6.2 ECENT DEVELOPMENT TO IMPROVE NAEFS SPP

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

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

[2]System Research Group Inc.,

[3]I.M. Systems Group

1. INTRODUCTION

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

Ensemble forecast is contaminated by system bias and random errors (Toth et. al, 2003; Wilks and Hamill, 2007). The Statistical Post Process (SPP) is used to reduce the forecast bias, improve forecast reliability, and enhance the ensemble forecast skill. The NAEFS SPP includes bias-correction and downscaling. The bias correction in the current NAEFS SPP is mainly first moment adjustment by applying an adaptive Kalman Filter (KF) technique with most recent forecast/observation information (Cui et al., 2013). There is no second moment adjustment or correction.

The KF algorithm was developed by the NWS at National Center of 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 storing a large amount of sample dataset once initialized, which meets the requirement of daily operational run. The operational statistical verification since 2006 reveals the NAEFS product is significantly enhanced by the decaying bias-correction method. However, the method is, sometimes, failed during spring and fall transition seasons for long lead time.

---

* Corresponding author address: Hong Guan, NOAA *NWS* Environmental Modeling Center, College Park, MD. email: Hong.Guan@noaa.gov

Recently, we have tested the method to improve NAEFS 1st moment correction. In order to improve the performance of bias correction, 26-year GEFS reforecast has been applied in additional to current operational NAEFS bias correction process. Several different methods have been examined to optimize the usage of past 26-year reforecast information. Several forecast elements are investigated in this study, such as surface temperature, 500hPa geopotential height and et al, but we will mainly focus on the discussion of surface temperature up to 16 days.

2. MODEL and DATA

The current operational GEFS version (v9.01) was implemented on February 14, 2012 at 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 parameterization (Zhu et al. 2007). The model is run at a horizontal resolution of T254 (~55km) for the first 8 days and T190 (~70km) for the later 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 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 only including 11 members (1 control member and 10 perturbation members). The model was only run at 00UTC cycle for the 10 perturbation members. The data set used here was bilinear interpolated to 1°x1° latitude and longitude grids from the native resolution. The data in GRIB2 format for 00UTC cycle is available since 1985 (+28 years). We use a subset of the data from 1985 to 2010 (26 years), obtained from NOAA/ESRL. More detail description on model and dataset can be found in Hamill et al. (2013).

3. METHOD

a. Bais estimation

In this study, the bias (***b***) for each lead-time ***t*** (24-hour interval up to 384 hours), each grid point (*i, j*) is defined as the difference of best analysis $a_{i,j}(t_0)$ and forecast $f_{i,j}(t)$ at the same valid time t_0 which is up on latest available analysis.

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

b. Decaying average method

The detail of the decaying average method can be found in Cui et al., (2012). Here we introduce its basic equation. Decaying average bias $B^p_{i,j}$(t) is updated by considering prior period bias and current bias by using decaying average (or KF method) with a weight coefficient (*w*) equal to 0.02.

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

c. Application of reforecast (or hindcast)

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

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

d. Bias correction

To remove lead-time dependent bias on model grid, new (or bias corrected) forecast *F* is generated by applying decaying-averaged bias ($B^p_{i,j}$) and reforecast bias ($B^h_{i,j}$) to raw forecast ($f_{i,j}$) at each grid point (*i,j),* for each lead time (*t*), and each parameter.

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

where *r* is the correlation coefficient estimated by linear regression from the joint samples (ensemble mean and analysis). To avoid storing large dataset, the mean values used in computing *r* were generated from decaying averaging with a weight of 0.10. The relative contribution of reforecast and decaying-averaged bias was quantified by the r^2. For the two spatial cases of *r*=0 and *r*=1, the equation represent the reforecast bias correction and decaying bias correction, respectively.

f. Methodology of verification

The calibration of ensemble forecast system is evaluated by means of mean forecast error, mean absolute forecast error, root means square error (RMSE), and continue ranked 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 value the CRPS, the better the probabilistic system is in terms of both reliability and resolution.

4. Results

We calibrate 2-m temperature in 2009 and 2010 using prior 24-year (1985-2008) and 25-year bias (1985-2009), respectively. We also calibrate 500hpa height in 2009 using 24-year bias. A preliminary check shows it is very hard to improve the forecast skill of 500hpa height, possibly due to its relative small bias or insensitivity. Thus, our focus will be on the calibration of 2-m temperature. We explore the sensitivity of the calibration on the number of training years by using the bias from most recent 2 (2008-2009), 5 (2005-2009), 10 (2000-2009), and 25 (1985-2009) years of training data, and evaluated using the last year (2010) validation. We compare the calibration with the two training-data window (1 day and 31 days) around the corresponding date in each of the training years (25 years). The impact of sample interval on the calibration is estimated by comparing verification scores with a sample interval of 1 day and 7 days within a window of 31 days. Finally, we apply reforecast information to the operational GEFS product of NCEP.

4.1 Calibrating 2010 forecast using 25-year training dataset

Figure 1 shows the verification for 2-m temperature over the Northern Hemisphere for 4 seasons. We present here a comparison of the results of the raw ensemble forecast (ERAW) and two calibrated forecasts (Ebc2% and Erf). The Ebc2% and Erf denote the bias-corrected forecast with decaying averaging and reforecast method, respectively. The GEFS model is apparently under-dispersed for all seasons and lead times (see Fig.1a, 1c, 1e, and 1g). The raw ensemble forecast (black lines) has a cold bias in winter (Fig.1b) and autumn (Fig.1h). Conversely, a warm bias is prevalent in spring (Fig.1d) and summer (Fig.1f). These biases are almost completely corrected by the Erf method (green lines). The corrected bias is closer to zero and corresponding absolute error and RMSE are also smaller than the raw ensembles, hinting the effectiveness of the calibration methods in reducing the system error of the ensemble forecast. A decaying method (Ebc2%) also does good job for the

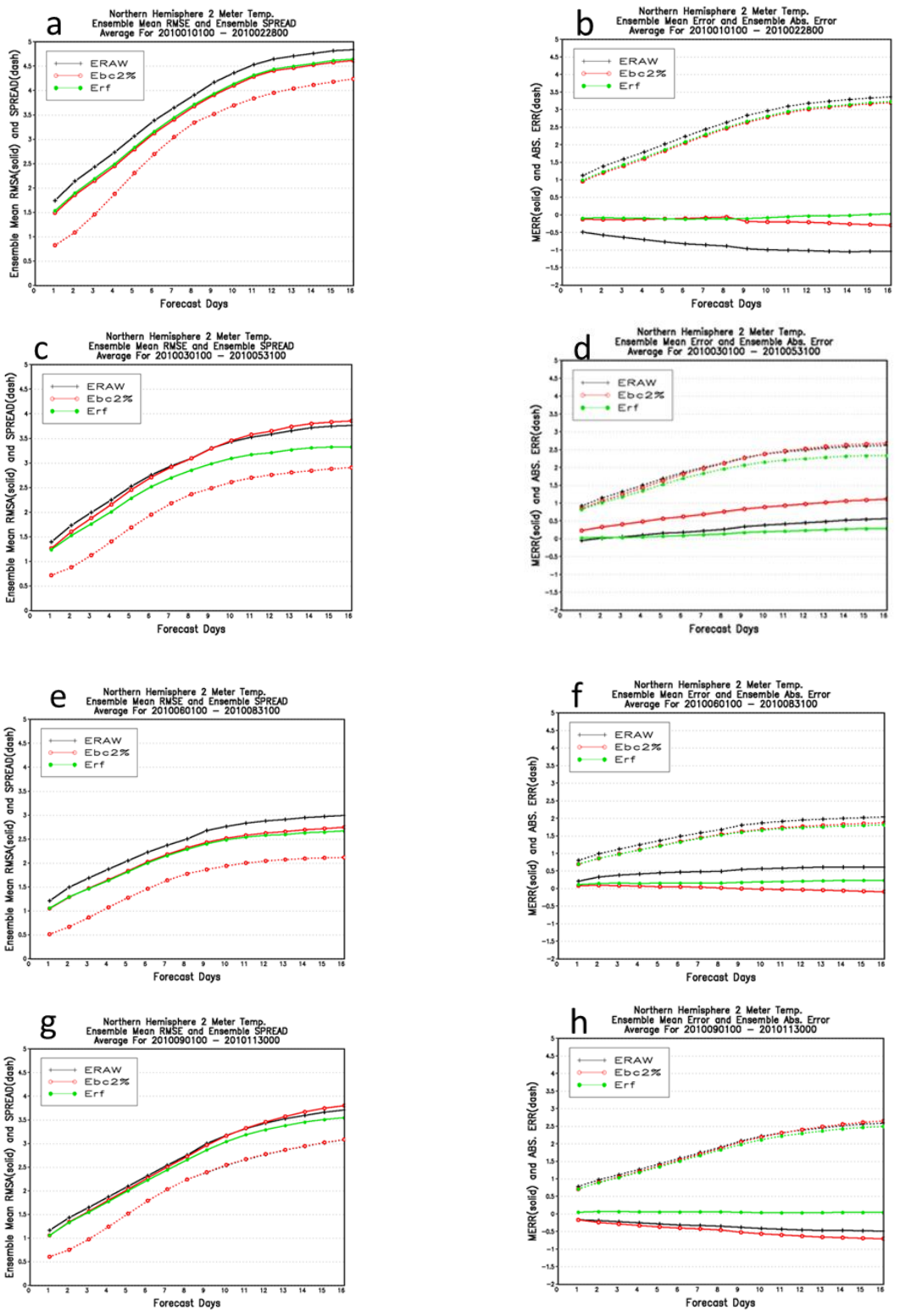


Fig.1. Ensemble mean RMSE (solid lines of left panels), spread (dash lines of left panels), error (solid lines of right panels), and absolute error (dash lines of right panels) of 2-m temperature averaged over the Northern Hemisphere for the 4 seasons of 2010. ERAW (black lines), Ebc2% (red lines), and Erf (green lines) are the raw, decaying-bias corrected, and reforecast-bias corrected ensemble forecasts, respectively.

non-transitional seasons (winter and summer). However, the technique does not work well in all circumstances as pointed out in Cui et al. (2012). Fig. 1d and h reveal applying decaying method leads to a degradation of forecast accuracy during transition seasons. The maximum degradation occurs in spring. The absolute errors (Fig.1d and 1h) in the Ebc2% are larger than those of RAW forecast and the RMSEs (Fig.1c and 1g) are larger than those of RAW forecast only for long-lead time. To determine the underlying reason, we display the monthly evolutions of mean error and mean absolute error of 2-m temperature for four experiments over the Northern Hemisphere in Fig.2. Beside the above three experiments, the result for the decaying method with a weight of 10% is also added in the comparison. We note a persistent cold bias (black lines) in the winter (January and February). In the beginning of the spring (March), the cold bias becomes smaller, eventually, turns to warm bias in April. In the two winter months, the Ebc2% performance is very similar to the Erf, yielding a more accurate forecast than the raw ensemble forecast. This is due to the ensemble forecast error being relatively consistent among the non-transitional months. The 2% decaying averaging is using most recent 50-60 days of bias information (Cui et al., 2012) with the highest weight for the latest information. The Ebc2% method is failed in March and April, when error characteristics experience a dramatic change within a period of ~50-60 days. In April, the Ebc2% uses cold bias, accumulated from winter and early spring, to calibrate warm bias in spring. This outdated information leads to degrading forecast (i.e. an increased 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 general, the Erf has obvious advantage over the Ebc2% and Ebc10%. The Ebc10% is slightly better than the Ebc2%.

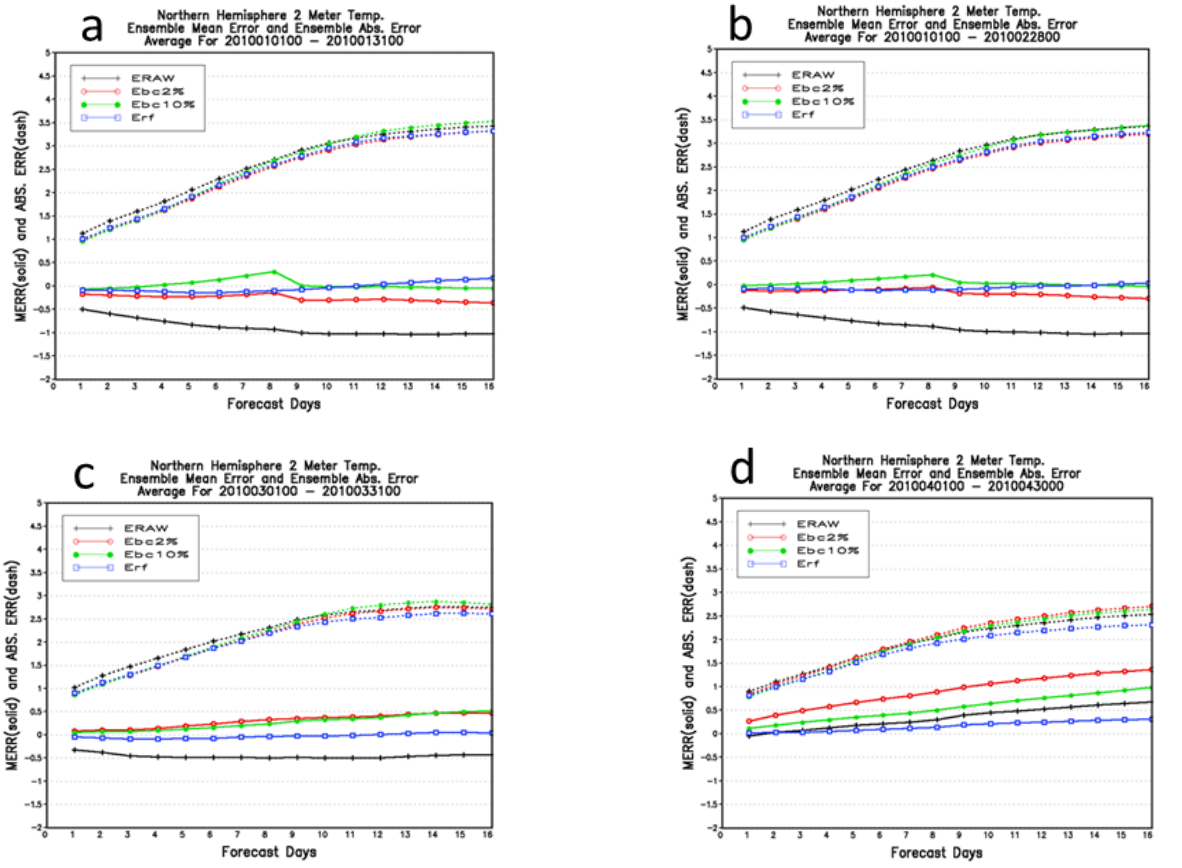
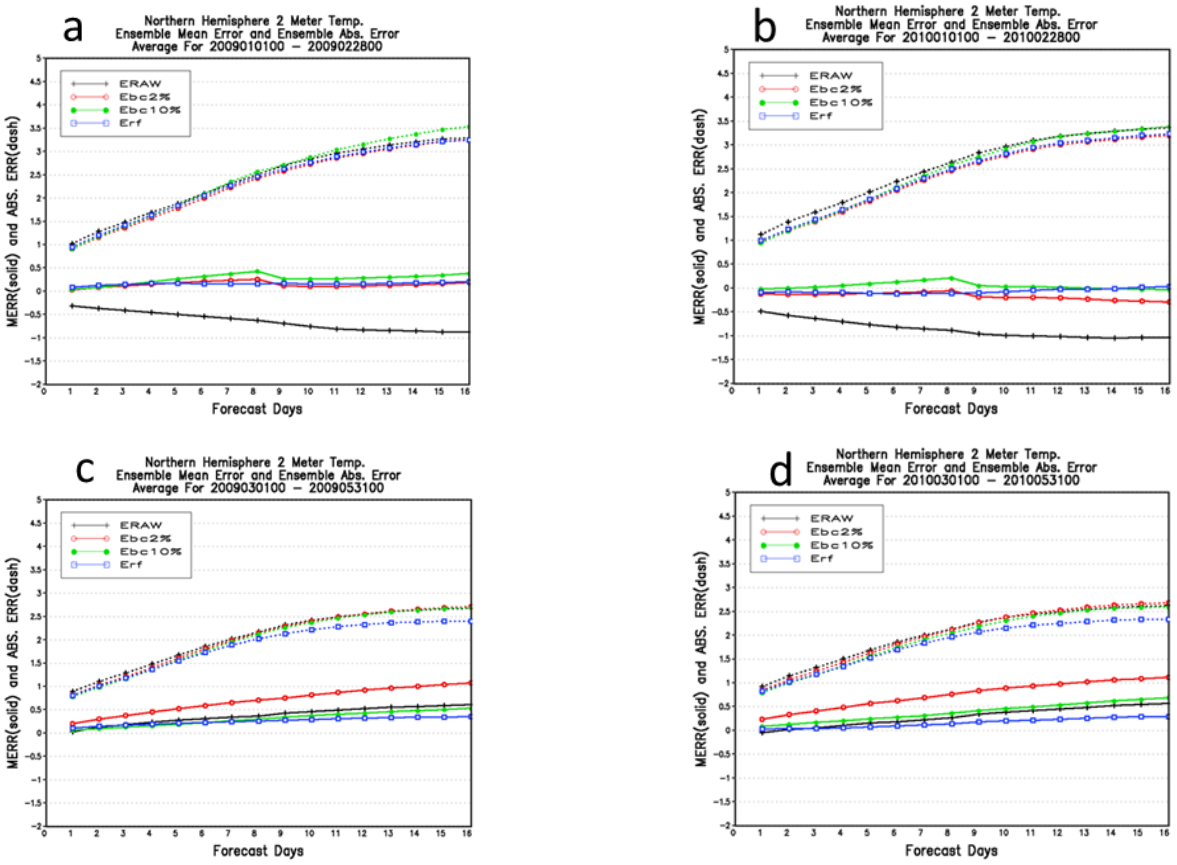


Fig. 2 Mean errors (solid lines) and mean absolute errors (dash lines) of 2-m temperature over the Northern Hemisphere for January (a), February (b), March (c), and April (d), 2010. ERAW, Ebc2%, Ebc10%, and Erf are the raw forecast, decaying-bias corrected forecasts with two weights (2% and 10%), and reforecast-bias corrected ensemble forecast, respectively.

4.2 Comparison between 2009 and 2010

The notable improvement in the accuracy of T2m forecast by the Erf is impressive. The key question is if this improvement only occurs particularly for 2010. To answer this, we also calibrate the 2009 forecast and compare the results to 2010. The data prior to the validation year (2009) were used to train the reforecast-bias correction algorithm.

Figure 3 compares the mean error and mean absolute error of 2-m temperature between 2009 and 2010 for the Northern Hemisphere. The performance in 2009 is, qualitatively, very similar to that of 2010. The winter cold bias and summer warm bias in the raw ensemble can also be seen in 2009. The Ebc2%, again, improve the forecast in the non-transition seasons for all lead times but not in the other two seasons, when the Ebc2% tends to degrade the long-lead time forecasts. The Erf improves the ensemble forecast over the Ebc2% for almost all lead times and seasons as noted in 2010. The biases for all seasons are, again, handily removed by the Erf.



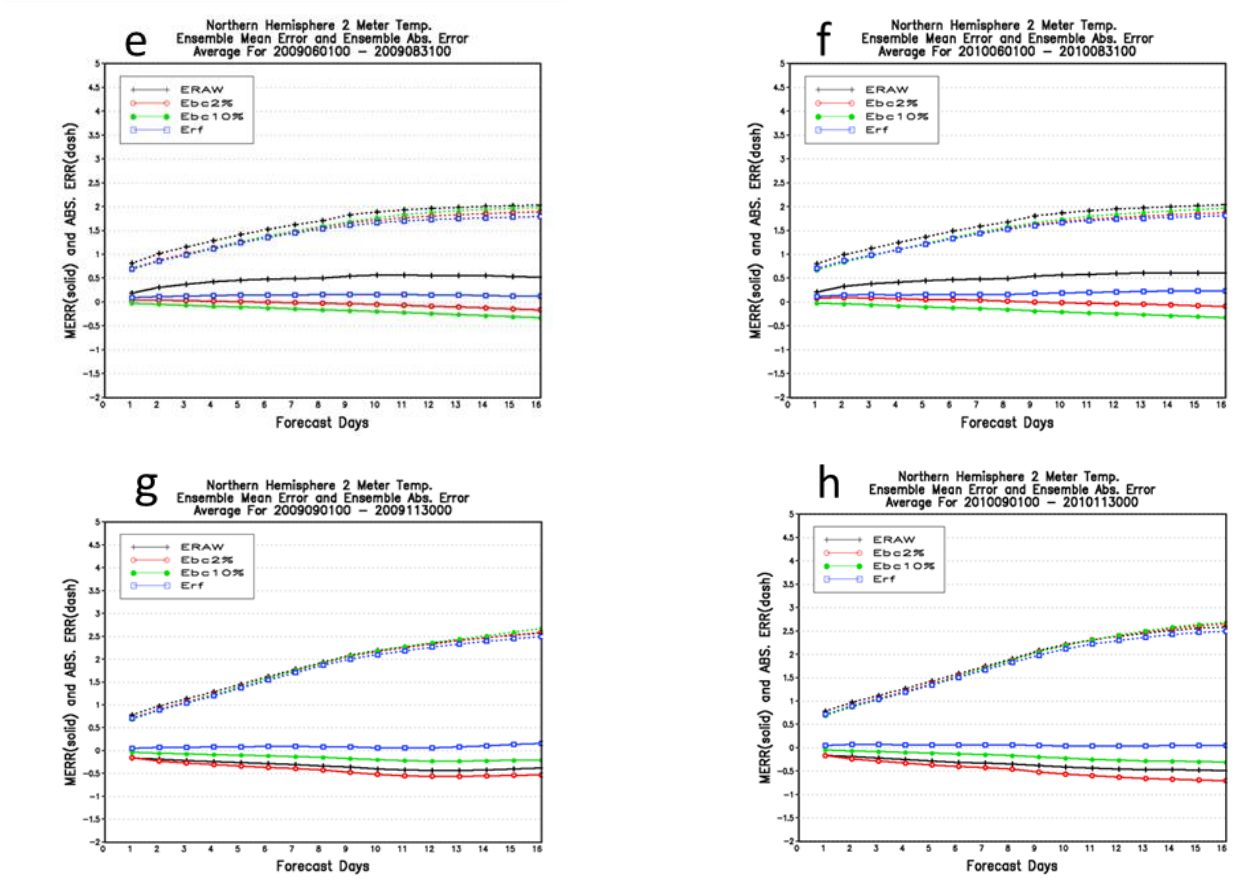


Fig.3 Comparisons of mean errors (solid lines) and mean absolute errors (dash lines) of 2-m temperature over the Northern Hemisphere between 2009 and 2010 for winter (a), spring (b), summer (c) and autumn (d). ERAW, Ebc2%, Ebc10%, and Erf are the raw forecast, decaying-bias corrected forecasts with two weights (2% and 10%), and reforecast-bias corrected ensemble forecast, respectively.

4.3 Calibration using various training samples

The CRPS of forecasts from the RAW (black lines) ensemble and calibrated ensembles (color lines) with training sample of various sizes are displayed in Fig.4. Figure 4a and 4b examine the sensitivity of forecast skill on the number of sample year and interval day, respectively. All calibrated forecasts demonstrate a better performance than the raw forecast. The difference among the calibrated forecasts is relatively small. The scores for 10 and 25-years with a 31-day window are very similar, slightly better than the other smaller training samples (Fig. 4a), suggesting that ten-year dataset is large enough to obtain most usable skill. 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.4b) and both are better than the result with single data (red line) from each year. Therefore, the 25-year weekly training dataset is a good option to reduce computational expense and keep desiring skill. These results are consistent with the findings of previous researchers (Hamill et al., 2004, Hagedorn et al., 2008), although they have used different model or GEFS version.

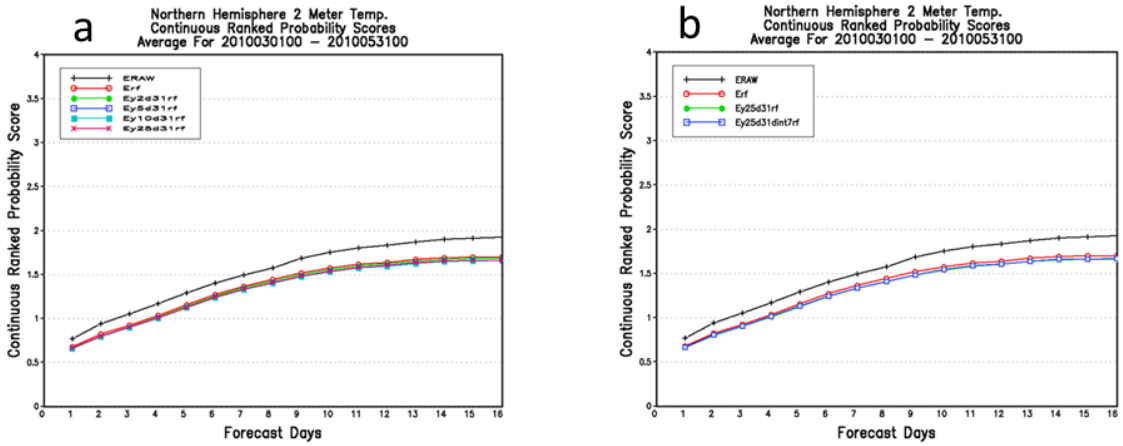


Fig. 4 CRPS of 2-m temperature averaged from 1 Mar 2010 to 31 May 2010 over the Northern Hemisphere. ERAW is the raw ensemble forecast. Erf is the reforecast-bias corrected ensemble forecast with historical data at exact forecast date. Ey2d31rf, Ey5d31rf, Ey10d31rf, and Ey25d31rf in Fig4a are the reforecast-bias corrected ensemble forecasts with historical data span a time window of 31 days centered at forecast day for the most recent 2, 5, 10, 25 years, respectively. Ey25d31rf and Ey25d31int7rf in Fig. 4b are the reforecast-bias corrected ensemble forecasts with historical data covered a time window of 31 days centered at forecast day. The frequency of data sample for Ey25d31rf and Ey25d31int7rf of Fig 4b are 1 and 7 days, respectively.

4.4. Using reforecast to improve NCEP bias-corrected product

Having seen the remarkable value of using reforecast information, we now combine the reforecast method with the operational decaying method together, aimed to provide an option of improving forecast accuracy for transitional season. Fig. 5 displays the change of r^2 with forecast lead-time averaged over the Northern Hemisphere for 4 seasons. The r^2 denotes the square of correlation coefficient between ensemble mean and analysis. Forecast ability declines as forecast lead time increases. There are no significant differences of r^2 among seasons. The r^2 values are slightly smaller in summer than other seasons for short lead times.

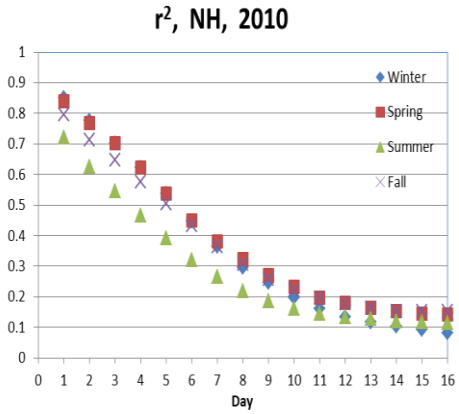


Fig. 5 The change of square of correlation coefficient with forecast lead time for the 4 seasons of 2010.

Figure 6 displays the time series of RMSE for the ERAW, Ebc2% and ER2 for 24 and 240-hr forecasts. The Ebc2% and ER2 represent the bias-corrected forecast with decaying method and decaying-reforecast combined method, respectively.

For 24-hr forecast (Fig.6a), the RMSE in Ebc2% is smaller than the raw forecast almost all times. Including reforecast bias-correction (ER2) does not change the forecast accuracy too much since the weight of reforecast is small at this short lead-time forecast. For 240-hr (Fig. 6c and d) forecast, the Ebc2% does not always improve the forecast with a significant degradation in spring. Our results agree with those in Cui et al. (2012), who found that decaying-averaging method mainly works well for the first few days. It is very clear that adding reforecast information improve the forecast accuracy for all time with maximum benefits on April, May, and June.

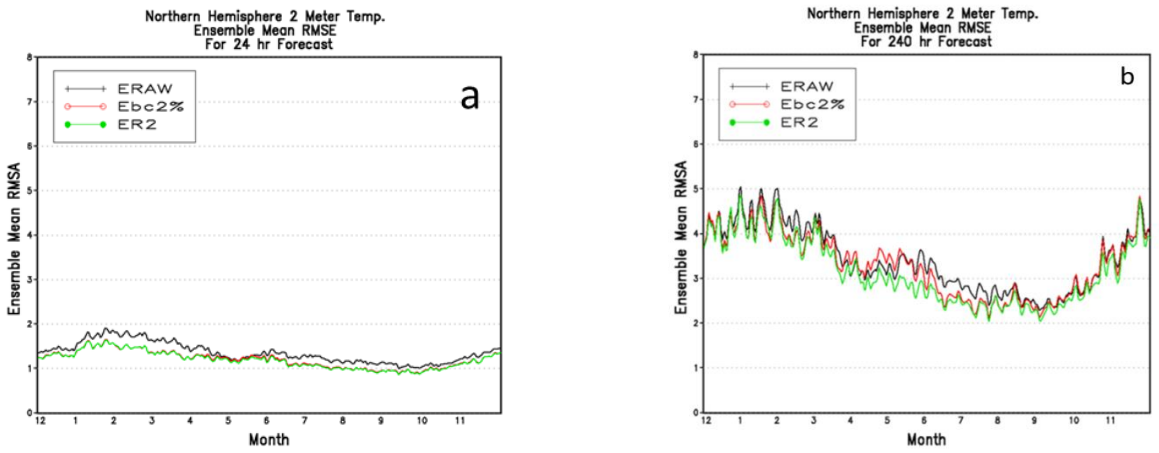


Fig. 6 RMSE of NH 2-m temperature for the 24hr (a) and 240hr (b) forecasts of the 2009–2010 winter. ERAW, Ebc2%, and ER2 denote the raw, decaying-bias corrected, and decaying-reforecast bias-corrected ensemble forecast, respectively.

5. CONCLUSION

In this paper, we develop the method to improve NAEFS 1st moment correction using GEFS reforecast dataset.

We use 24-year and 25-year GEFS reforecast bias information to calibrate 2009 and 2010 forecast, respectively. We found that the forecast of 2-m temperature is strongly biased for the Northern Hemisphere with a cold bias in winter and warm bias in summer. The bias is mostly removed by reforecast method. Decaying method improves the forecast skill in winter and summer as good as reforecast method, but it degrades the long-lead forecast during transitional seasons due to dramatic change in bias characteristic.

It is very difficult to improve the forecast skills for 500hPa height. This is possibly due to less bias or insensitivity of this variable to bias correction.

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

Bias and its seasonal variation are model-dependent. Whether the improvement found here will occur for the new GEFS version need to be confirmed in the future.

6. REFERENCES

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei and Y. Zhu, 2005: A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems, *Mon. Wea. Rev.,* Vol. **133**, 1076-1097.

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., 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.

Hagedorn, R., T. M. Hamill and J. S. 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., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

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, Y. Zhu and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565.

Hersbach, H., 2000: Decomposition on the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

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.

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

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proceedings, ECMWF Workshop on Predictability,* ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.].

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 ensembleforecasts.

Forecast Verification: A Practitioner's Guide in Atmospheric Science, I. T. Jolliffeand D. B. Stephenson, Eds., John Wiley and Sons, 137–163.

Wang, X. and Bishop, C. H., 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.

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

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability, *Advance in Atmospheric Sciences, Vol. 22,* No.6, 781-788.

Zhu Y., G. Iyenger, Z. Toth, S. M. Tracton, and T. Marchok, 1996: "Objective evaluation of the NCEP global ensemble forecasting system", AMS conference proceeding.

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", AMS conference proceeding.