Improvement of Statistical Post-processing Using GEFS Reforecast Information

Hong Guan[1,2*], Bo Cui[1,3], and Yuejian Zhu[1]

[1]Environmental Modeling Center/NCEP/NWS/NOAA, College Park, MD

[2]System Research Group Inc., Colorado Springs, CO

[3]I. M. Systems Group, Inc., College Park, MD

To be submitted to *Weather and Forecasting*

* *Corresponding author address:*

Dr. Hong Guan,

Environmental Modeling Center/NCEP/NWS/NOAA

5830 University Research Court

College Park, MD 20740

E-mail: Hong.Guan@noaa.gov

Abstract

The National Oceanic and Atmospheric Administration (NOAA)/Earth System Research Laboratory (ESRL) generated a multi-decadal (1985-current) 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 00UTC. This GEFS version has a strong cold bias for winter and warm bias for summer in the Northern Hemisphere. Although the operational Kalman Filter (KF) 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, we use a 24-year (1985-2008) and 25-year (1985-2009) reforecast bias 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 KF method during transition seasons. A long training period (at least 5 years) 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-year, 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 of improving forecast accuracy for transition seasons, we add reforecast information 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 in April, May, and 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 multi-center, multi-model ensemble forecast system to estimate the uncertainty of weather forecasts and 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 Canadian Meteorological Center 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 post-processing 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 (Raftery et al., 2005; Wilson et al., 2007); non-homogeneous Gaussian regression (Gneiting et al., 2005); Gaussian ensemble dressing (Roulston and Smith, 2003; Wang and Bishop, 2005; Bishop and Shanley, 2008); a simple analog technique (Hamill et al., 2004; 2006; 2013); and adaptive Kalman Filter (KF) (Cheng and Steenburg, 2007; 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 Processor (SPP). The current SPP

includes bias-correction and downscaling. The bias correction is mainly a first moment adjustment by applying an adaptive KF to accumulate the decaying averaging bias (Cui et al., 2012). 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 ensemble forecasts. This method is fast and does not need to store a large amount of sample data once initialized, which meets the requirements of 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 spring and fall transition seasons for long lead time forecasts.

Recently, NOAA/ ESRL generated an ensemble reforecast dataset using the 2010 version of GEFS. This multi-decadal dataset has been applied to precipitation calibration, diagnosis of the ability of GEFS to forecast uncommon phenomenon, and the initialization of regional reforecasts (Hamill, 2013). In this study, we use a 26-year 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 to the decaying method is highlighted in Section 5. The summary and 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-hour intervals up to 384 hours 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.

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

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, 005, 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 4 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.

$$B^p_{i,j} = (1-w) \times B^p_{i,j}(t-1) + w \times b_{i,j}(t) \quad (2)$$

c. Reforecast method

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 multi-year (N) reforecast ensemble.

$$B_{i,j}^{h} = \frac{\sum_{k=1}^{N} b_{i,j,k}(t)}{N} \quad (3)$$

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 ($B_{i,j}^{p}(t)$) and the reforecast bias ($B_{i,j}^{h}(t)$) to the raw forecast ($f_{i,j}(t)$) at each grid point (i,j), for each lead time (t), and each parameter.

$$F_{i,j} = f_{i,j}(t) - r^2 \times B_{i,j}^{p}(t) - (1-r^2) \times B_{i,j}^{h}(t) \quad (4)$$

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 relative contribution of the reforecast and decaying-average bias was quantified by r^2. The high correlation indicates that the model can well capture the temporal variation in the 2-m temperature analysis during the most recent period. This implies that bias feature is 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 equation represents 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, mean absolute forecast error, root mean square error (RMSE), and continuous rank probability score (CRPS). The CRPS score is frequently used for evaluating the performance of probabilistic forecasts (Zhu et al., 2008; Glahn et al., 2009; Friederichs and Thorarinsdottir, 2012). The lower the CRPS, the better the probabilistic system is in terms of reliability and resolution.

3. Model and reforecast data

The current operational GEFS version (v9.01) was implemented on February 14, 2012 at the National Centers for Environmental Prediction (NCEP). It consists of 21 members (one control member and 20 perturbation members) and is run 4 times daily (00, 06, 12, and 18 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 condition uses the ensemble transform technique (ETR, Wei et al., 2008). The model uncertainty is estimated using the stochastic tendencies (STTP) method (Hou et al., 2008).

The reforecast data was 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 00 UTC cycle for the 10 members. The control member was run at both 00 and 12 UTC. The dataset used here was bilinearly interpolated to 1°x1° latitude and longitude grids from the native resolution. These data are available from 1985 forward (+29 years). We use a subset of the data from 1985 to 2010 (26 years), 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-meter temperature errors over the Northern Hemisphere for 120-hr and 240-hr forecasts are displayed in Fig. 1. It is evident that there is a warm bias for April to August (warm season), while a cold bias is prevalent for the rest of year (cold season). The sharpest error change occurs between March and May with a change rate of ~0.5 and 0.6°C/month for the 120-hr and 240-hr forecasts, respectively. The large change in bias during the spring season can make it difficult to do bias correction with the current decaying average post- processing 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-year periods is relatively small. We did not find a significant improvement in forecast skill from the late 80's to the most recent year. The bias curve for the last 5-year period (2005-2009) shifts only slightly in the direction of positive bias from the first 5-year period (1985-1989). 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 semi-annual change in bias over the continents of the Northern Hemisphere (NH) 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 2K. During the cold season (Fig. 2a), the maximum negative bias also exceeds 1K. 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 sub-tropics, 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-year (1985-2008) and 25- year biases (1985-2009), respectively. We also calibrate the 500 hPa height for 2009 using the 24-year bias, but a preliminary check shows that it is very hard to improve the forecast skill of this variable, possibly due to 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 (2008-2009), 5 (2005-2009), 10 (2000-2009), and 25 (1985-2009) years 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 years). The impact of 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 7^{th} day from the starting date (Jan. 1, 1985) of the reforecast. For each year, the daily and weekly sampling creates 31and 4-5 datasets, respectively. Finally, we apply reforecast information to the NCEP operational GEFS product.

4.1 Calibrating the 2010 forecasts using the 25-year training dataset

Figure 3 shows the verification for 2-m temperature over the Northern Hemisphere for the 4 seasons. For the 2009/2010 winter season, only January and February are included in the

verification in order to keep the same sample size. We present a comparison of the results of the raw ensemble forecast (RAW) and two calibrated forecasts (Ebc2% and Erf). The Ebc2% and Erf denote the decaying average and reforecast bias-correction methods, respectively. A weight of 2% is used for the decaying method.

The GEFS model is under-dispersed for all seasons and lead times (Figs. 3a, c, e, and g). Our focus here is on the 1st moment adjustment. Improvement for the 2nd 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 good job in the non-transitional 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). Figs. 3d and h 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, c, e, and g, the simple bias correction methods do not change ensemble spread since the bias of the ensemble mean is applied to each ensemble member.

The mean errors in the Ebc2% are larger than those of the RAW forecast in the spring and autumn (Fig. 3d and 3h). 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 result from the decaying method with a weight of 10% is also added to the comparison. We note a persistent cold bias 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. In the two winter months, the performance by the Ebc2% is very similar to the Erf, yielding a more accurate forecast than the raw ensembles. This is due to the ensemble forecast error being relatively consistent during the non-transitional 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 forecast in March and April, when error characteristics change dramatically within a period of ~50-60 days. In April, the 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 weren't used to compute the error as their valid date hasn't passed yet. During the transition seasons, the Erf has an obvious advantage over the Ebc2% and Ebc10%, particularly for the long lead forecasts. The Ebc10% is slightly better than the 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).

4.2 Comparison between 2009 and 2010

The improvement in the accuracy of 2-m temperature forecasts by the 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 of 2010. The cold bias in winter and autumn and warm bias in spring and summer can also be seen in 2009 (Fig.7). The Ebc2%, again, improves the forecast in the non-transition seasons for all lead times but does not improve the forecast in the other two seasons, when the Ebc2% tends to degrade the longer lead time forecasts. The Erf improves the ensemble forecasts over the Ebc2% in transition seasons as noted in 2010. The biases for all seasons are, again, mostly removed by the Erf. However, the extent to which the Erf can improve RMSE is slightly different. The improvement in winter and autumn for 2009 is a little less than for 2010.

4.3 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 results for the RMSE are very similar to the CRPS (not shown). Figures 8a and b 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 years with a 31-day window are very similar, slightly better than the other smaller training samples (Fig. 8a), suggesting that the five-year dataset is large enough. The CRPS of the forecasts from the calibration with the 25-year weekly dataset (blue line) and 25-year 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-days to 61-days (not shown) does not bring any obvious change. Therefore, the 25-year 31-day weekly training dataset is a good option to reduce 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.

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 in 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 other seasons for short lead times.

Figure 10 shows the time series of RMSE for RAW, Ebc2%, and ER2 for the 24 and 240-hr forecasts of 2010. Ebc2% and ER2 represent the bias-corrected forecast with the decaying method and decaying-reforecast-combined method, respectively. For the 24-hr 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 too much since the weight of reforecast is small at this short lead-time (see Fig. 9). For the 240-hr 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 in the end of January where the reforecast degrades the operational bias-corrected product. The combined method leads to a maximum improvement in April, May, and 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.

6. Conclusions

In this paper, we develop a method to improve the NAEFS 1st-moment correction using a 26-year GEFS reforecast dataset. We use 24-year and 25-year 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. The bias is mostly removed by the reforecast method. The decaying method improves the forecast skill in winter and summer as well as the reforecast method, but it degrades long-lead forecasts during transition seasons due to dramatic changes in the bias characteristics.

Several different methods have been examined to optimize the usage of the past 25-year reforecast information. This is important considering limited computing resources. Based on the sensitivity tests for different reforecast samples, we found that the 25-year weekly training dataset is a good option to reduce computational expense while maintaining desired skill.

To provide an option for improving forecast accuracy for transition seasons, we add reforecast information into the current operational bias-correction method. The relative contribution of the two methods is quantified using a correlation coefficient between the ensemble mean and 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, May, and June.

The current work and previous studies (Hamil, 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 reforecast very difficult because creating reforecast dataset needs huge computer resources. Hamill et al. (2014) is making a great effort to find the most valid configuration of the real-time GEFS reforecast runs. This would make a calibration using the reforecast feasible in operations.

Acknowledgements:

We thank NOAA's Earth System Research Laboratory for 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 English.

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.

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

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.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410.

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

Glahn, B., 2014: Determining an optimal decay factor for bias-correcting MOS temperature and dewpoint forecasts, *Wea. Forecasting*, 29, 1076–1090.

Glahn, B., 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.

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.

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

Hamill, T. M. et al., 2014: A recommended reforecast configuration for the NCEP global ensemble forecast system. [White paper], Retrieved from http://www.esrl.noaa.gov/psd/people/tom.hamill/White-paper-reforecast-configuration.pdf.

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

Hamill, T. M. and J. S. Whitaker, 2007: Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2- meter temperatures using reforecasts. *Mon. Wea. Rev.*, 135, 3273-3280.

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.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, 87, 33-46.

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

Hamill T. M., 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.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Evaluation of the impact of the stochastic perturbation schemes on global ensemble forecast. Preprints, 19th Conference on Probability and Statistics, New Orleans, LA, Amer. Meteor. Soc.

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.

Roulston, M. S. and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, 55A, 16-30.

Saha, S. and co-authors, 2010: The NCEP climate forecast system reanalysis. *Bull. Amer. Meteor. Soc.*, 91, 1015–1057.

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

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

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts.

Wang, X. and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Q. J. R. Meteorol. Soc.*, 131, 965–986.

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.

Whitaker, J. S., F. Vitart, and X. Wei, 2006: Improving week two forecasts with multi-model re-forecast ensembles. *Mon. Wea. Rev*., 134, 2279-2284.

Wilks, D. S. and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev*., 135, 2379-2390.

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.

Zhu Y., R. Wobus, M. Wei, B. Cui, and Z. Toth, 2007: March 2007 NAEFS upgrade. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.]

Zhu Y. and Z. Toth, 2008: Ensemble based probabilistic verification, Preprints, the 19th conference on predictability and statistics, 20-24 January 2008, New Orleans, Louisiana, *Amer. Meteor. Soc.*.

Figure Captions:

Figure 1. Errors in valid 2-m temperature forecasts averaged over 5-year periods for the NH during the reforecast period for 120-hr (a) and 240-hr (b) projections. Black lines inidicate the errors for the 1985-1990 period, red lines for 1990-1994, green lines for 1995-1999, blue lines for 2000-2004, and light blue lines for 2005-2009. Thick black lines are bias=0 lines.

Figure 2. Global 2-m temperature error averaged over 25-year reforecast period for 120-hr forecasts during the cold season (1 Jan – 15 Mar and 16 Aug – 31 Dec) (a) and warm season (16 Mar – 15 Aug) (b).

Figure 3. Ensemble mean RMSE (solid lines, left panels), spread (dashed lines, left panels), error (solid lines, right panels), and absolute error (dashed lines, right panels) of 2-m temperature averaged over the Northern Hemisphere for the 4 seasons of the year 2010. Eraw, Rbc2%, and Erf are the raw (black lines), decaying-bias-corrected (red lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.

Figure 4. Mean error (solid lines) and mean absolute error (dashed lines) of 2-m temperature averaged over the Northern Hemisphere for (a) January, (b) February, (c) March, and (d) April, 2010. Eraw, Rbc2%, Rbc10%, and Erf are the raw (black lines), decaying-bias-corrected using a 2% weight (red lines), decaying-bias-corrected using a 10% weight (green lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.

Figure 5. Ensemble mean RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the Southern Hemisphere for spring (a) and autumn (b) of the year 2010. Eraw, Rbc2%, and Erf are the raw (black lines), decaying-bias-corrected (red lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.

Figure 6. RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the Northern Hemisphere for the 4 seasons of 2009. Eraw, Ebc2%, Ebc10%, and Erf are the raw (black lines), decaying-bias-corrected with the two weights (2% red lines and 10% green lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.

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

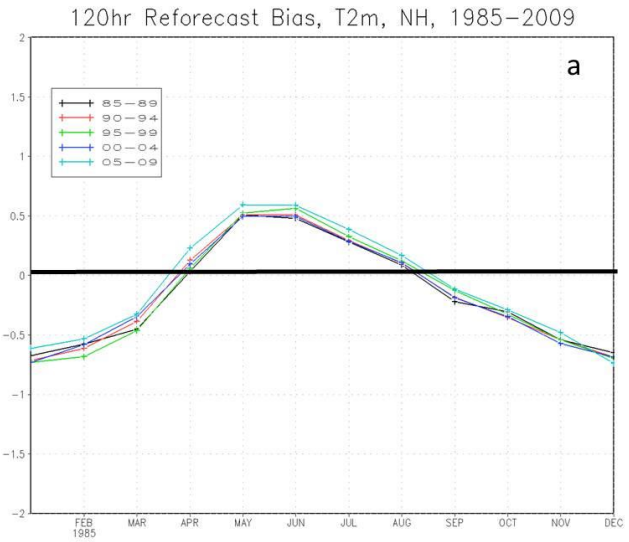
Figure 8. CRPS of 2-m temperature averaged from 1 Mar 2010 to 31 May 2010 over the Northern Hemisphere. ERAW is the raw ensemble forecast (black lines). Erf is the reforecast-bias-corrected ensemble forecast with historical data at the exact forecast date (red lines). Ey2d31rf, Ey5d31rf, Ey10d31rf, and Ey25d31rf in Fig. 8a 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 years (green line), 5 years (blue), 10 years (cyan), and 25 years (magenta). Ey25d31rf and Ey25d31int7rf in Fig. 8b are the reforecast-bias-corrected ensemble forecasts using all 25 years of historical data, covering a time window of 31 days centered on the forecast day. The frequency of data samples for Ey25d31rf and Ey25d31int7rf of Fig. 8b are 1 day (green line) and 7 days (blue line), respectively.

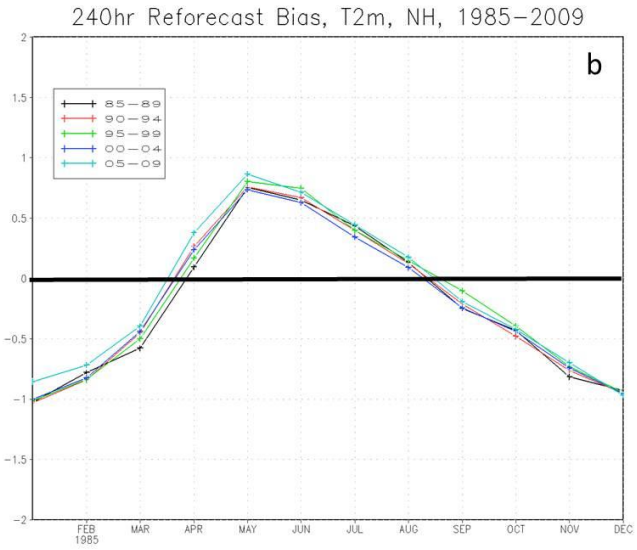
Figure 9. The change in the square of the correlation coefficient between the ensemble mean and analysis with forecast lead time for the 4 seasons of 2010.

Figure 10. RMSE of 2-m temperature averaged over the Northern Hemisphere for 24hr (a) and 240hr (b) forecasts between Dec 2009 and Nov 2010. ERAW, Ebc2%, and ER2 denote the raw forecast (black lines), decaying-bias-corrected forecast (red lines), and decaying-reforecast-bias-corrected ensemble forecast (green lines), respectively.

Figure 11. RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the Northern Hemisphere for the 4 seasons of 2010. Eraw, Ebc2%, and Ey2531dint7rf, and ER2 are the raw (black lines), decaying-bias-corrected (red lines), reforecast-bias-corrected (green) and decaying-reforecast-bias-corrected ensemble forecast (blue lines), respectively. In the Ey2531dint7rf, historical data spans a time window of 31 days, centered on the forecast day for the full 25 years with a sample frequency of 7 days.



1



2

3

4 Figure 1. Errors in valid 2-m temperature forecasts averaged over 5-year periods for the NH
5 during the reforecast period for 120-hr (a) and 240-hr (b) projections. Black lines inidicate the
6 errors for the 1985-1990 period, red lines for 1990-1994, green lines for 1995-1999, blue lines
7 for 2000-2004, and light blue lines for 2005-2009. Thick black lines are bias=0 lines.

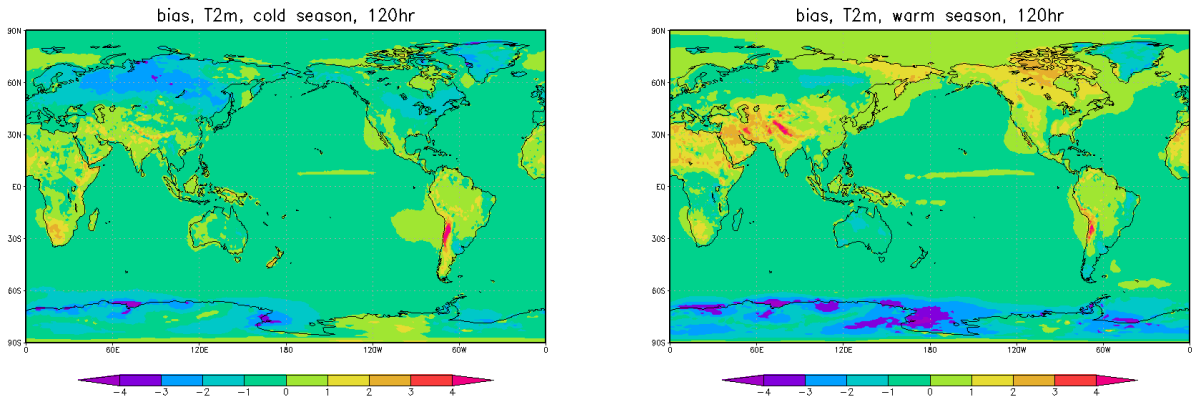
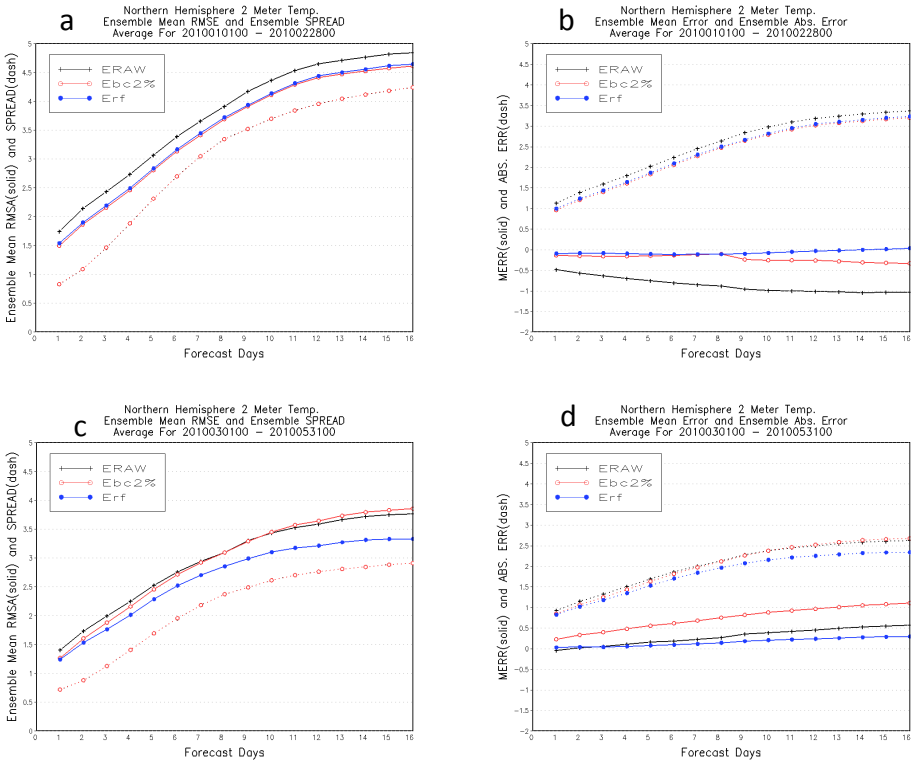


Figure 2. Global 2-m temperature error averaged over 25-year reforecast period for 120-hr forecasts during the cold season (1 Jan – 15 Mar and 16 Aug – 31 Dec) (a) and warm season (16 Mar – 15 Aug) (b).



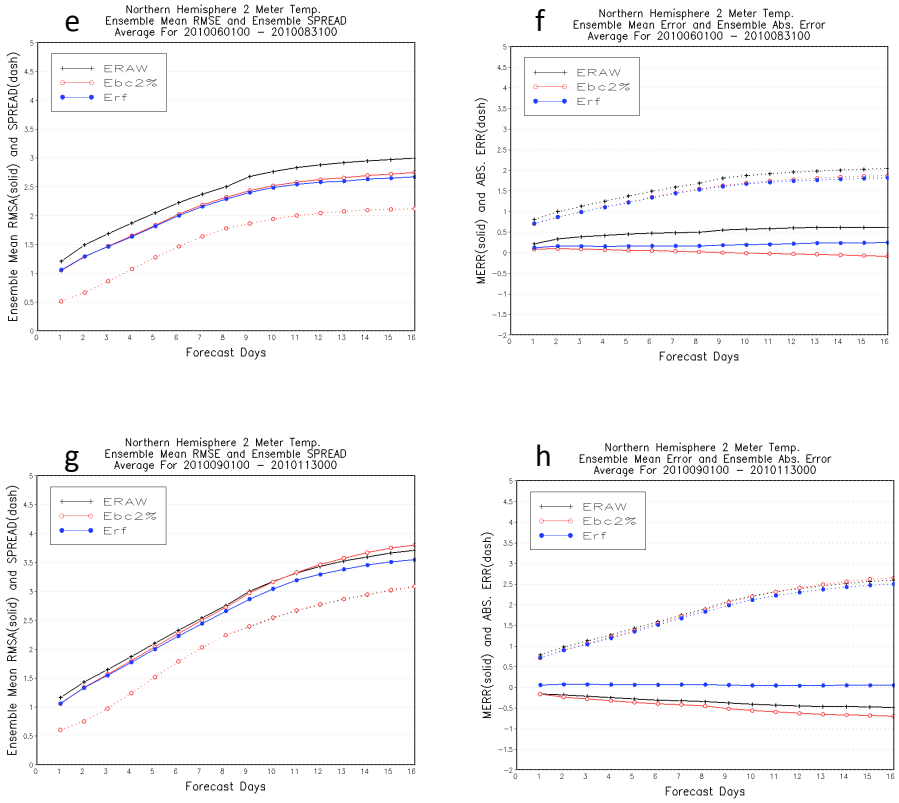
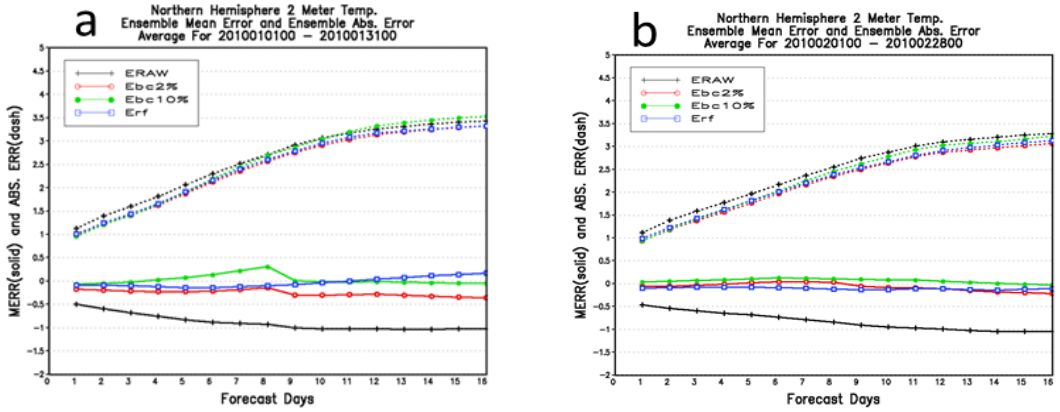


Figure 3. Ensemble mean RMSE (solid lines, left panels), spread (dashed lines, left panels), error (solid lines, right panels), and absolute error (dashed lines, right panels) of 2-m temperature averaged over the Northern Hemisphere for the 4 seasons of the year 2010. Eraw, Rbc2%, and Erf are the raw (black lines), decaying-bias-corrected (red lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.



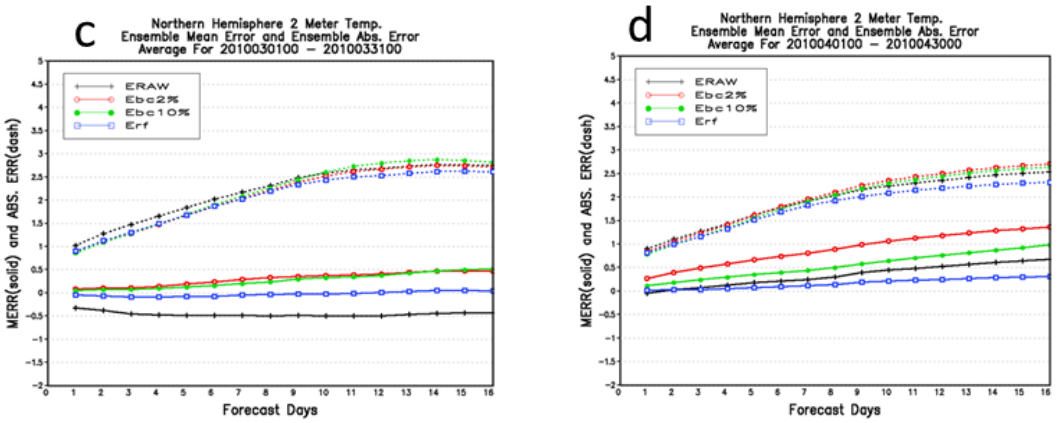


Figure 4. Mean error (solid lines) and mean absolute error (dashed lines) of 2-m temperature averaged over the Northern Hemisphere for (a) January, (b) February, (c) March, and (d) April, 2010. Eraw, Rbc2%, Rbc10%, and Erf are the raw (black lines), decaying-bias-corrected using a 2% weight (red lines), decaying-bias-corrected using a 10% weight (green lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.

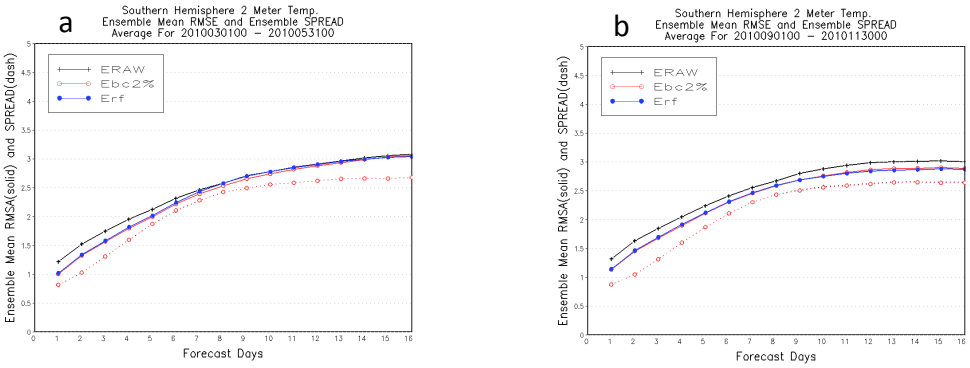


Figure 5. Ensemble mean RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the Southern Hemisphere for spring (a) and autumn (b) of the year 2010. Eraw, Rbc2%, and Erf are the raw (black lines), decaying-bias-corrected (red lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.

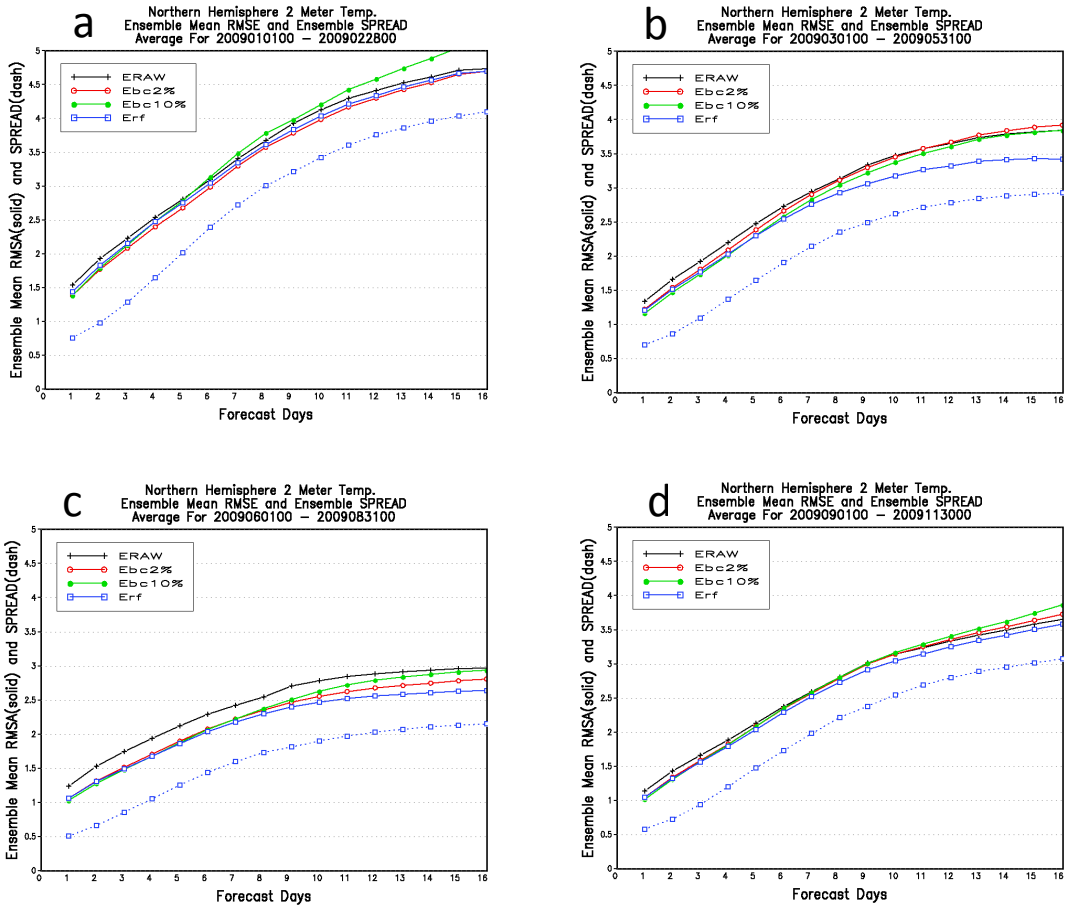
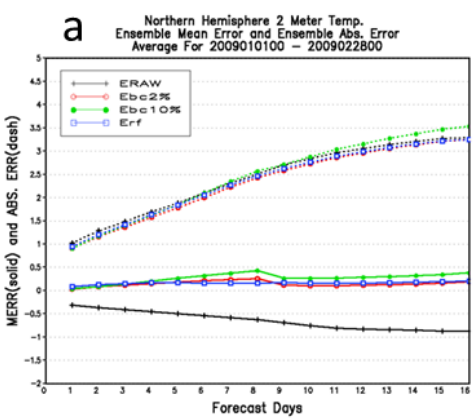
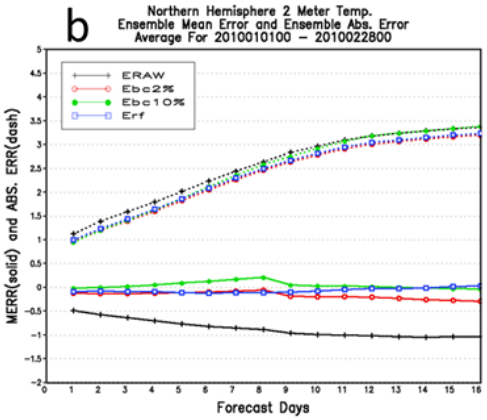
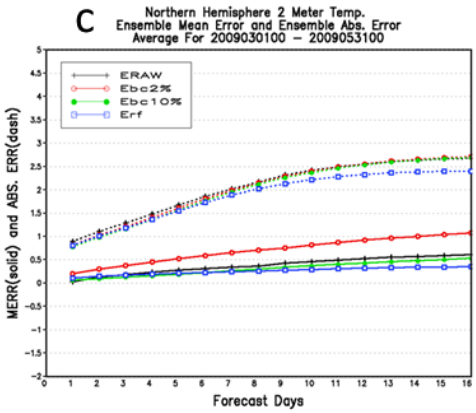


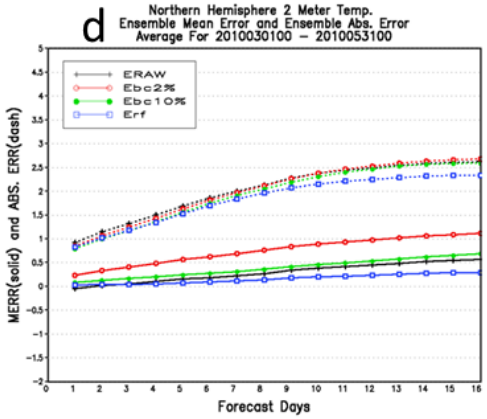
Figure 6. RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the Northern Hemisphere for the 4 seasons of 2009. Eraw, Rbc2%, and Erf are the raw (black lines), decaying-bias-corrected with the two weights (2% red lines and 10% green lines), and reforecast-bias-corrected ensemble forecasts (blue lines), respectively.


a
Northern Hemisphere 2 Meter Temp.
Ensemble Mean Error and Ensemble Abs. Error
Average For 2009010100 – 2009022800
ERAW
Ebc2%
Ebc10%
Erf
MERR(solid) and ABS. ERR(dash)
Forecast Days

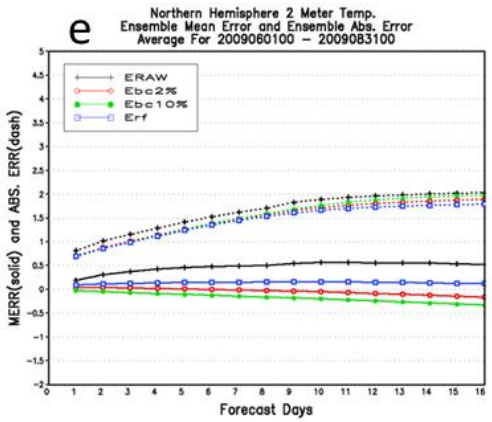

b
Northern Hemisphere 2 Meter Temp.
Ensemble Mean Error and Ensemble Abs. Error
Average For 2010010100 – 2010022800
ERAW
Ebc2%
Ebc10%
Erf
MERR(solid) and ABS. ERR(dash)
Forecast Days

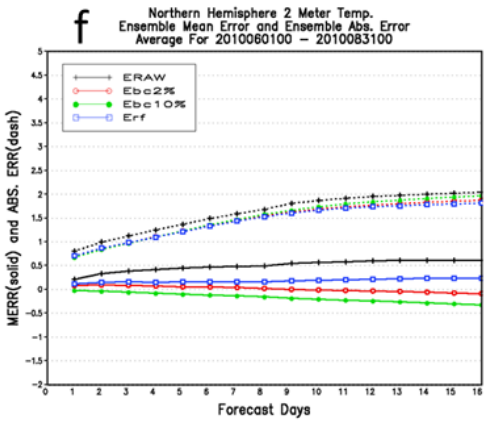
1


c
Northern Hemisphere 2 Meter Temp.
Ensemble Mean Error and Ensemble Abs. Error
Average For 2009030100 – 2009053100
ERAW
Ebc2%
Ebc10%
Erf
MERR(solid) and ABS. ERR(dash)
Forecast Days


d
Northern Hemisphere 2 Meter Temp.
Ensemble Mean Error and Ensemble Abs. Error
Average For 2010030100 – 2010053100
ERAW
Ebc2%
Ebc10%
Erf
MERR(solid) and ABS. ERR(dash)
Forecast Days

2


e
Northern Hemisphere 2 Meter Temp.
Ensemble Mean Error and Ensemble Abs. Error
Average For 2009060100 – 2009083100
ERAW
Ebc2%
Ebc10%
Erf
MERR(solid) and ABS. ERR(dash)
Forecast Days


f
Northern Hemisphere 2 Meter Temp.
Ensemble Mean Error and Ensemble Abs. Error
Average For 2010060100 – 2010083100
ERAW
Ebc2%
Ebc10%
Erf
MERR(solid) and ABS. ERR(dash)
Forecast Days

3

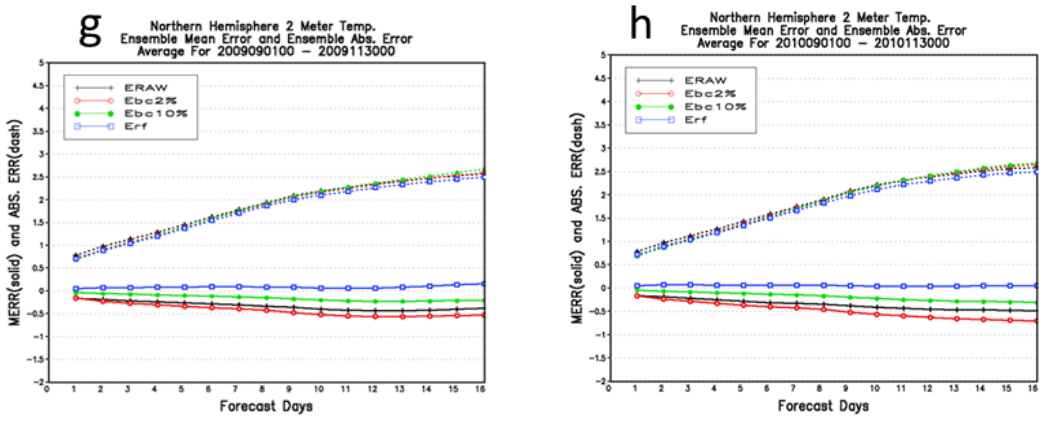
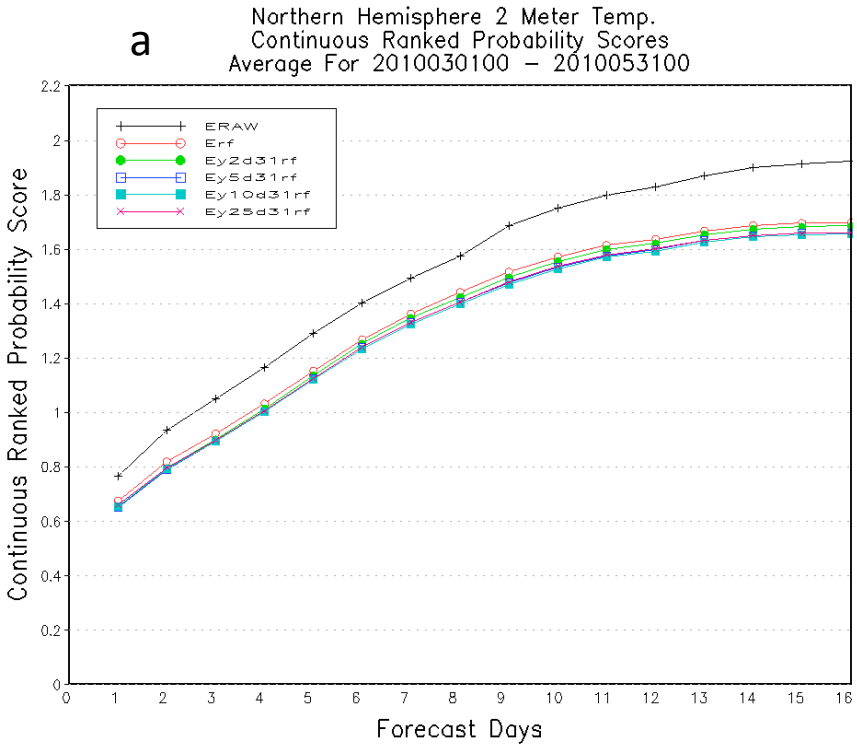


Figure 7. Comparisons of mean errors (solid lines) and mean absolute errors (dashed lines) of 2-m temperature over the Northern Hemisphere between 2009 (left) and 2010 (right) for winter (a-b), spring (c-d), summer (e-f) and autumn (g-h). ERAW, Ebc2%, Ebc10%, and Erf are the raw forecast (black lines),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



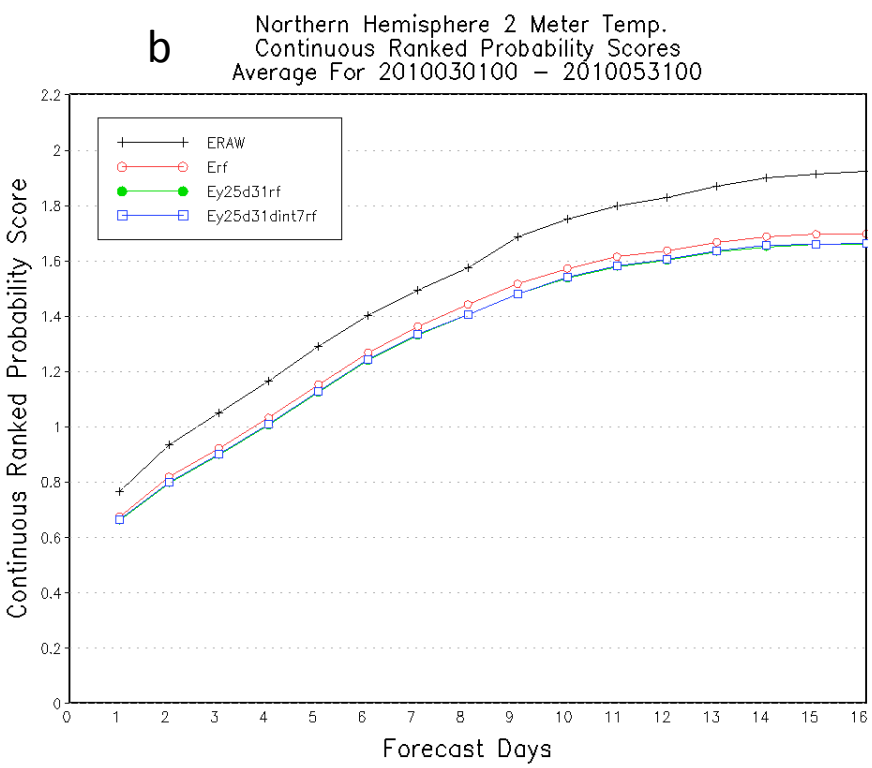


Figure 8. CRPS of 2-m temperature averaged from 1 Mar 2010 to 31 May 2010 over the Northern Hemisphere. ERAW is the raw ensemble forecast (black lines). Erf is the reforecast-bias-corrected ensemble forecast with historical data at the exact forecast date (red lines). Ey2d31rf, Ey5d31rf, Ey10d31rf, and Ey25d31rf in Fig. 8a 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 years (green line), 5 years (blue), 10 years (cyan), and 25 years (magenta). Ey25d31rf and Ey25d31int7rf in Fig. 8b are the reforecast-bias-corrected ensemble forecasts using all 25 years of historical data, covering a time window of 31 days centered on the forecast day. The frequency of data samples for Ey25d31rf and Ey25d31int7rf of Fig. 8b are 1 day (green line) and 7 days (blue line), respectively.

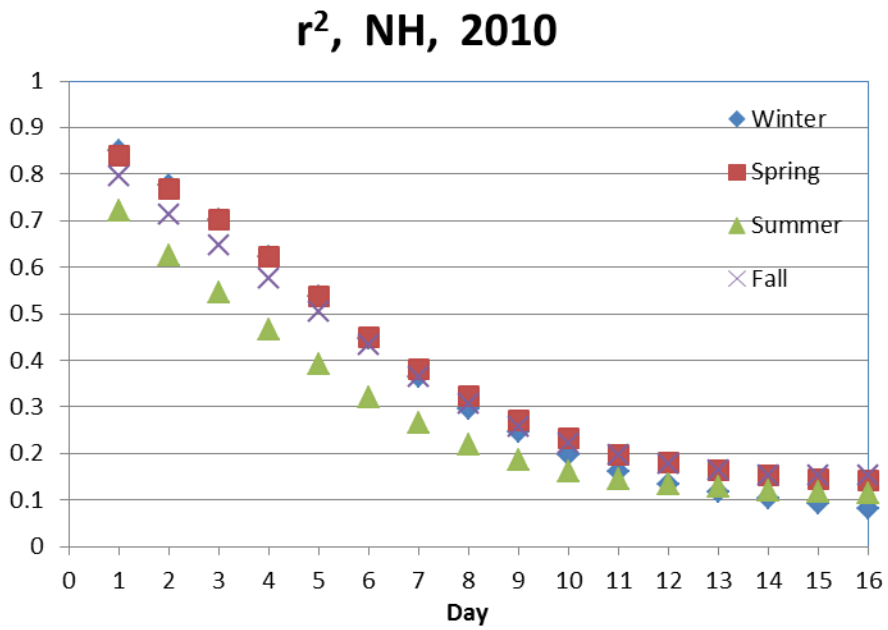


Figure 9. The change in the square of the correlation coefficient between the ensemble mean and analysis with forecast lead time for the 4 seasons of 2010.

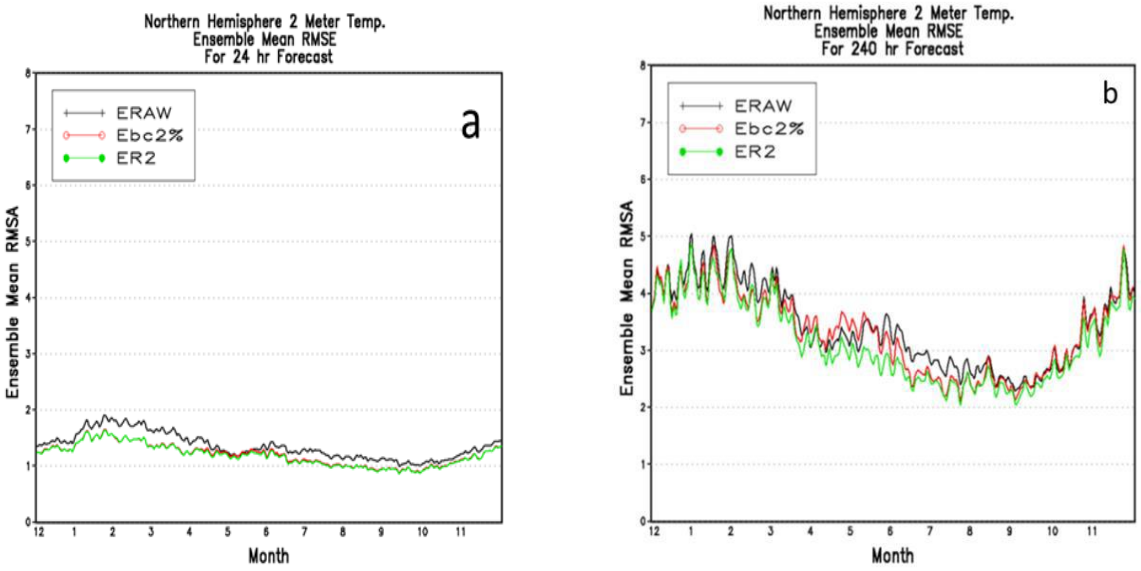


Figure 10. RMSE of 2-m temperature averaged over the Northern Hemisphere for 24hr (a) and 240hr (b) forecasts between Dec 2009 and Nov 2010. ERAW, Ebc2%, and ER2 denote the raw forecast (black lines), decaying-bias-corrected forecast (red lines), and decaying-reforecast-bias-corrected ensemble forecast (green lines), respectively.

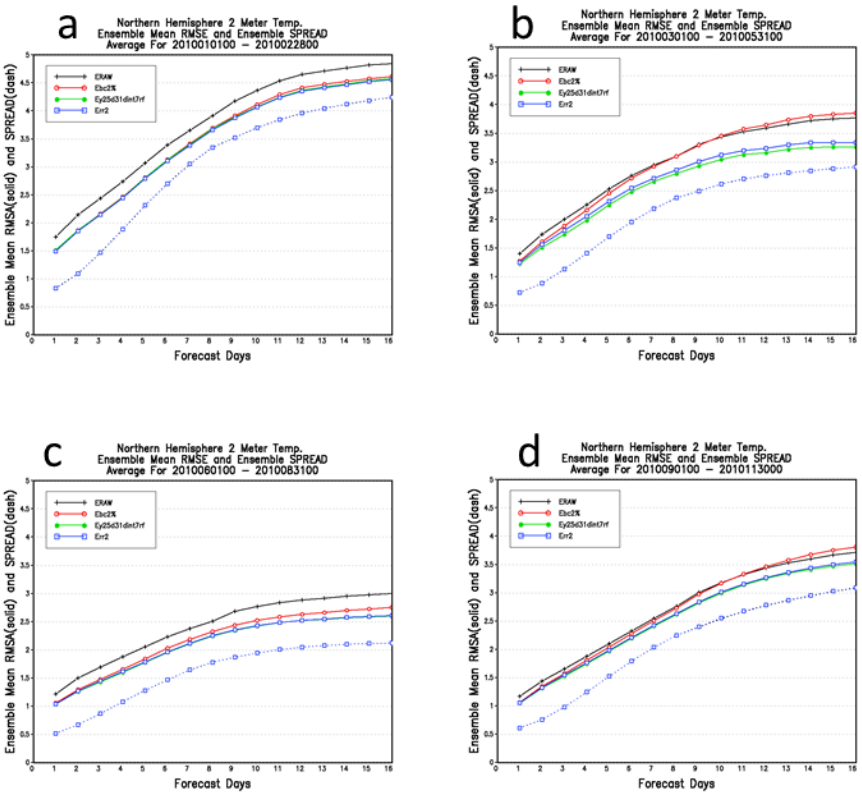


Figure 11. RMSE (solid lines) and spread (dashed lines) of 2-m temperature averaged over the Northern Hemisphere for the 4 seasons of 2010. Eraw, Ebc2%, and Ey2531dint7rf, and ER2 are the raw (black lines), decaying-bias-corrected (red lines), reforecast-bias-corrected (green) and decaying-reforecast-bias-corrected ensemble forecast (blue lines), respectively. In the Ey2531dint7rf, historical data spans a time window of 31 days, centered on the forecast day for the full 25 years with a sample frequency of 7 days.