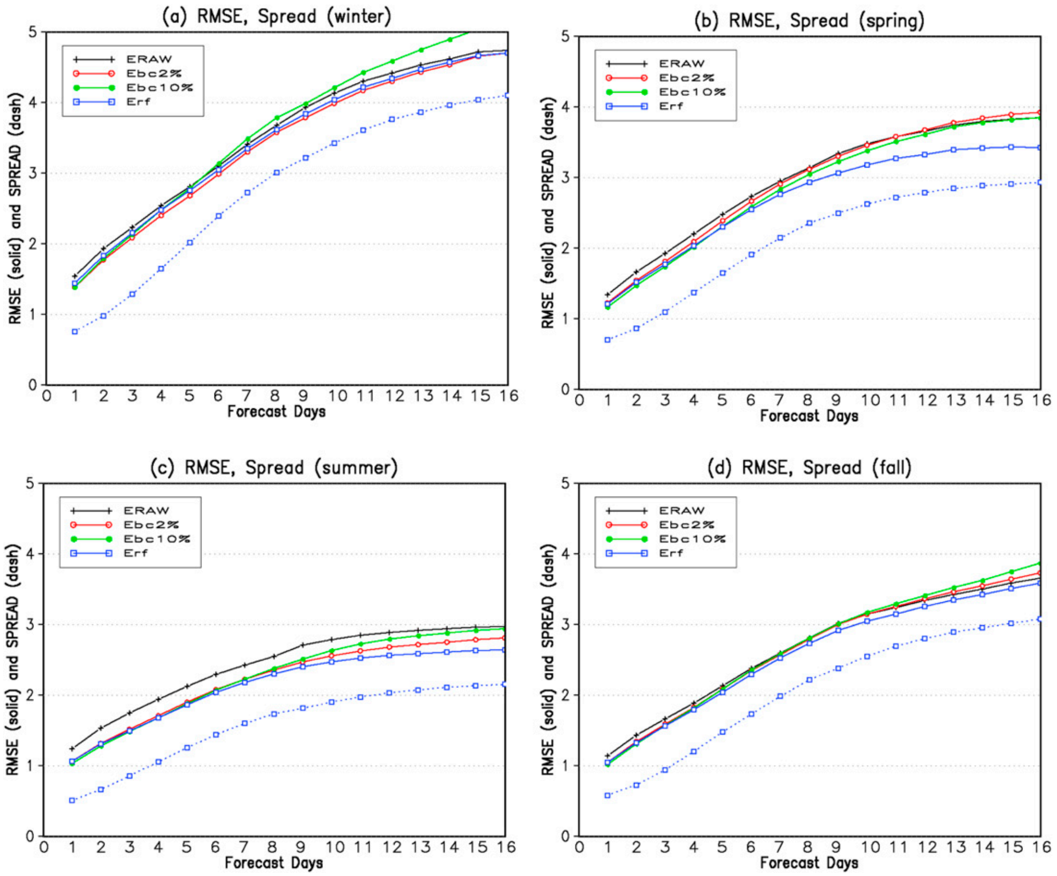

FIG. 6. RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the NH during the four seasons in 2009. In the legends, ERAW, Ebc2%, Ebc10%, and Erf represent the raw (black lines), decaying-bias-corrected with the two weights (2% red lines and 10% green lines), and reforecast-bias-corrected ensemble (blue lines) forecasts, respectively.

better than Ebc2% since it uses more recent error information.

Unlike in the Northern Hemisphere, the decaying method in the Southern Hemisphere does not degrade the forecast skill in the spring and autumn transition seasons as illustrated in Fig. 5. The performance of the reforecast method is very similar to the decaying method. This is likely due to less seasonal variation of model bias because of the ocean (Fig. 2).

b. Comparison between 2009 and 2010

The improvement in the accuracy of the 2-m temperature forecasts by Erf for 2010 is impressive. The key question is whether this improvement is unique to the year 2010. To answer this question, we also calibrate the 2009 forecast and compare the results to 2010. The data prior to the validation year (2009) are used to train the reforecast-bias-correction algorithm.

Figure 6 shows the RMSE and spread of 2-m temperature for 2009 for the Northern Hemisphere. Figure 7 provides the comparisons of mean error and mean absolute error between 2009 and 2010. The performance in 2009 is, qualitatively, very similar to that in 2010. The cold bias in winter and autumn and the warm bias in spring and summer can also be seen in 2009 (Fig. 7). The Ebc2% method, again, improves the forecast in the nontransitional seasons for all lead times but does not improve the forecast during the other two seasons, when Ebc2% tends to degrade the forecasts, particularly for the longer-lead-time forecasts. The Erf approach improves the ensemble forecasts over Ebc2% in transition seasons, as noted in 2010. The biases for all seasons are, again, mostly removed by Erf. However, the extent to which Erf can improve RMSE is slightly different. The improvement in winter and autumn for 2009 is slightly less than for 2010.

c. Calibration using various training samples

The CRPS of forecasts from the raw ensemble (black line) and calibrated ensembles (color lines) with training samples of various sizes are displayed in Fig. 8. The

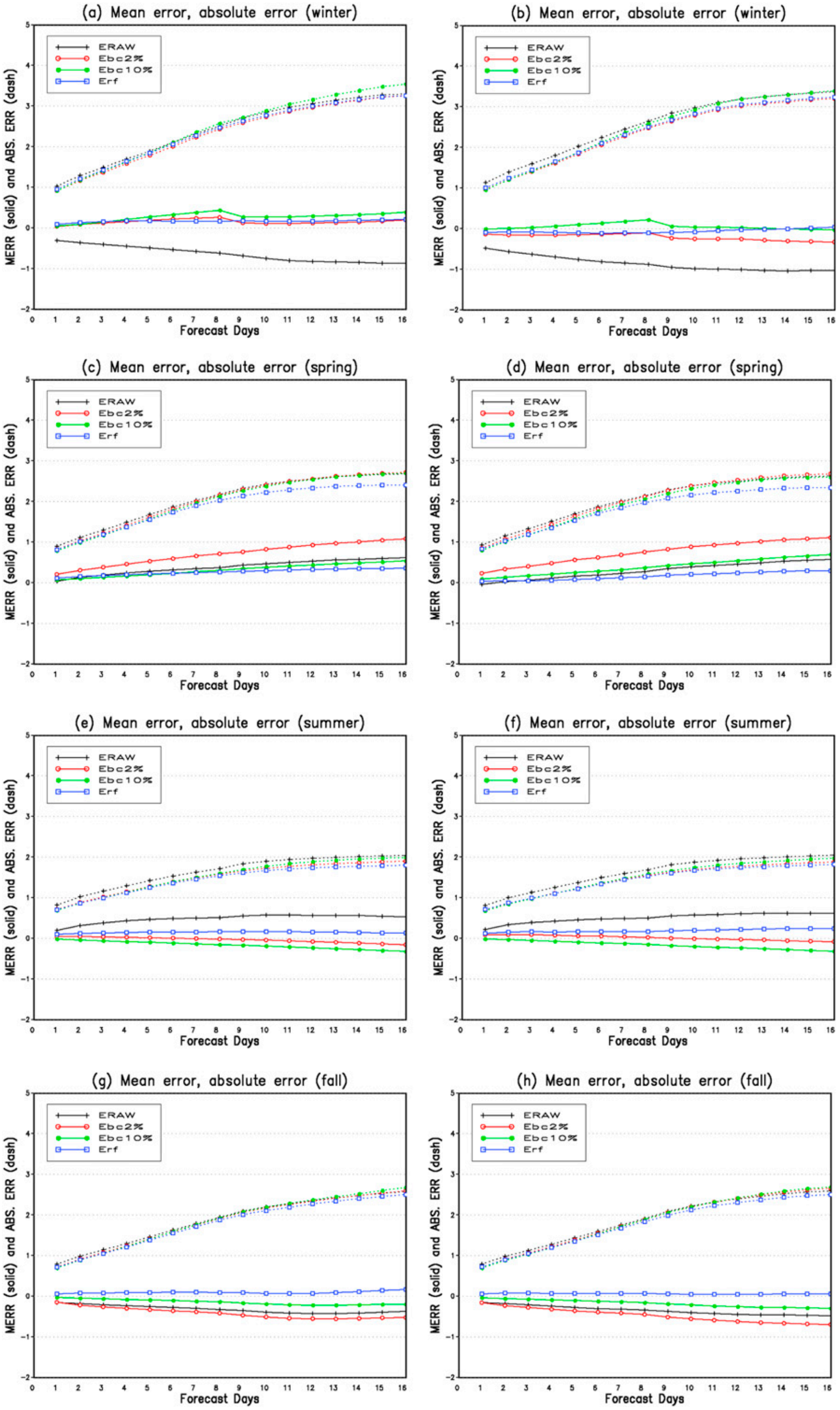

FIG. 7. Comparisons of mean errors (solid lines) and mean absolute errors (dashed lines) of 2-m temperature over the NH between (left) 2009 and (right) 2010 for (a),(b) winter; (c),(d) spring; (e),(f) summer; and (g),(h) autumn. In the legends, ERAW, Ebc2%, Ebc10%, and Erf represent the raw (black lines), decaying-bias-corrected with the two weights (2%, red lines; 10%, green lines), and reforecast-bias-corrected ensemble (blue lines) forecasts, respectively.

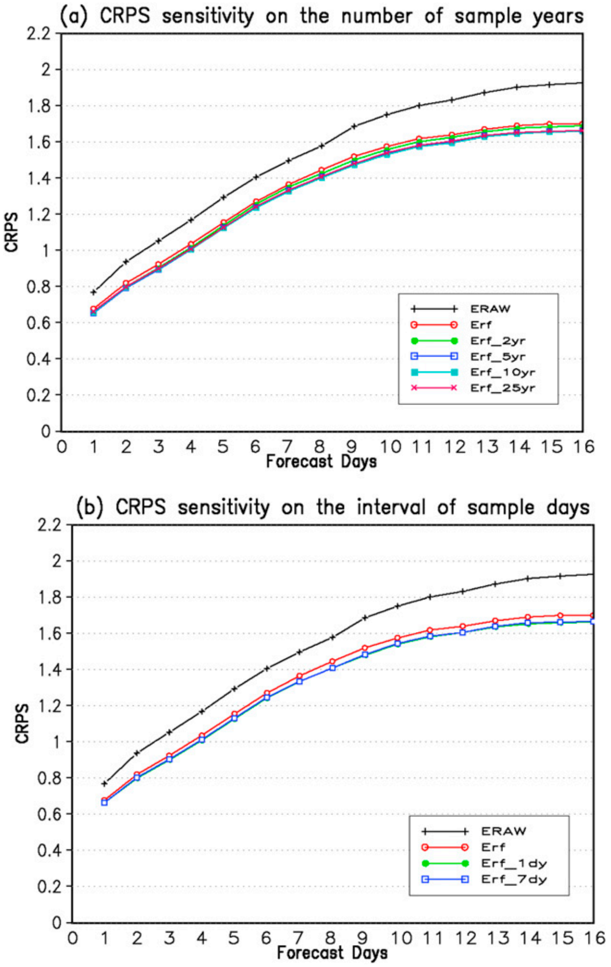

FIG. 8. CRPS of 2-m temperature averaged from 1 Mar to 31 May 2010 over the NH. In the legends, ERAW is the raw ensemble forecast (black lines) and Erf is the reforecast-bias-corrected ensemble forecast with historical data at the exact forecast date (red lines). In (a), Erf_2yr, Erf_5yr, Erf_10yr, and Erf_25yr are the reforecast-bias-corrected ensemble forecasts with historical data spanning a time window of 31 days, centered on the forecast day for the most recent 2 (green line), 5 (blue line), 10 (cyan line), and 25 yr (magenta line), respectively. In (b), Erf_1dy and Erf_7dy are the reforecast-bias-corrected ensemble forecasts using all 25 yr of historical data, covering a time window of 31 days centered on the forecast day. The frequencies of the data samples for Erf_1dy and Erf_7dy in (b) are 1 day (green line) and 7 days (blue line), respectively.

results for RMSE are very similar to those of CRPS (not shown). Figures 8a and 8b examine the sensitivity of forecast skill to the number of sample years and interval days, respectively. All calibrated forecasts demonstrate better performance than the raw forecast. The difference among the calibrated forecasts is relatively small with only a small degradation for each shorter period. The scores for 5, 10, and 25 yr with a 31-day window are very similar, slightly better than the other smaller training samples (Fig. 8a), suggesting that the 5-yr dataset is large enough to cover a wide range of weather types or scenarios. The CRPS of the forecasts from the calibration with the 25-yr weekly dataset (blue line) and 25-yr daily dataset (green line) within a 31-day window are almost identical (Fig. 8b) and both are better than the result using a single data value from each year (red line). A further increase in window size from 31 to 61 days (not shown) does not bring any obvious change. Therefore, the 25-yr, 31-day weekly training dataset is a good option for reducing computational expense while maintaining desired skill. These results are consistent with the findings of previous researchers (Hamill et al. 2004; Hagedorn et al. 2008), although they used different model or GEFS versions.

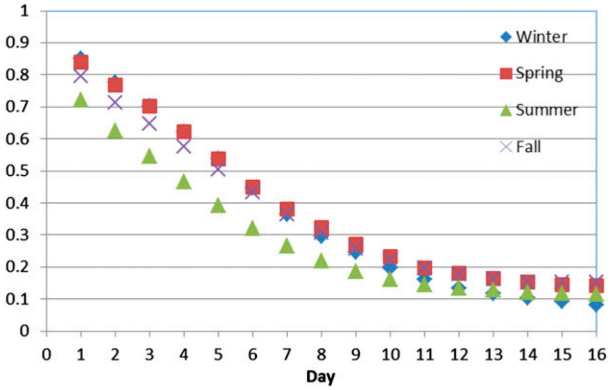

FIG. 9. The change in the square of the correlation coefficient between the ensemble mean and analysis with forecast lead time for the four seasons during 2010.

5. Using the reforecast to improve the NCEP bias-corrected product

Having seen the remarkable value of using reforecast information, we now combine the Erf with the operational Ebc2% method, aimed at providing an option for improving forecast accuracy during transition seasons. Figure 9 displays the change in r^2 with forecast lead time, averaged over the Northern Hemisphere for the four seasons of 2010. The r^2 denotes the square of the correlation coefficient between the ensemble mean and analysis. Forecast ability declines as forecast lead time increases. The r^2 values are slightly smaller in summer than during other seasons for short lead times. Almost equal weight is given to the two methods for ~day-5 forecasts [see Eq. (4) and Fig. 9]. Thereafter, the decaying method is expected to become less powerful (Cui et al. 2012) as the reforecast method starts playing a more important role.

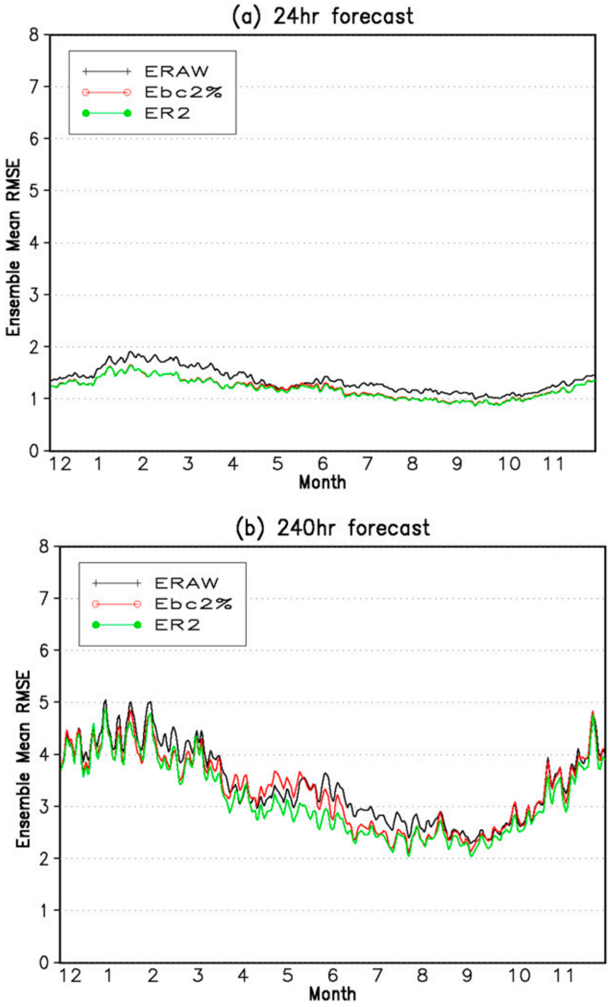

FIG. 10. RMSE of 2-m temperature averaged over the NH for (a) 24- and (b) 240-h forecasts between December 2009 and November 2010. In the legends, ERAW, Ebc2%, and Er2 denote the raw (black lines), decaying-bias-corrected (red lines), and decaying-reforecast-bias-corrected ensemble (green lines) forecasts, respectively.

Figure 10 shows the time series of RMSE for ERAW, Ebc2%, and Er2 for the 24- and 240-h forecasts of 2010. Ebc2% and Er2 represent the bias-corrected forecasts with the decaying method and combined decaying–reforecast method, respectively. For the 24-h forecast (Fig. 10a), the Ebc2% RMSE is smaller than the raw forecast for the majority of the period. Including the reforecast bias correction (Er2) does not change forecast accuracy noticeably since the weight of the reforecast is small at this short lead time (see Fig. 9). For the 240-h forecast, Ebc2% does not always improve the forecast, but shows a significant degradation in the forecast during the spring season. Our results agree with those in Cui et al. (2012), who found that the decaying averaging method mainly works well for the first few days. It is also very clear that the combined method performs better than the decaying average method, except at the end of January, when the reforecast degrades the operational bias-corrected product. The combined method leads to a maximum improvement in April–June.

Figure 11 displays the corresponding seasonal-average RMSE and spread. The result using the reforecast bias correction is added into the comparisons to see if there is any gain from using the decaying average rather than just the reforecast. For the transition seasons, the reforecast correction always gives the best performance. The combined method slightly degrades the reforecast-corrected forecast. In summer and winter, in general, the results are very similar between the reforecast and combined methods. Although the combined method beats the decaying method most of the time (Fig. 10), we still think the decaying method is a valuable or operationally applicable method. Here, we try to find a practical method of improving the forecast accuracy (RMSE) and reliability (bias free). Based on our study, large reforecast samples (at least 5 yr) will provide relatively ideal calibrated results (although there may not be enough cases for rare events). Figure 10 gives a maximum benefit from our study. In our real application, it is impossible to offer full ensemble reforecasts for many years (because of limited resources). In most cases, we can only provide a very limited number of reforecast samples; therefore, the decaying method (short hindcast training) is still a valuable (or operationally applicable) method. For the limited reforecast sample (25-yr, 1-day span), the reforecast method is not as good as decaying for the winter season (Fig. 3a).

6. Conclusions

In this paper, we develop a method for improving the NAEFS first-moment correction by using a 26-yr GEFS reforecast dataset. We use 24- and 25-yr GEFS reforecast bias information to calibrate 2009 and 2010 forecasts, respectively. We found that the forecast of 2-m temperature is strongly biased in the Northern Hemisphere, with a cold bias in the cold season and a warm bias in the warm season. Most of the bias is removed by the reforecast method. The decaying method improves the forecast skill in winter and summer, as does the reforecast method, but it degrades long-lead forecasts during transition seasons because of dramatic changes in the bias characteristics.

Several different methods have been examined to optimize the usage of the past 25-yr reforecast information.

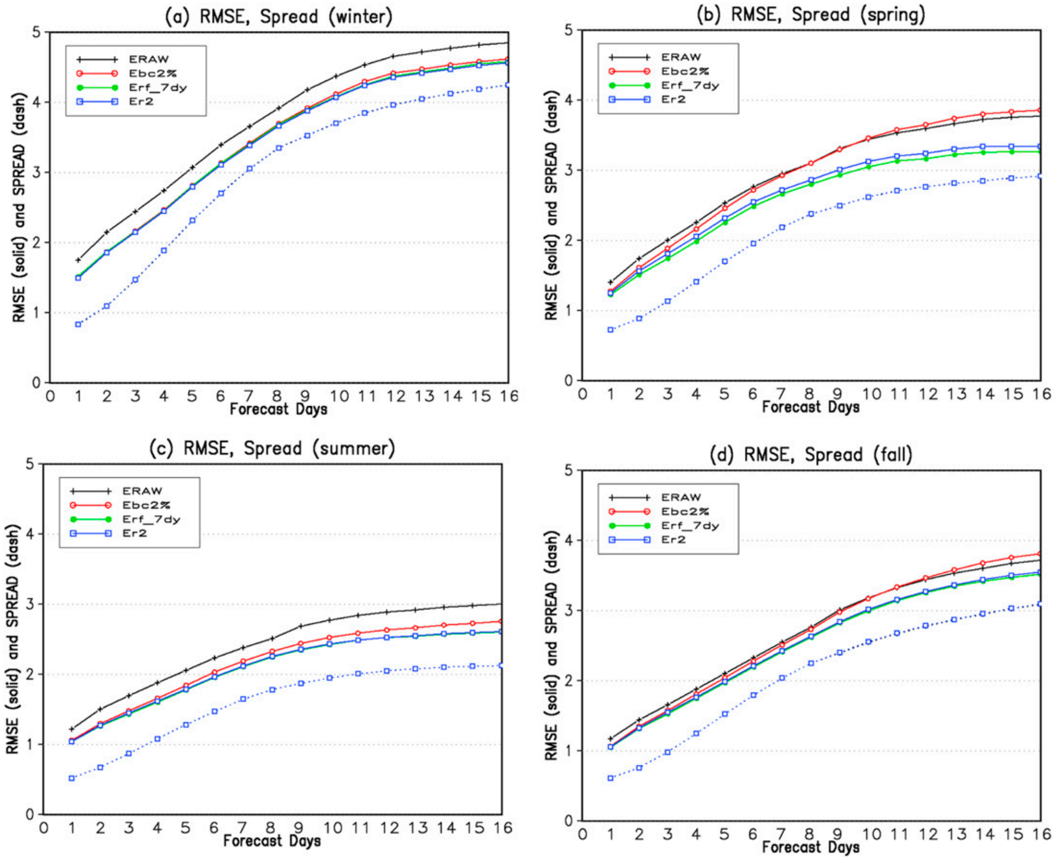

FIG. 11. RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the NH during the four seasons in 2010. In the legends, ERAW, Ebc2%, Erf_7dy, and Er2 represent the raw (black lines), decaying-bias-corrected (red lines), reforecast-bias-corrected (green), and decaying-reforecast-bias-corrected ensemble forecasts (blue lines), respectively. In the Erf_7dy simulation, the historical data span a time window of 31 days, centered on the forecast day for the full 25 yr with a sample frequency of 7 days.

This is important considering the limited computing resources. Based on the sensitivity tests for different reforecast samples, we found that the 25-yr weekly training dataset is a good option for reducing computational expense while maintaining the desired skill.

To provide an option for improving forecast accuracy during transition seasons, we add reforecast information into the current operational bias-correction method. The relative contribution of the two methods is quantified by using a correlation coefficient between the ensemble mean and the analysis. In general, the combined method performs better than the decaying average method except at the end of January. The maximum improvement occurs in April–June.

The current work and previous studies (Hamill et al. 2013) demonstrate the important value of using reforecast information to improve forecast skill. However, bias and its seasonal variation are model dependent. Whether the improvement found here will occur in the new GEFS version needs to be confirmed in the future. Frequent model upgrades make calibration using reforecasting very difficult because creating reforecast datasets requires huge computer resources. Hamill et al. (2014) are working toward finding the most valid configuration of the real-time GEFS reforecast runs. This would make a calibration using the reforecast method feasible in operations.

Acknowledgments. We thank NOAA's Earth System Research Laboratory for the reforecast dataset. We also acknowledge the helpful advice and commentary of Walter Kolczynski (NCEP) and Mike Charles (NCEP). Special appreciation goes to Mary Hart (NCEP) for editing the English in this paper.

REFERENCES

Bishop, C. H., and K. T. Shanley, 2008: Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon. Wea. Rev.,* **136,** 4641–4652, doi:10.1175/2008MWR2565.1.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.,* **137,** 1655–1665, doi:10.1175/2008MWR2682.1.

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, doi:10.1175/2007WAF2006084.1.

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.

Friederichs, P., and T. Thorarinsdottir, 2012: Forecast verification scores for extreme value distributions with an application to peak wind prediction. *Environmetrics,* **23,** 579–594, doi:10.1002/env.2176.

Glahn, B., 2014: Determining an optimal decay factor for bias-correcting MOS temperature and dewpoint forecasts. *Wea. Forecasting,* **29,** 1076–1090, doi:10.1175/WAF-D-13-00123.1.

——, M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.,* **137,** 246–268, doi:10.1175/2008MWR2569.1.

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, doi:10.1175/MWR2904.1.

Hagedorn, R., T. M. Hamill, and S. J. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.,* **136,** 2608–2619, doi:10.1175/2007MWR2410.1.

——, R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.,* **138,** 1814–1827, doi:10.1002/qj.1895.

Hamill, T. M., and J. S. Whitaker, 2007: Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Mon. Wea. Rev.,* **135,** 3273–3280, doi:10.1175/MWR3468.1.

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46, doi:10.1175/BAMS-87-1-33.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast data set. *Bull. Amer. Meteor. Soc.,* **94,** 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

——, and Coauthors, 2014: A recommended reforecast configuration for the NCEP global ensemble forecast system. NOAA White paper, 24 pp, accessed 9 June 2015. [Available online at http://www.esrl.noaa.gov/psd/people/tom.hamill/White-paper-reforecast-configuration.pdf.]

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Evaluation of the impact of the stochastic perturbation schemes on global ensemble forecast. *Proc. 19th Conf. on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc. [Available online at https://ams.confex.com/ams/88Annual/webprogram/Paper134165.html.]

Monache, L. D., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to postprocess numerical weather predictions. *Mon. Wea. Rev.,* **139,** 3554–3570, doi:10.1175/2011MWR3653.1.

——, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.,* **141,** 3498–3516, doi:10.1175/MWR-D-12-00281.1.

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, doi:10.1175/MWR2906.1.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus,* **55A,** 16–30, doi:10.1034/j.1600-0870.2003.201378.x.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.,* **91,** 1015–1057, doi:10.1175/2010BAMS3001.1.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330, doi:10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319, doi:10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

——, 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.

Wagner, J., and B. Glahn, 2010: Ensemble MOS forecasts from multiple models. Preprints, *20th Conf. on Probability and Statistics in the Atmospheric Sciences,* GA, Amer. Meteor. Soc., 7.5. [Available online at https://ams.confex.com/ams/pdfpapers/156479.pdf.]

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.,* **131,** 965–986, doi:10.1256/qj.04.120.

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, doi:10.1111/j.1600-0870.2007.00273.x.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2d ed. Academic Press, 627 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390, doi:10.1175/MWR3402.1.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385, doi:10.1175/MWR3347.1.

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.

Zhu, Y., and Z. Toth, 2008: Ensemble based probabilistic verification. Preprints, *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.]

——, R. Wobus, M. Wei, B. Cui, and Z. Toth, 2007: March 2007 NAEFS upgrade. NOAA/NCEP/EMC, accessed 9 June 2015. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.]