Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation

THOMAS M. HAMILL

NOAA/Earth System Research Laboratory, Boulder, Colorado

RENATE HAGEDORN

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

JEFFREY S. WHITAKER

NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 9 October 2007, in final form 10 December 2007)

ABSTRACT

As a companion to Part I, which discussed the calibration of probabilistic 2-m temperature forecasts using large training datasets, Part II discusses the calibration of probabilistic forecasts of 12-hourly precipitation amounts. Again, large ensemble reforecast datasets from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Global Forecast System (GFS) were used for testing and calibration. North American Regional Reanalysis (NARR) 12-hourly precipitation analysis data were used for verification and training. Logistic regression was used to perform the calibration, with power-transformed ensemble means and spreads as predictors. Forecasts were produced and validated for every NARR grid point in the conterminous United States (CONUS). Training sample sizes were increased by including data from 10 nearby grid points with similar analyzed climatologies. "Raw" probabilistic forecasts from each system were considered, in which probabilities were set according to ensemble relative frequency. Calibrated forecasts were also considered based on three amounts of training data: the last 30 days of forecasts (available for 2005 only), weekly reforecasts during 1982–2001, and daily reforecasts during 1979–2003 (GFS only).

Several main results were found. (i) Raw probabilistic forecasts from the ensemble prediction systems' relative frequency possessed little or negative skill when skill was computed with a version of the Brier skill score (BSS) that does not award skill solely on the basis of differences in climatological probabilities among samples. ECMWF raw forecasts had larger skills than GFS raw forecasts. (ii) After calibration with weekly reforecasts, ECMWF forecasts were much improved in reliability and were moderately skillful. Similarly, GFS-calibrated forecasts were much more reliable, albeit somewhat less skillful. Nonetheless, GFS-calibrated forecasts were much more skillful than ECMWF raw forecasts. (iii) The last 30 days of training data produced calibrated forecasts of light-precipitation events that were nearly as skillful as those with weekly reforecast data. However, for higher precipitation thresholds, calibrated forecasts using the weekly reforecast datasets were much more skillful, indicating the importance of large sample size for the calibration of unusual and rare events. (iv) Training with daily GFS reforecast data provided calibrated forecasts with a skill only slightly improved relative to that from the weekly data.

1. Introduction

This paper continues to examine how reforecasts, datasets of prior forecasts from the same model run operationally, can be used to improve the calibration of probabilistic ensemble weather forecasts. *Calibration* here refers to the statistical adjustment of numerical forecasts to produce probabilistic forecasts that are as sharp as possible while remaining reliable (Wilks 2006, 258–259; Gneiting et al. 2007).

Hagedorn et al. 2008 (hereafter Part I) considered the problem of calibrating 2-m temperature forecasts from a newly available reforecast dataset from the Eu-

Corresponding author address: Dr. Thomas M. Hamill, NOAA/Earth System Research Laboratory, Physical Sciences Division R/PSD1, 325 Broadway, Boulder, CO 80305.
E-mail: tom.hamill@noaa.gov

DOI: 10.1175/2007MWR2411.1

ropean Centre for Medium-Range Weather Forecasts (ECMWF). Part I showed that dramatic improvements in probabilistic 2-m temperature forecast skill were possible even when calibrating the version of the ECMWF forecast model operational in the autumn of 2005, which was a much more high-resolution, skillful, and less biased forecast model than the 1998 Global Forecasting System (GFS) model used in previous reforecast experiments (Hamill et al. 2004, 2006; Hamill and Whitaker 2006, 2007; Whitaker et al. 2006; Wilks and Hamill 2007). However, the long, 20-yr ECMWF reforecast training dataset was not needed for the successful calibration of short-term temperature forecasts. Calibration based on a relatively short time series of the most recent forecasts and observations produced forecasts of comparable skill, although longer-lead forecasts were better calibrated with the 20-yr reforecasts' increased training sample size.

Gaining large improvements from calibration using a short training dataset is a particularly encouraging result because an extensive set of reforecasts is computationally expensive to produce. The significant computational expense of using long reforecast data must be justified by very large improvements in forecast skill, improvements larger than would be obtained by, say, increasing the model resolution or improving the realism of radiation calculations. Arguably, such gains were not realized with short-range temperature forecasts.

Unfortunately, short training datasets may not be as useful for precipitation forecast calibration as they were for the short-term temperature forecast calibration (Hamill et al. 2006, their Fig. 7). Precipitation accumulated over short periods tends to have a positively skewed distribution, with many zero events, fewer light-precipitation events, and very few heavy-precipitation events. Should today's mean forecast indicate heavy rainfall, a small training sample of prior forecasts may be dominated by the zero-rainfall or light-rainfall forecast events and thus be unhelpful. A longer time series of old forecasts and observations may be needed to provide enough similar events.

This article, then, reconsiders precipitation forecast calibration, this time including results from the new ECMWF reforecast dataset. Some important questions to be answered include: (i) Are large improvements in precipitation forecast skill and reliability possible through a reforecast-based calibration of ECMWF ensemble forecasts, as they were with the older GFS model forecasts, even though the ECMWF model is more skillful and less biased? (ii) How much additional benefit can be obtained from calibrating with a large reforecast dataset compared to calibrating with a brief time series of forecasts and observations from the recent past? (iii) Can training sample size be enlarged artificially by agglomerating data from locations with similar characteristics, thereby decreasing the need for an even larger set of reforecasts?

Below, section 2 reviews the datasets used in this experiment. Section 3 describes the calibration methodology and the methods for evaluating forecast skill, section 4 provides results, and section 5 presents conclusions.

2. Forecast and observational datasets used

a. 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 United States (CONUS) were used, and precipitation was accumulated over 12-hourly periods ending at 0000 and 1200 UTC. Precipitation analyses from this dataset were derived from an objective analysis of 24-hourly rain gauge data that was then temporally disaggregated into 3-hourly analyses based on nearby hourly rain gauge data. Orographic detail was inferred using techniques described in Daly et al. (1994). The NARR precipitation analysis can be expected to be less accurate in regions where rain gauge data was sparse (say, the intermountain western United States). Other deficiencies of this precipitation analysis dataset were noted in West et al. (2007).

b. ECMWF forecast data

The forecast data used here were the same as in Part I, except that 12-hourly accumulated precipitation forecast data valid at 0000 and 1200 UTC were used instead of 2-m temperature forecasts. The ECMWF reforecast dataset consisted 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 dataset were 1982–2001, and the initial conditions were provided by the 40-yr ECMWF Re-Analysis (ERA-40; Uppala et al. 2005). 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 member forecast was run to a 10-day lead, although because of the comparatively rapid decay of precipitation forecast skill, only forecasts to a 6-day lead will be considered here. The 1° gridded forecast data were bilinearly interpolated to the NARR grid at points within the CONUS.

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, although the initial analyses were provided by the operational four-dimensional variational analysis system (Mahfouf and Rabier 2000) rather than the three-dimensional variational analysis system used in ERA-40. The 2005 daily data permit experiments comparing calibration using a short training dataset of prior forecasts with calibration using the reforecasts.

c. GFS forecast data

The GFS reforecast dataset, 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 a 15-day lead for every day from 1979 to the present. Forecasts were started from 0000 UTC initial conditions, and forecast information was archived on a 2.5° global grid. GFS forecast accumulated precipitation was also bilinearly interpolated to the NARR grid at 12-hourly intervals. For most of the experiments 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 were also extracted for 1 July to 1 December 2005.

3. Forecast calibration and validation methodologies

a. Calibration with logistic regression

Logistic regression analysis (e.g., Agresti 2002, chapter 5) will be used as the general method of forecast calibration. The nonhomogeneous Gaussian regression technique used in Part I was not useful for precipitation, for which forecast distributions are usually non-Gaussian. Given an unknown observed amount O (the predictand), the precipitation threshold T, and model-forecast predictors $x_1, \ldots, x_p$, logistic regression analysis determines the parameters $\beta_0, \ldots, \beta_p$ to fit a predictive equation of the form

$$P(O > T) = 1.0 - \frac{1.0}{1.0 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}. \quad (1)$$

There are many other possible precipitation forecast calibration methodologies; see Hamill and Whitaker (2006) for a comparison of logistic regression with analog techniques using GFS reforecasts or Sloughter et al. (2007) for a Bayesian model averaging approach. Here, logistic regression was chosen because (i) it was a standard method, with readily understood characteristics and algorithms available from off-the-shelf software, (ii) it permitted unusual predictors and variable sample weights to be incorporated readily, and (iii) 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 because the ECMWF reforecast data consist 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 rather than a full probability density function.

Much experimentation was performed to determine, given an ensemble of precipitation forecasts, what would be the ideal predictors in the logistic regression formulation. Because we would prefer to focus in this paper on differences between the ECMWF and GFS forecasts and the effects of training sample size, we will not document the results of testing different potential predictors. These indicated (not shown) that small improvements were possible with a careful choice of predictors, but model and sample size were bigger factors in determining forecast skill.

The two chosen predictors to be used in all subsequent regression analyses are the ensemble mean, raised to the quarter power, and the ensemble spread, also raised to the quarter power, where the spread denotes the standard deviation of the ensemble about its mean. The power transformation has the effect of allowing the precipitation forecast data to be more normally distributed. Letting $\bar{x}^f$ denote the forecast mean and σ^f denote the spread, the regression takes the form

$$P(O > T) = 1.0 - \frac{1.0}{1.0 + \exp[\beta_0 + \beta_1(\bar{x}^f)^{0.25} + \beta_2(\sigma^f)^{0.25}]}. \quad (2)$$

Previously, Sloughter et al. (2007) used a one-third-power transformation in their precipitation calibration, and Hamill and Whitaker (2006) used a one-half-power transformation in the logistic regression of daily precipitation forecasts. For this application to 12-hourly accumulated precipitation, the reliability of high-probability forecasts was improved slightly with the one-quarter-power transformation compared to one-half- or no-power transformation. The more skewed the distribution, the smaller the appropriate exponent

for the power transformation, and the 12-h accumulated precipitation tested here is more highly skewed than previous work with 24-h accumulated precipitation.

The logistic regression algorithm used in this study allowed for training samples to be weighted individually. Through experimentation, we determined that weighting the samples with higher forecast mean precipitation improved the reliability of the forecasts, especially in situations in which high probabilities were issued. Consequently, we chose a weight w for a particular sample based on the relationship of its ensemble-mean forecast to the precipitation threshold in question:

$$w = \begin{cases} 1.0, & \text{if } \bar{x}^f + 0.01 > T, \text{ or} \\ 0.1 + 0.9 \times \exp[-1.0 \times |\log_{10}(\bar{x}^f + 0.01) - \log_{10}(T)|], & \text{if } \bar{x}^f + 0.01 \le T. \end{cases} \quad (3)$$

This produced a weighting function of the form shown in Fig. 1, with higher weights for samples with larger forecast precipitation amounts. The form of this weight is somewhat arbitrary; the important aspect was simply to provide more weight to the heavier forecast precipitation events.

For the temperature forecast calibration in Part I, acceptable results were sometimes produced even with limited training data. For the precipitation forecasts considered here, however, even with logistic regression, a robust training dataset 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 especially large training dataset. Here, a logistic regression analysis was run using Eqs. (2) and (3). 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 yr × 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. (Notice the patch of near-zero probabilities in western Nebraska as one example.) 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.

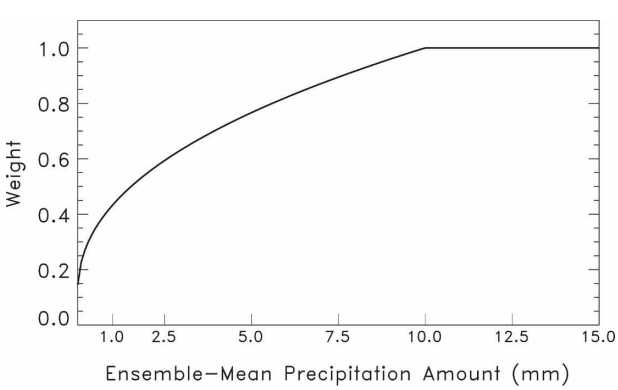

FIG. 1. Weighting function for logistic regression at the 10-mm threshold.

A different grid point should provide a suitable "analog," and its data could be used to enlarge the training sample size if that grid point had (i) a similar observed (O) climatological cumulative density function (CDF), (ii) a similar forecast (F) cumulative density function, (iii) a similar predictive relationship between forecast and observed (e.g., similar F–O correlations), and (iv) independent errors from the original grid point. Unfortunately, with ECMWF reforecasts available only once per week, and given the non-Gaussian, intermittent nature of precipitation, finding analog locations meeting criteria (ii) and (iii) above was difficult. The forecast CDFs were noisy given the limited sample size, and F–O relationships were sometimes misdiagnosed from a few unrepresentative cases. However, it was possible to use the long record from the NARR to determine supplemental locations that at least had similar observed climatologies and that were distant enough from each other to have quasi-independent forecast errors. For all of the regression analysis results presented hereafter from the ECMWF reforecasts (and GFS reforecasts, unless otherwise noted), the training data for a grid point were supplemented by training data from 10 other supplemental analog grid points with similar observed climatologies.

The specific procedure for finding 10 analog grid points was as follows: first, the climatological probability of 24-h accumulated precipitation exceeding 1, 2.5, 5, 10, 25, and 50 mm was calculated at each NARR grid

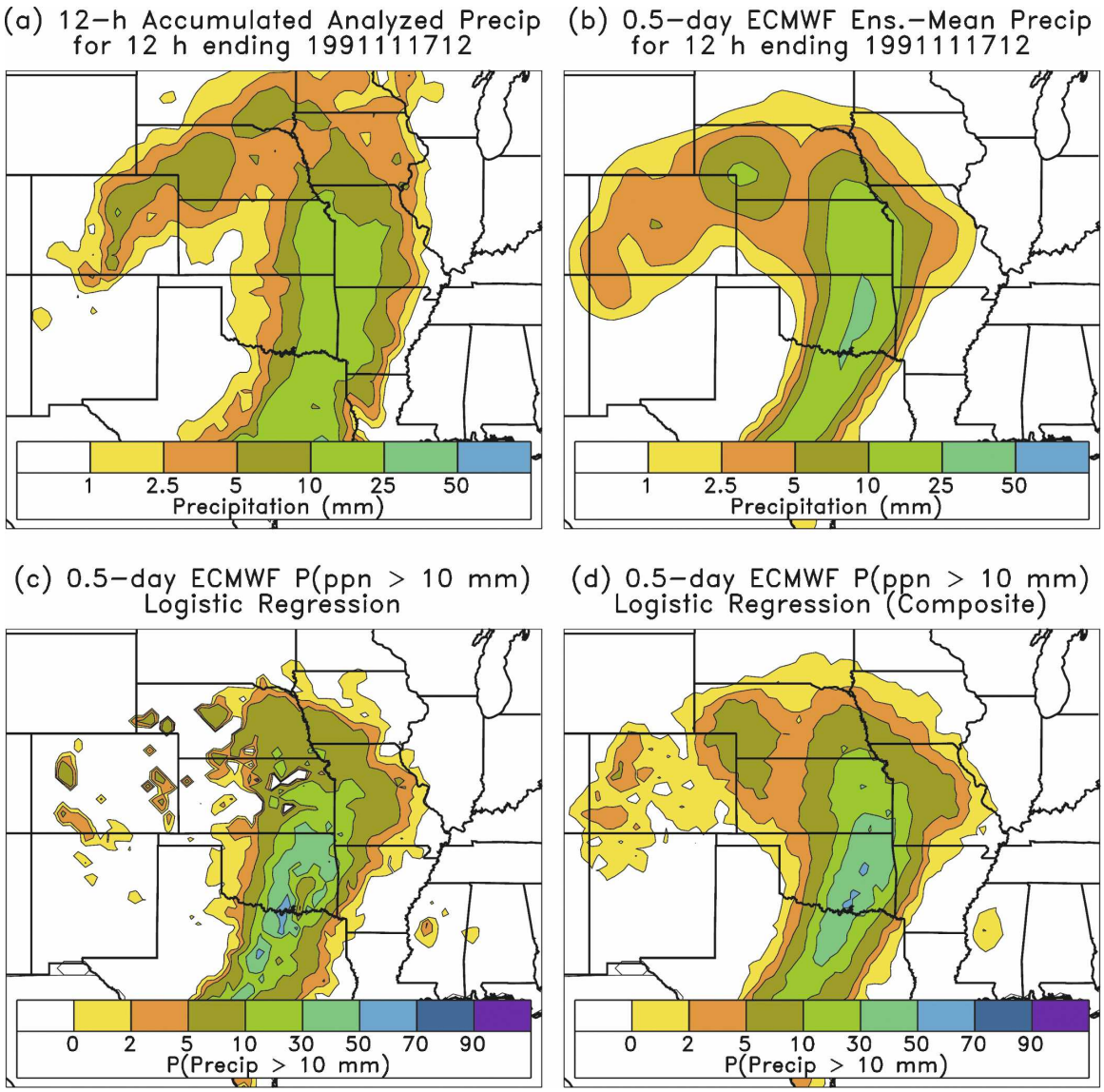

FIG. 2. (a) NARR precipitation analysis of 12-h accumulated precipitation for the 12-hourly period ending 1200 UTC 17 Nov 1991, (b) 0–12-h ECMWF ensemble-mean forecast of accumulated precipitation, (c) probability of greater than 10-mm precipitation in this period using a logistic regression in which each grid point's data are treated independently, and (d) same as (c), but here the logistic regression training data include forecasts and observations from 10 other locations that have similar observed precipitation climatologies for this day of the year.

point for each day of the year. The climatological probabilities were based on NARR data from 1979–2003 and ±30 days around the date of interest. For a given grid point i, only other grid points within a radius of 25 grid points (~800 km) were considered as potential analogs. The D_n statistic [Wilks 2006, Eq. (5.15)] was calculated at all grid points within this radius, with the test statistic for a grid point j within the radius defined by

$$D_n(j) = \max_T |F_j(T) - F_i(T)|. \quad (4)$$

Here, $F(T)$ denotes the CDF value at a threshold T, and the six precipitation thresholds noted above were considered. This test statistic thus indicates the maximum difference among CDFs when examined over all precipitation amounts. We then searched for other grid points that had small test statistics (i.e., similar CDFs). All grid points less than 3.5 grid points' distance from the grid point of interest were excluded from consideration, under the assumption that nearby grid points were likely to provide nonindependent data. Next, grid points were ordered from lowest to highest test statis-

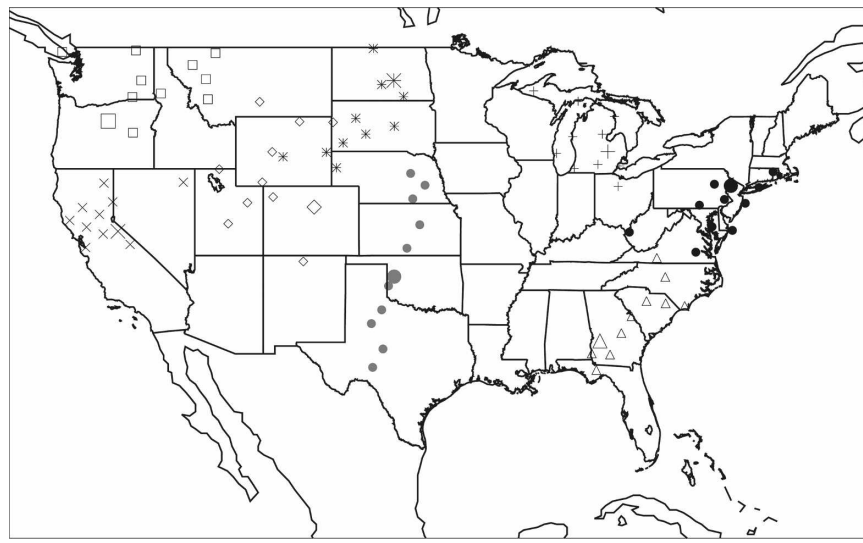

FIG. 3. Selected analog locations for increasing training sample size in the logistic regression analysis. Grid points are sought that have a similar climatology to a location of interest. The locations of interest are denoted by the large symbols, and 10 other grid points that have similar climatologies are denoted by the corresponding smaller symbols. When training logistic regression at a grid point with a large symbol, observations and forecasts are used both at this location and at the locations of the small symbols.

tics. The location with the lowest test statistic became the first analog grid point. All grid points less than 3.5 grid points' distance from this analog location were then also excluded from further consideration. The process was then repeated, finding the grid point with the next-lowest test statistic, excluding grid points around it, and so on until the locations for 10 analogs were determined. Figure 3 shows sample analog locations for several grid points around the CONUS.

b. Experiments performed

Probabilistic forecasts were evaluated from five sources, namely (i) *ECMWF raw* forecasts, whose probabilities were set directly from the relative frequency in the ensemble [e.g., if 5 of 15 forecasts indicate greater than 5 mm, $P(O > 5 \text{ mm}) = \frac{1}{3}$], (ii) *GFS raw* forecasts, (iii) *ECMWF-calibrated* logistic regression forecasts, based on the logistic regression method described above, (iv) *GFS-calibrated* forecasts, and (v) *multimodel-calibrated* forecasts. Here, the predictors for the multimodel forecasts were weighted linear combinations of the ECMWF and GFS forecasts. Given the greater skill of the ECMWF forecasts, the weights were arbitrarily set to 0.75 for ECMWF and 0.25 for the GFS. A more sophisticated weighting such as that performed in Part I is conceivable, but it was not attempted here. Partly this was because the weighting in Part I assumed that forecast errors were normally distributed, an assumption that cannot be made here.

Calibrated forecasts based on three amounts of training data were considered. The primary results were based on the cross-validated, 20-yr, weekly reforecast datasets (e.g., 1982 forecasts were calibrated with 1983–2001 forecasts and observations). Unlike the temperature calibration in Part I, where forecast bias was assumed to be seasonal and the training data were limited to the few dates surrounding the week of interest, for precipitation calibration the full training data were lumped together. For example, when calibrating the 1 September forecasts, the training data included all reforecast dates available, from 1 September–1 December. This was done under the assumption that an increased training sample size was more beneficial for precipitation forecasts than the possible degradation from not accounting for seasonal changes in precipitation forecast bias between September and December. To facilitate a direct comparison of GFS and ECMWF forecasts, the GFS reforecast data were subsampled to 20-yr, September–December weekly data from 1982–2001.

The second amount of training data was 30 days; that is, forecasts from the most recently available 30 days of forecasts were used for training. Note, however, that to be consistent with operational practice in which training data can be utilized only when observations become available, longer-lead forecasts used older training data than shorter-lead forecasts. For example, a 6-day lead forecast used training data that was actually data from

day −35 to day −6, whereas a 1-day lead forecast used training data from day −30 to day −1. Daily samples of GFS and ECMWF forecasts from the same model version as the reforecasts were only available in 2005, so the comparison of calibration from weekly and 30-day training datasets is limited to fall 2005 data.

The last training dataset size, available only for GFS forecasts, was the full reforecast. Here, rather than a 20-yr weekly sample between 1 September and 1 December, a 25-yr (1979–2003) daily sample between these dates was utilized for training data. Forecasts were trained with data from the adjacent 2 months and the current month. For example, September forecasts were trained with August, September, and October precipitation forecasts and analyses.

c. Forecast validation techniques

1) Reliability diagrams

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 estimated from a 1000-member block bootstrap sample [following Efron and Tibshirani (1993) and Hamill (1999), and similar to the method in Bröcker and Smith (2007)]. Each case day was considered as a separate block of fully independent data in the bootstrap, which was justifiable with samples 1 week apart. 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 the top of the forecast frequency of usage.

2) Brier skill score

Following Hamill and Juras (2006), the standard method for computing Brier skill score (BSS) was adapted so that positive skill was not inappropriately attributed to the forecasts simply because of variations in climatological event probabilities among the samples. The modification to the standard method of BSS computation was relatively straightforward: the overall forecast sample was divided into subgroups in which the climatological event probability was approximately homogeneous, the BSS was calculated for each subgroup, and the final BSS was calculated as a weighted average of the subgroups' BSS. For precipitation, there were NC = 10 subgroups, each with a more narrow range of climatological uncertainty in each subgroup. Let $\overline{\mathrm{BS}}^{f}(s)$ denote the average Brier score (Wilks 2006, chapter 7) of forecasts populating the sth subgroup, let $\overline{\mathrm{BS}}^{c}(s)$ denote the average Brier score of the climatological reference forecast in this subgroup, and let $u(s)$ be the fraction of samples from the sth subgroup. Then the overall BSS is calculated as

$$\mathrm{BSS} = \sum_{s=1}^{\mathrm{NC}} u(s)\left[1 - \frac{\overline{\mathrm{BS}}^{f}(s)}{\overline{\mathrm{BS}}^{c}(s)}\right]. \tag{5}$$

(For more details, please see Hamill and Whitaker 2007.) Also, as with the reliability curves, a 1000-member block bootstrap procedure was used to quantify the uncertainty in the skill score estimates. Note that with these daily samples, at short leads the forecast errors can be considered independent from one day to the next (Hamill 1999, his Table 3). However, we have not verified independence for longer-lead forecasts, so the block bootstrap may slightly underestimate the width of the confidence intervals.

4. Results

a. Forecast reliability with weekly training data

Figure 4 provides 1-, 3-, and 5-day reliability diagrams at the 5-mm threshold for ECMWF's raw forecasts, validated over all forecasts (20 yr × 14 weekly reforecasts for all grid points in the CONUS). Figure 5 provides the same, but for GFS raw forecasts subsampled to the dates of the ECMWF reforecasts. ECMWF raw forecasts were slightly more reliable than GFS raw forecasts, although both were notable more for the lack of reliability than for its presence. Inset BSS 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 had 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 2002) that a finer discretization of the forecast in time and space

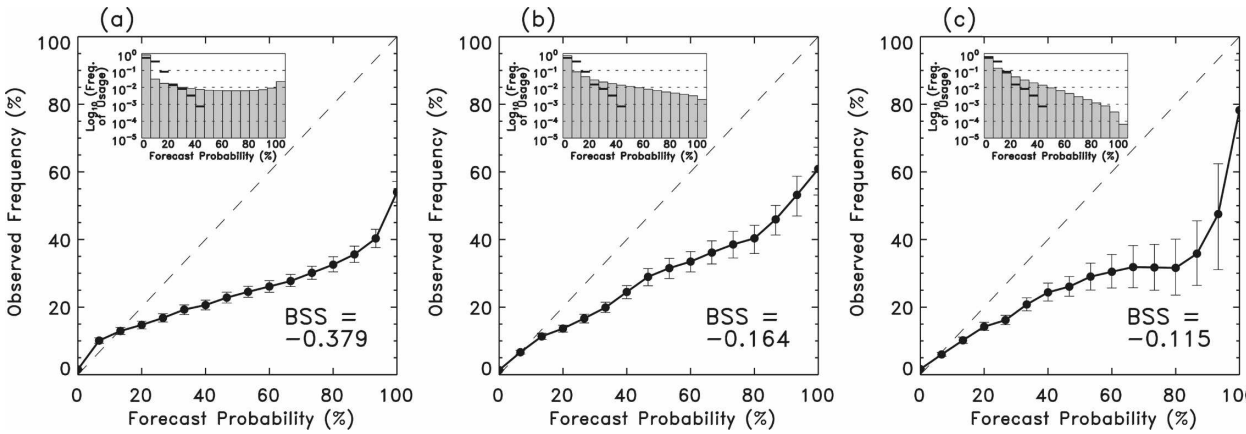

FIG. 4. Reliability of 5-mm ECMWF raw forecasts at (a) 1-, (b) 3-, and (c) 5-day leads. Plotted confidence intervals show the 5th and 95th percentiles as determined through block bootstrap resampling techniques. The inset histogram denotes frequency of forecast usage of each probability bin. Solid lines plotted on the histogram denote the climatological frequency of usage of each probability bin.

decreased the apparent predictability or skill of a forecast. Somewhat improved reliability and skill were evident when the verification data were accumulated instead at the forecasts' 1° gridbox scale. Also, the low skill was partly due to the use of Eq. (5), which was much more stringent in assigning skill than the conventional method of calculation of the BSS (Hamill and Juras 2006). Commonly, a reliability curve that was halfway between perfect reliability and flat, such as Fig. 4a, would be assigned near-zero skill [see the interpretation of the related "attributes diagram" in Hsu and Murphy (1986) and in Wilks (2006), p. 292]. However, implicit in the definition of skill in such diagrams is that the associated reference climatological probability is the same for all forecast samples. When a reliability diagram is in fact composed of forecast samples that are associated with a mixture of reference climatological probabilities, subzero skill is possible with such a curve.[1] With these reliability diagrams, forecast samples were taken from grid points across the

[1] In fact, one could conceive of a degenerate case of a "perfect" reliability diagram but with a 0.0 BSS using Eq. (5). Suppose half the samples populating the diagram had a forecast probability of 1.0 and half had 0.0, and each forecast was perfectly sharp and perfectly reliable. But suppose all the sample data for the 1.0 forecast probability had a climatological event probability of 1.0 as well, and all the data for the 0.0 probability had a climatological event probability of 0.0. Then the forecast, although perfect, is no better than climatology; thus, the BSS of Eq. (5) would assign these forecasts zero skill (Hamill and Juras 2006). This highlights a difficulty with such skill metrics: when climatology is a very good forecast on most occasions (e.g., heavy rainfall in the desert: the climatological forecast of low probability is a good one nearly all the time), establishing forecast skill relative to the climatology is especially difficult.

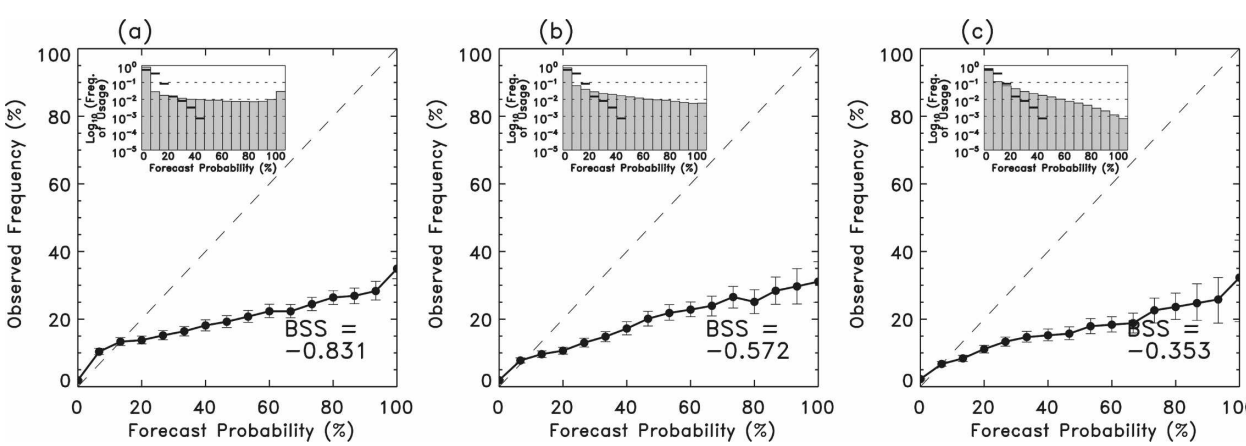

FIG. 5. Same as Fig. 4, but for 5-mm GFS raw forecasts.

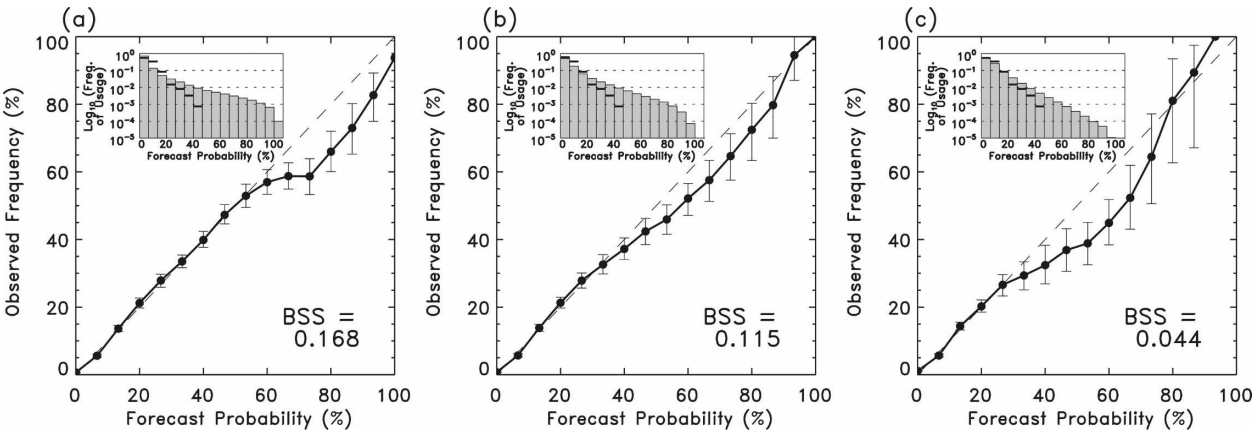

FIG. 6. Same as Fig. 4, but for 5-mm calibrated ECMWF forecasts.

CONUS that were associated with a distribution of climatological probabilities (this distribution was plotted as a horizontal bar over the top of the forecast distribution in the reliability diagrams).

Figures 6 and 7 show the reliability diagrams for calibrated ECMWF and GFS forecasts, respectively. There was a dramatic improvement in reliability at all leads relative to the raw forecasts, although 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 forecasts 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.[2] 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). 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.

Interestingly, multimodel-calibrated forecasts (Fig. 8) did not produce any noticeable improvement in skill

[2] A common technique for the analysis of sources of skill is a decomposition of the forecast Brier score into components describing the reliability, resolution, and uncertainty (e.g., Wilks 2006, p. 284). Implicit in this decomposition, however, is the assumption that all samples have the same underlying climatological event probability, an assumption that is violated here.

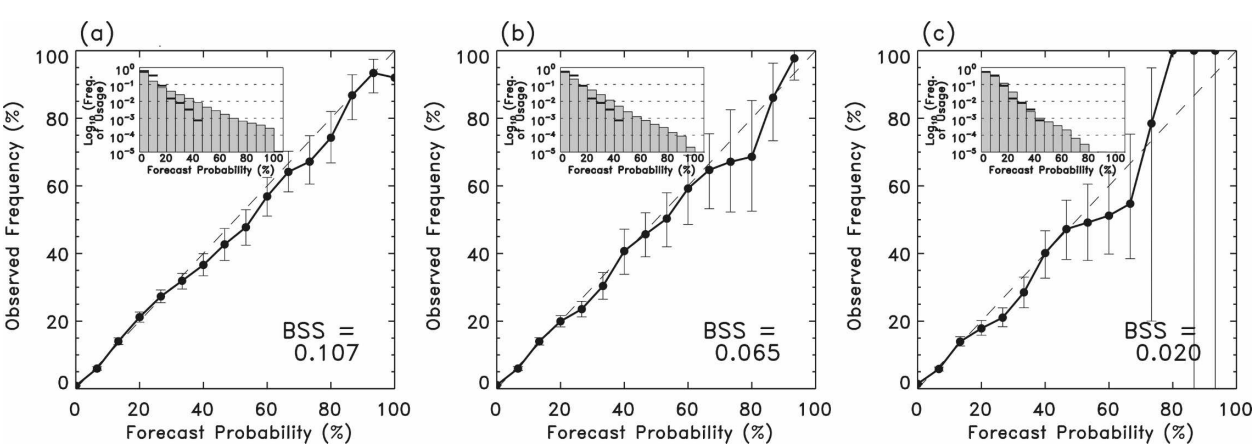

FIG. 7. Same as Fig. 4, but for 5-mm calibrated GFS forecasts.

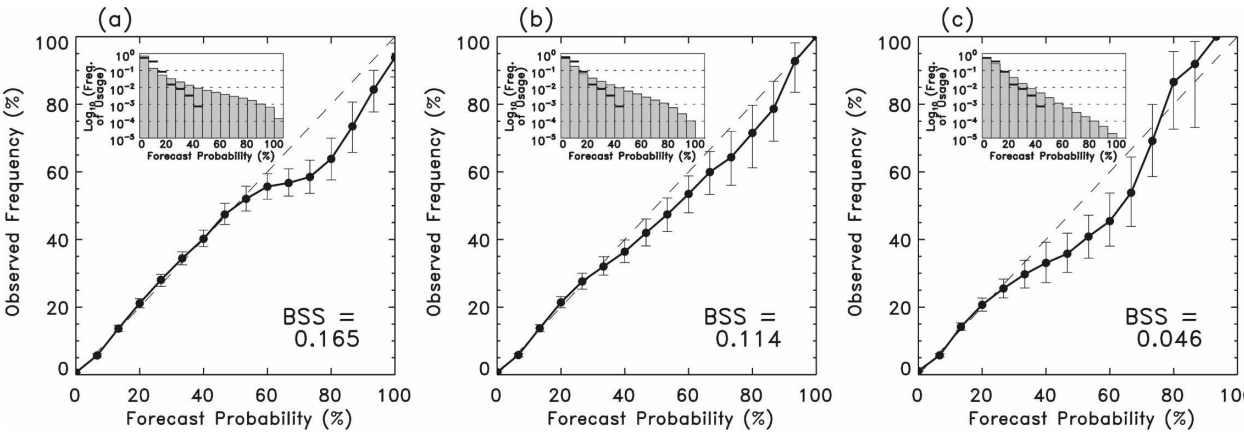

FIG. 8. Same as Fig. 4, but for 5-mm calibrated multimodel forecasts.

relative to the ECMWF-calibrated forecasts, unlike the temperature results in Part I. The differences between the multimodel 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 might have slightly improved the skill of the calibrated multimodel forecasts.

For brevity, we have chosen not to show reliability diagrams at other tested thresholds. Generally, the forecasts were more reliable at the lower thresholds and less so at the higher ones. High probabilities were issued much less frequently at the high thresholds, so the confidence intervals plotted over the reliability diagrams were much larger for the high probabilities.

b. Brier skill scores with weekly training data

Figure 9 shows BSS as a function of forecast lead and threshold for raw and calibrated forecasts. Calibrated multimodel 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 10 demonstrates in part why this occurs. Here, the ECMWF forecast characteristics changed diurnally, tending to overforecast significant rainfall events during 1200–0000 UTC. This overforecast bias was much less pronounced at 0000–1200 UTC. GFS forecasts had an even more pronounced daytime overforecast bias. Also, the diurnal oscillations in skill were large, in part because of the use of Eq. (5) for calculating the BSS. The conventional method of BSS calculation lessens weights on samples that have a small climatological event frequency relative to those with a more moderate event frequency, but Eq. (5) provides approximately equal weight to all samples. Because of this, in climatologically dry locations a small diurnal change in the precipitation bias may result in a large change in forecast skill, and the negative impact will be felt more fully using Eq. (5).

Several other characteristics of the BSS for the raw forecasts can be noted in Fig. 9. Especially at the higher thresholds, forecast skill actually was negative at early leads and increased somewhat with forecast lead, a counterintuitive result. This was caused primarily by 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. 4 and 5. The BSS heavily penalized these unrealistically sharp forecasts at the early leads, especially in the new method of calculation [Eq. (5)].

The characteristics of greatest interest in Fig. 9 are the skill differences between ECMWF and GFS forecasts and the amount of skill improvement resulting from forecast calibration. As with temperature in Part I, 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 very small bootstrap confidence intervals.

Positive skill was noted in both ECMWF and GFS-calibrated forecasts at 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.

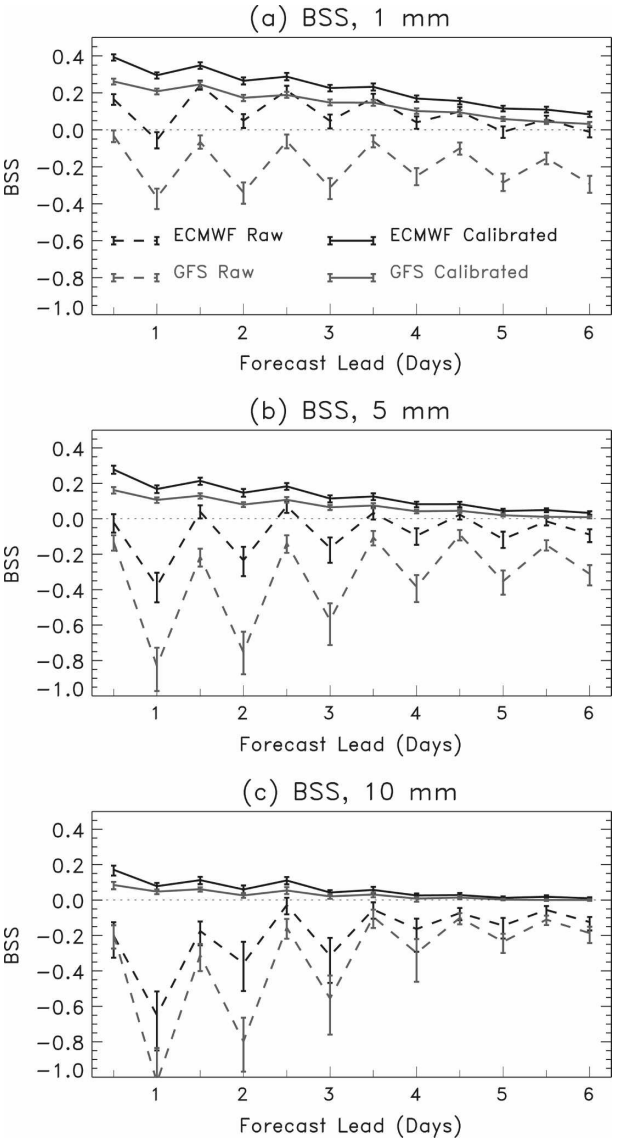

FIG. 9. Forecast BSS at (a) 1-, (b) 5-, and (c) 10-mm precipitation thresholds for ECMWF and GFS raw and calibrated forecasts (multimodel-calibrated forecasts are not plotted; they were statistically indistinguishable from ECMWF-calibrated forecasts). Plotted confidence intervals show the 5th and 95th percentiles as determined through block bootstrap resampling techniques.

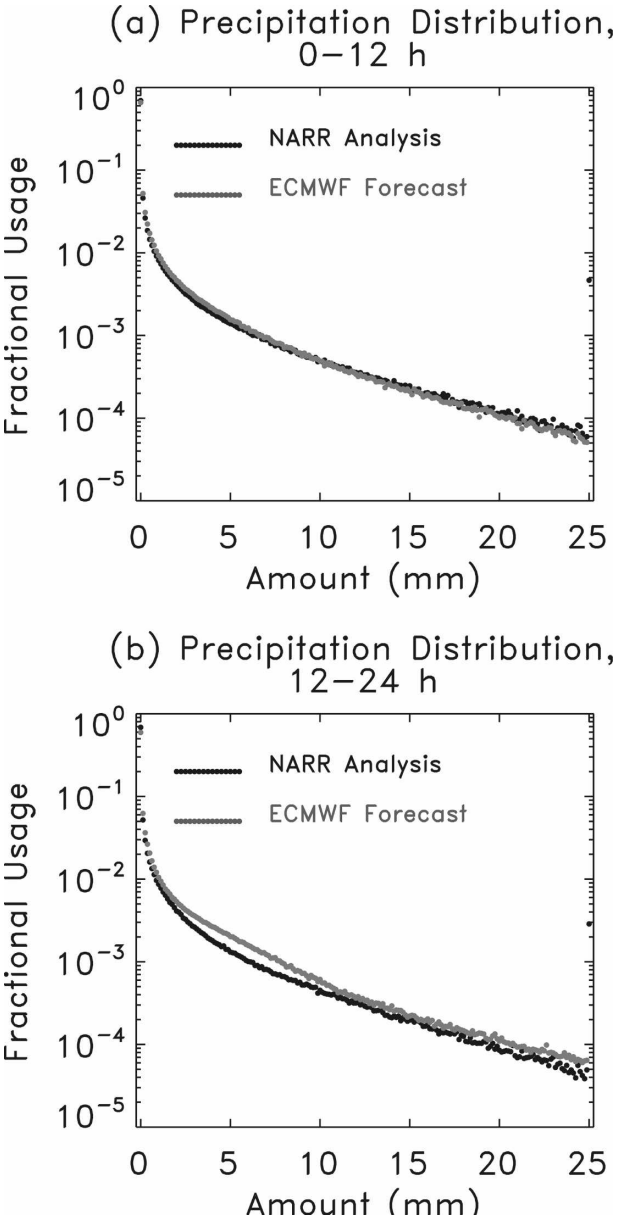

FIG. 10. Distribution of the fractional usage of precipitation amounts from the NARR and ECMWF raw forecasts for (a) 0–12-h and (b) 12–24-h forecasts.

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.

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

With the temperature forecasts in Part I, a 30-day training dataset provided a skill increase at short leads that was nearly equivalent to the skill increase when using the 20-yr weekly reforecast datasets. We return to

consider whether short training datasets are similarly adequate for precipitation forecast calibration. 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 (the last available 30 days), weekly (the once-weekly, 20-yr reforecast dataset), and full (for the GFS, 25 yr of once-daily September–November reforecasts and observations) training data. Weekly and 30-day training datasets used the compositing technique whereby training data was supplemented from 10 other grid points with similar analyzed climatologies.

Figure 11 shows the positive impact of the weekly training datasets. Although the degradation of skill was not particularly large at the 1-mm threshold, at 5 and especially 10 mm the degradation of forecast skill with the 30-day training dataset relative to the weekly dataset was quite large. At 10 mm, the improvement from using weekly ECMWF reforecasts compared to 30-day data 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 greatly improve in skill when daily samples were used (without calibrating using the analog locations) relative to the GFS weekly data using the analog locations. This suggests that for this application, daily reforecasts may not be necessary.

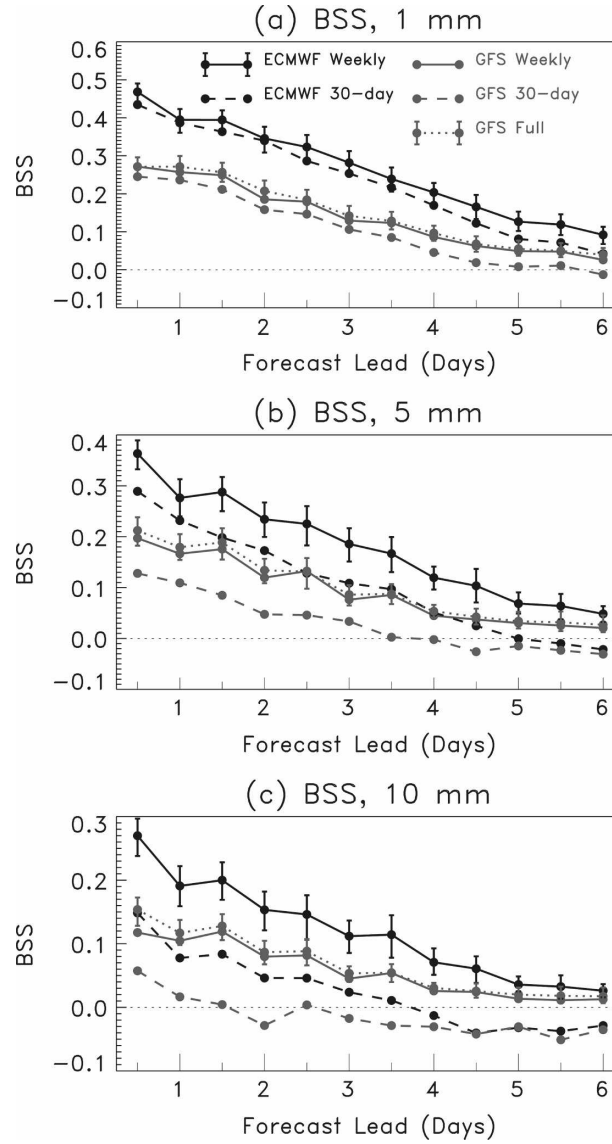

FIG. 11. BSS of daily forecasts from 1 Sep–1 Dec 2005 for (a) 1-, (b) 5-, and (c) 10-mm thresholds. Error bars indicate the 5th and 95th percentiles of the ECMWF weekly and GFS full resampled distribution of scores. Error bars for other forecasts were similar in magnitude.

5. Conclusions

This article considered the calibration of probabilistic short-term precipitation forecasts and how much training data were needed from a stable model and data assimilation system to produce an effective calibration. Two sources of forecasts were considered: ensemble forecasts from a T255, 2005 version of ECMWF's ensemble prediction system and GFS forecasts from a T62, 1998 version. ECMWF reforecasts were available once every week from 1982 to 2001, 1 September to 1 December. Daily forecast data were also available in the fall of 2005. GFS reforecast data were available every day from 1979 to the present and could be subsampled to the dates of the ECMWF reforecasts to facilitate comparison. Here, 12-hourly NARR data were used for training and validation.

Precipitation forecast calibration, as this article has shown, is very different in character from temperature forecast calibration. For temperature calibration in Part I, a small training dataset was adequate for calibration of short-lead forecasts, although longer-lead forecasts were better calibrated with more training data. With precipitation, calibration using a small 30-day training dataset improved forecast skill much less than calibration using the 20-yr weekly reforecasts. The difference was greater at higher precipitation thresholds; that is, the rarer the event, the more training data were needed. This result confirms a similar result with GFS reforecasts (Hamill et al. 2006, their Fig. 7).

Another important result from this study is 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 precipitation forecasts (Hamill et al. 2006; Hamill and Whitaker 2007) might be attributable to the low baseline set by the now-outdated 1998 GFS. The 2005 ECMWF system, arguably, is still representative of circa 2007–08 systems for many other operational forecast centers, given ECMWF's substantial lead in probabilistic forecast skill (Buizza et al. 2005). Hence, here and in Part I, we have shown the usefulness of large reforecast datasets, in particular for the calibration of forecasts of heavy precipitation and longer-lead temperature forecasts.

Taken together, Parts I and II indicate that large improvements in forecast skill and reliability are possible through the use of reforecasts, even with a modernized forecast model. For heavy precipitation or long-lead temperature forecasts, the extra training data were especially valuable. What these articles have not discussed, however, is just what the optimal reforecast configuration should be. How many members should be in the reforecast ensemble? How many years? Should a reforecast be performed every day, every third day, or every week? Hamill et al. (2004), reinforced by the positive results from weekly samples here, found that for these particular applications, weekly data over a period of decades were adequate. For other applications such as hydrologic flood forecasting, however, daily samples may still be preferable. Our unpublished results have suggested that much of the benefit can be obtained from a 5-member reforecast. We hope to examine these issues in future work.

Acknowledgments. Publication of this manuscript was supported by a NOAA THORPEX grant.

REFERENCES

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

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting,* **22,** 651–661.

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

Daly, C., R. P. Neilson, and D. L. Philips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.,* **33,** 140–158.

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

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 436 pp.

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

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration, and sharpness. *J. Roy. Stat. Soc.,* **69B,** 243–268.

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

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

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

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.,* **134,** 3209–3229.

——, and ——, 2007: Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Mon. Wea. Rev.,* **135,** 3273–3280.

——, ——, and X. Wei, 2004: Ensemble reforecasting: 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.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

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. Roy. Meteor. Soc.,* **126,** 1171–1190.

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

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

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

West, G. L., W. J. Steenburgh, and W. Y. Y. Cheng, 2007: Spurious grid-scale precipitation in the North American Regional Reanalysis. *Mon. Wea. Rev.,* **135,** 2168–2184.

Whitaker, J. S., X. Wei, and F. Vitart, 2006: Improving week-2 forecasts with multimodel 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.