Development of Statistical Post-processor for NAEFS

Bo Cui[1], Yuejian Zhu[2], Zoltan Toth[3], Dingchen Hou[2]

[1]FIT at Environmental Modeling Center, NCEP/NWS

[2]Environmental Modeling Center, NCEP/NWS

[3]Global Systems Division, ESRL/OAR

(Submitted to Weather and Forecasting)

Corresponding address:

Dr. Bo Cui

Environmental Modeling Center

NCEP/NWS/NOAA

5200 Auth Road,

Camp Springs, MD 20746

301-763-8000 ext 7594

E-Mail: Bo.Cui@noaa.gov

Abstract

In the NAEFS framework, a set of statistical post processing techniques has been designed and applied for both the NCEP and CMC ensembles. Bias correction and statistical downscaling are the two most important approaches among them. The bias correction entails the statistical correction of ensemble lead-time-dependent systematic errors. The statistical downscaling aims to compensate for the lack of details of model spatial/temporal resolution. These are two separate and independent types of processing. The correction method was implemented operationally at both centers (NCEP and CMC) in 2006. The statistical downscaling method was implemented at NCEP in 2007. Current post-processor produces outputs on 1°x1° latitude/longitude grids for 49 variables and on a 5x5 km grid for four fields covering the CONUS: surface pressure, 2-meter temperature, and 10-meter wind components (u and v).

This study describes the NAEFS post-processing system with a focus on the downscaling method. After applying the statistical post-processing, the quality and value of the ensemble forecast undergo a significant increase. The NAEFS product has reduced mean absolute errors by 4+ days and improved probabilistic skills by 8+ days. NAEFS has a better global skill than a single NCEP or CMC ensemble. Downscaling contributes more to ensemble performance than do bias correction and the multi-model ensemble.

1. Introduction

1.1 North American Ensemble Forecast System

The North American Ensemble Forecast System (NAEFS) is a joint project involving the US National Weather Service (NWS), the Meteorological Service of Canada (MSC) and the National Meteorological Service of Mexico (NMSM). An agreement of corporation between United States-Canada-Mexico, became official in November 2004. The goal of NAEFS is to exchange and combine the American and Canadian ensemble forecast outputs s in real time, producing probabilistic forecasts from the combined ensemble that are consistent over North America. The NAEFS was established in 2005 at the National Centers for Environmental Prediction (NCEP) of US NWS as an operational multi-center ensemble forecast system. In addition to the raw ensemble forecast data exchanges between NCEP and MSC, a statistical post-processing system has been established for NAEFS that includes bias correction, dual resolution, statistical downscaling techniques, etc. This paper presents a study on the post-processing system designed and applied within the NAEFS.

1.2 Systematic error (bias) correction

Compared to deterministic Numerical Weather Prediction (NWP), the ensemble approach provides more possibilities of extending in lead time for prediction of weather events. However, as accurate as they might be, ensemble predictions do not provide absolutely

reliable information. In addition to deficiencies in ensemble system design, all ensemble forecast systems suffer from systematic errors associated with imperfect numerical forecast models and analysis schemes. Some uncertainties inevitably remain and cause inaccurate forecasts. The negative effects include (a) lead-time-dependent systematic errors (forecast bias) due to imperfect numerical modeling techniques and (b) finite spatial and temporal resolution due to limited computational resources.

The first type of error is mainly due to limitation of model physics, causing "model drift", i.e. lead-time-dependent systematic error. The predictions become substantial drift away from the observations (Toth and Pena 2007). In an ensemble system, such "model drift" greatly affects ensemble measures of reliability (Atger Frédéric 2003; Zoltan et al. 2006b). How can we reduce lead-time-dependent systematic error to improve ensemble statistical reliability? The benefits of ensemble calibration have been examined and verified in different ways (Atger Frédéric 2003; Cui et al. 2006; Hamill and Whitaker 2007). At NCEP, a decaying average bias correction method was adopted. This simple correction method was implemented into operations of both centers (NCEP and Canadian Meteorological Centre, CMC) in 2006 (Cui et al. 2006, 2011). Verifications show that individual ensemble members benefit from this simple de-biasing method particularly in the short range. The calibrated operational ensemble has improved probabilistic performance for all measures out to day 5.

1.3 Statistical downscaling

The second type of forecasting deficiency is associated with imperfect analysis that comes from the finite spatial/temporal resolution of the observations. For a better-quality ensemble, simply increasing model resolution gives limited benefit to ensemble statistical resolution, and its application is constrained by computational resources. However, reliable skillful high-resolution ensemble predictions are desired by many forecast users. The demand for important surface variables, such as 2-meter temperature is continuously increasing due to the development of more and more end-user applications in hydrology, agronomy, energy, etc. To bridge the product gap between the low-resolution and high-resolution ensemble probabilistic predictions, global or low resolution ensemble output can be treated with "downscaling" techniques to maximize its utility. There are a variety of ways to perform the downscaling process, such as dynamical methods (Leung et al. 2004; Wang et al. 2004), physically based methods (Leung and Ghan 1995, 1998), statistical methods (Benestad 2004; Maurer and Duffy 2005, Antolik 2006) and combination methods. At NCEP, a regime-dependent statistical downscaling method has been adopted. It became operational in NAEFS in December 2007. This downscaling process generates National Digital Guidance Database (NDGD) 5-km resolution products for the Continental United State (CONUS). There are four variables for this application: 2-meter temperature, surface pressure, and 10-meter wind (u and v components). The downscaled products include probabilistic forecasts (at the 10%, 50%, and 90% levels), ensemble mean, spread and mode - twice per day out to 16 days.

1.4 Statistical Post-processor for NAEFS

In the NAEFS operational context, bias correction and downscaling are done as two separate steps. Both methods statistically adjust ensemble forecasts to improve statistical reliability and resolution. However, the objective of and the procedures for the two methods are different. According to Toth et al. (2006a), reliability and resolution are the two main and independent attributes of forecast systems. In principle, reliability can be improved through statistical post-processing techniques so that the calibrated forecasts follow the distribution of ensuing verifying observations. Bias correction is applied on a coarser model grid. It improves the reliability of each ensemble prediction system component, NCEP/GEFS and CMC/GEFS. Verifications show that it also slightly improves the accuracy of the predicted ensembles and thus the probabilistic resolution of the forecasts (Zhu and Toth 2008; Candille et al. 2010).

However, unlike reliability, resolution cannot be improved via statistical bias correction. Statistical resolution reflects a forecast system's ability to distinguish between different future events in advance. Statistical resolution can be improved only through the modification of the forecast scheme with additional knowledge about the temporal evolution of the observed system (Toth et al. 2006b). The purpose of statistical downscaling is to enhance information missing on coarse spatial scales as an attempt to improve statistical resolution.

Bias correction is beneficial for getting feedback on systematic errors for model development. Its application is inexpensive on the model grid. Ideally, systematic errors are reduced and statistical reliability is enhanced while random errors are not increased to maintain statistical resolution. On the other hand, downscaling helps to enhance "spatial" resolution but is not dependent on lead time. Such interpolation can be applied in various spaces, in time and across variables. It is needed partly due to computational resource limitations. The downscaling process is not directly related to the choice of numerical model, is cheap for data transition and easy for model upgrades in operations.

The NAEFS post-processing system consists of a set of techniques, of which bias correction and downscaling are the two most important. Other post-processing methods include dual resolution, multi-ensemble combination techniques, etc. All these post-processing methodologies as applied in the operational NAEFS are introduced in section 2. Section 3 briefly introduces the verification package used to evaluate NCEP/GEFS and NAEFS which chooses tools described in Zhu and Toth (2008). The performances of the ensemble post-processing system are then quantitatively compared in section 4. Investigations are focused on the effects of the downscaling on each GEFS component (NCEP and MSC) and on the multi-ensemble NAEFS. Also, the improvements due to the downscaling method and multi-ensemble approach are compared in this section. Finally, section 5 contains a discussion and some conclusions.

2. Data set and methodology

The ensemble configurations of NCEP/GEFS and CMC/GEFS are different. The operational NCEP/GEFS system configuration is described in Toth et al. 2004. The 20-member ensembles are produced at T126L28 horizontal (90km) and vertical resolution. The perturbations of the initial conditions are from Ensemble Transform with Rescaling (ETR) technique (Wei et al. 2008). The operational CMC ensemble is described in Charron et al. (2010). A single dynamical core, i.e. the Global Environmental Multiscale (GEM) model is used to produce the ensembles. The CMC's multi-parameterization approach and stochastic perturbations are used in order to sample model errors for the 20 members of the ensemble. NCEP ensembles run with 20 ensemble members per cycle plus one control at 00UTC, 06UTC, 12UTC, and 18UTC. The CMC ensemble runs at 00UTC and 12UTC. All runs go out to 384 hours at 6-hour intervals. All forecast data are interpolated to 1°x1° latitude/longitude resolution.

The NAEFS multi-ensemble is an un-weighted combination of the NCEP and CMC ensembles. But the generation of NAEFS products consists not only of simple combinations of the two ensembles. A set of post-processing steps are applied to the two ensembles, which includes bias correction, dual resolution, multi-ensemble combination, probabilistic forecast generation and statistical downscaling. The first four techniques are performed on forecasts on 1° × 1° grid. Two items are developed for NAEFS on the 1° × 1° forecasts: probabilistic forecast (10%, 50%, 90%, ensemble mean, mode and spread) and climate anomaly forecasts. The last post-processing step, i.e. the statistical

downscaling generates NDGD 5km resolution products for the CONUS from 1° × 1° NAEFS forecasts.

The bias correction or ensemble calibration has been described in previous work (Cui et al. 2006; 2011). It is a correction of the first moment of the ensemble. First the difference between a ensemble mean (NCEP) or the individual members (CMC) forecast and the corresponding analysis is computed. Then this difference is multiplied by a weight (currently 2% in operations) and added to the previous day's cumulative bias value. Such bias estimation is applied independently to every 6-hour time step of the forecast. For each grid-point, each ensemble forecast is adjusted by the difference between the above two fields (Cui et al. 2011). Note that only the variables, such as height, temperature and wind components, for which the errors are assumed is normal distributed, are bias corrected. There are 49 variables included in recent NAEFS upgrade in March 2011 (Zhu and Cui 2011). This section will present the other 3 post-processing methodologies one by one.

2.1 Dual resolution technique

The dual resolution technique is designed for and applied to the NCEP ensemble only and its application is accompanied by the bias correction. Dual resolution technique combines a limited set of NCEP/GEFS ensembles and high resolution forecasts from the NCEP Global Forecast System (GFS). NCEP/GFS is a high- resolution deterministic forecast. It is an unfrozen system and major changes have been implemented frequently (http://www.emc.ncep.noaa.gov/GFS/doc.php). Historical statistics shows the

NCEP/GFS (T254L64) performs consistently better than lower resolution forecasts such as the ensemble control at T126L28 resolution up to 120 hours (Zhu et al. 2005; more GFS performance statistics are available online at http://www.emc.ncep.noaa.gov/gmb/STATS_vsdb/). In order to take advantage of both the GFS and NCEP ensembles, the NCEP/GEFS are adjusted toward GFS forecasts for forecast lead time up to 180 hours. Higher weights are chosen at short lead time and decrease with lead time.

Before applying the dual resolution technique, GFS forecast $F^{gfs}_{i,j}(t)$, NCEP ensemble control forecast $F^{ctl}_{i,j}(t)$ and each NCEP/GEFS member $F^{k}_{i,j}(t)$ are first bias corrected with the same process (Cui et al. 2011). In order to combine the bias corrected $F^{gfs}_{i,j}(t)$, $F^{ctl}_{i,j}(t)$ and $F^{k}_{i,j}(t)$ for the first 180 hours, a cosine weighting function $w_{gfs}(t)$ has been used to weight the difference between $F^{gfs}_{i,j}(t)$ and $F^{ctl}_{i,j}(t)$ with highest weight for the GFS at short lead time in such a way that it smoothly approaches the ensemble forecast as the lead-time approaches 180 hours (Figure 1). Here is the formulation for each ensemble forecast:

$$F^{k*}_{i,j}(t) = F^{k}_{i,j}(t) - w_{gfs}(t) \left(F^{gfs}_{i,j}(t) - F^{ctl}_{i,j}(t) \right) \qquad (k=1,2,\ldots n) \qquad (1)$$

Where $w_{gfs}(t) = (1+ \cos (t))/ 2$, t represents forecast hours from 0 to 180. Figure 1 illustrates that the GFS has higher weights at short lead time that decrease with lead time from 1 to 0 at 180 hour. The dual resolution technique began to be operationally performed in 2007 at NCEP.

2.2 CMC ensemble adjustment before multi-ensemble combination

The generation of the NAEFS joint ensemble from NCEP and CMC ensembles is not a simple grouping together of the calibrated forecasts. Due to the different data assimilation systems, there are systematic differences between the NCEP and CMC analyses. Such systematic differences exist not only in initial conditions but accompany every time step of the forecast. It is not proper to remove these analysis differences by use of the usual calibration methods. A simple and practical method is designed and applied at NCEP. The differences between the NCEP and CMC analyses are computed and updated daily with a 30% decay weight for the latest differences. Then CMC ensemble is adjusted individually by removing the accumulated analysis differences prior to the merging of NCEP and CMC ensembles. The reason for adjusting the CMC ensemble toward the NCEP analysis is that the products generated at NCEP are provided to the US NWS. The NWS users are familiar with NCEP analysis characteristics and tendencies. The impacts of different weights for adjustments on the multi-ensemble performance are not investigated in this study. Further research on multi-ensemble combinations is needed.

2.3 NAEFS probabilistic forecasts and forecast anomaly

After a set of post-processing steps, i.e. bias correction, dual resolution, CMC ensemble adjustment, NAEFS combines the CMC and NCEP ensembles to create probabilistic forecast products on the 1° × 1° resolution grid. The L-moment method has been introduced in probabilistic forecast generation (Zhu and Cui 2007). In particular, the

Generalized Extreme-Value Distribution (GEV) has been assumed to assimilate the ensemble forecast distribution. In general, GEV has properties similar to those of the common Gamma-3 distribution, Pearson Type 3 distribution and so on. The 10%, 50% (median), 90% probability forecast are produced by the L-moment method. The mean and spread of NAEFS are taken straight from the multi-center ensembles. The ensemble mode comes from the approximated formulation: mode = 3*median – 2*mean. Delivery of these products to users started in 2007 at NCEP.

The forecast anomalies for the ensemble mean are also generated for 19 selected bias-corrected ensemble variables. The anomalous values are the differences between climatological mean from Climate Data Assimilation System (CDAS, Kalnay et al. 1996) and bias corrected ensemble mean. The corrected ensemble means have been adjusted by considering the systematic difference between CDAS reanalysis and current NCEP Global Data Assimilation System (GDAS) analysis.

2.4 Statistical downscaling methodology

Statistical downscaling is the last post-processing step in the NAEFS context. It was implemented at NCEP in December 2007 and is currently generating NDGD 5km resolution products for CONUS from 1° ensemble forecasts. There are four variables for this application: 2-meter temperature, surface pressure, and 10-meter u and v wind components. The downscaled products include probabilistic forecasts (10%, 50%, and

90%), ensemble mean, spread and mode - twice per day out to 16 days. The statistical downscaling process follows three steps.

(a). Choosing truth (reference)

The Real Time Mesoscale Analysis (RTMA) system generates CONUS-scale real-time hourly analyses on the National Digital Forecast Database (NDFD) grid at NCEP (Pondeca et al. 2007). As a 5-km resolution analysis with high quality, RTMA is chosen as the truth or reference for the statistical downscaling process.

(b). Getting the Downscaling Vector DV^{5km}

This step is used to establish the relationship between the coarse and fine resolution analysis. We assume that the differences between the low resolution NCEP GDAS (Global Data Assimilation System) analysis and high resolution RTMA analysis are systematically reproducible and remain contant throughout the ensuing forecasts. The time mean difference between low and high resolution analysis is used as to define the analysis uncertainty. $GDAS^{5km}$ is the GDAS analysis on the 5km grid interpolated from 1° resolution data. DV^{5km} is the downscaling vector, defined as the difference between $GDAS^{5km}$ and $RTMA^{5km}$ at the same valid time. The DV^{5km} are updated daily by applying a decaying averaging algorithm.

$$DV^{5km}(t_0) = (1 - w)\, DV^{5km}(t_{-1}) + w\,(GDAS^{5km}(t_0) - RTMA^{5km}(t_0)) \qquad (2)$$

Where t_0 is the latest analysis valid time that can be at 00UTC, 06UTC, 12UTC and 18UTC, respectively. t_{-1} is previous analysis cycle valid at the same time as t_0 . The DV^{5km} is calculated for each individual grid point and there are 4 downscaling vectors DV^{5km} each day. w is the decay weighting coefficient. The value w=0.3 has been used in operations which is mainly using information from the past 3-4 days. The decaying average algorithm is used because it is communicated rapidly. It is simple and does not require a buffer of recent samples, which is of practical importance for real-time operation.

(c). Downscaling the forecasts DF^{5km}

In order to get the downscaled forecast, DF^{5km}, bias-corrected ensemble forecasts at 5km BF^{5km} are generated through bilinear interpolation. BF^{5km} can be either one ensemble member or probabilistic forecasts such as ensemble mean and mode. DV^{5km} is subtracted from BF^{5km} at the same analysis time.

$$DF^{5km}(t) = BF^{5km}(t) - DV^{5km}(t_0) \qquad (3)$$

Where t is the forecast time, and t_0 is chosen to have the same valid analysis time as t. There are four DV^{5km} available - at 00UTC, 06UTC, 12UTC and 18UTC. Downscaled products are generated following the above step.

The downscaling strategy does not require large amounts of computational resources. This approach can be easily applied to different model or ensemble forecast output.

Downscaled products from NCEP/GEFS and NAEFS for CONUS respectively have been available since December 2007. In this paper, the bias-corrected and downscaled NAEFS forecast are denoted as NAEFS final products.

3. Verification procedure

There is an ensemble-based probabilistic forecast verification system at NCEP (Zhu and Toth 2008). This system was developed for NCEP/GEFS in the 1990s and recently upgraded and applied to NAEFS. To assess the benefits of the post-processing techniques, several probabilistic measures and traditional evaluation methods are used from the NCEP/GEFS and NAEFS verification system (Zhu 2004; Zhu and Toth 2008). The statistics include mean bias, ensemble Root Mean Square Error (RMSE), ensemble spread, mean absolute errors (MAE), Relative Operational Characteristics (ROC) and Continuous Ranked Probability Skill Score (CRPS). Diagnoses on regional areas are also performed. The RMSE of the ensemble mean measures the distance between forecasts and analyses (or observations). Ensemble spread is calculated by measuring the deviation of ensemble forecasts from their mean (Zhu 2005). The statistic CRPS measures the reliability and resolution of probabilistic forecasts and therefore evaluates ensemble performances (Toth et al. 2003; Zhu and Toth 2008). The lower the CRPS score, the better the probabilistic system is by being both reliable and exhibiting high resolution. The CRPS score is one of the most important measures for evaluating the performance of probabilistic forecasts.

In addition to the verification statistics as the basic measurements of ensemble performance, one operational product similar to the NAEFS downscaled forecasts is used

for comparison. The Meteorological Development Laboratory (MDL) of NOAA's NWS has been producing a gridded Model Output Statistics (GMOS) forecast guidance system in recent years (Dallavalle and Glahn 2005; Ruth et al. 2009). The MDL/GMOS populates a 5-km grid covering CONUS with elements needed for weather forecast grids. In the light of results from downscaling techniques, NAEFS and MDL/GMOS both perform verification based on RTMA grids (De Pondeca et al. 2007). The goal of this comparison is not to provide an exhaustive validation of the performance of the different products, but meant to complement verification of NAEFS probabilistic forecasts to show the potential benefit of downscaling techniques.

4. Results

The decaying average algorithm is chosen for both bias correction and statistical downscaling. The first issue for their application is to choose a decaying weight coefficient. A previous study (Cui et al. 2006, 2011) has shown that a 2% weight is an optimal option for bias correction. Subsection 4.1 will briefly show the performance of bias correction and dual resolution. Subsection 4.2 will verify how decaying weights affect the downscaling process.

4.1 Performance of bias correction and dual resolution

Figure 2 gives an example of the performance of bias correction. It shows NCEP/GEFS 2-meter temperature 120-hour forecasts (ensemble mean) and forecast errors before and after

calibration. The initial time is 0000UTC 7 October, 2006. Most of systematic errors are removed and the improvement of the ensemble is noticeable. Though in some areas the bias signs change from positive to negative, error magnitudes decrease greatly, indicating the effectiveness of bias correction.

The performance of the dual resolution technique is estimated in this subsection. The improvement depends on the variables. For example, there is significant improvement for Northern Hemisphere (NH) 2-meter temperature. Figure 3 shows one month of statistics. E14s denotes the NCEP 14 global ensemble raw forecast, E14sb is the bias- corrected forecast, and E14hbbc is the ensemble with both bias-correction and dual resolution technique application . Both GFS and ensemble use a 2% weight in calibration (Cui et al. 2011). Apparently the bias-corrected ensemble with dual resolution adjustment gives significant improvement for short lead time up to 180 hours compared to the ensemble with bias correction only. E14hbbc and E14sb are similar to each other after 180 hours, corresponding to the zero weight applied to the GFS forecast after 180 hours. The bias-corrected ensemble, E14sb has a better score than the raw ensemble. The contribution from dual resolution is even lager than the bias correction.

4.2 Statistical downscaling performance

Figure 4 shows sensitivity tests of decaying weights for the downscaling technique. All seven curves are domain-averaged bias (absolute values) over CONUS. The effects of the downscaling treatment are clearly displayed. The seven curves are divided into two

clusters, one with the downscaling process and one without. The gefs_raw curve is from the NCEP raw ensemble, and the gefs_bc is from NCEP bias-corrected ensemble. Both are bilinearly interpolated values on NDGD grids and have no downscaling processing. Their biases are much bigger than those of the five downscaled ensembles, indicating that the downscaling processing can effectively reduce the forecast errors on fine grids. Bias correction contributes to the improvement of the ensemble but has weak impact compared to the downscaling process. Around 70% of forecast errors are reduced at all lead time (gefs_bcds_10%). The effect of downscaling on the low resolution analysis is also displayed. It is notable that the biases are significantly reduced at 00hr after downscaling. Diagnosis of the analyses also shows that most differences between low resolution analysis and RTMA analysis disappear or become smaller after the downscaling process (not shown), suggesting that low resolution GDAS analysis is made to look more similar to high resolution RTMA through downscaling. It provides a possibility that the downscaling process is able to create fine resolution information based on coarse resolution fields, i.e., to predict high resolution analysis from low resolution analysis.

Among the 5 downscaled ensembles, discrepancies are also discernible. The downscaled ensemble with 10% weight is the best. The 5% and 10% weight curves are relatively close at all lead times. Though the room available for improvement becomes smaller when the decay weight increases from 2% to 10%, it was suggested that we experiment with higher weights. Downscaled ensembles with decay weights of 20%, 30% and 50% were created and compared (not shown). A weight of 30% was finally adopted as an optimal choice and was used in the operational statistical downscaling process for NAEFS in 2007.

4.3. NAEFS downscaled product performance

These subsections describe the comparison between NAEFS final products and raw ensemble outputs. NAEFS products come from multi-ensembles system with ensemble size increased from 20 to 40 members. Considering its multi-ensemble characteristics, one hopes that the NAEFS will have a better global skill than a single NCEP/GEFS or CMC/GEFS. In order to show the multi-ensemble results, the comparison is also performed with the single NCEP and CMC ensemble.

Figure 5 gives an example of a 24-hour forecast of 2-meter temperature from the NCEP raw ensemble mean and NAEFS final product. The NAEFS can account for the complex influence of land surface heterogeneities. There is more detailed information added to the forecasts of NAEFS than the raw ensemble.

Figure 6 compares mean absolute errors (MAE) of 12-hour forecasts of 2-meter temperature for September 2007 with 3 ensembles. The three images are for the NCEP raw ensemble, the NCEP bias-corrected and downscaled ensemble, and the NAEFS final product. All the MAE is calculated with respect to the RTMA analysis and use the same shading scales. Most biases are reduced due to the bias correction and downscaling processes (Figure 6a and 6b), especially over western high topography areas. The effects of multi-ensembles are also shown through the comparisons between downscaled NCEP and downscaled NAEFS. The domain- averaged MAE values for the 3 images are 1.999, 1.161

and 1.002, indicating that the combined NAEFS has more advantage than the single NCEP/GEFS system.

Figure 7 shows the MAE change with forecast lead time for four ensemble systems. They are the NCEP raw ensemble (NCEP_raw), NCEP bias corrected and downscaled ensemble (NCEP_drbcds,), CMC bias corrected and downscaled ensemble (CMC_bcds), and NAEFS final products. The MAE of 2-meter temperature is shown to decrease mostly after the downscaling process is applied to the ensembles (Figure 7a). Both NCEP and CMC have very comparable ensemble forecast systems. The final NAEFS product gains +4 days compared to NCEP raw forecast. The result for the u-component of the 10-meter wind also indicate the fact that an ensemble size increase from 20 to 40 members can improve the skill of the multi-ensemble system compared to its components (Figure 7b).

Figure 8 displays ensemble RMSE and ensemble spread for the four ensembles as in Figure 7. The downscaled forecasts (NAEFS, NCEP_drbcds, CMC_bcds) have reduced ensemble RMSE compared with the raw and bias-corrected ensemble (NCEP_raw). The NAEFS has the smallest RMSE at all lead times with RMS errors reduced by around 12% - 35%. NAEFS has the largest spread, especially for long lead time. The NCEP_raw and NCEP_drbcds have identical ensemble spread. The reason for this is that the bias correction of NCEP/GEFS is designed only to reduce their systematic errors for the first moment. Higher moments such as ensemble spread are not impacted.

The reliability and resolution of NAEFS products are also investigated. Figure 9 shows the CONUS performance (CRPS) of four ensemble systems. CRPS is used to measure the distance of the ensemble's distribution from the true distribution. The CRPS score shows that the downscaled NAEFS gains 8 days in score as compared to the NCEP raw ensemble.

4.4. NAEFS and MDL/GMOS products

This subsection compares downscaled NAEFS and MDL/GMOS products. The comparison shown for 2-meter temperature only. The period is from September 5 2007 to September 30 2007. Both MDL/GMOS and downscaled NAEFS products are gridded guidance and the verifications are against RTMA. As already mentioned, the goal is not to present an exhaustive comparison, but to show the general characteristics of statistical downscaling in this study. Figure 10 shows verifications of NAEFS and MDL/GMOS for 2-meter temperature. Results indicate that downscaled NAEFS have smaller MAE than GMOS from day 1 through day 8.

To examine the extent of the improvement obtained from downscaling, comparisons of the NAEFS with MDL/GMOS in four NWS CONUS regions (southern US, western US, central US and eastern US) were performed. For NAEFS, the extent of MAE improvement in western areas is better than that of the other three regions. There was approximately a 15%- 30% of reduction of MAE errors as compared to MDL/GMOS. For the other three regions, NAEFS is better than MDL/GMOS from day 1 to 5, and becomes comparable to MDL/GMOS for the remaining three days. The western areas have complex terrain.

Downscaling technique is effective in producing high- resolution products that represent grid analysis very well with smaller MAE. On the other hand, most elements of GMOS are created by analyzing all available MOS station forecasts (Glahn et al. 2008). The relative high MAE of GMOS comes from station observation uncertainty. It is not a surprising result because the downscaling process is trained through RTMA analysis and GMOS are trained by observations. In current data assimilation system, it is difficult to find a reference to satisfy all the products verification. The evaluations and comparisons among several products help us to understand the characteristics of a new technique.

5. Summary and Future Plan

In the NAEFS, a set of statistical post processing techniques has been designed and applied for both NCEP/GEFS and CMC/GEFS in operation. Among them, the most important approaches are the bias correction and statistical downscaling. The bias correction is used to statistically correct ensemble lead-time- dependent systematic errors. The statistical downscaling aims to compensate the lack of details of model spatial/temporal resolution. These are two separate and independent processing techniques. Both are intended to improve ensemble prediction reliability and resolution. The correction method was implemented operationally at both centres (NCEP and CMC) in 2006. The statistical downscaling method was implemented at NCEP in 2007. Today's post-processor produces outputs on 1°x1° latitude/longitude grids for 49 variables and on 5x5 km grid for four fields covering the CONUS: surface pressure, 2-meter temperature, and 10-meter wind components (u and v). The post-processor uses equal weights and dual resolution methods

when combining a limited set of NCEP ensembles and high resolution GFS forecasts. The downscaled probabilistic forecast products are for NECP/GEFS and NAEFS. After applying the statistical post processing, the quality and value of ensemble forecast, including both its reliability and resolution, show a significant increase.

This study has provided a discussion of the positive and negative effects of the NAES post-processing system with a focus on the downscaling method. The statistical downscaling method is a process of deriving information from high resolution analysis and inserting variability into coarse forecasts. A decay weight of 0.3 is chosen in operations for accumulating analysis differences between RTMA and GDAS. Downscaling helps enhance "spatial" resolution and is not dependent on lead time. There is more forecast detail in the downscaled forecast while around 10% - 30% of RMS errors are reduced. The variability of the reliability of downscaled ensemble forecasts is investigated. CRPSS show that the downscaled and bias-corrected ensemble forecasts have been improved compared with the raw ensembles. NAEFS product has reduced mean absolute errors by 4+ days and improved probabilistic skills by 8+ days. NAEFS has a better global skill than a single NCEP/GEFS or CMC/GEFS. Downscaling contributes more to ensemble performance than bias correction and the multi-ensemble approach.

There is great need for bias-corrected and downscaled forecasts among the ensemble user groups. Downscaling offers the chance to transfer more information into the low resolution forecasts for valuable surface variables at low cost. The design and application of downscaling are simple that offer the potential to easily incorporate successful

downscaling methods into operational prediction systems for different regions, models or ensemble outputs. Once the RTMA is available, the statistical downscaling method can be applied to other regions such as Alaska, Hawaii, Puerto Rico and Guam. More new variables are also easily added. New downscaling techniques were implemented recently in December 2010 for the Alaska region. In addition to the four variables in current CONUS operations, four new important and high-demand surface variables have been included: wind speed and direction, maximum and minimum temperature. New downscaling methods for the four new variables have been developed (to be discussed in another paper) and probabilistic forecasts (10%, 50%, and 90%), ensemble mean, mode and spread are produced, four times per day out to 16 days.

One goal of the statistical downscaling is to provide high-resolution probabilistic products for as many NDGD elements as possible. The probabilistic guidance should be accurate, reflect high-resolution terrain, and provide good forecast continuity. NAEFS enhancements and implementations are still in progress. NAEFS plan to upgrade CONUS products with four additional variables as in Alaska and two new variables (2-meter dew point temperature, 2-meter relative humility). New downscaling methods are being developed and are expected to be implemented in NCEP operations in 2012. NAEFS is providing more guidance of user-relevant variables on the fine scale.

Acknowledgments: The authors are grateful to Valery Dagostaro and Kathy Gilbert for helpful discussions and providing data. Thanks also to Mr. Peter Caplan for his careful proofreading.

References

Antolik, M.,2006: Gridded MOS at MDL: A Downscaling Tool. *The 3rd NCEP/NWS Ensemble User Workshop*. November 1 -3, 2006, Laurel, MD.

Atger Frédéric, 2003: Spatial and Interannual Variability of the Reliability of Ensemble-Based Probabilistic Forecasts: Consequences for Calibration. *Mon. Wea. Rev*., 131, 1509- 1523.

Benestad R. E., 2004: Empirical-statistical downscaling in climate modeling. Eos, Trans. *Amer. Geophys. Union*, 85, 417–422.

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

Candille, G., S. Beauregard, and N. Gagnon, 2010: Bias Correction and Multiensemble in the NAEFS Context or How to Get a "Free Calibration" through a Multiensemble Approach. *Mon. Wea. Rev.*, **138**, 4268–4281.

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

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System. *Mon. Wea. Rev*., 138, 1877-1901.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, D. Unger, and S. Beauregard, 2006: The Trade-off in Bias Correction between Using the Latest Analysis/Modeling System with a Short, versus an Older System with a Long Archive. *The First THORPEX*

International Science Symposium. Deccember 6-10, 2004, Montréal, Canada, World Meteorological Organization, P281-284.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2011: Bias correction for global ensemble forecast. Submitted to Weather and Forecasting, in press.

Dallavalle, J. P., and B. Glahn, 2005: Toward a gridded MOS system. Preprints, *21st Conference on Weather Analysis and Forecasting*, Washington, DC, Amer. Meteor. Soc., 13B.2.

De Pondeca, M. S. F. V., G. S Manikin, S. Y. Park, D. F. Parrish, W. S. Wu, G. DiMego, J. C. Derber, S. Benjamin, J. D. Horel, S. M. Lazarus, L. Anderson, B. Colman, G. E. Mann, and G. Mandt, 2007: The development of the real time mesoscale analysis system at NCEP. Preprints, 23rd Conference on Interactive Information Processing Systems, San Antonio, TX, Amer. Meteor. Soc., P1.10.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

_____, and D. P. Ruth, 2003: The new digital forecast database of the National Weather Service. *Bull. Amer. Meteor. Soc.*, **84**, 195-201.

_____, K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2008: The gridding of MOS. *Wea. and Forecasting*, in press.

Hamill M. T 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.

Kalnay E., M Kanamitsu, R Kistler, W Collins, D Deaven, L Gandin, et al. 1996: The NCEP/NCAR 40-year reanalysis project, *Bull. Amer. Meteor. Soc*., 77, 437-470, 1996.

Leung L. R., and S. J. Ghan, 1995: A subgrid parameterization of orographic precipitation. *Theor. Appl. Climatol*., 52, 95–118.

Leung L. R., and S. J. Ghan, 1998: Parameterizing subgrid orographic precipitation and surface cover in climate models. *Mon. Wea. Rev.*, 126, 3271–3291.

Leung L. R., Y. Qian, X. Bian, W. M. Washington, J. Han, and J. O. Roads, 2004: Mid-century ensemble regional climate change scenarios for the western United States. *Climate Change*, 62, 75–113.

Maurer E. P., and P. B. Duffy, 2005: Uncertainty in projections of streamflow changes due to climate change in California. *Geophys. Res. Lett.*, 32, L03704, doi:10.1029/2004GL021462.

Ruth, D. P., B. Glahn, V. Dagostaro, K. Gilbert, 2009: The Performance of MOS in the Digital Age. *Wea. Forecasting*, 24, 504–519.

Toth, Z., 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 & Sons, Ltd., 137-163.

Toth, Z., Y. Zhu, and R. Wobus, 2004: March 2004 Upgrades of the NCEP global ensemble forecast system. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html]

Toth, Z., O. Talagrand, and Y. Zhu, 2006a: The attributes of forecast systems: A framework for the evaluation and calibration of weather forecasts. In Palmer, T.N.

and Hagedorn, R., editors, Predictability of Weather and Climate. Cambridge University Press, pp. 584-595.

Toth, Z., P. Schultz, S. Mullen, J. Demargne and Y. Zhu, 2006b: Completing the forecast: assessing and communicating forecast uncertainty. *3rd NCEP/NWS Ensemble User Workshop*. 31 Oct - 2 Nov, 2006 in Washington DC

Toth, Z., and M. Pena, 2007: Data assimilation and numerical forecasting with imperfect models: The mapping paradigm. *Physica* D, 230, 146-158.

Wang Y., L. R. Leung, J. L. McGregor, D.-K. Lee, W.-C. Wang, Y. Ding, and F. Kimura, 2004: Regional climate modeling: Progress, challenges, and prospects. *J. Meteor. Soc.* Japan, 82, 1599–1628.

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.

Zhu, Y., 2004: Probabilistic forecasts and evaluations based on a global ensemble prediction system, *World Scientific Series on Meteorology of East Asia, Vol. 3* - Observation, Theory, and Modeling of Atmospheric Variability, 277-287

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

Zhu, Y., B. Cui, and Z. Toth, 2007a: December 2007 upgrade of the NCEP global ensemble forecast system (NAEFS). [Available online at http://wwwt.emc.ncep.noaa.gov/gmb/yzhu/imp/i200711/IMP_PLAN_final_v08_brief.pdf]

Zhu, Y., and Z. Toth, 2008: Ensemble Based Probabilistic Forecast Verification. Preprints, *19th Conference on Probability and Statistics*. New Orleans, Louisiana, Amer. Meteor. Sco. 2.2

Zhu, Y. and Z. Toth, R. Wobus, Hou, D. and B. Cui, 2010: February 2010 upgrade of the NCEP global ensemble forecast system (NAEFS). [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html]

Zhu, Y., and B. Cui, 2011: NAEFS upgrade science review. [Available online at http://www.emc.ncep.noaa.gov/gmb/yzhu/html/imp/201103_imp.html]

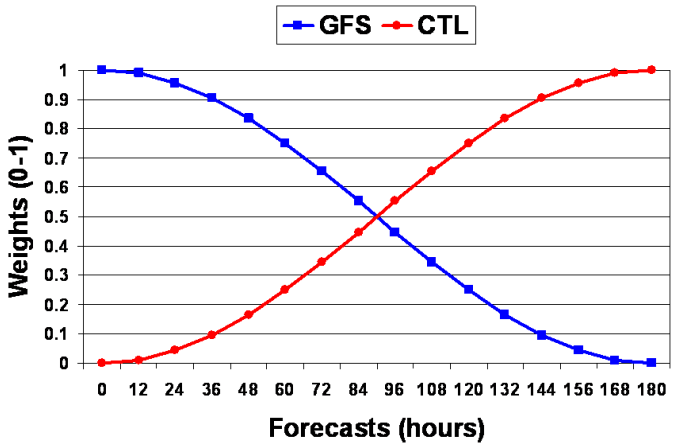

Figure 1. Cosine weighting function $w_{gfs}(t)$ used to weight NCEP high resolution forecast GFS and ensemble control forecast CTL.

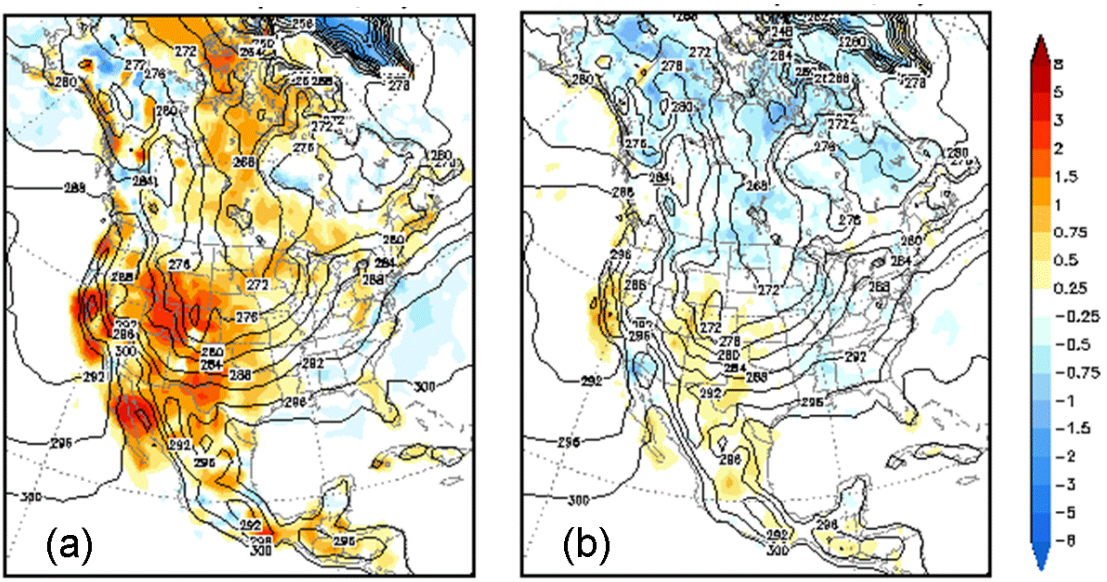

Figure 2. Example of 2-meter temperature 120 hour forecasts (K, contour) and forecast errors (shaded) for (a) NCEP raw ensemble mean and (b) NCEP bias corrected ensemble mean. The initial times are 0000UTC 7 October, 2006.

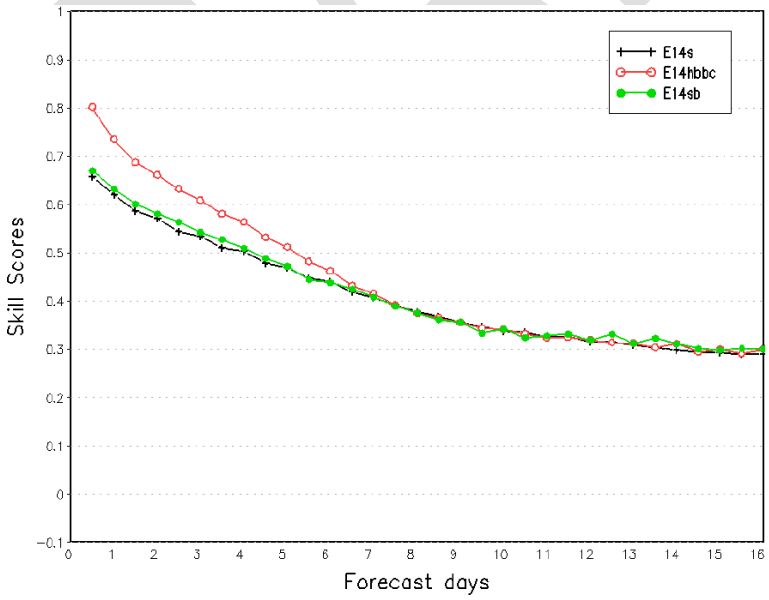

Figure 3. ROC areas (from 0 to1) for the NCEP 14 global ensemble raw forecast (E14s), compared to the bias corrected forecast (E14sb) and the dual resolution bias corrected forecast (E14hbbc) for NH 2-meter temperature for one month from May 13, 2007 to June 15, 2007.

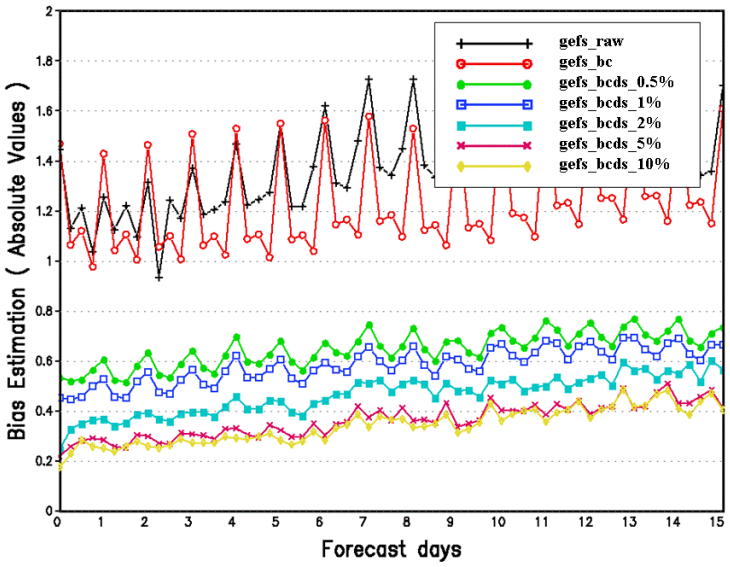

Figure 4. 3 months averaged ensemble mean forecast errors of 2-meter temperature ending at 2007030500 for NCEP raw ensemble (gefs_raw), bias corrected ensemble (gefs_bc), bias corrected and downscaled ensemble with weights of 0.5%, 1%, 2%, 5% and 10% (gefs_bcds_0.5%, gefs_bcds_1%, gefs_bcds_2%, gefs_bcds_5%, gefs_bcds_10%).

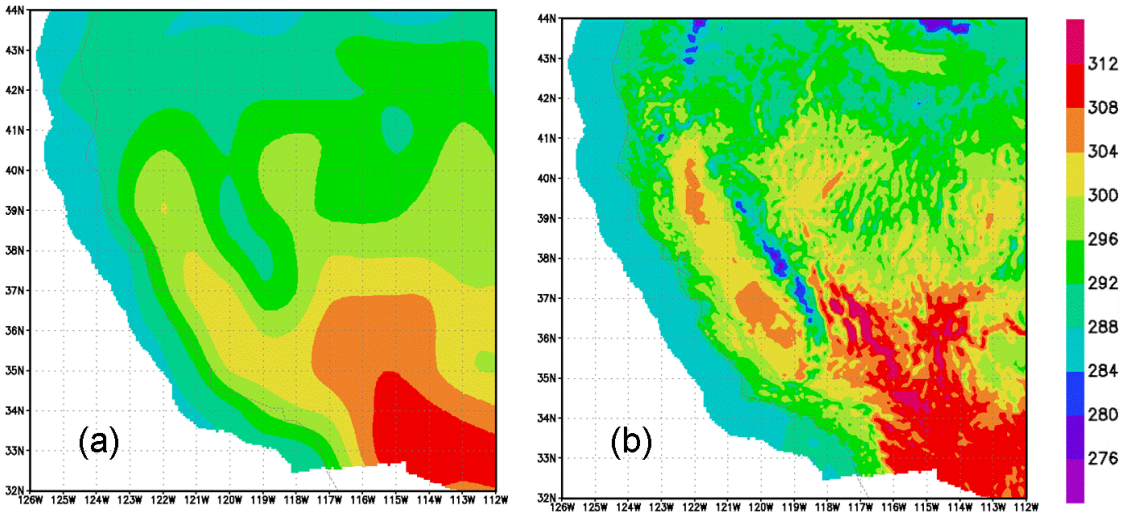

Figure 5. Example of 2-meter temperature 24 hour forecast (K): (a) NCEP raw ensemble mean and (b) NAEFS bias corrected and downscaled ensemble mean.

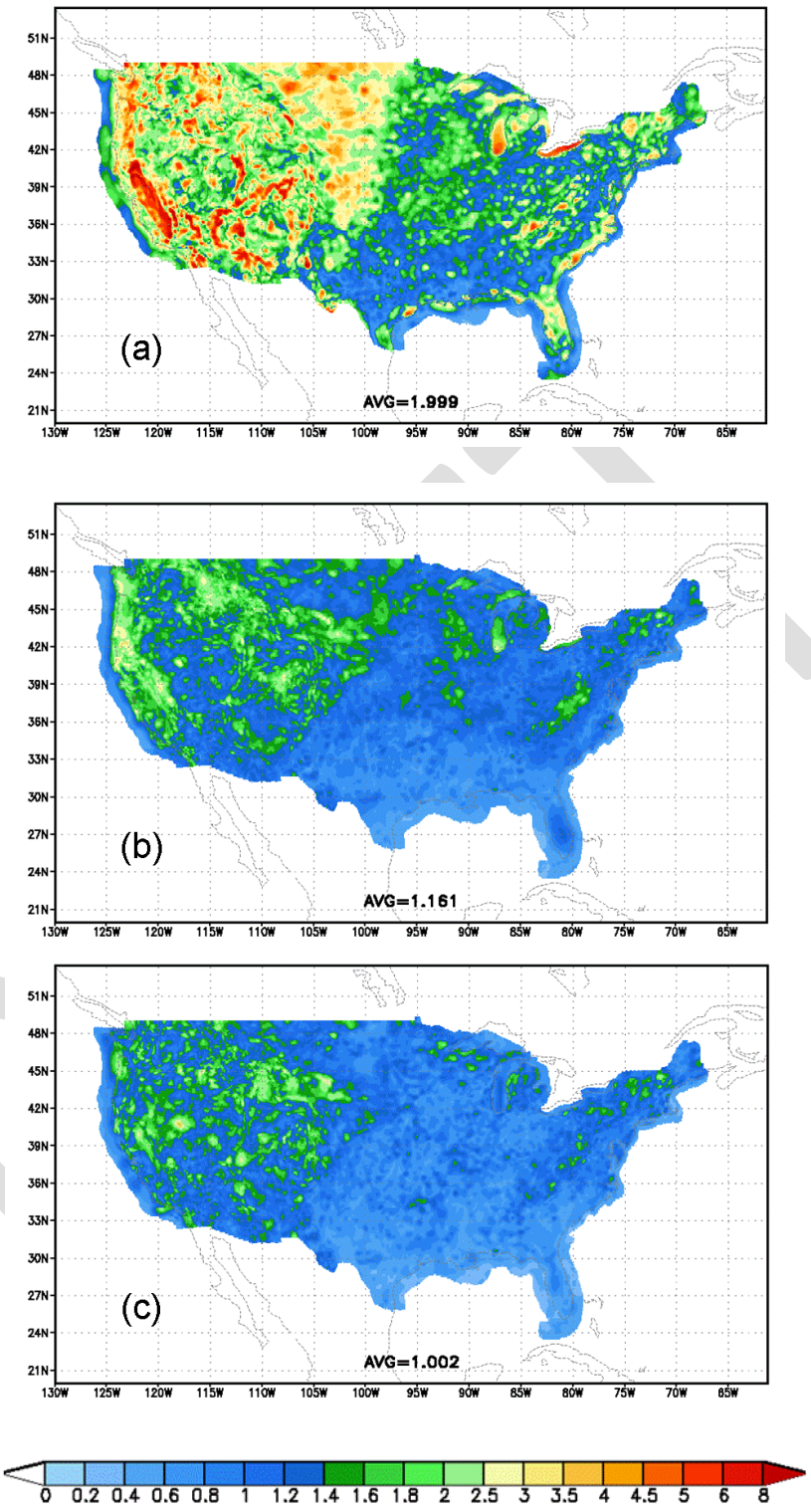

Figure 6. Mean absolute error of 2-meter temperature 12 hour forecasts with respect to RTMA for CONUS for September 2007: (a) NCEP raw ensemble forecasts, (b) NCEP bias corrected and downscaled ensemble forecasts, and (c) NAEFS final forecasts.

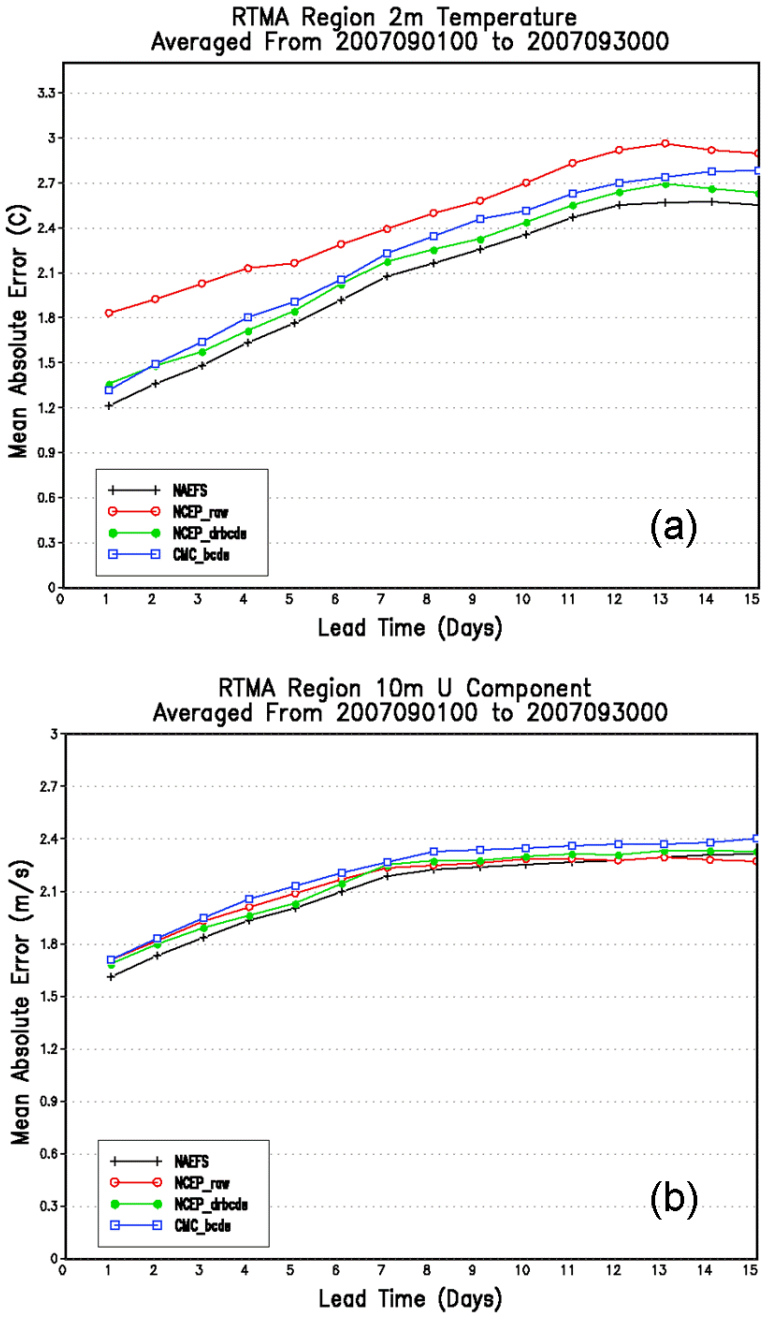

Figure 7. CONUS MAE of September 2007 for NAEFS final products (NAEFS), NCEP raw ensemble (NCEP_raw), NCEP bias corrected and downscaled ensemble (NCEP_drbcds), CMC bias corrected and downscaled ensemble (CMC_bcds) for (a) 2-meter temperature, and (b) 10-meter wind u component.

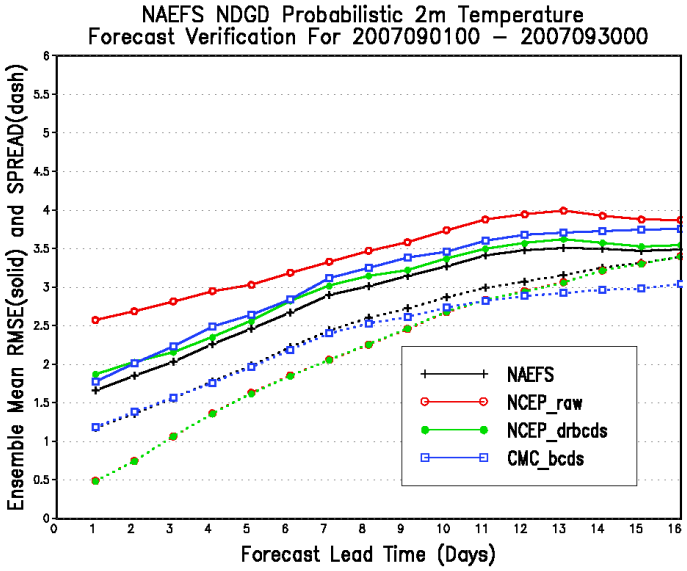

Figure 8. As in Fig.7, but for the ensemble RMSE (solid) and ensemble spread (dashed) of 2-meter temperature

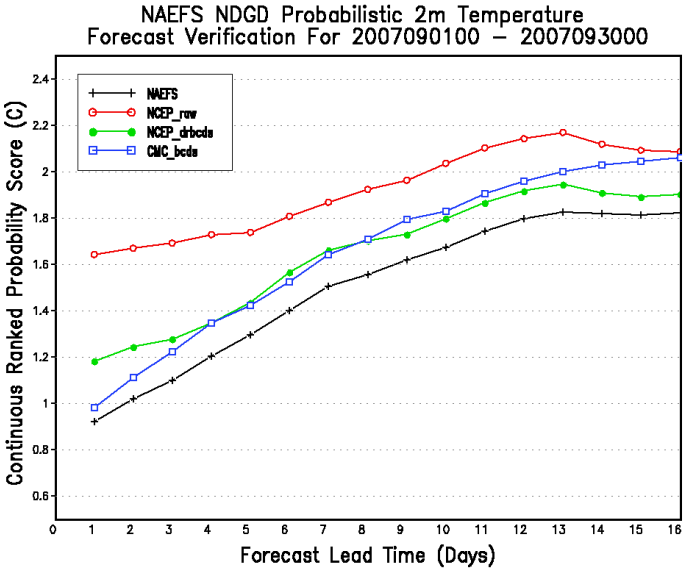

Figure 9. As in Fig.7, but for the CRPS of 2-meter temperature.

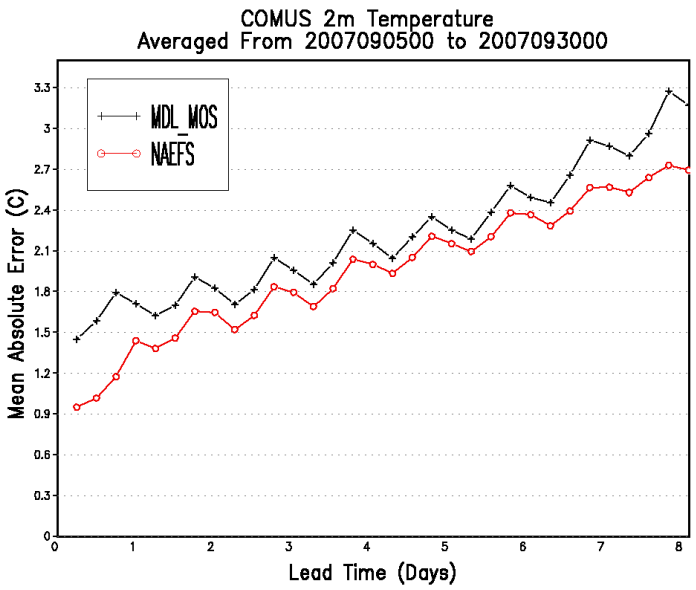

Figure 10. Mean absolute errors of 2-m temperature for September 2007. Forecasts are from MDL/GMOS forecast and NAEFS final products and verified against RTMA.