Exploring ensemble forecast calibration issues using reforecast data sets

Thomas M. Hamill[(1)] **and Renate Hagedorn**[(2)]

[(1)] *NOAA Earth System Research Lab, Boulder, Colorado, USA 80303*
Tom.hamill@noaa.gov ; http://esrl.noaa.gov/psd/people/tom.hamill

[(2)] *ECMWF, Reading, Berkshire, England RG2 9AX*
renate.hagedorn@ecmwf.int ; http://www.ecmwf.int

Abstract

Issues related to the calibration of probabilistic 2-meter temperature and 12-hour accumulated precipitation forecasts were explored using reforecast data sets from a 1998, T62 version of the NCEP Global Forecast System (GFS) and a 2005, T255 version of the ECMWF model (version 29r2). The main results were: (1) ECMWF "raw" forecasts (probabilities from the ensemble relative frequency) were not particularly skillful when measured using skill scores that avoid attributing skill to variations in climatology. GFS raw forecasts were even worse. (2) GFS calibrated forecasts were typically more skillful than ECMWF raw forecasts, despite the use of a low-resolution, old model. (3) Despite using a newer, better model, ECMWF calibrated forecasts were also greatly improved from the reforecast calibration and exceeded the skill of GFS calibrated forecasts. (4) For temperature, most of the improvement was a result of the correction of gross model bias. (5) Small training data sets were adequate for light precipitation events and short-range temperature forecasts, but the large training data sets afforded by reforecasts were particularly helpful for improving longer-lead temperature forecasts and precipitation forecasts at high thresholds.

1. Introduction

A series of recent articles have introduced the use of reforecasts for the calibration of a variety of probabilistic weather-climate forecast problems, from week-2 forecasts (Hamill et al. 2004; Whitaker et al. 2006) to short-range precipitation forecast calibration (Hamill et al. 2006, Hamill and Whitaker 2006) to forecasts of approximately normally distributed fields such as geopotential and temperature (Wilks and Hamill 2007, Hamill and Whitaker 2007) to streamflow predictions (Clark and Hay 2004). The reforecast data set used was a reduced resolution, T62, 28-level, circa-1998 version of the Global Forecast System (GFS) from the National Centers for Environmental Prediction. Fifteen-member forecasts were available to 15 days lead for every day from 1979 to current. With a stable data assimilation and forecast system, the systematic errors of the forecast could be readily diagnosed and corrected. Calibration using reforecasts were able to adjust the forecasts to achieve substantial improvements in the skill and reliability of the forecasts, commonly to levels competitive with or exceeding those achieved by current-generation ensemble forecast systems without calibration.

The GFS model version used in these reforecast studies is now ~10 years out of date, and the reforecasts and real-time forecasts from it are run at a resolution far less than that used currently at operational weather prediction centers. Arguably, the dramatic improvement from the use of reforecasts may be due in large part to the substantial deficiencies of this forecast modeling system. Would the calibration of a modern-generation ensemble forecast system similarly benefit from the use of reforecasts?

Recently, ECMWF produced a limited reforecast data set with a model version that was operational in the last half of 2005. They produced a 15-member reforecast once weekly from 1 September to 1 December, over a 20-year period from 1982 to 2001. Each forecast was run to 10 days lead using a T255, 40-level version of the ECMWF global forecast model. During the past decade, ECMWF global ensemble forecasts have consistently been the most skillful of those produced at any national center (e.g., Buizza et al. 2005), so calibration experiments with this model may be representative of the results that other centers may obtain with reforecasts over the next 5 years or so.

This data set allows us to ask and answer questions about reforecasts that were not possible with only the GFS data set. Some relevant questions include: (1) how does an old GFS model forecast that has been statistically adjusted with reforecasts compare with a probabilistic forecast estimated directly from the state-of-the-art ECMWF ensemble forecast system? (2) If this state-of-the-art system could also be calibrated using its own reforecast, would there still be substantial benefits from the calibration, or would they be much diminished relative to the improvement obtained with the older GFS forecast model? (3) Is a calibrated, multi-model combination more skillful than that provided solely by the ECMWF system? (4) How much of the benefits of calibration in a state-of-the-art model can be obtained using only a short time series of past forecasts and observations? Large improvements from calibration using a short training data set would be a particularly encouraging result, for an extensive set of reforecasts is computationally expensive to produce. The significant computational expense of a long reforecast data must be justified by very large improvements in forecast skill from its usage, improvements larger than would be obtained by, say, increasing the model resolution of improving the realism of radiation calculations.

This extended abstract summarizes the work of two recently submitted journal articles, Hagedorn et al. (2007) and Hamill et al. (2007). The reader is referred to these manuscripts for additional detail on the calibration issues and the underlying methods discussed here.

2. Verification and reforecast data sets

2.1. ECMWF forecast data

The ECMWF reforecast data set consists of a 15-member ensemble reforecast computed once weekly from 0000 UTC initial conditions for the initial dates of 1 September to 1 December. The years covered in the reforecast data set were from 1982 to 2001. The model cycle 29r2 was used, which was a spectral model with triangular truncation at wavenumber 255 (T255) and 40 vertical levels using a sigma coordinate system. Each forecast was run to 10 days lead. The 15 forecasts consisted of an ERA-40 reanalysis initial condition (Uppala et al. 2005) plus 14 perturbed forecasts generated using the singular-vector methodology (Molteni et al. 1996; Barkmeijer et al. 1998, 1999). While data was available to cover the entire globe, for this study the model forecasts were extracted on a 1-degree grid from 135 to 45 degrees west longitude and 15 to 75 degrees north latitude. This covered the conterminous US (CONUS) and most of Canada. For temperature observations, described below, the 1-degree gridded forecasts were bilinearly interpolated at 0000 UTC and 1200 UTC to the observation locations. For 12-h accumulated precipitation, the forecasts were also interpolated to the 32-km precipitation analysis grid in the CONUS. The precipitation data set is also described below.

In addition, the operational ECMWF 0000 UTC forecasts were extracted in 2005 for every day from 1 July to 1 December. These forecasts used the same model version that was used to produce the reforecasts, though the initial analyses were provided by the operational 4-dimensional variational analysis system (Mahfouf and Rabier 2000) rather than the 3-dimensional variational analysis system used in ERA-40. The 2005 daily data permitted experiments comparing calibration using a short training data set of prior forecasts with calibration using the reforecasts.

2.2. GFS forecast data

The GFS reforecast data set, more completely described in Hamill et al. (2006), was also utilized here. The underlying forecast model was a T62, 28 sigma-level, circa-1998 version of the GFS. Fifteen-member forecasts are available to 15 days lead for every day from 1979 to current. Forecasts were started from 0000 UTC initial conditions, and forecast information was archived on a 2.5-degree global grid. GFS forecast accumulated precipitation was also bi-linearly interpolated to station locations and the precipitation analysis grid at 12-hourly intervals. For most of the experiments to be described here, the GFS reforecasts were extracted from 1982-2001 at the weekly dates of the ECMWF reforecast, to facilitate comparison. Daily GFS forecast data was also extracted for 1 July to 1 December 2005.

2.3. Two-meter temperature observations

0000 UTC and 1200 UTC 2-meter temperature observations were extracted from the National Center for Atmospheric Research (NCAR) data set DS472.0. Only observations that were within the domain of the ECMWF reforecast data set as described above were used. Additionally, only the stations that had 96 percent or more of the observations present over the 20-year period were utilized. A plot of these 439 station locations is provided in Fig. 1.

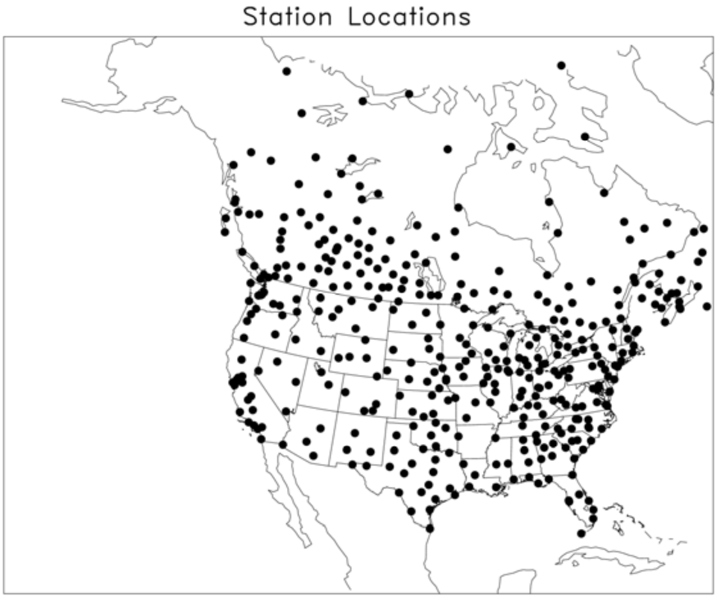

Figure 1: Station locations for the verification of 2-meter temperature forecasts.

2.4. Precipitation analyses

The reference for verification and training was the North American Regional Reanalysis (NARR) precipitation analysis (Mesinger et al. 2006), archived on a ~32-km Lambert-conformal grid covering North America and adjacent coastal waters. Only data over the conterminous US (CONUS) was used, and precipitation was accumulated over 12-hourly periods ending at 0000 UTC and 1200 UTC.

3. Calibration and validation methods

3.1. Calibration of 2-meter temperatures with non-homogeneous Gaussian regression

Non-homogeneous Gaussian regression (NGR; Gneiting et al. 2005, Wilks and Hamill 2007) was used for the calibration of 2-meter temperature forecasts. NGR is an extension to conventional linear regression. It was assumed that there may be information about the forecast uncertainty provided by the ensemble sample variance (Whitaker and Loughe 1998). However, due to the limited number of members and other system errors, the ensemble sample variance may not properly estimate by itself the forecast uncertainty. Accordingly, the regression variance was allowed to be non-homogeneous (not the same for all values of the predictor), unlike linear regression. In this implementation of NGR, the mean forecast temperature and sample variance interpolated to the station location were predictors, and observed 2-meter temperature at station locations were the predictands. We assumed that stations had particular regional forecast biases sometimes distinct from those at nearby stations. Hence, the training did not composite the data, i.e., the fitted parameters at Atlanta were determined only from Atlanta forecasts and not from a broader sample of locations around and including Atlanta.

To describe NGR more formally, let $\sim\mathcal{N}(\alpha, \beta)$ denote that a random variable has a Gaussian distribution with mean α and variance β. Let $\bar{x}_{ens}$ denote the interpolated ensemble mean and s^2_{ens} denote the ensemble sample variance. Then NGR estimated regression coefficients *a*, *b*, *c*, *d* so as to fit $\mathcal{N}(a+b\,\bar{x}_{ens},\, c+d\,s^2_{ens})$. Following Gneiting et al. (2005), the four coefficients were fit iteratively to minimize the continuous ranked probability score (e.g., Wilks 2006).

In all temperature calibration experiments using the weekly reforecast data, cross validation was utilized in the regression analysis. The year being forecast was excluded from the training data, e.g., 1983 forecasts were trained with 1982 and 1984-2001 data. Also, because biases can change with the seasons, the full September-December data was not used as training data. Rather, only the 5 weeks centered on the date of interest were used, e.g., when training for 15 September, the training data was comprised of 1, 8, 15, 22, and 29 September forecasts. For dates at the beginning and end of the reforecast, a non-centered training data set was used; for example, the training dates for 1 September were 1, 8, and 15 September. Unless otherwise noted, the GFS reforecast data was sub-sampled to the same weekly dates of the ECMWF training data set.

The methodology for application of NGR to multi-model forecasts is somewhat more complex; please see Hagedorn et al. (2007) for details.

3.2. Precipitation forecast calibration with logistic regression

Logistic regression analysis (e.g., Agresti 2002, Chapter 5) will be used as the general method of forecast calibration. Non-homogeneous Gaussian regression techniques (Hagedorn et al. 2007) were not useful for precipitation, where forecast distributions are usually non-Gaussian. Here, logistic regression was chosen because: (1) it was a standard method, with readily understood characteristics and algorithms available from off-the-shelf software, (2) it permitted unusual predictors and variable sample weights to be incorporated readily, and (3) relative to methods like the analog technique, the logistic regression was expected to perform better when sample size was relatively limited. This may be helpful here since the ECMWF reforecast data consisted of once-weekly reforecasts, more infrequent than the daily GFS reforecasts used in prior studies. One disadvantage of logistic regression is that output provides probabilities for one threshold, not a full probability density function

Given an unknown analyzed amount *O* (the predictand), the precipitation threshold *T*, and model-forecast predictors $\bar{x}^f$ (the forecast mean) and σ^f (the forecast spread), the logistic regression takes the form

$$P(O>T) = 1.0 - \frac{1.0}{1.0 + \exp\left\{\beta_0 + \beta_1\left(\bar{x}^f\right)^{0.25} + \beta_2\left(\sigma^f\right)^{0.25}\right\}}. \tag{1}$$

Here, the predictors used power-transformed variables. Previously, Sloughter et al. (2007) used a 1/3-power transformation in their precipitation calibration, and Hamill and Whitaker (2006) used a ½ power transformation in the logistic regression of daily precipitation forecasts. The logistic regression algorithm used in this study allowed for training samples to be weighted individually. We chose to apply larger weights for the heavier forecast precipitation events, which improved the calibration of these events slightly. See Hamill et al. (2007) for more details.

For the precipitation forecasts considered here, a robust training data set will be shown to be crucial; heavy or even moderate precipitation may be a rare event at many locations, and a modest number of samples with other heavy precipitation events may be needed to generate trustworthy regression coefficients.

Figure 2 illustrates the potential benefits of an enlarging the training data set. Here, a logistic regression analysis was run using eq. (2). For a given forecast lead and a given grid point, the reforecast data at this lead and at this grid point for all other years and all dates were utilized as training data, 19 years × 14 dates = 266 samples. The precipitation analysis is shown in Fig. 2a. Despite the comparatively smooth ensemble-mean forecast (Fig. 2b), the subsequent forecast of probabilities from the logistic regression analysis (Fig. 2c) had more spatial structure than was warranted. After enlarging the training sample size by adding data from locations that had similar observed climatologies and repeating the logistic regression analysis (Fig. 2d, now using 11 times more samples), the probability forecasts had a much smoother spatial structure.

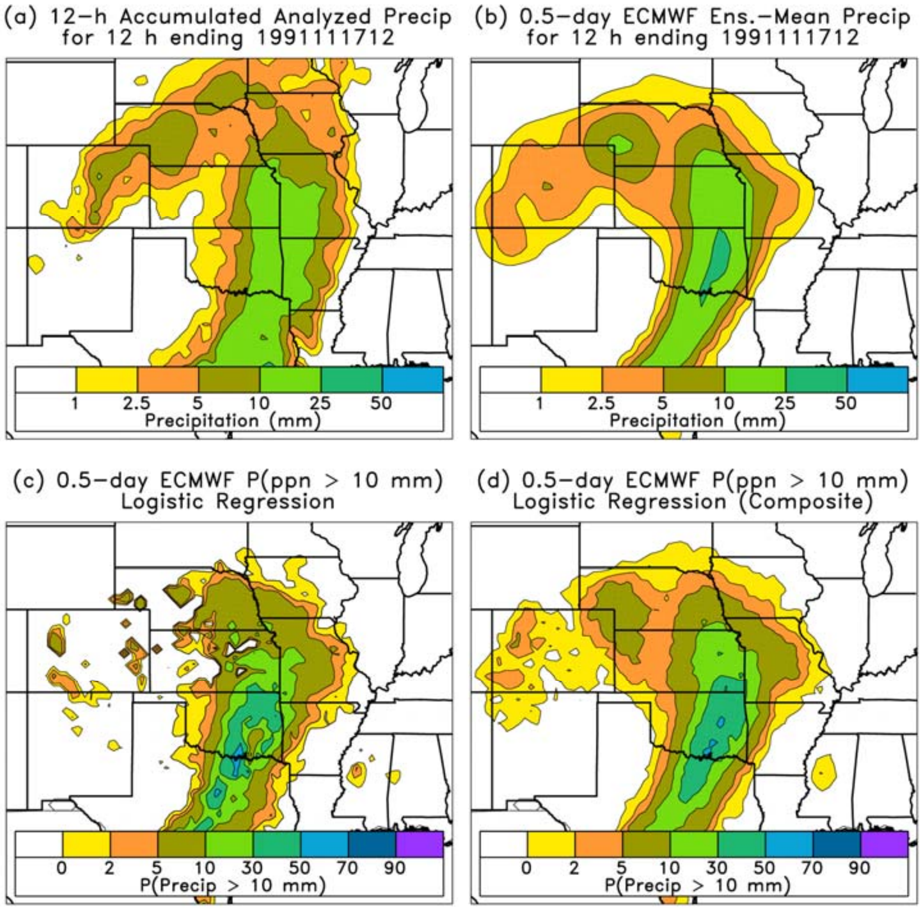
(a) 12–h Accumulated Analyzed Precip for 12 h ending 1991111712
(b) 0.5–day ECMWF Ens.–Mean Precip for 12 h ending 1991111712
1
2.5
5
10
25
50
Precipitation (mm)
(c) 0.5–day ECMWF P(ppn > 10 mm) Logistic Regression
(d) 0.5–day ECMWF P(ppn > 10 mm) Logistic Regression (Composite)
0
2
5
10
30
50
70
90
P(Precip > 10 mm)

Figure 2(a) NARR precipitation analysis of 12-h accumulated precipitation for the 12-hourly period ending 1200 UTC 17 November 1991. (b) 0-12 hour ECMWF ensemble-mean forecast of accumulated precipitation. (c) Probability of greater than 10 mm precipitation in this period using a logistic regression where each grid point's data is treated independently. (d) As in (c), but where the logistic regression training data includes forecasts and observations from 10 other locations that have similar observed precipitation climatologies for this day of the year.

The 10 additional locations used to supplement the logistic regression training data set used grid points that had a similar cumulative density function of observed precipitation but were spaced somewhat apart from each other to preserve some statistical independence of the samples. See Hamill et al. (2007) for more details.

3.3. Validation procedures

Reliability characteristics of the probabilistic temperature forecasts were diagnosed with rank histograms (Hamill 2001). When generating rank histograms for the "raw" unmodified, forecasts, random, normally distributed noise with a magnitude of 1.5 C was added to each member to account observation and representativeness error.

Forecast skill of temperature forecasts was evaluated with a revised version of the continuous ranked probability skill score (*CRPSS*) that followed the method described in Hamill and Whitaker (2007). As noted in Hamill and Juras (2006), the conventional method of calculating many verification metrics, including the *CRPSS*, can provide a misleadingly optimistic assessment of the skill if the climatological uncertainty varies among the samples. The verification metric may diagnose positive skill that can be attributed to a difference in the climatologies amongst samples rather than any inherent forecast skill. To ameliorate this, we followed the specific method outlined in Hamill and Whitaker (2007). The overall forecast sample was divided into subgroups where the climatological uncertainty was approximately homogeneous. The *CRPSS* was then determined for each subgroup, and the final *CRPSS* was determined as a weighted average of the subgroups' *CRPSS*. Here, there were *NC=8* subgroups, with a more narrow range of climatological uncertainty in each subgroup, and equal numbers of samples assigned to each subgroup. Let $\overline{CRPS}^{f}(s)$ denote the average forecast continuous ranked probability score (*CRPS*; Wilks 2006) for the *s*th subgroup, and $\overline{CRPS}^{c}(s)$ denote the average *CRPS* of the climatological reference forecast for this subgroup. Then the overall *CRPSS* was calculated as

$$CRPSS = \frac{1}{NC}\sum_{s=1}^{NC}\left(1-\frac{\overline{CRPS}^{f}(s)}{\overline{CRPS}^{c}(s)}\right) \tag{2}$$

The climatological mean and standard deviation were calculated using 5 weeks of centered data. For more details on the calculation of the alternative formulation of the *CRPSS*, please see Hamill and Whitaker (2007). Confidence intervals for assessing the statistical significance of differences between forecasts was done following the block bootstrap procedure outlined in Hamill (1999). Separate case days were treated as independent blocks of data. In general, the 5th and 95th percentiles of the resampled distribution were very narrow, ~0.02 skill units or less, and so were not plotted.

For precipitation verification, some enhancements to the standard reliability diagram (e.g., Wilks 2006, Chapter 7) were utilized here. Because high probability forecasts of heavy precipitation amounts were issued very infrequently, inset histograms for the frequency of usage were plotted on a log-10 scale, providing a better visualization of the distribution in the tails. Also, 5% and 95% confidence intervals were placed on the reliability curves, with the confidence intervals again estimated with a block. Another modification to the standard reliability diagram was the inclusion of a frequency of usage of the climatological probabilities for all forecast samples, plotted as a solid line over top of the forecast frequency of usage.

For the Brier Skill Score (*BSS*), calculations followed a procedure analogous to that for the *CRPSS* above. Again, see Hamill et al. (2007) for details.

4. Two-meter temperature forecast results

4.1. 20-year weekly training data

Figure 3 provides rank histograms for the ECMWF and GFS reforecasts. For the raw forecasts the common U shape was more pronounced at the short leads and slightly more pronounced for GFS forecasts than ECMWF forecasts. After calibration with NGR, the rank histograms were much flatter, though there still was some slight excess of population of the lowest rank.

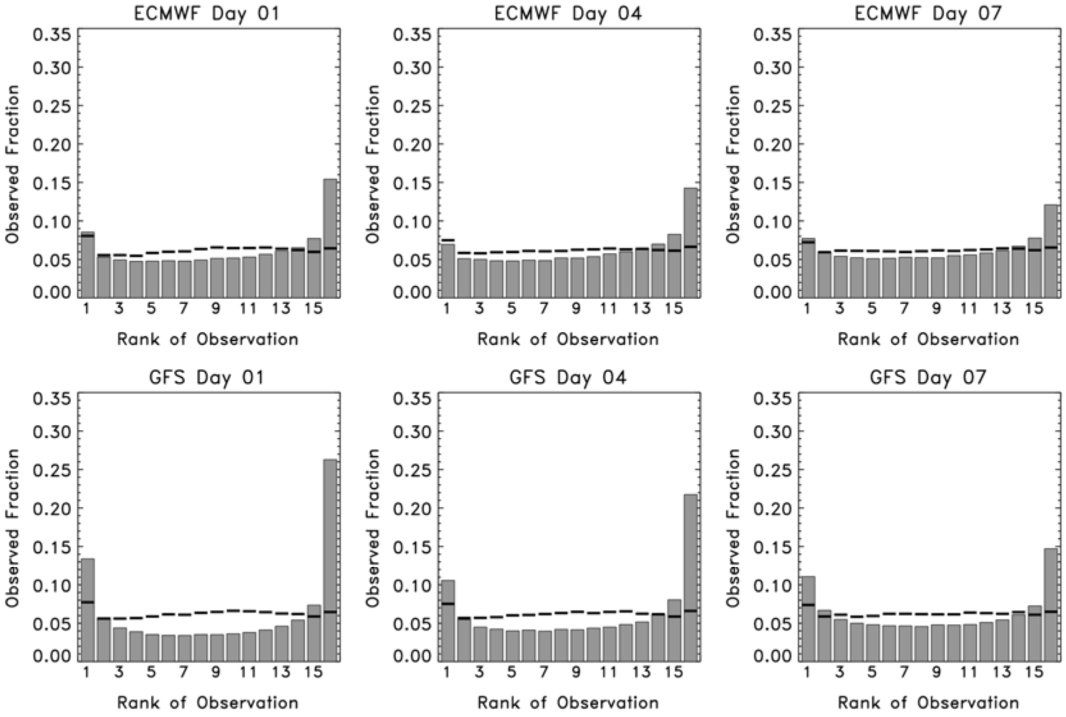

Figure 3: Rank histograms for 2-meter temperatures from ECWMF and GFS ensembles at 1, 4, and 7 days lead. Histograms denote the raw ensemble and solid lines the histogram derived from the pdf of the calibrated forecasts.

We now consider the overall *CRPSS* of the calibrated and uncalibrated forecasts in Fig. 4. Several main points can be made. First, the raw ECMWF forecasts were indeed more skillful than the GFS forecasts. Second, while the raw GFS forecasts had zero or negative skill relative to climatology, after statistical correction with NGR they exceeded the *CRPSS* of the raw ECMWF forecasts, demonstrating the large skill improvement that was possible with calibration. Third, even though the ECMWF model started with substantially greater skill than the GFS, it too benefited greatly from the statistical correction. Though improvements were not as large as with the GFS, a statistically modified 4-5-day ECMWF forecast had approximately the same *CRPSS* as did the raw 1-day forecast. Fourth, consider the multi-model NGR forecast. It consistently out-performed the calibrated ECMWF forecast by a small amount, indicating that there was some independent information provided by the older, less sophisticated *GFS*. Last, note that even at day 10 there is still some skill in the calibrated ECMWF and multi-model forecasts. If one considers averages over several days such as an 8-10 day average the skill increases above that of the averages of the skills at days 8, 9, and 10 (not shown). This is because some of the loss of skill is due to small errors in the timing of events.

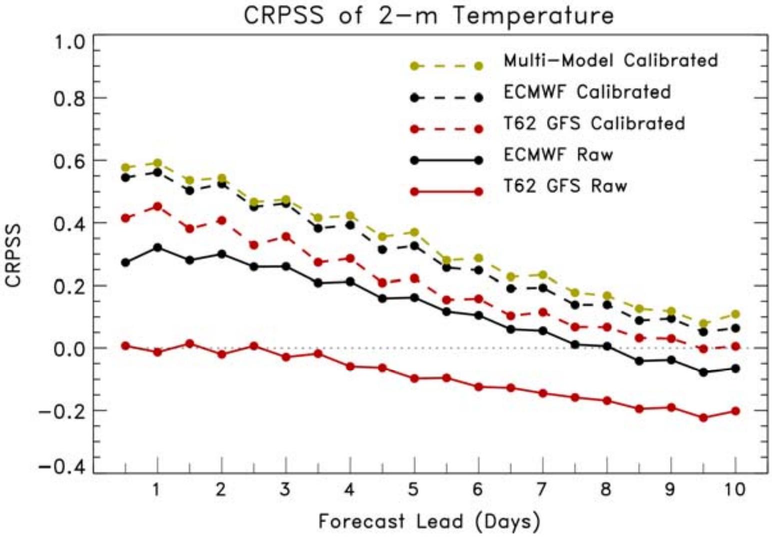

Figure 4. CRPSS of surface temperature forecasts from ECMWF and GFS models with and without calibration, and from calibrated multi-model forecast.

Figure 5 shows that a substantial fraction of the forecast improvement in each system can be attributed to a simple correction of model bias. The bias-corrected ensemble forecasts were generated by subtracting the mean bias (forecast minus observed) from each ensemble member in the training sample. Between 60 and 80 percent of the improvement in skill in the ECMWF forecasts can be attributed to this simple bias correction; the NGR added the remaining 20-40 percent through its regression-based correction, spread correction, and fitting of a smooth parametric distribution. Slightly less of the improvement was attributable to bias for the GFS ensemble.

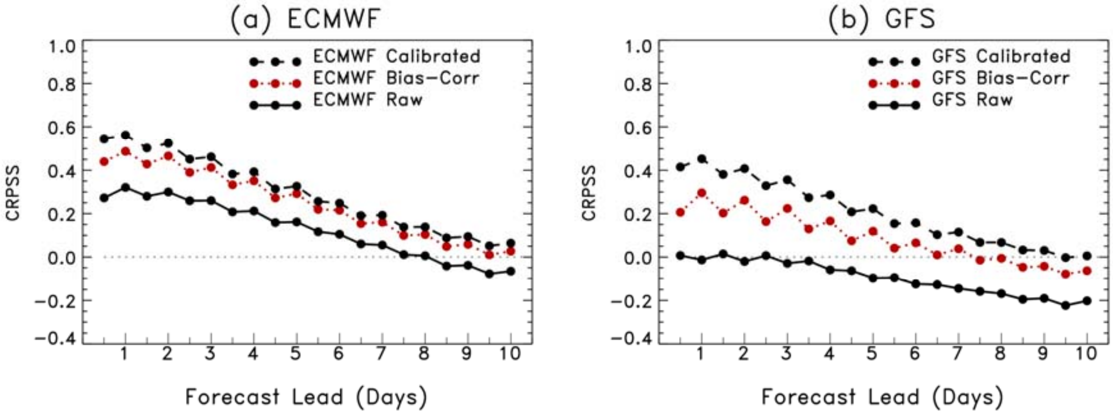

Figure 5: CRPSS including bias-corrected ensemble forecasts for (a) ECMWF forecasts, and (b) GFS forecasts.

4.2. Differences between 20-year weekly and 30-day daily training data sets

To facilitate a comparison of long and short training data sets, the ECMWF and GFS ensemble forecasts were also extracted every day for the period 1 July – 1 December 2005. This permitted us to examine the efficacy of a smaller training data set. Recent results (Stensrud and Yussouf 2005, Cui et al. 2006) have suggested that temperature forecast calibration may be able to be performed well even with a small number of recent forecasts. This may be because the ensemble forecast bias is relatively consistent and can be estimated with a small sample. Another possibility is that recent samples are more relevant for the statistical correction, with their more similar circulation regimes and land-surface states than data from other years.

Accordingly, we compared the calibration of forecasts using the prior 30 days as training data to calibration using the full reforecast training data set. Forecasts were compared for the period of 1 September – 1 December 2005. Non-homogeneous Gaussian regression was again used for the calibration. Figure 6 shows that at short forecast leads, the 30-day training data set provided approximately equal skill improvements relative to the 20-year training data set for the ECMWF model, and marginally less for the GFS. However, as the forecast lead increased, then the benefit of the longer training data set becomes apparent.

Why were more samples particularly helpful for the longer leads? We suggest that there were at least three contributing factors. First, the prior 30-day training data set was 9 days older for a 10-day forecast (training days -39 to -10) than for a 1-day forecast (training days -30 to -1). The second reason is that determining the bias to a pre-specified tolerance will require more samples at the long leads than at the short leads. At these long leads, the proportion of the error attributable to bias shrinks due to the rapid increase of errors due to chaotic error growth. The third reason was that the short-lead forecast training data sets were comprised of samples that tended to have more independent errors than the longer-lead training data sets. The ECMWF 1-day lagged correlation of forecast minus observed averaged over all stations (not shown) increased from around 0.2 at the early leads to 0.5 at the longer leads. Using the definition of an effective sample size *n*' (Wilks 2006, p. 144) with *n*=30, this indicated that the effective sample size was approximately 20 at the short leads and 10 at the longer leads. The once-weekly, 20-year reforecast data set should, in comparison, be comprised of samples that were truly independent of each other.

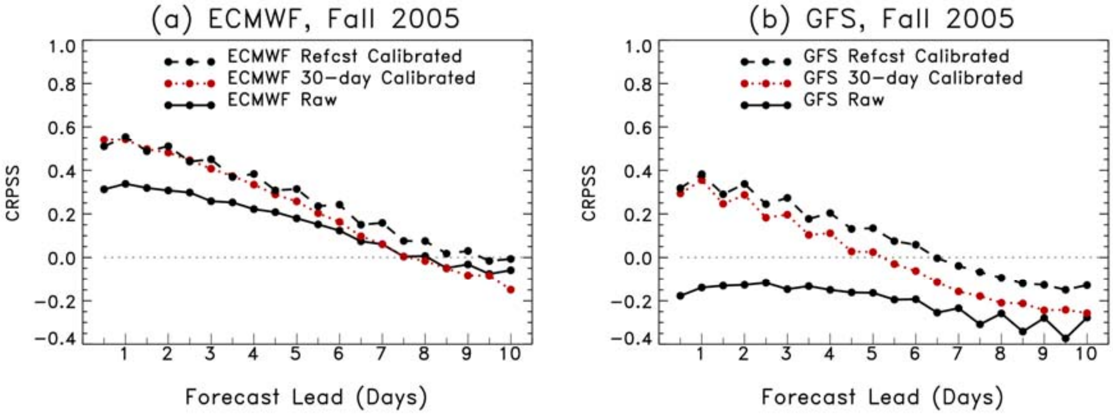

Figure 6: Comparison of CRPSS using 30-day and 20-year training data sets for the period 1 October – 1 December 2005. (a) ECMWF data, (b) GFS data.

5. 12-hour accumulated precipitation forecast results.

5.1. 20-year weekly training data.

Figure 7 provides 1, 3, and 5-day reliability diagrams at the 5-mm threshold for ECMWF's raw forecasts, validated over all forecasts (20 years × 14 weekly reforecasts for all grid points in the CONUS). Figure 8 provides the same, but for GFS raw forecasts sub-sampled to the dates of the ECMWF reforecasts. ECMWF raw forecasts were slightly more reliable than GFS raw forecasts, though both were notable more for the lack of reliability than its presence. Inset *BSS*es indicated that the forecasts were less skillful than the reference climatologies. Raw forecast skill was somewhat larger for lighter thresholds. The reasons why longer-lead forecasts have reduced negative skill will be discussed later.

The unreliability and in particular the low skill were worse than has been reported in some comparable studies (e.g., Eckel and Walters 1998, Mullen and Buizza 2001). This was due to several factors. First, the validation in this study was performed over a shorter temporal period (here, 12-h accumulations) and on a comparatively finer-resolution grid, 32 km. Previously, it was shown (e.g, Islam et al. 1993, Gallus 2001) that a finer discretization of the forecast in time and space decreased the apparent predictability or skill of a forecast. Somewhat improved reliability and skill were evident when the verification data were instead accumulated at the forecasts' 1-degree grid-box scale (not shown). Also, the low skill was partly due to calculating the BSS in a manner analogous to eq. (3), which was much more stringent in assigning skill than the conventional method of calculation of the *BSS* (Hamill and Juras 2006).

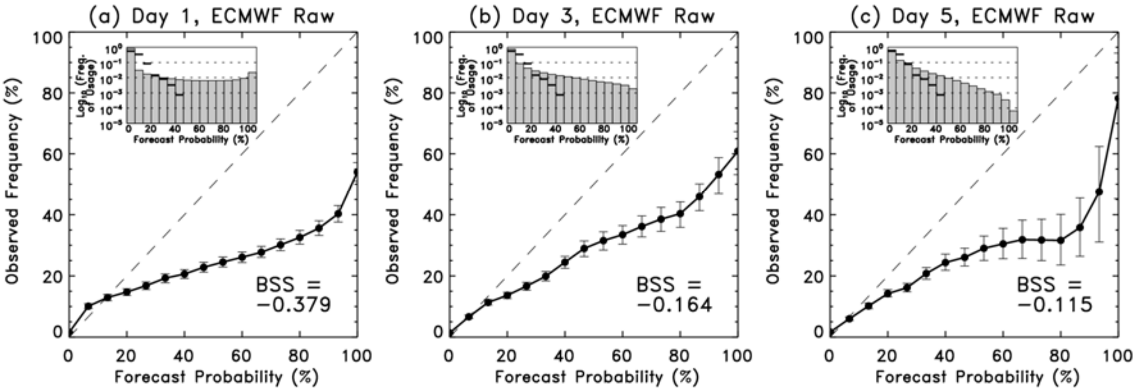

Figure 7: Reliability of 5-mm ECMWF raw forecasts at 1, 3, and 5-day leads. Overplotted confidence intervals provide 5th and 95th percentiles of determined through block bootstrap resampling techniques. Inset histogram denotes frequency of forecast usage of each probability bin. Solid lines overplotted on histogram denote the climatological frequency of usage of each probability bin.

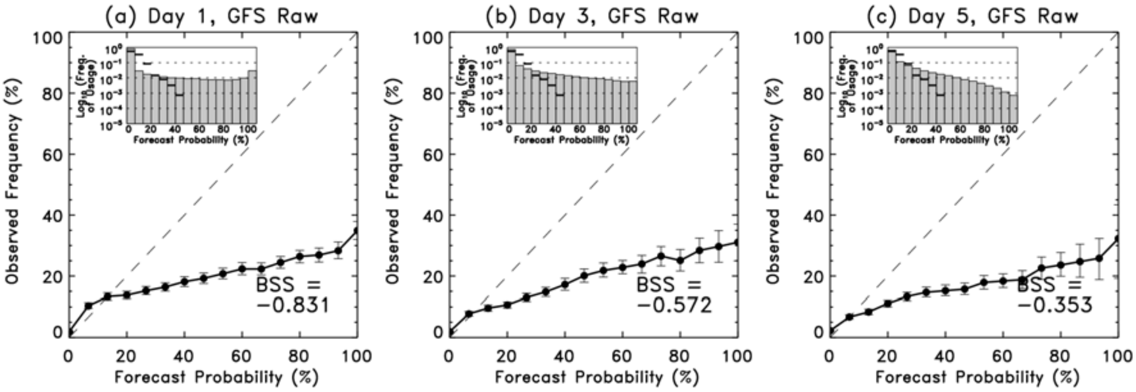

Figure 8: As in Fig. 7, but for GFS raw forecasts.

Figures 9 and 10 show the reliability diagram for calibrated ECMWF and GFS forecasts, respectively. There was a dramatic improvement in reliability at all leads relative to the raw forecasts, though sharpness was greatly lessened; high-probability forecasts in particular were not issued nearly as frequently. The *BSS* was improved dramatically at all leads. ECMWF calibrated forecasts were consistently higher in skill than GFS calibrated forecasts. However, in some instances such as for the day-1 forecasts, GFS forecast appeared to be slightly more reliable than ECMWF forecasts. However, by comparing each figure's inset frequency of usage histograms, it was apparent that the ECMWF calibrated forecasts were somewhat sharper, issuing higher probability forecasts and thus deviating from the climatological distribution more often. When the calibrated ECMWF forecasts issued high-probability forecasts, the event typically occurred, as judged from the reliability curves. Consequently, ECMWF forecasts had lower Brier Scores (and higher *BSS*es). Note also that the GFS calibrated day-5 forecast had a frequency of usage distribution very similar to that of climatology, reflected in a *BSS* near zero. The calibrated GFS forecasts, finding little forecast signal with this model, regressed to the local climatological probabilities.

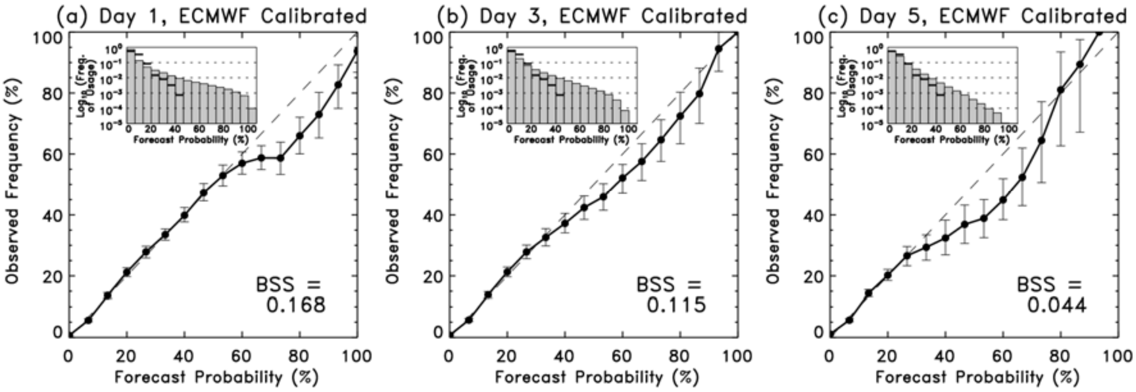

Figure 9: As in Fig. 7, but for calibrated ECMWF forecasts.

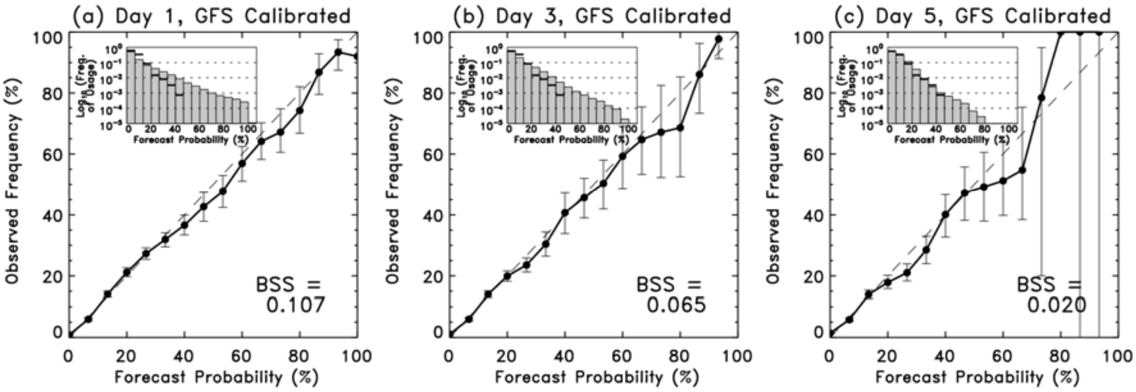

Figure 10: As in Fig. 7, but for calibrated GFS forecasts.

Interestingly, multi-model calibrated forecasts (not shown) did not produce any noticeable improvement in skill relative to the ECMWF calibrated forecasts, unlike the temperature results. The differences between the multi-model and ECMWF skill consistently lay within the 5th and 95th percentiles of the bootstrapped skill score distribution (shown in the next section), indicating that the differences were not statistically significant. It is possible that a more careful combination of ECMWF and GFS forecast data may have slightly improved the skill of calibrated multi-model forecasts.

Figure 11 shows *BSS*es as a function of forecast lead and threshold for raw and calibrated forecasts. Calibrated multi-model forecasts were not plotted, but differences at all leads were statistically insignificant relative to ECMWF calibrated forecasts. Several interesting characteristics of the forecast can be noted. First, both ECMWF and GFS raw forecasts oscillated in forecast skill, exhibiting higher skill for 0000 – 1200 UTC forecasts and lower skill for 1200 – 0000 UTC forecasts. Figure 12 demonstrates why this occurs. Here, the ECMWF forecast characteristics changed diurnally, tending to over-forecast significant rainfall events during 1200 – 0000 UTC. This over-forecast bias was much less pronounced at 0000-1200 UTC. GFS forecasts had an even more pronounced daytime over-forecast bias.

Several other characteristics of the *BSS*es for the raw forecasts can be noted in Fig. 11. Especially at the higher thresholds, forecast skill actually was negative at early leads and increased somewhat with forecast lead, a counter-intuitive result. This was due primarily to the lack of spread (i.e., greater sharpness) in shorter-lead ensemble forecasts and the larger spread in longer-lead forecasts, as shown in the inset frequency of usage histograms from Figs. 7 and 8. The *BSS* heavily penalized these unrealistically sharp forecasts at the early leads, especially using the new method of calculation.

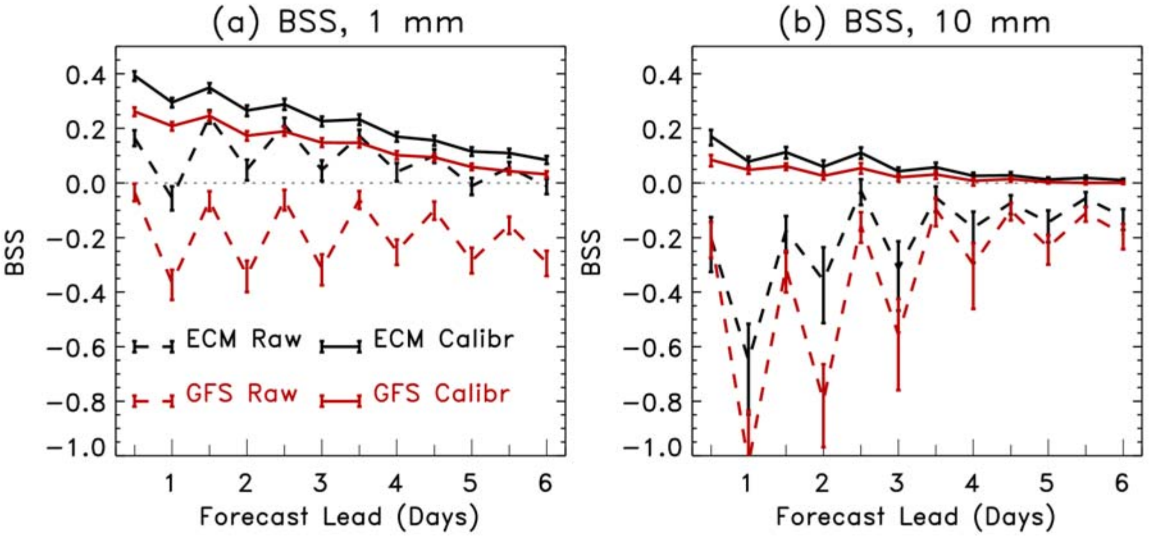

Figure 11: Forecast Brier skill scores at (a) 1 mm, (b) 5 mm, and (c) 10-mm precipitation thresholds for ECMWF and GFS raw and calibrated forecasts (multi-model calibrated forecasts are not plotted; they were statistically indistinguishable from ECMWF calibrated forecasts). Overplotted confidence intervals provide 5th and 95th percentiles determined through block bootstrap resampling techniques.

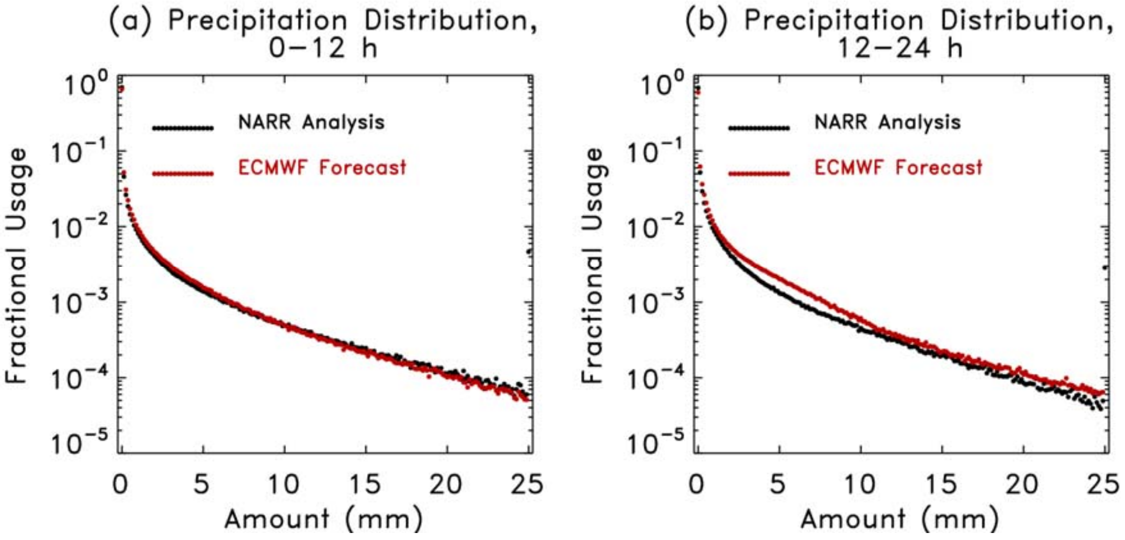

Figure 12: Distribution of the fractional usage of precipitation amounts from the North American Regional Reanalysis and ECMWF raw forecasts for (a) 0 -12 hour forecasts, and (b) 12-24 hour forecasts.

The characteristics of particular interest in Fig. 11 are the skill differences between ECMWF and GFS forecasts and the amount of skill improvement resulting from forecast calibration. As with temperature, ECMWF raw forecasts were more skillful than GFS raw forecasts, but skill was relative here; both were uniformly unskillful at the 10-mm threshold. ECMWF calibrated forecasts were significantly more skillful than GFS calibrated forecasts, judging from the small bootstrap confidence intervals.

Positive skill was noted in both ECMWF and GFS calibrated forecasts at nearly all leads, with a large reduction in the amplitude of the diurnal fluctuations in forecast skill relative to the raw forecasts. These forecasts were much more skillful than the raw forecasts at all leads. Comparing the skill at 1 mm, a 3- to 3.5-day ECMWF calibrated forecast was as skillful as a 1.5-day raw forecast, an approximately 2-day increase in forecast lead. Similar comparisons at other thresholds and forecast leads provided even more optimistic estimates of the skill improvement from calibration. A major conclusion from this study, then, is that the benefits of statistical calibration demonstrated previously with the GFS precipitation reforecasts (Hamill et al. 2006, Hamill and Whitaker 2006) are still evident with the much-improved ECMWF model. Forecast calibration still dramatically improved the forecasts from a state-of-the-art forecast model from 2005.

5.2. Comparison of skill using full, weekly, and 30-day training data

Forecast skill was evaluated for calibrated forecasts every day between 1 September and 1 December 2005. Skill was evaluated using three amounts of training data described in section 3b, the “30-day” training data (the last available 30 days), “weekly” (the once-weekly, 20-year reforecast data set), and “full” (for the GFS, 25 years of once-daily September-October-November reforecasts and observations). Weekly and 30-day training data sets used the compositing technique whereby training data was supplemented from 10 other grid points with similar analyzed climatologies.

Figure 13 shows the positive impact of the weekly training data sets. While the difference was not tremendously large at the 1-mm threshold, at 10 mm the degradation of forecast skill with the 30-day training data set relative to the weekly data set was quite large. At 10 mm, the improvement from using weekly ECMWF reforecasts compared to 30-day was at least 1.5 days of increased forecast lead time; a 2-day weekly calibrated ECMWF forecast was as skillful as a 0.5-day 30-day calibrated ECMWF forecast.

Interestingly, GFS forecast calibration did not appear to improve in skill when daily samples were used (without calibrating using the analog locations) relative to the GFS weekly using the analog locations. This suggests that daily reforecasts may not be necessary for this particular application.

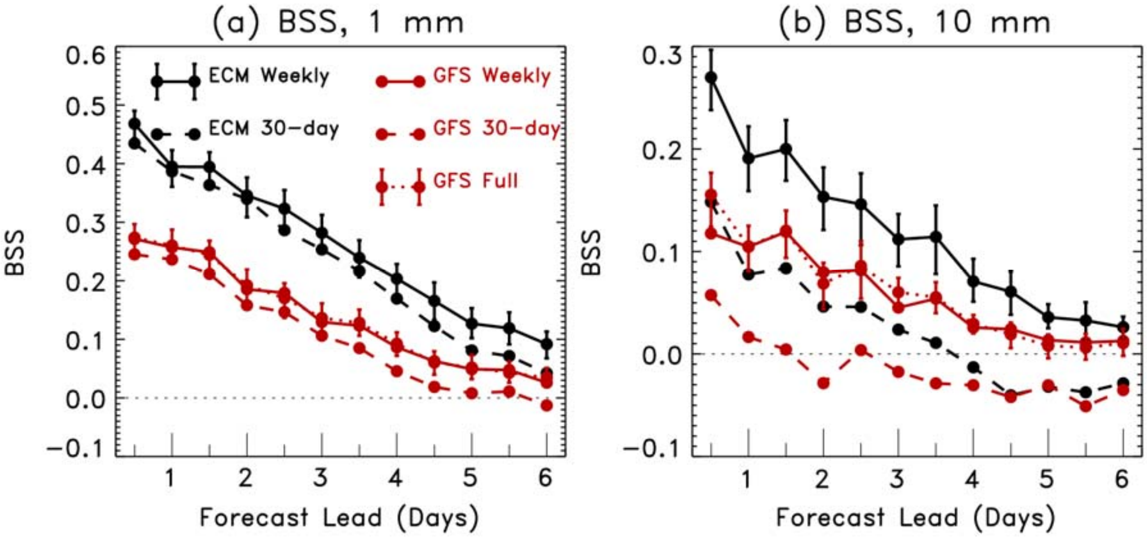

Figure 13: BSS of daily forecasts from 1 Sep 2005 – 1 Dec 2005, for (a) 1-mm threshold, (b) 5-mm threshold, and (c) 10-mm threshold. Error bars indicate the 5th and 95th percentiles of the ECMWF weekly and GFS full resampled distribution of BSSes. Error bars for other forecasts were similar in magnitude.

6. Conclusions

This study has shown that calibration with reforecasts substantially benefited even a higher-resolution, improved forecast model. Arguably, the beneficial results obtained from the reforecast-based calibration of GFS forecasts might be attributable to the low baseline set by the now aged 1998 GFS. The 2005 ECMWF system, arguably, is still representative of circa 2007-2008 systems for many other operational forecast centers, given ECMWF’s substantial lead in probabilistic forecast skill (Buizza et al. 2005). Hence, this abstract has shown the usefulness of large reforecast data sets, in particular for the calibration of forecasts of heavy precipitation and longer-lead temperature forecasts.

7. References

Agresti, A., 2002: *Categorical Data Analysis*. Wiley-Interscience, 710 pp.

Barkmeijer, J., M. van Gijzen, and F. Bouttier, 1998: Singular vectors and estimates of the analysis error covariance metric. *Quart. J. Royal Meteor. Soc*., **124**, 1695-1713.

------------, R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc*., **125**, 2333-2351.

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

Clark, M.P. and L.E. Hay (2004): Use of medium-range weather forecasts to produce predictions of streamflow. J. Hydrometeor., 5, 15-32.

Cui, B., Toth, Z., Zhu, Y., Hou, D., Unger, D., Beauregard, S., 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. *Proceedings, First THORPEX International Science Symposium*. December 6-10, 2004, Montréal, Canada, World Meteorological Organization, 281-284.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132-1147.

Gallus, W. A., Jr., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*, **17**, 1296-1302.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble Model Output Statistics and minimum CRPS estimation. *Mon. Wea. Rev*., **133**, 1098-1118.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: 2-meter temperature. *Mon. Wea. Rev*., submitted. Available at www.cdc.noaa.gov/people/tom.hamill/ecwmf_reforecast_temp.pdf.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

------------, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev*., **129**, 550-560.

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

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

------------, and ------------, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev*., **134**, 3209-3229.

------------, and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc*., **132**, 2905-2923.

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

------------, R. Hagedorn, and J. S. Whitaker, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, submitted. Available at www.cdc.noaa.gov/people/tom.hamill/ecmwf_refcst_temp.pdf .

Islam, S., R. L. Bras, and K. A. Emanuel, 1993: Predictability of mesoscale rainfall in the tropics. *J. Appl. Meteor.*, **32**, 297-310.

Mahfouf, J.-F., and F. Rabier, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. II: Experimental results with improved physics. *Quart. J. Royal Meteor. Soc.*, **126**, 1171-1190.

Mesinger, F., and coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343-360.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Royal Meteor. Soc.*, **122**, 73-199.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **129**, 638-663.

Sloughter, J. M., Raftery, A. E., Gneiting, T., and Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209-3220.

Stensrud, D.J., and N. Yussouf, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.*, **12**, 217-230.

Uppala, S. M., and coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Royal Meteor. Soc.*, **131**, 2961-3012.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble-mean skill. *Mon. Wea. Rev.*, **126**, 3292-3302.

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

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

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