Precipitation Calibration Based on Frequency Matching Method (FMM)

Yuejian Zhu[1*] and Yan Luo[1,2]

1. *Environmental Modeling Center/NCEP/NWS/NOAA, College Park, MD*
2. *I. M. Systems Group, Inc. College Park, MD*

To be submitted to *Weather and Forecasting*

* *Corresponding author address:*

Yuejian Zhu,
Environmental Modeling Center/NCEP/NWS/NOAA
5830 University Research Court
College Park, MD 20740
E-mail: Yuejian.Zhu@noaa.gov

Abstract

A post-processing technique is employed to correct model bias for precipitation fields in real time based on a comparison of the frequency distributions of observed and forecast precipitation amounts. Essentially, a calibration is made by defining an adjustment to the forecast value in such a way that the adjusted cumulative forecast distribution over a moving time window dynamically matches the corresponding observed distribution accumulated over a domain of interest, e.g., the entire contiguous United States (CONUS), or different River Forecast Center (RFC) regions in our cases. In particular, the Kalman Filter method is used to catch the flow-dependence and bias information. Calibration is done on a point-wise basis for a specified domain. Using this unique technique, the calibration of precipitation forecasts for the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) was implemented into operations in May 2004. To further satisfy various users, especially for hydro-meteorological and short-range weather forecast applications, a recent upgrade to the May 2004's implementation has been made. It includes application of bias correction for higher resolution forecasts with better analysis, and construction of a cumulative frequency distribution based on each RFC region instead of the entire CONUS domain to take realistic regional climate features into account. This study focuses on one degree spatial and 6-hour temporal resolution out to a 384 hour (about 16-day) forecast to provide detailed information on precipitation events, using the newly developed NCEP Climatology-Calibrated Precipitation Analysis (CCPA) as the proxy for the truth. Mean forecast errors and skill are evaluated with respect to CCPA over the CONUS and each

RFC for the period 2009-2010 for 6-hour accumulations at one-degree spatial resolution. From this study it was found that this frequency matching algorithm substantially improves NCEP GFS/GEFS model precipitation forecast biases over a wide range of forecast amounts and produces more realistic precipitation patterns. Moreover, this approach improves the forecast prediction skill measured by most verification scores. In addition, the skill of probabilistic quantitative precipitation forecast (PQPF) has been also improved by applying this method to the individual GEFS ensemble members.

1. Introduction

There are many important applications that require a more accurate quantitative precipitation forecast (QPF) and probabilistic quantitative precipitation forecast (PQPF). One of these applications is the daily forecast. The QPF and ensemble based PQPF forecast products were implemented into NCEP operations in the late 1990's (Zhu et al. 1998, Zhu 2005). A better calibrated PQPF could benefit the short- and medium-range forecasts, and extend the forecast predictability (Eckel and Walters 1998, Zhu and Toth 1999). There are several studies on calibration of PQPF; some focus on the methodology (Krzysztofowicz and Sigrest 1998; Christopher et al. 2008) and others use reforecast information (Hamill et al. 2002, Fundel et al. 2009). An analog method has been developed by using large samples of reforecasts (Hamill and Whitaker 2006), and is experimentally run at ESRL and provides additional guidance for NCEP Weather Prediction Center (WPC) forecasters. Another important application of this method is for down-stream applications. Water management decisions are crucially dependent on forecast information regarding the possible future evolution of precipitation. On the other hand, hydrologic models need accurate precipitation forecasts from numerical weather prediction (NWP) as forcing inputs. Therefore, a realistic representation of the precipitation field in forecasts is very important. However, many studies have demonstrated systematic biases in the model precipitation products due to model deficiencies. It has long been recognized that model precipitation uncertainty affects the accuracy of hydrologic modeling (Demargne et al. 2013), because the performance of distributed, physically based hydrologic models depends greatly on the quality of the precipitation input data. For both these reasons, post-processing techniques have been developed and applied to reduce these biases in the precipitation. Many studies have demonstrated some success with precipitation forecasts through statistical post-processing. For instance, Yuan et al. (2007) applied an artificial neural network as a postprocessor to calibrate Probabilistic Quantitative Precipitation Forecasts (PQPF) from the NCEP Regional Spectral Model (RSM) ensemble forecast system. Voison et al. (2010) described two bias correction methods with spatial disaggregation (BCSD) and an analog technique for downscaling and calibrating errors from ensemble precipitation forecasts. In this study we developed a method for precipitation calibration in real time called the “frequency matching method”. Basically the methodology employed here is a statistical adjustment based on cumulative frequency distributions of forecast and observed precipitation amounts. Two steps are undertaken in calibration with frequency matching. As first, it requires an observation dataset at the same spatial and temporal resolution as the model forecast output and a reasonable number of days of prior forecasts to construct their respective cumulative frequency distributions for forecasts and observations. In addition, it introduces a time moving window for sampling appropriate historical bias information that makes the cumulative frequency distribution of forecasts match that of the observations. The second step in this method makes use of the cumulative frequency distributions of observations and forecasts, in this way a frequency match is performed between prior observations and forecasts. The resulting correction factor is applied to adjust a target forecast value at each grid point and each grid point is treated individually.

In this paper, we first briefly review the background of the 4 May 2004 implementation at NCEP. The 2004 implementation was first developed as a pioneer version of precipitation calibration with frequency matching for application in precipitation forecasts with 24 hour accumulations at 2.5 degree resolution (Zhu and Toth 2004). Just as with any other numerical weather prediction (NWP) model, Quantitative Precipitation Forecasts (QPF) from the Global Forecast System (GFS) at NCEP suffer from biases due to model deficiencies. Probabilistic Quantitative Precipitation Forecasts (PQPF) based on the Global Ensemble Forecast System (GEFS) at NCEP are biased as well due to imperfections in the model and ensemble formation. Typically, model precipitation bias is dependent on the model version, lead time and location. In most cases, small amounts of precipitation are over-forecasted while large amounts are under-forecasted. By calibrating each member of the ensemble based on verification statistics accumulated over the continental US (CONUS), the bias in QPF (first moment) is practically eliminated, and the PQPF (second moment) is substantially improved. By following the approach of the 2004 implementation with timely availability of higher resolution model output and a better analysis, named the Climatology-Calibrated Precipitation Analysis (CCPA, Hou et al. 2013), we pursue a similar application at one degree resolution and every 6-hours out to 384 hours (about 16 days) globally.

To provide a better proxy of the truth for the precipitation field over CONUS at high spatial and temporal resolutions, CCPA has been developed and evaluated at NCEP by Hou et al. (2013). The dataset takes advantage of the higher climatological reliability of the CPC dataset (Xie et al. 2010) and the higher temporal and spatial resolution of the Stage IV dataset (Lin and Mitchell 2005). Thus, CCPA is reliable and quality controlled, with a high spatial and temporal resolution. It is available as 6 hour accumulations from 2002 onwards. The CCPA data are first produced on the 4 km HRAP (Hydrologic

Rainfall Analysis Project) grid, the same as the NCEP Stage IV over CONUS, as a primary product and then interpolated to 1, 0.5, 0.125 degree and NDGD (5km) grids by a volume conservation scheme as by-products. The 1 degree CCPA is applied in this study as it exactly matches the model output grid.

We continue to investigate here the method that applies to the NCEP GFS/GEFS precipitation model output with CCPA. Then we analyze aspects of the bias correction of ensemble precipitation forecasts, including precipitation forecast skill and reliability. Our objective is to produce bias-corrected precipitation ensemble forecasts through post processing for near real time forecast applications.

The reminder of the paper is organized as follows. Section 2 describes the frequency matching method for precipitation calibration. Section 3 reviews the background of 2004 implementation. A few cases to demonstrate the success of this method will be presented. Section 4 applies and evaluates the bias correction approach for higher resolutions using CCPA, and in the last section we present our conclusions with suggestions for future work that will further improve the calibration of precipitation.

2. Methodology

A systematic difference (or 'bias') between forecast and observed precipitation amounts can be progressively removed using information provided by observations. In this study, the bias information can be estimated through comparing forecast and observed precipitation frequency distributions. The general frequency matching method proceeds as follows. First, we conduct a bias assessment by constructing a cumulative distribution function (CDF) for the preceding forecast and corresponding observed precipitation amounts. Given a set of precipitation thresholds in ascending order, the CDF is calculated as the count of numbers of grid points over a given domain where the forecast or observed precipitation values exceed a threshold. The CDF is updated with the Kalman filter method, which is similar to the bias correction method in the NAEFS (Northern American Ensemble Forecast System) (Cui et al. 2011, 2013), expressed as:

$$\overline{\mathbf{CDF}}_{i,j} = (\mathbf{1-W}) * \overline{\mathbf{CDF}}_{i,j-1} + \mathbf{W} * \mathbf{CDF}_{i,j} \qquad (1)$$

where $\overline{\mathbf{CDF}}_{i,j}$ is the decaying averaged CDF at threshold i for day j, while $\overline{\mathbf{CDF}}_{i,j-1}$ is the prior decaying averaged CDF for day j-1. $\mathbf{CDF}_{i,j}$ is the newly counted CDF at threshold i for day j. **W** is the decaying weight between 0 and 1, defined simply by an approximated time moving window nd (nd cannot equal zero).

$$\mathbf{W} = 1 / nd \qquad (2)$$

Here a time moving window (or decaying weight) is chosen to make a weighted average of these CDFs over the domain depending on how far it is from the target forecast day, which is illustrated in Figure 1. The higher the weight the faster the decaying speed (which indicates there is a higher weight on the most recent data and less on the oldest data) and vice versa. Our strategy is to specify prior forecast days (an approximated time moving window, or decaying weight) for each grid point and each lead time as a pool for sampling appropriate historical bias information from forecasts and observations. For instance, a 50-day window (W=0.02) means training data is accumulated over the most recent 50-day period with the most weight on the most recent data (See Figure 1 for W=0.02). Thus, the idea behind the adaptive method is to catch the dynamic flow-dependence and statics of observations. The time moving window (or decaying weight) can be tuned from short (or large) to long (or small) times (weights) to ensure the best performance of the method. In our adaptation of the frequency matching method, there are two ways to construct CDFs for forecasts and observations. We call the CDF based on the whole CONUS domain the CONUS CDF, and the CDF based on each RFC region (See Figure 2) is the RFC CDF. For each grid point within a specific domain (e.g., CONUS or any RFC) and for each forecast lead time, the observed and forecast CDFs are derived using the same time moving window (or decaying weight). To be useful for applications, this method needs to handle the initial CDFs, which is termed spin-up.

Second is the bias adjustment. In order to keep the spatial and temporal coherence of a forecast as similar as possible to that of the observation, we match the cumulative frequency distribution of the forecast to that of the observation using a frequency matching algorithm. Here the updated CDFs from Equation (1) form cumulative frequency distributions. As illustrated in Figure 3, according to this pair of distributions, for a raw forecast value ("RAW") we find and assign an observed value that has the same frequency within a given domain as the forecast value to the correspondent calibrated forecast ("CAL"). Consequently, the bias information is estimated based on the paired and updated CDFs for the forecast and corresponding observed values. For example, in Figure 3 in the case of a CDF (forecast) greater than the CDF (observed), model precipitation tends to be over-forecasted, so to match the frequency a correction factor of less than one will be expected to reduce a forecast value. In doing this matching process, linear interpolation is applied twice in real calculations to derive a correction factor for each grid point. Mathematically, given an array of thresholds T_1, T_2, …, T_n in ascending order as the abscissas, an array of observed CDFs O_1,O_2, …, O_n and an array of forecast CDFs F_1,F_2, …, F_n as ordinates, an array of calibrated thresholds T_1^*, T_2^*, …, T_n^* are derived through the first linear interpolation. Consequently, the forecast CDF F_i^* at T_i^* (i=1,…, n) is equal to observed CDF O_i at T_i (i=1,…, n) just as in what we call frequency matching. That is:

$$O_1(T_1) = F_1^*(T_1^*),$$

$$O_2(T_2) = F_2^*(T_2^*),$$

$$\ldots\ldots$$

$$O_n(T_n) = F_n^*(T_n^*).$$

Next, a correction factor is calculated as the ratio of a calibrated threshold to its related threshold, i.e., $Ri = T_i^* / T_i$, i=1,…, n. Once again, given an array of thresholds T_1, T_2, …, T_n as the abscissas and the array of correction factors R_1, R_2, …, R_n as ordinates, for a forecast value ("RAW") at any grid point a correction factor ("r"), the ratio of a calibrated forecast value ("CAL") to its corresponding raw forecast value at a grid point, is derived by linear interpolation. Then the correction factor ("r") is applied to the raw forecast value ("RAW") to compute the final calibrated forecast value ("CAL= r * RAW"). This correction is applied to each model grid point which implies that the correction is a function of forecast value. No adjustment of a zero precipitation forecast value is made in order to prevent an unrealistic negative precipitation value due to interpolation.

This calibration technique with frequency matching should work with any model output as long as observations are available and are processed to be at model grid points. However, our experience with this technique indicates some important considerations must be addressed. That is, precaution must be taken about the selections of thresholds and number of decay days, particularly when the CDF is calculated for each RFC rather than the CONUS because there will be a much smaller sample size as implied from Table 1. For example, an insufficient amount of non-zero sample data is very likely to cause more than two equal values of zero as CDFs for adjacent highest thresholds, though this situation is not allowed in this method as it may lead to a failure in the interpolation. To deal with this problem, selecting a reasonable range of thresholds is necessary to produce non-equal CDF values. Another solution is choosing a proper number of decaying days. If the number of decaying days is too small it will be problematic since there will not be sufficient sample data, especially when dry-climate regions experience a long duration drought. Therefore, there is an inevitable trade-off as to the number of decaying days when tuning for optimal calibration performance. It is believed that potential difficulties in CDF construction in dry regions are related to the small number of days with precipitation, imposing a practical challenge to this method. When the above statistical deficiencies and operational limitations are avoided, the method should be computationally realistic and feasible for real-time implementation.

3. Background review

In the 2004 implementation (Zhu and Toth 2004) the calibration system was designed to apply a bias correction globally to all 00 UTC forecasts, including high and low resolution control forecasts and all ensemble member forecasts for 24 hour amounts at 2.5 degree resolution. The operational NCEP GFS/GEFS forecast system runs four times per day (00, 06, 12 and 18 UTC) and produces 1 degree global ensemble precipitation forecast products for 6-hour accumulations. It contains twenty-two ensemble members - a high resolution GFS run, low resolution GEFS control run and ten pairs of perturbed runs using the ET method (Wei et al. 2006, 2008). Once generated, precipitation forecasts from the 00 UTC cycle only are processed into 24 hour accumulations and aggregated to 2.5 degree resolution prior to bias correction. Technical information about NCEP's latest GEFS ensemble forecast system is available online (Zhu et al. 2012)

Bias assessment is approached separately for the GFS high resolution and ensemble control (low resolution) forecasts at each lead time to save computational time. Data are sampled from prior forecasts and observations with a 30-day average of the whole CONUS domain as the cold start sampling. Later, the corresponding decaying weight used is 1/30. The observations with 24 hour accumulations come from the US RFC rain gauge network with about 10,000 observation station reports after re-gridding to the common 2.5 degree model grid. A set of thresholds of 0.2, 2.0, 5.0, 10.0, 15.0, 25.0, 35.0, 50.0, 75.0 mm/day were carefully selected for the 24 hour accumulation amount to ensure no failure of interpolation in the calibration procedure, as detailed in Section 3. The bias assessment based on CONUS CDF may be applied to the global domain, when assuming that the bias information over CONUS is much the same as over other parts of the globe, which may not be an optimum application. This application can be improved when global precipitation observations become available in real time. The calibration system runs once daily at the 00 UTC cycle and typically the daily runs are completed within a minute in real time on a supercomputer.

The evaluation period for this implementation was chosen to be 1 Dec. 2000 ─ 28 Feb. 2001. Comparisons of the calibrated forecast against the raw forecast in terms of some scores were made and are shown in Figure 4. Figure 4(a) presents equitable threat scores (ETS) and bias scores at the 2.0mm threshold for each forecast lead time. Figure 4(b) provides the 36-60 hour reliability diagram at the 2.5mm threshold, validated for all grid points in the CONUS. The calibrated forecast shows a remarkably improved bias score over CONUS at all thresholds. Not only is the bias reduced, the post-processing through frequency matching helped increase the probabilistic forecast skill, such as with the Brier score (not shown). There was a much reduced PQPF (mean) bias in the calibrated forecasts, indicating a dramatic improvement in reliability relative to the raw forecasts. The reliability curve approaches the diagonal line, which indicates that the biases in PQPF were removed to some degree. The Brier score was also improved dramatically at all lead times. In general, these calibrated forecasts were much more skillful than the raw forecasts at all lead times.

4. Applications and evaluations

In this section, we expand on earlier work to upgrade the calibration system and make it capable of bias correction at higher temporal and spatial resolutions. More specifically, we use the current application at 1 degree resolution with 6 hour accumulations. This section describes how the higher resolution precipitation forecasts are calibrated so that their cumulative frequency distribution matches that of the observations.

The operational NCEP GFS/GEFS 6-hourly precipitation forecast (up to a 384 hour lead time) has a spatial resolution of one degree latitude and longitude and runs up to real time. There is a high resolution GFS run, a low resolution GEFS control run and 20 ensemble members for each forecast. Unlike in the 2004 implementation, here all 6-hourly one-degree forecasts for the four cycles are directly bias corrected with respect to the gridded precipitation analysis CCPA at the same resolution as the forecasts. To be more realistic and better capture regional climate regimes, 12 RFC CDFs are derived for each lead time to construct cumulative frequency distributions. For each category of the 9 thresholds (0.2, 1, 2, 3.2, 5, 7, 10, 15, 25 mm/6hr), a CDF is calculated as the number of grid points over each RFC where the forecasts or observed precipitation amounts are greater than the threshold. Again, to reduce the computational burden, we only derive one set of CDFs from the high resolution GFS run and another set of CDFs from the low resolution GEFS control run. Then the latter set of CDFs is applied to the 20 ensemble members since all of them are low resolution forecasts from the same forecast model, resulting in 2 rather than 22 sets of CDFs per lead time per threshold per RFC region. In each bias correction run, there are a total 1536 forecast-observation CDF pairs for 64 forecast lead times for the low high resolution runs and 768 pairs for 30 forecast lead times for the high resolution runs, summed for a total of 9 thresholds and 12 RFC regions. Because there is no high quality global precipitation analysis available, the CDF of CONUS (is sum of 12 CDFs for all the RFCs) is using for a grid point outside of CONUS, therefore the quality of precipitation calibration for these areas is limited.. Bias information is sampled with an approximate 50-80 day moving time window, and thus the decaying weight selected is 0.02. The bias correction is applied four times per day to each 6-hourly forecast at each grid point globally and to each forecast lead time independently.

The operational forecasts initialized daily at 00 UTC from 1 March 2009 through 28 February 2010 will be assessed. These forecasts produced with the same modeling suite were used to produce the calibrated forecasts. Both sets of forecasts will be examined out to 384 hours with precipitation accumulation output available every 6 hours. Although the method we developed can apply to global forecasts, in this study our evaluation domain is the CONUS, which allows evaluations of this method using the one degree CCPA dataset. The evaluation focuses on the biases and skill levels of the calibrated ensemble precipitation forecasts with respect to raw forecasts. We analyze several examples and present some verification statistics. The verification statistics will be stratified by either lead time or threshold.

Figure 5 shows one application of this calibration for the high resolution GFS forecast for selected forecast lead times (78-hr, 84-hr, 90-hr and 96-hr). The comparison is of 6 hourly accumulated precipitation (mm) initialized at 00UTC 24 January 2010 for the raw GFS forecast (left), calibrated forecast (middle) and observation (CCPA, right). Apparently the GFS over-forecasted for the CONUS in general, and the calibrated forecast reduced the forecast amount accordingly. Figure 6 shows the ensemble PQPF (same time period) for the 0.254mm/6 hours threshold where raw ensemble PQPF is on the left, calibrated PQPF (CPQPF) is in the middle and the observation is on the right. The forecast area of the PQPF is reduced; the quantity (value) of PQPF is smaller in the calibrated PQPF, which matches better with the observations.

To demonstrate the benefits from this calibration, several different scores have been presented for the seasonal and yearly averages. The bias scores and ETS for CONUS for the period of 1 December 2009 - 28 February 2010 are shown in Figure 7 and Figure 8. Figure 7 (a) is for the 0-6 hour forecast bias of the different thresholds, and Figure 7 (b) shows forecast lead times out to 180 hours for greater than 0.2mm/6 hours. The numbers above the thresholds in Figure 7 (a) indicate the sample size of the one by one degree forecast box we have verified. Overall, the bias is reduced and ETS is increased in the calibrated forecasts for both the GFS and GEFS control for all lead times, and the improvement of ETS tends especially to be more effective for shorter lead times. Similar improvements in bias scores and ETS are also observed for the RFC regions, such as the MBRFC and NERFC shown in Figure 9 and Figure 10, respectively, although they exhibit slightly larger diurnal variability.

The RMSE (root mean square error) and ABSE (absolute error) of CONUS for the period of 1 March 2009 – 28 February 2010 (one year) for every 6-hr accumulated precipitation forecast are shown in Figures 11 and 12. Figure 11 is for the GFS forecast and Figure 12 is for the GEFS control forecast. Based on this year of statistics, RMSE is reduced significantly for the GFS, but not for the (lower resolution) GEFS control. This difference might be related to the model resolutions and model versions (the operational GFS model version is slightly different from GEFS for this period due to different implementation times). In particular, the higher resolution model produces larger errors compared to the lower resolution model due to resolution and forecast sharpness (Figure 11 and Figure 12). It may be better to separately verify the forecast intensity and pattern (or position). The results could be different if different verification methods are applied, such as MODE (the Method for Object-Based Diagnostic Evaluation; Davis et al. 2006a, 2006b). However, the RMSEs are very similar after calibration for both the higher and lower resolution model forecasts. Meanwhile, for this one year of statistics, ABSE is reduced for both the GFS and GEFS control at all lead times.

For the ensemble forecast, RMSE and ABSE of the ensemble mean, ensemble spread and CRPS (Continuous Ranked Probability Score; Zhu and Toth 2008) have been calculated for the period of 1 March 2009 – 28 February 2010 and displayed in Figure 13. This is one year verification against CCPA for every 24-hr accumulated precipitation forecast. The results indicate that 1) the RMSE is marginally reduced (similar to the ensemble control in Figure 11) and the ABSE for ensemble mean is significantly reduced; 2) CRPS is improved; and 3) ensemble spread is increased for longer lead time forecasts. The improved spread and CRPS could be explained as a by-product of the frequency matching method. The algorithm not only matches the precipitation frequency (reducing the bias), but also adjusts the amount of precipitation forecasted by each ensemble member (adjusting the distribution). A comparison of the Brier scores between raw and calibrated forecasts is also shown in Figure 14. The Brier score is negatively oriented, which means the smaller the score value the better the results. As expected, the score is reduced after bias correction (dotted curves) for all lead times.

5. Conclusions and future plans

The frequency matching method is developed and applied to the NCEP QPF and PQPF forecasts for the first precipitation calibration since 2004. The latest version will be implemented in 2013 for finer temporal (every 6 hours out to 16 days) and spatial (1*1 degree) resolutions. The prior CDFs of the forecast and observation can be easily generated from the GFS/GEFS precipitation forecast and CCPA through applying the Kalman filter method (or decaying average). The performance of this method has been investigated with respect to one year of operational GEFS precipitation products. Results show that model bias has been effectively reduced and some skill scores have been improved in the calibrated forecasts. The good performance of the bias correction is obviously due to the fact that it can dynamically catch systematic model biases in most cases. Another attractive advantage of this method is that it saves a significant amount of both computer and human resources. Unlike other statistical post-processing methods, it is not heavily reliant on a huge amount of data for model bias training, so it takes up much less disk space on the computer systems and is able to update the model bias when a model is upgraded as well.

One important issue is the validity of the frequency matching method. The method used in this study is based on a certain knowledge of model bias information drawn from past verification statistics. Remember that as mentioned in Section 2, this method is not perfect as it is unable to make adjustments to areas that have no precipitation; therefore, this kind of dry bias can never be removed, though this is also the case with other traditional precipitation bias correction methods. Generally, model forecasts include two kinds of errors, intensity error and pattern errors. This method appears to have a positive impact on intensity error dominated cases. However, it has a neutral or negative impact on pattern error dominated cases (Figure 12), causing a poorer sampling of bias information. In this case bias is reduced at the expense of an increase in random error. Further investigation is needed to fully understand the performance of this method and to determine where and when it has a significantly positive impact and the usefulness of the calibrated products.

In this study the decaying average weight is constantly selected as 0.02 (except for the 2000-2001 application) for all lead times. Actually the decaying average weights really depend on experiments which could range from 0.01 to 0.5. In general, the weight is varied for different forecast lead times; a larger weight is good for short lead times which can catch up quick moving systems and a smaller weight is more favorable for long lead time forecasts (not shown). Therefore, choosing an optimum weight for each lead time could be a constructive way to improve the calibration system in the future. Meanwhile, the weight is varied for geographical locations and seasons. There are two improvements we are expecting to validate through future study. One is an optimum weight, which will need large samples for experiments. The weights should be a function of lead time, location and season. The second is a down-scaling process to produce a much finer resolution forecast (5km and 2.5km resolutions).

Acknowledgements:

The authors thank Drs. Bo Cui and Dingchen Hou and the other members of ensemble and post processing team at EMC/NCEP for helpful suggestions during this work. The authors also thank Mr. Binbin Zhou for useful comments. The first author gratefully acknowledges Drs. Zoltan Toth and Stephen Lord for their long time support of this calibration work. We also thank Ms. Mary Hart (NCEP) and Dr. Glenn White for editing our English.

References:

Cui, B., Z. Toth, Y. Zhu and D. Hou, 2012: Bias Correction for Global Ensemble Forecast. *Weather and Forecasting* Vol. 27 396-410

Cui, B., Y. Zhu, Z. Toth and D. Hou, 2013: Statistical Post Process for NAEFS. Summited to *Weather and Forecasting.*

Davis, C.A., B.G. Brown, and R.G. Bullock, 2006a: Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134, 1772-1784.

Davis, C.A., B.G. Brown, and R.G. Bullock, 2006b: Object-based verification of precipitation forecasts, Part II: Application to convective rain systems. *Monthly Weather Review*, 134, 1785-1795.

Demargne, J., L. Wu, S. Regonda, J. Brown, H. Lee M. He, D-J Seo, R. Hartman, M. Fresch, and Y. Zhu, 2013: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. Submitted to *Bulletin of American Meteorological Society*.

Eckel, F. A., and M. K. Walters, 1998: Calibrated Probabilistic Quantitative Precipitation Forecasts Based on the MRF Ensemble, *Weather and Forecasting*, 13, 1132-1147

Fundel, F., A. Walser, M. A. Liniger, C. Frei, and C. Appenzeller, 2009: Calibrated Precipitation Forecasts for a Limited-Area Ensemble Forecast System Using Reforecasts, *Monthly Weather Review*, 138, 176-189

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application, *Monthly Weather Review*, 134, 3209-3229

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation, *Monthly Weather Review*, 136, 2620-2632

Hou, D., M. Charles, Y. Luo, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, D. J. Seo, M. Pena and B. Cui, 2013: Climatology-Calibrated Precipitation Analysis at Fine Scales: Statistical Adjustment of STAGE IV towards CPC Gauge-Based Analysis, *Journal of Hydrometeorology* (in press).

Krzysztofowicz, R. and A. A. Sigrest, 1999: Calibration of Probabilistic Quantitative Precipitation Forecasts, *Weather and Forecasting*, 14, 427-442

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV Hourly Precipitation Analyses: Development and Applications. Preprints, *19th Conf. on Hydrology*, San Diego, CA, *Amer. Meteor. Soc.*, 1.2. [Available online at http://ams.confex.com/ams/pdfpapers/83847.pdf]

Christopher, C. P., A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, 2008: Calibration of Probabilistic Forecasts of Binary Events, *Monthly Weather Review*, 137 1142-1149

Voisin N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and Downscaling Methods for Quantitative Ensemble Precipitation Forecasts. *Weather and Forecasting*, 25, 1603-1627.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H. H. Juang, 2007: Calibration of Probabilistic Quantitative Precipitation Forecasts with an Artificial Neural Network. *Weather and Forecasting*, 22, 1287-1303.

Xie P, M. Chen M and W. Shi, 2013: CPC unified gauge analysis of global daily precipitation. (to be submitted)

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, C. H. Bishop, X. Wang, 2006: Ensemble Transform Kalman Filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus* 58A, 28-44

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 59A, 62-79

Zhu, Y., Z. Toth, E. Kalnay, and S. Tracton, 1998: Probabilistic Quantitative Precipitation Forecasts based on the NCEP Global Ensemble. Preprints, 12th Conf. on Numerical Weather Prediction, Phoenix, AZ, *Amer. Meteor. Soc.*, 286–289.

Zhu, Y. and Z. Toth, 1999: Calibration of Probabilistic Quantitative Precipitation Forecast, Preprints of the 17th AMS Conference on Weather Analysis and Forecasting 13-17 September 1999, Denver, CO, *Amer. Meteor. Soc.*

Zhu, Y. and Z. Toth, 2004: May 2004 Implementation of a QPF Bias-Correction Algorithm, [Available online at: http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html]

Zhu, Y., 2005: Ensemble forecast: A New Approach to Uncertainty and Predictability, *Advance in Atmospheric Sciences*, Vol. 22, No. 6, 781-788

Zhu, Y. and Z. Toth, 2008: Ensemble based Probabilistic Forecast Verification, Preprints, 19th Conference on Predictability and Statistics, 20-24 January 2008, New Orleans, Louisiana, *Amer. Meteor. Soc.*

Zhu, Y., D. Hou, M. Wei, R. Wobus, J. Ma, B. Cui and S. Moorthi, 2012: [Available online at: http://www.emc.ncep.noaa.gov/gmb/yzhu/html/imp/201109_imp.html]

Table 1. Total grid point counts of CONUS and each RFC at 1 degree spacing.

index	*RFC*	*count*
1	CNRFC	67
2	CBRFC	83
3	MBRFC	152
4	ABRFC	56
5	WGRFC	75
6	NCRFC	105
7	LMRFC	51
8	OHRFC	47
9	NERFC	31
10	MARFC	21
11	SERFC	61
12	NWRFC	95
Total	CONUS	844

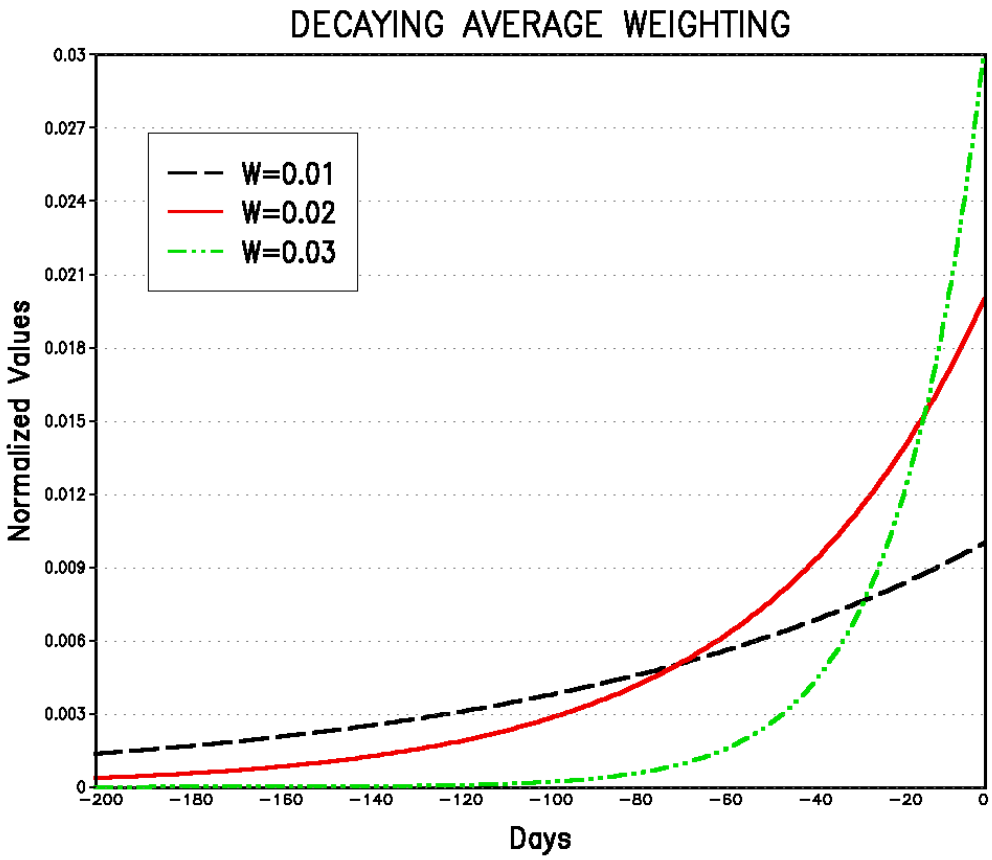


Figure 1. Decaying averaged weight as a function of preceding days (weighting function for decaying average of preceding days). The dashed curve denotes a weight beginning with a maximum value of 0.01 at day 0. The solid curve denotes a weight beginning with a maximum value of 0.02 and the dotted curve denotes a weight beginning with a maximum value of 0.03 at day 0. All curves gradually approach zero depending how far away the preceding days are from day 0. The larger the weight at day 0 the faster the decaying speed, which indicates greater weight on the most recent data and less on the oldest data.

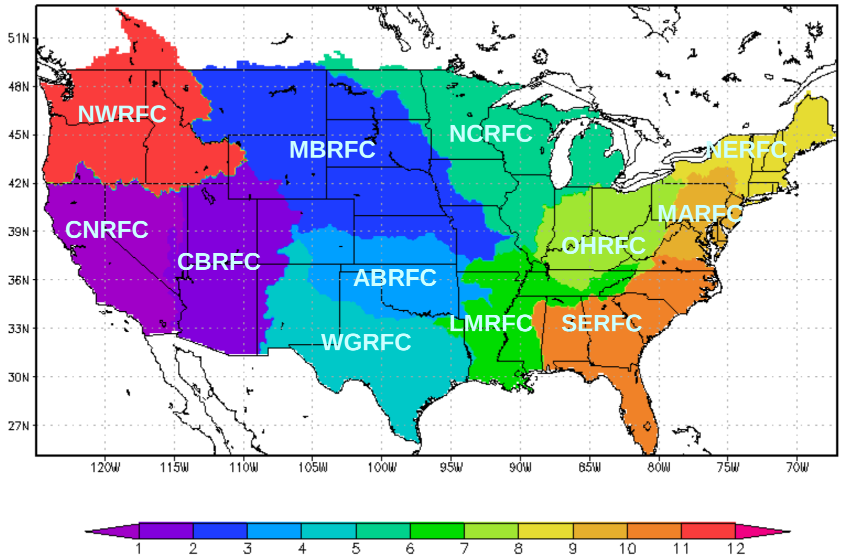


Figure 2. The domains of the twelve River Forecast Centers (RFC). Note that the CCPA analysis covers the twelve RFCs over the Contiguous United States (CONUS).

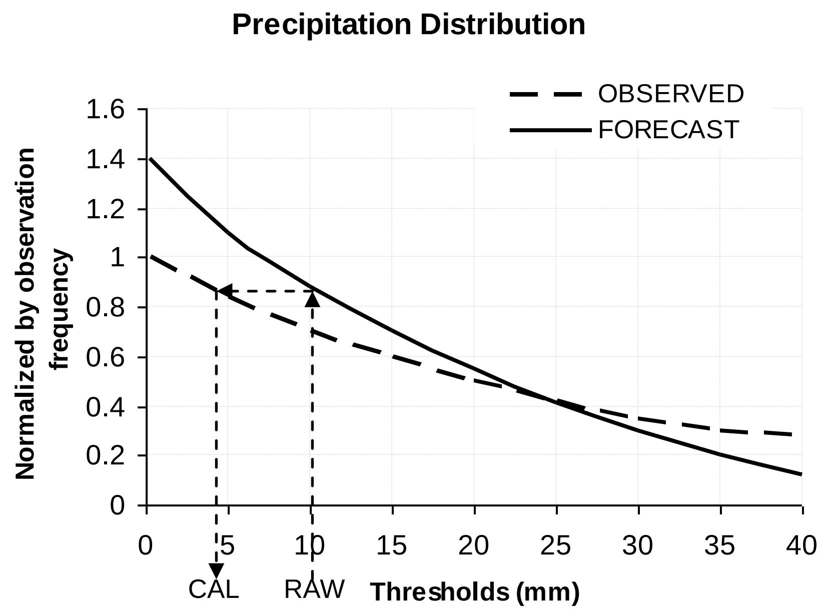
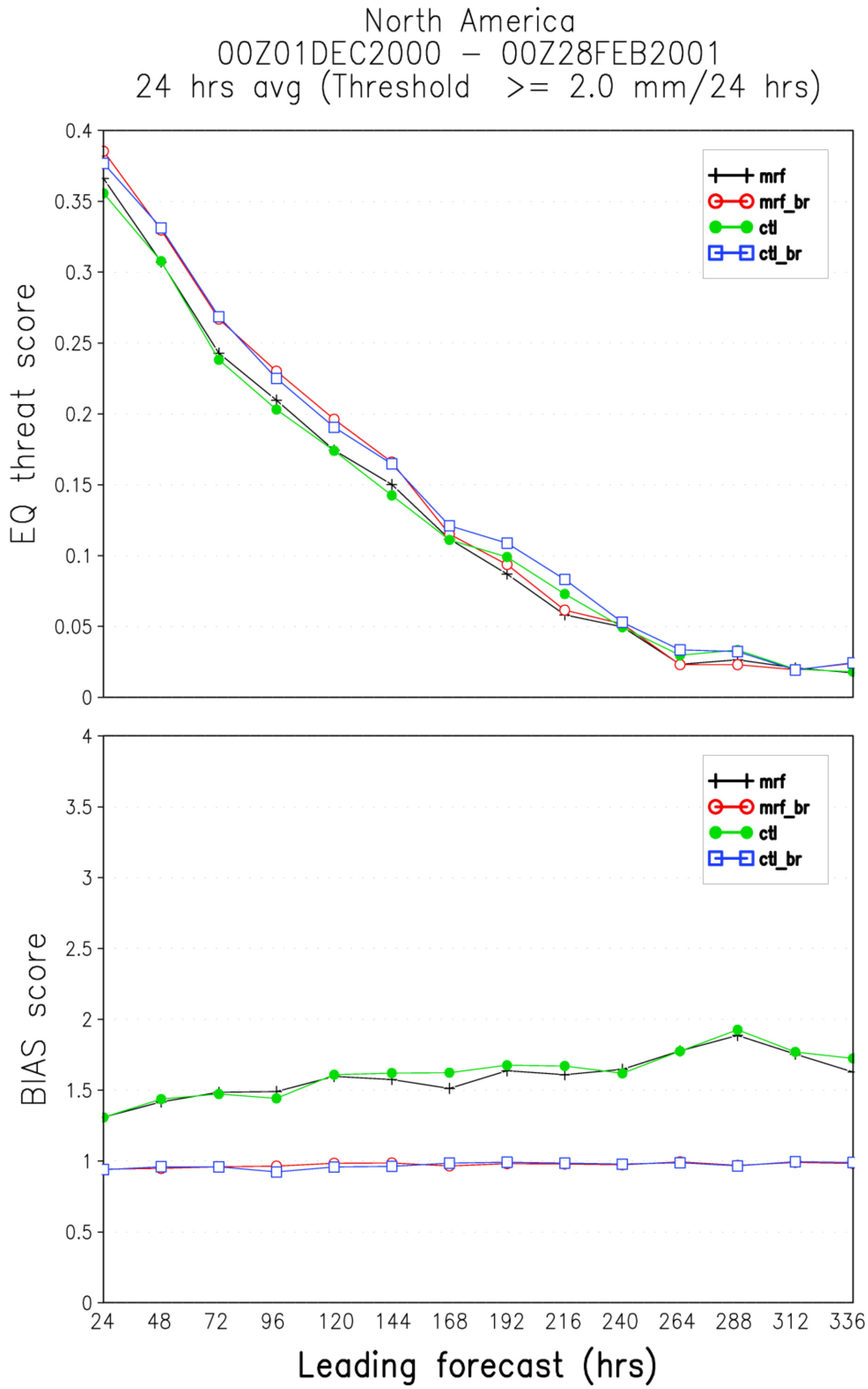


Figure 3. Schematic of the frequency matching algorithm demonstrated as precipitation distributions normalized by observation frequency varying with threshold. The dashed line is for observed precipitation and the solid line is for forecast precipitation. See text for details.

1 4 (a)



North America
00Z01DEC2000 – 00Z28FEB2001
24 hrs avg (Threshold >= 2.0 mm/24 hrs)
EQ threat score
BIAS score
mrf
mrf_br
ctl
ctl_br
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
4
3.5
3
2.5
2
1.5
1
0.5
0
24 48 72 96 120 144 168 192 216 240 264 288 312 336
Leading forecast (hrs)

2

4(b)

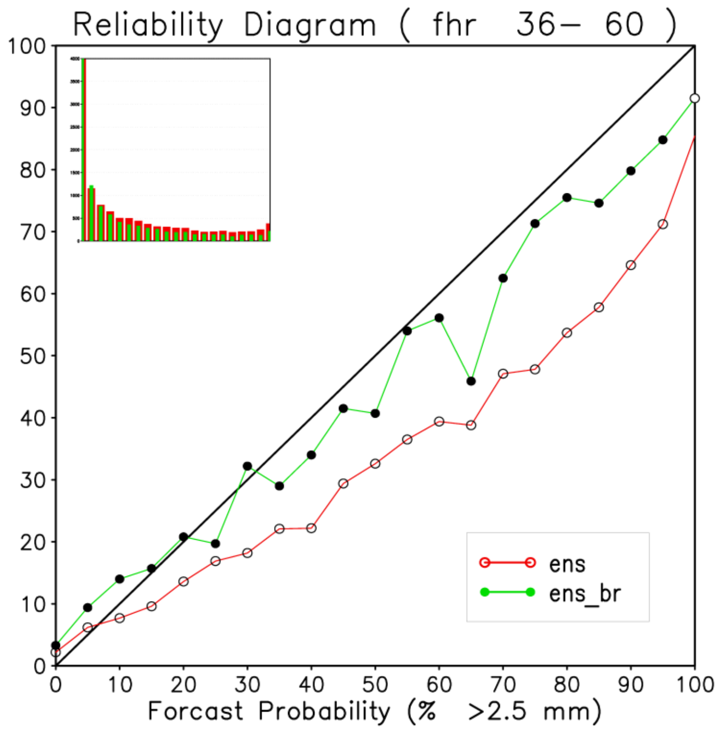


Figure 4. Examples from the 2004 implementation. Results were selected for the period of 1 December 2000 - 28 Feburary 2001. (a) Averaged Equitable Threat Score (ETS) and bias scores of the raw GFS (mrf) and GEFS control (ctl) forecasts and their calibrated forecasts at a threshold of 2.0mm/day. (b) Reliability of the 2.5mm/day GEFS raw (ens, red) and calibrated (ens_br, blue) forecasts at 36-60 hour lead time. The inset histogram denotes the frequency of forecast usage of each probability bin.

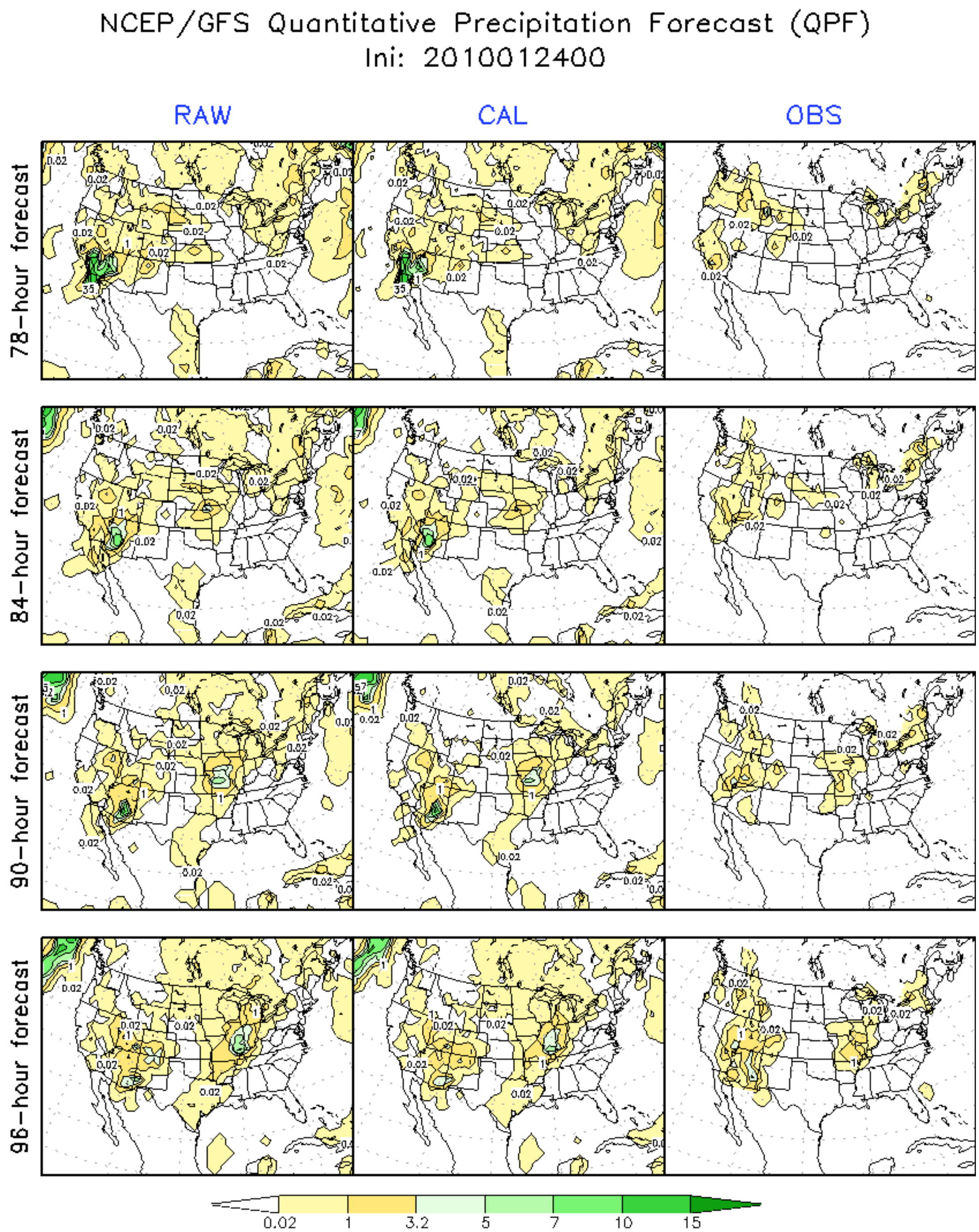


Figure 5. Comparisons of 6 hourly accumulated precipitation (mm) initialized at 0000 UTC 24 January 2010 from the raw GFS forecasts (left column) and calibrated forecasts (middle column) against the CCPA product (right column) that are valid at corresponding time periods.

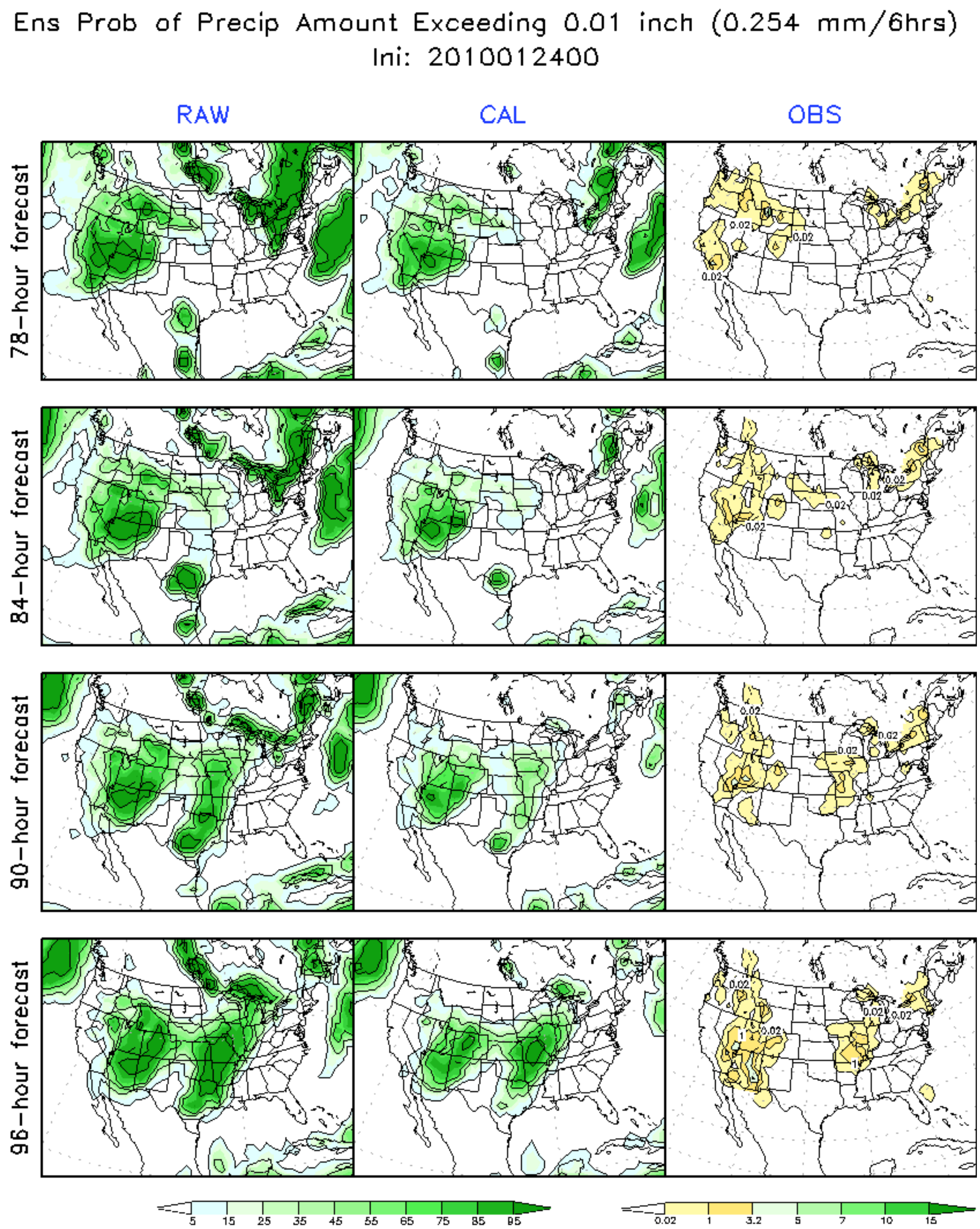
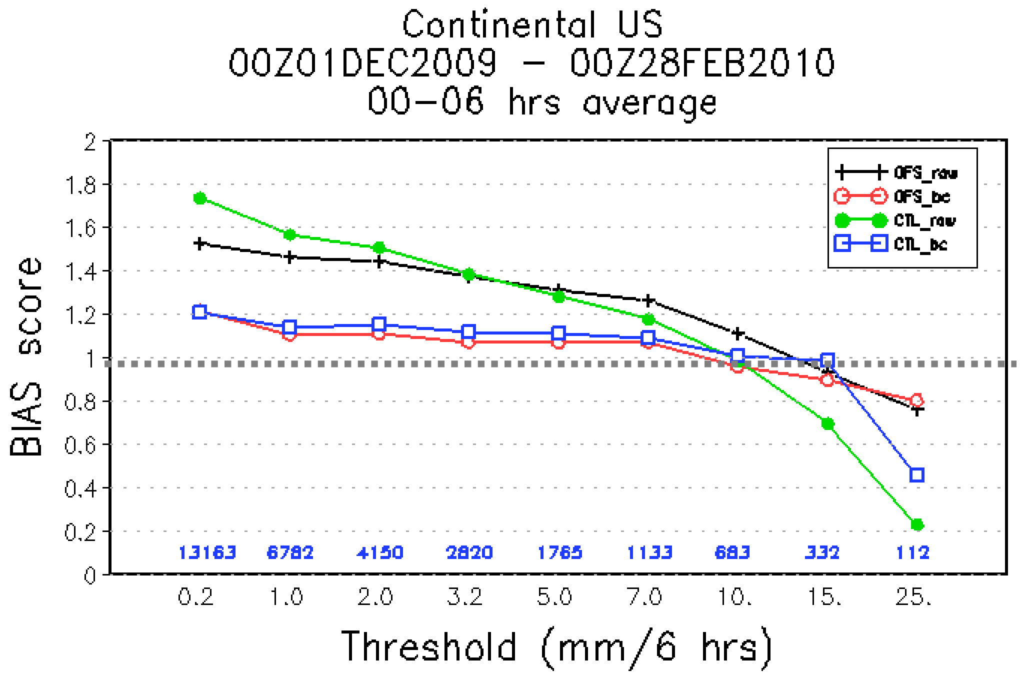


Figure 6. GEFS probabilities (left), calibrated probabilities (middle) of the 6-hr precipitation amount exceeding 0.01 inch initialized at 00 UTC 24 January 2010, and CCPA precipitation estimates (right) for 6-h precipitation that are valid at the corresponding time periods.

(a)



(b)

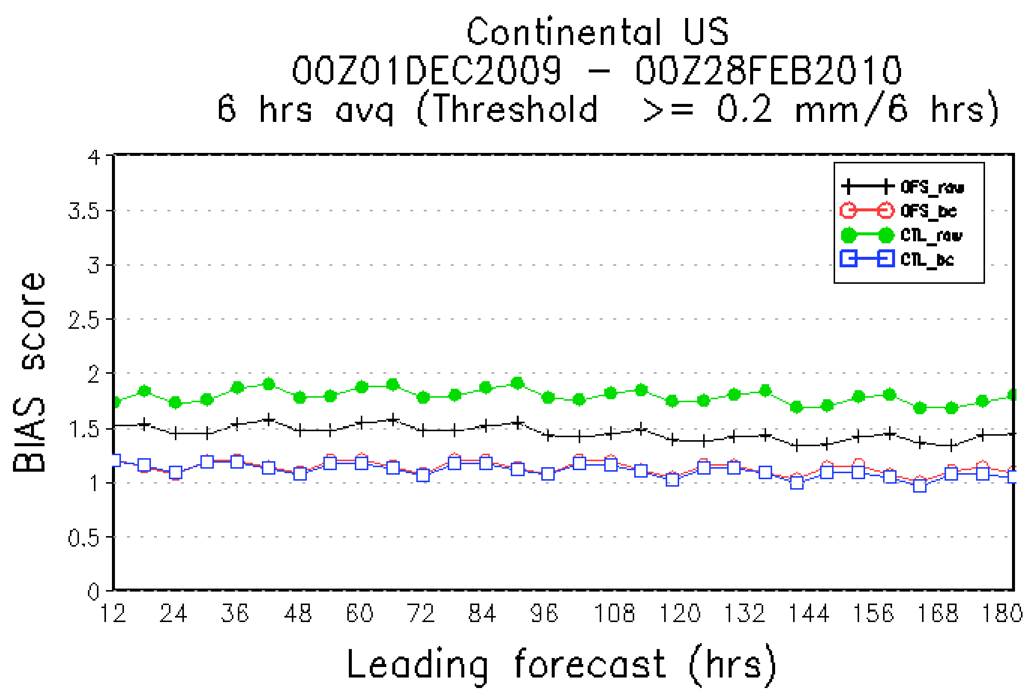
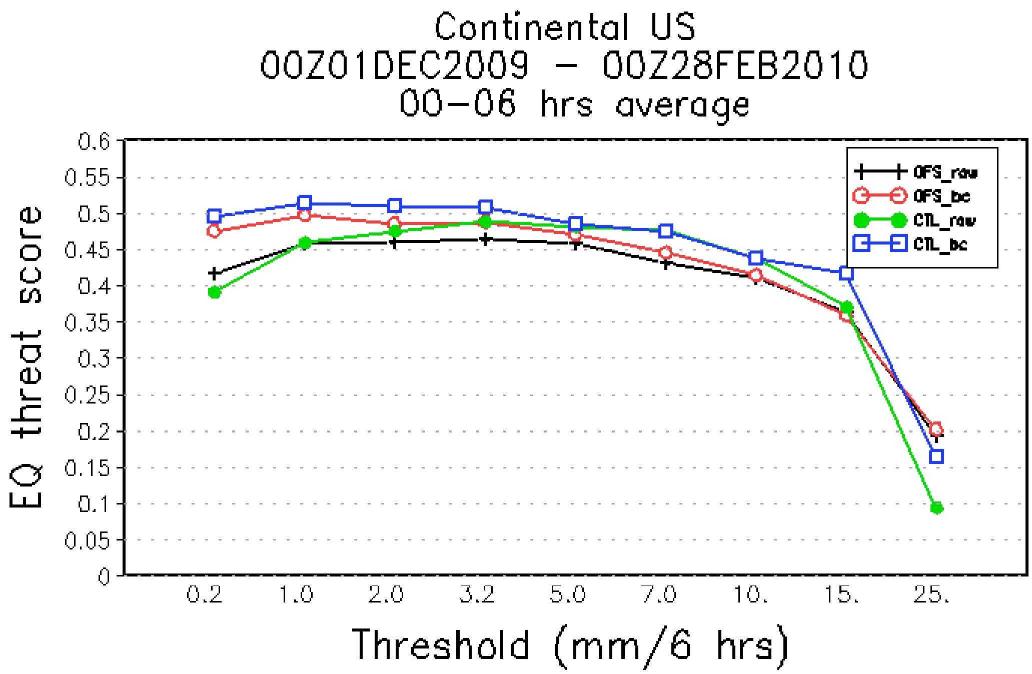


Figure 7. Bias scores of raw forecasts (GFS-black; GEFS/CTL-green) and calibrated forecasts (GFS-red; GEFS/CTL-blue) with increasing lead times for 6-h precipitation averaged between 1 December 2009 and 28 February 2010 (a) as a function of threshold

and (b) at a 0.2mm threshold. The numbers in the plot above x-axis are total number of cases verified.

(a)



(b)

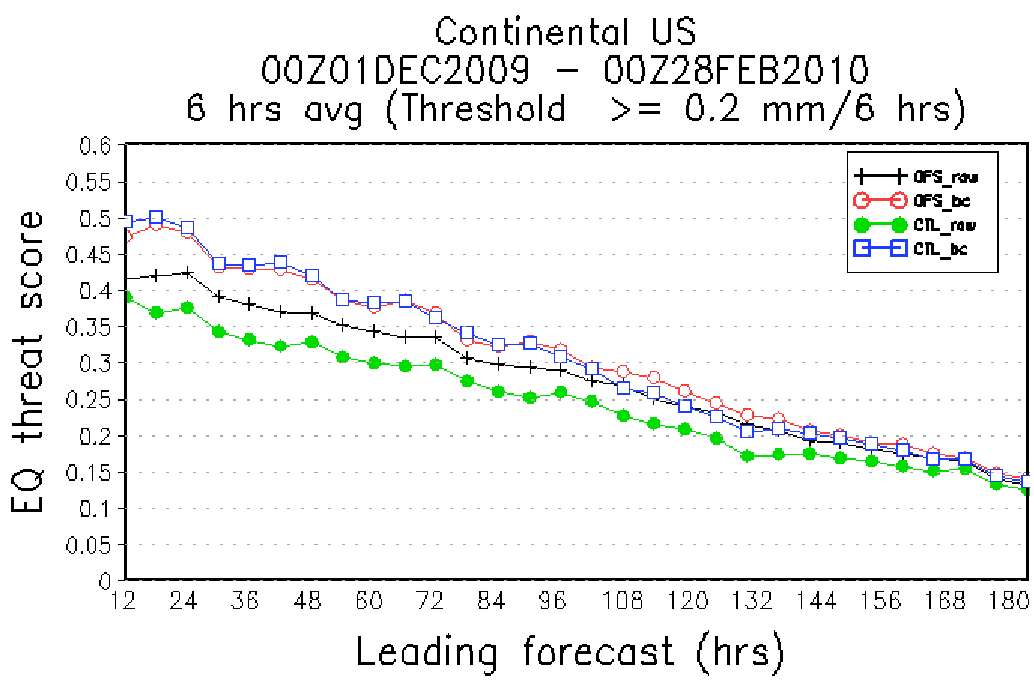


Figure 8. Same as Figure 7, but for Equitable Threat Scores (ETS).

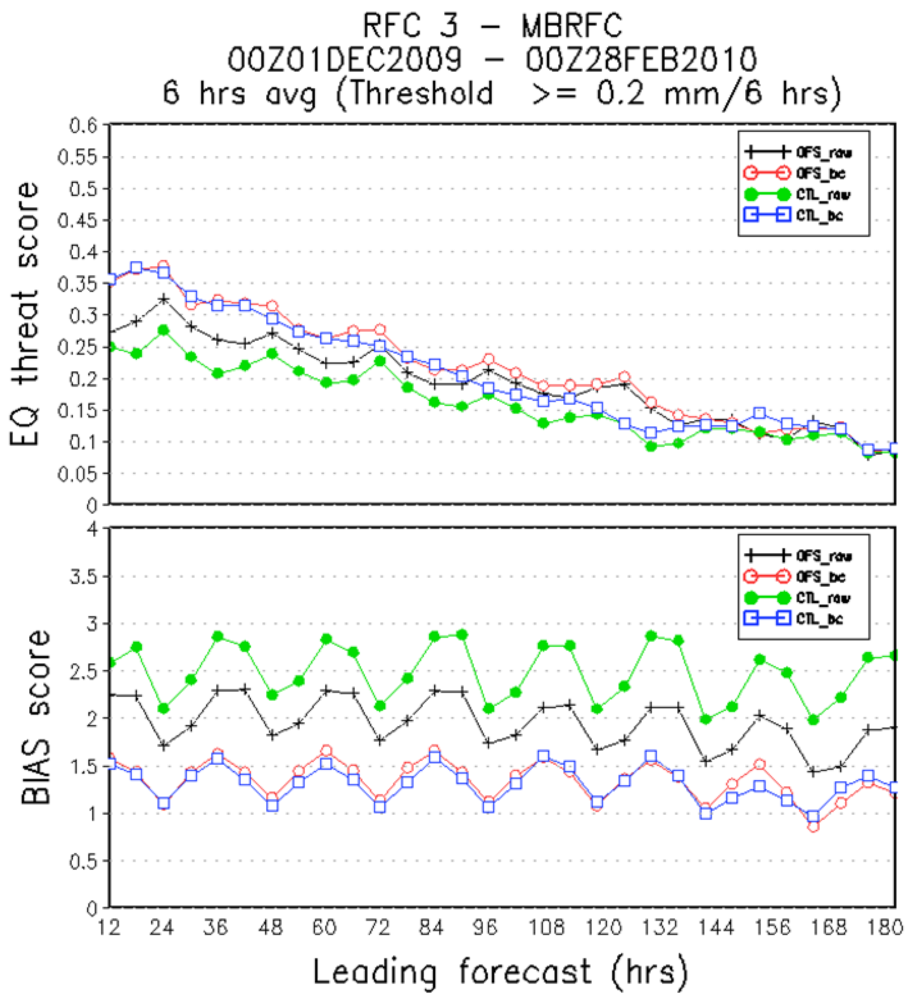
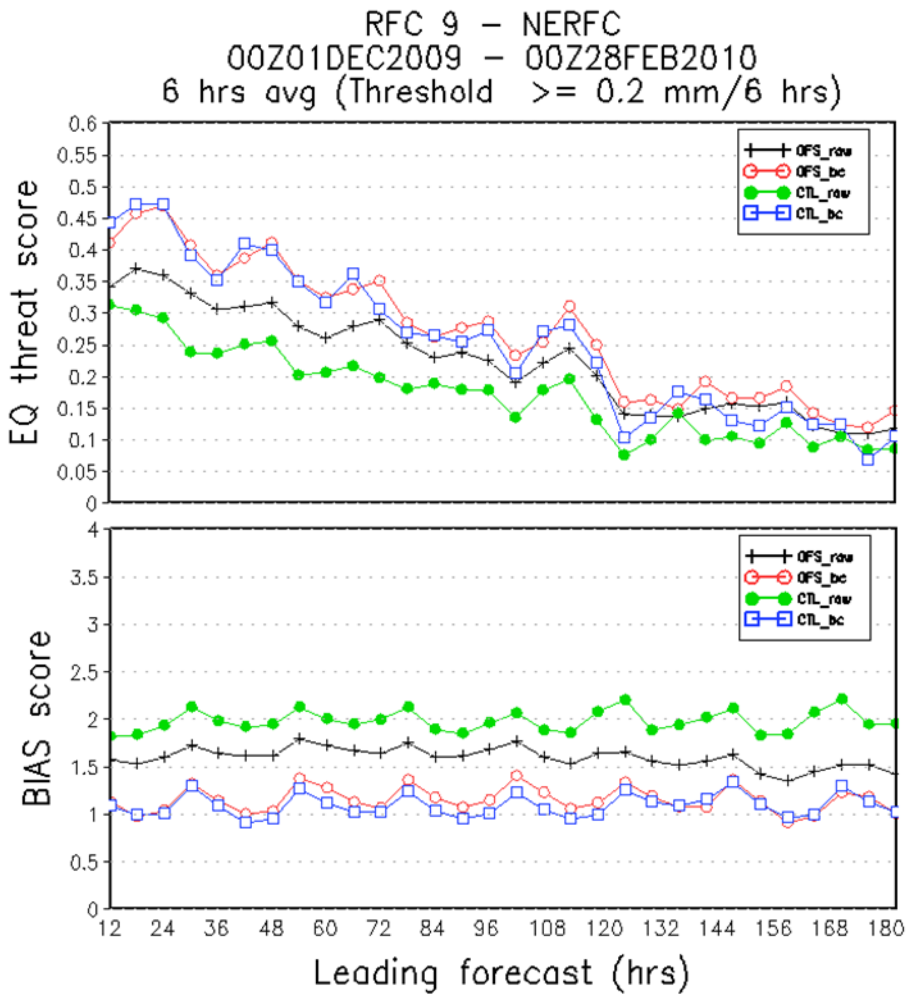


Figure 9. Comparison of raw forecasts (GFS-black; GEFS/CTL-green) and calibrated forecasts (GFS-red; GEFS/CTL-blue) with increasing lead times for 6-h precipitation averaged between 1 December 2009 and 28 February 2010 and analyzed for the MBRFC region for (a) Equitable Threat Score and (b) Bias score at a 0.2mm threshold.



1
2 Figure 10. Same as Figure 9, but for the NERFC region.
3

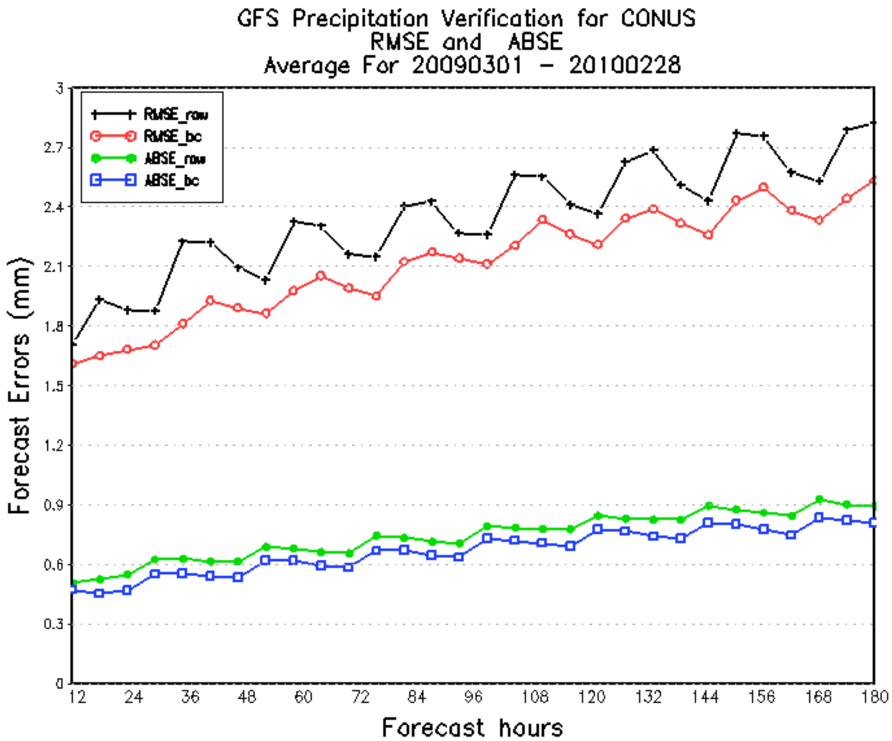


Figure 11. RMSE with increasing lead times for 6-h precipitation from the GFS high resolution raw forecasts (black) and calibrated forecasts (red); and ABSE with increasing lead times for 6-h precipitation from the GFS raw forecasts (green) and calibrated forecasts (blue).

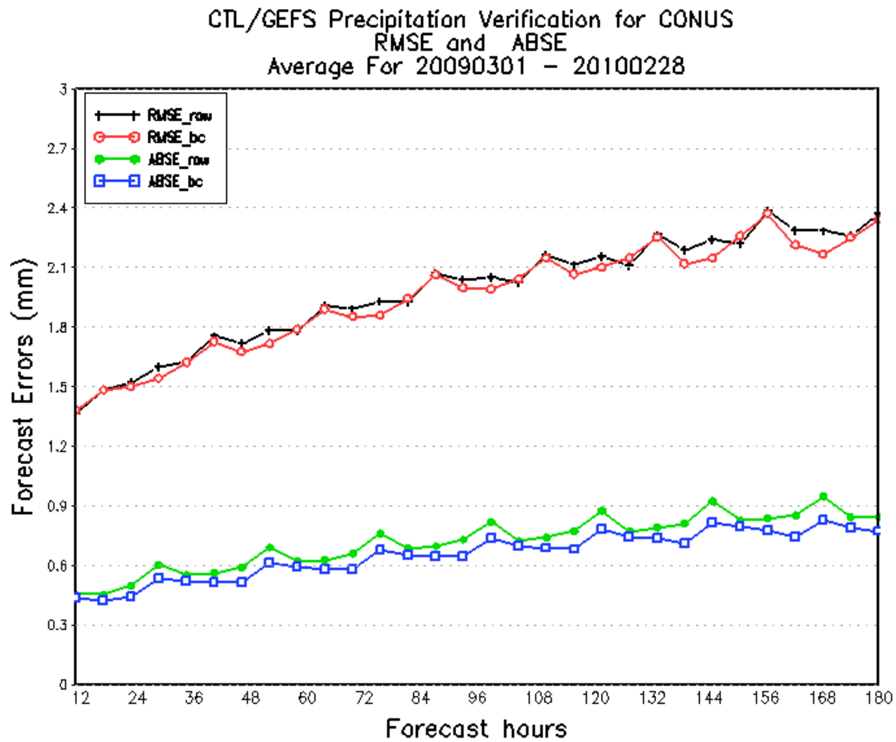


Figure 12. Same as Figure 11, but for GEFS/CTL forecasts.

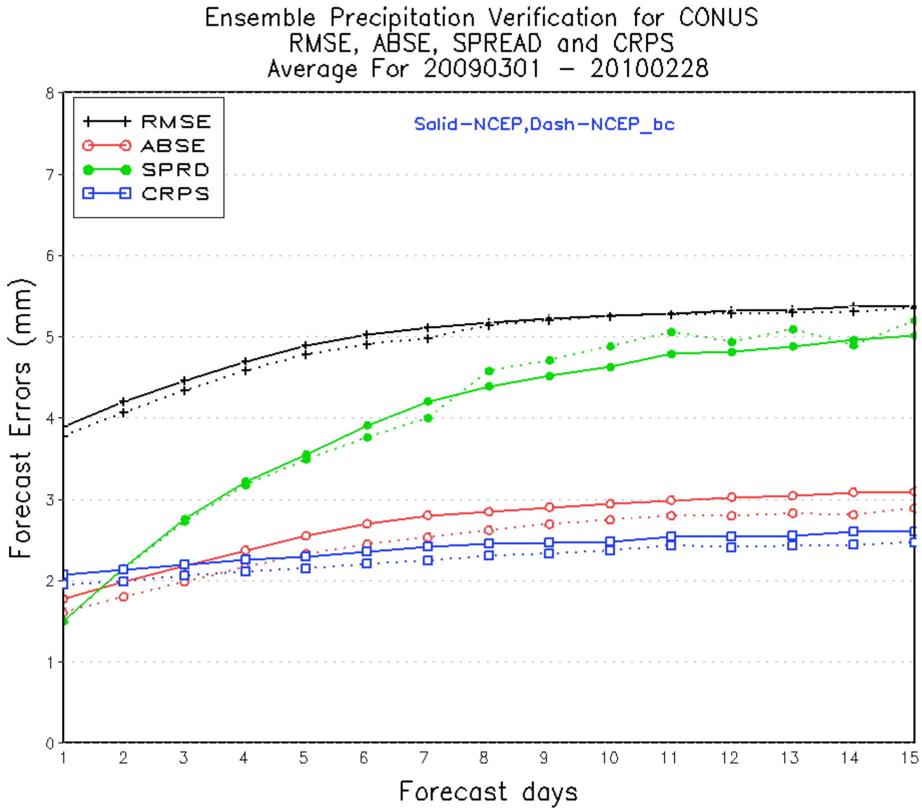


Figure 13. The RMSE (black), ABSE(red), SPRD (green) and CRPS (blue) with increasing lead times for 24-h precipitation from the GEFS ensemble mean (RMSE and ABSE) and ensemble members (spread and CRPS) raw forecasts (solid) and calibrated forecasts (dotted).

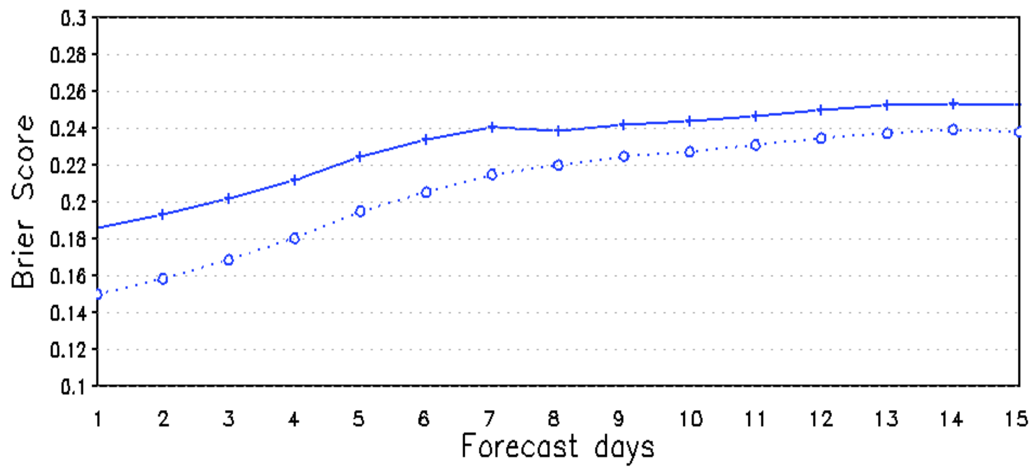


Figure 14. The Brier score at a 0.2mm threshold with increasing lead times for 24-h precipitation from the GEFS ensemble mean raw forecast (solid) and calibrated forecasts (dotted).