Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures

RENATE HAGEDORN

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

THOMAS M. HAMILL AND JEFFREY S. WHITAKER

NOAA/Earth System Research Laboratory, Boulder, Colorado

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

ABSTRACT

Recently, the European Centre for Medium-Range Weather Forecasts (ECMWF) produced a reforecast dataset for a 2005 version of their ensemble forecast system. The dataset consisted of 15-member reforecasts conducted for the 20-yr period 1982–2001, with reforecasts computed once weekly from 1 September to 1 December. This dataset was less robust than the daily reforecast dataset produced for the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS), but it utilized a much higher-resolution, more recent model. This manuscript considers the calibration of 2-m temperature forecasts using these reforecast datasets as well as samples of the last 30 days of training data. Nonhomogeneous Gaussian regression was used to calibrate forecasts at stations distributed across much of North America. Significant observations included the following: (i) although the "raw" GFS forecasts (probabilities estimated from ensemble relative frequency) were commonly unskillful as measured in continuous ranked probability skill score (CRPSS), after calibration with a 20-yr set of weekly reforecasts their skill exceeded that of the raw ECMWF forecasts; (ii) statistical calibration using the 20-yr weekly ECMWF reforecast dataset produced a large improvement relative to the raw ECMWF forecasts, such that the ~4–5-day calibrated reforecast-based product had a CRPSS as large as a 1-day raw forecast; (iii) a calibrated multimodel GFS/ECMWF forecast trained on 20-yr weekly reforecasts was slightly more skillful than either the individual calibrated GFS or ECMWF reforecast products; (iv) approximately 60%–80% of the improvement from calibration resulted from the simple correction of time-averaged bias; (v) improvements were generally larger at locations where the forecast skill was originally lower, and these locations were commonly found in regions of complex terrain; (vi) the past 30 days of forecasts were adequate as a training dataset for short-lead forecasts, but longer-lead forecasts benefited from more training data; and (vii) a small but consistent improvement was produced by calibrating GFS forecasts using the full 25-yr, daily reforecast training dataset versus the subsampled, 20-yr weekly training dataset.

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 dataset used was a reduced-resolution, T62, 28-level, circa-1998 version of the Global Forecast System (GFS) from the National Centers for Environmental Prediction (NCEP). Fifteen-member forecasts were available to 15-day leads for every day from 1979 to the present. With a stable data assimilation and forecast system, the systematic errors of the forecast could be readily diagnosed and corrected. Calibrations using reforecasts were able to adjust the forecasts to achieve substantial improvements in their skill and reliability, commonly to levels competitive with or exceeding those achieved by cur-

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: 10.1175/2007MWR2410.1

rent-generation ensemble forecast systems without calibration.

The GFS model version used in these reforecast studies is now ~10 yr 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, the European Centre for Medium-Range Weather Forecasts (ECMWF) produced a more limited reforecast dataset 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-yr period from 1982 to 2001. Each forecast was run to a 10-day 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 yr or so.

This dataset allows us to ask and answer questions about reforecasts that were not possible with only the GFS dataset. Some relevant questions include: (i) 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? (ii) 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? (iii) Is a calibrated, multimodel combination more skillful than that provided solely by the ECMWF system? (iv) How much of the benefit of calibration in a state-of-the-art model can be obtained using only a short time series of past forecasts and observations?

This article will consider the problem of the calibration of probabilistic calibration of 2-m temperature forecasts. A companion article (Hamill et al. 2008) will discuss the calibration of 12-hourly accumulated precipitation forecasts. The calibration problems for each are unique; as will be shown, temperature forecasts tend to have more Gaussian errors and substantial improvements can be obtained with relatively short training datasets. Calibration of nonnormally distributed precipitation is more difficult, and larger samples tend to be needed to calibrate the more rare events.

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. ECMWF forecast data

The ECMWF reforecast dataset 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 dataset 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 a 10-day lead. The 15 forecasts consisted of a 40-yr ECMWF Re-Analysis (ERA-40) 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). Although data are available to cover the entire globe, for this study the model forecasts were extracted on a 1° grid from 15° to 75°N and 45° to 135°W, covering the conterminous United States and most of Canada. From this 1° grid, forecasts were bilinearly interpolated to the observation locations, described below.

In addition, the ECMWF 0000 UTC forecasts in the year 2005 were extracted for every day from 1 July to 1 December. These additional data permit experiments comparing short training datasets with the reforecasts. The 2005 forecasts were initialized with the operational four-dimensional variational data assimilation (4DVAR) system (Mahfouf and Rabier 2000), rather than the three-dimensional variational data assimilation (3DVAR) analysis of ERA-40.

b. GFS forecast data

The GFS reforecast dataset, more completely described in Hamill et al. (2006), was utilized here. It utilizes a T62, 28-sigma-level, circa-1998 version of the GFS. Fifteen-member forecasts are available to 15-day leads 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 data were also bilinearly interpolated to surface observation locations. For most of the experiments described here, the GFS reforecasts were subsampled to the dates of the ECMWF reforecast dataset to permit ease of comparison. However, some experi-

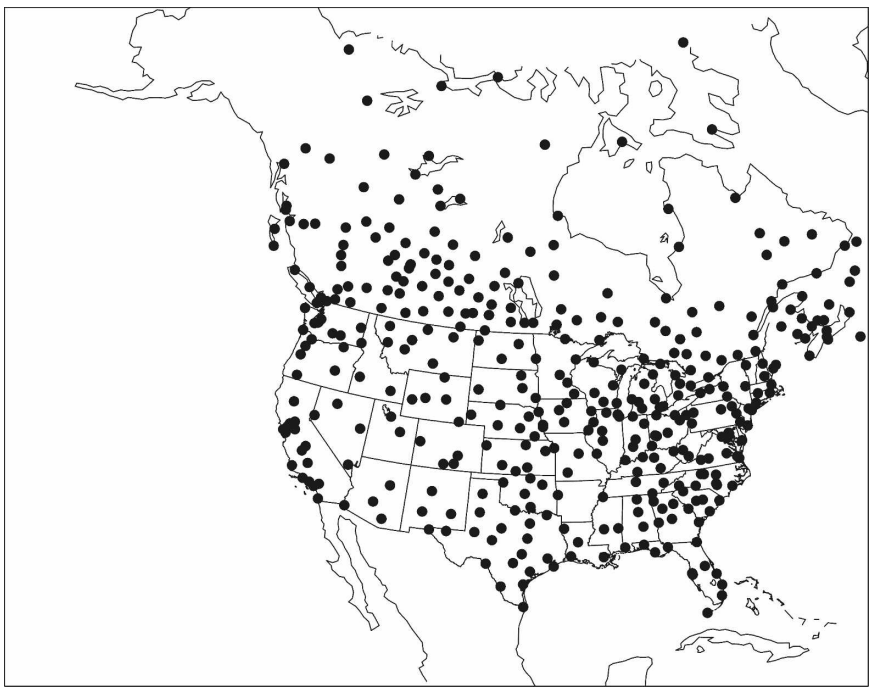

FIG. 1. Station locations where probabilistic 2-m temperature forecasts are evaluated.

ments utilized 25-yr (1979–2003) daily samples of reforecast training data.

c. *Two-meter temperature observations*

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

3. Calibration and validation methodologies

a. *Calibration with nonhomogeneous Gaussian regression*

Many methods may be used for the calibration of 2-m temperature forecasts; among those in the recent literature are rank histogram techniques (Hamill and Colucci 1998; Eckel and Walters 1998), ensemble dressing (Roulston and Smith 2003; Wang and Bishop 2005), Bayesian model averaging (Raftery et al. 2005), logistic regression (Hamill et al. 2006), analog techniques (Hamill and Whitaker 2007), and nonhomogeneous Gaussian regression (Gneiting et al. 2005). Wilks and Hamill (2007) provide an intercomparison of several of these techniques. In the intercomparison, nonhomogeneous Gaussian regression was determined to be more skillful than or nearly as skillful as the other candidate techniques. Accordingly, we shall use it as the calibration technique of choice here.

Nonhomogeneous Gaussian regression (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, because of the limited number of members and system errors, the ensemble sample variance by itself may not properly estimate the forecast uncertainty. Accordingly, the regression variance was allowed to be nonhomogeneous (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 the observed 2-m temperatures 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; that is, 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 $\sim N(\alpha, \beta)$ de-

note that a random variable has a Gaussian distribution with mean α and variance β. Let $\bar{x}_{\text{ens}}$ denote the interpolated ensemble mean and s^2_{ens} denote the ensemble sample variance. Then NGR estimates regression coefficients a, b, c, and d to fit $N(a + b\bar{x}_{\text{ens}}, c + ds^2_{\text{ens}})$. When $d = 0$, there is no spread–error relationship in the ensemble, and the resulting distribution resembles the form of linear regression, with its constant-variance assumption. Following Gneiting et al. (2005), the four coefficients are fit iteratively to minimize the continuous ranked probability score (CRPS; e.g., Wilks 2006).

In all experiments using the weekly reforecast data, cross validation was utilized in the regression analysis. The year being forecast was excluded from the training data; for example, 1983 forecasts were trained with 1982 and 1984–2001 data. Also, because biases can change with the seasons, the full set of September–December data was not used as training data. Rather, only the 5 weeks centered on the date of interest were used; thus, when training for 15 September, the training data comprised the 1, 8, 15, 22, and 29 September forecasts. For dates at the beginning and end of the reforecast, a noncentered training dataset was used; for example, the training dates for 1 September were 1, 8, and 15 September. Unless otherwise noted, the GFS reforecast data were subsampled to the same weekly dates of the ECMWF training dataset. However, some later experiments include a comparison with forecasts trained using daily GFS reforecast data from 1979–2003.

A slightly more complicated version of NGR was used for production of a calibrated multimodel ECMWF/GFS forecast. The first step was to perform a linear regression analysis of each model's ensemble-mean forecast against the observations separately for each forecast lead time. The result was an equation to predict the lowest root-mean-square error (rmse) forecast from each system's raw ensemble-mean forecast. Denote this corrected mean forecast as $\bar{x}_{\text{EC}}(k, l)$ from the ECMWF model for the kth of K training samples and lth of L locations, and similarly $\bar{x}_{\text{GPS}}(k, l)$ for the GFS. Denote the deviation of the ith of m ECMWF members from its mean as $x^i_{\text{EC}}{}'(k, l)$, and similarly $x^i_{\text{GPS}}{}'(k, l)$ for the GFS. Let D^2_{EC} denote the average squared difference between the regression-corrected ECMWF ensemble-mean forecast and observations; thus,

$$D^2_{\text{EC}}(l) = \frac{1}{K}\sum_{k=1}^{K}[\bar{x}_{\text{EC}}(k, l) - o(k, l)]^2, \qquad (1)$$

where $o(k, l)$ is the observation. The squared difference for the GFS, $D^2_{\text{GFS}}(l)$, is similarly defined.

We now seek to determine a multimodel weighted mean forecast and sample variance, providing larger weights to the forecasts with the smaller squared differences.

The weight to apply to the ECMWF forecasts [Daley 1991, his Eq. (2.2.3)] is defined as

$$W_{\text{EC}}(l) = \frac{D^2_{\text{GFS}}(l)}{D^2_{\text{GFS}}(l) - D^2_{\text{EC}}(l)}, \qquad (2)$$

and $W_{\text{GFS}} = 1.0 - W_{\text{EC}}$. A weighted multimodel ensemble mean was calculated as

$$\bar{x}_{\text{MM}}(k, l) = W_{\text{EC}}(l)\bar{x}_{\text{EC}}(k, l) + W_{\text{GFS}}(l)\bar{x}_{\text{GFS}}(k, l), \qquad (3)$$

and a weighted multimodel ensemble variance was calculated as

$$s^2_{\text{MM}} = W_{\text{EC}}(l)\frac{\sum_{i=1}^{m}[x^i_{\text{EC}}{}'(k, l)]^2}{m - 1} + W_{\text{GFS}}(l)\frac{\sum_{i=1}^{m}[x^i_{\text{GFS}}{}'(k, l)]^2}{m - 1}. \qquad (4)$$

These multimodel means and sample variances are then input into the NGR to produce the regression coefficients *a, b, c*, and *d*. A given forecast day's ensemble forecasts were processed using the same procedure as the training data [Eqs. (2)–(4)] to produce a multimodel mean and sample variance, and the regression coefficients were applied to determine the parameters of the fitted NGR distribution.

b. Validation procedures

1) RANK HISTOGRAMS

Reliability characteristics of the probabilistic 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 for observation and representativeness errors (Hamill 2001, his section 3c). The choice of 1.5°C was somewhat arbitrary but was generally consistent with the observation errors assigned to surface data in data assimilation schemes (Parrish and Derber 1992). Probably somewhat less random error should be added to the ECMWF forecasts than to the GFS forecasts because the ECMWF grid spacing is smaller, lessening the representativeness error; nonetheless, the random error was set the same for both forecasts.

Rank histograms assess the rank of the observed

relative to ensemble member forecasts; that is, the observed rank is relative to discrete samples from a probability density function (PDF) rather than the PDF itself. How then can the rank histogram be used to assess the reliability of a fitted PDF? We used the following approach, motivated by the probability integral transform (Casella and Berger 1990, p. 52). The original ensembles were comprised of $m = 15$ members, so we constructed 15 sample members where the value of the ith fitted member was defined as $x_{\text{fit}}(i) = q_{i/(m+1)}$, the $i/(m + 1)$th quantile of the fitted distribution. The m-constructed ensemble members defined the boundaries between $m + 1$ equally probable bins under the null hypothesis that the observed value was a random draw from the same underlying distribution as the ensemble.

Then $x_{\text{fit}}(i)$ was remapped from the $i/(m + 1)$th quantile $q^N_{i/(m+1)}$ of a standard normal distribution. Specifically, given the coefficients *a, b, c,* and *d* that define the fitted forecast for this sample, then

$$x_{\text{fit}}(i) = q^N_{i/(m+1)}(c + ds^2_{\text{ens}}) + (a + b\bar{x}_{\text{ens}}) \qquad (5)$$

[Wilks 2006, his Eq. (4.25)]. The rank of the observed value relative to $x_{\text{fit}}(1), \ldots, x_{\text{fit}}(m)$ was computed, and the process was repeated for all forecast samples to generate the rank histogram. Because the underlying fitted distribution was determined by training against real, imperfect observations, there was no need to perturb the ensemble members with observation noise, as was done with the raw ensemble.

2) Spread, error, and fractional bias

Ideally, an ensemble forecast system ought to have a similar magnitude of its spread and rmse (e.g., Whitaker and Loughe 1998). Plots of averages of these quantities are shown later, where the ECMWF's model spread at a given lead time σ_{EC} is defined as

$$\sigma_{\text{EC}} = \left\{ \frac{1}{KL} \sum_{l=1}^{L} \sum_{k=1}^{K} [x^i_{\text{EC}}{}'(k, l) + \varepsilon(k, l)]^2 \right\}^{1/2}, \qquad (6)$$

where $\varepsilon(k, l) \sim N[0, (1.5)^2]$. That is, the spread calculated here is calculated from ensemble perturbations from the ensemble mean plus a random realization of noise, sampled from a normal distribution with zero mean and a standard deviation of 1.5°C. This is presumed to represent the observation error, as done previously with the rank histograms. The rmse, RMS_{EC}, is defined as

$$\text{RMS}_{\text{EC}} = \left\{ \frac{1}{KL} \sum_{l=1}^{L} \sum_{k=1}^{K} [\bar{x}_{\text{EC}}(k, l) - o(k, l)]^2 \right\}^{1/2}. \qquad (7)$$

The fractional bias BF_{EC} is used to diagnose how much of the ensemble-mean forecast error can be attributed to bias, as opposed to random error. It is defined as

$$\text{BF}_{\text{EC}} = \left| \frac{\sum_{l=1}^{L} \sum_{k=1}^{K} [\bar{x}_{\text{EC}}(k, l) - o(k, l)]}{\sum_{l=1}^{L} \sum_{k=1}^{K} [|\bar{x}_{\text{EC}}(k, l) - o(k, l)|]} \right|. \qquad (8)$$

The spread (σ_{GFS}), error (RMS_{GFS}), and fractional bias (BF_{GFS}) of the GFS forecasts are similarly defined.

3) Continuous ranked probability skill score

Calculation of a revised version of the continuous ranked probability skill score (CRPSS) 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 among samples rather than to any inherent forecast skill. Here we followed the specific method outlined in Hamill and Whitaker (2007) to ameliorate this problem. The idea was simple: divide the overall forecast sample into subgroups where the climatological uncertainty was approximately homogeneous, determine the CRPSS for each subgroup, and then determine the final CRPSS 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 $\overline{\text{CRPS}}^f(s)$ denote the average forecast CRPS (Wilks 2006) for the sth subgroup, and $\overline{\text{CRPS}}^c(s)$ denote the average CRPS of the climatological reference forecast for this subgroup. Then the overall CRPSS is calculated as

$$\text{CRPSS} = \frac{1}{\text{NC}} \sum_{s=1}^{\text{NC}} \left[1 - \frac{\overline{\text{CRPS}}^f(s)}{\overline{\text{CRPS}}^c(s)} \right]. \qquad (9)$$

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 were computed following the block bootstrap procedure outlined

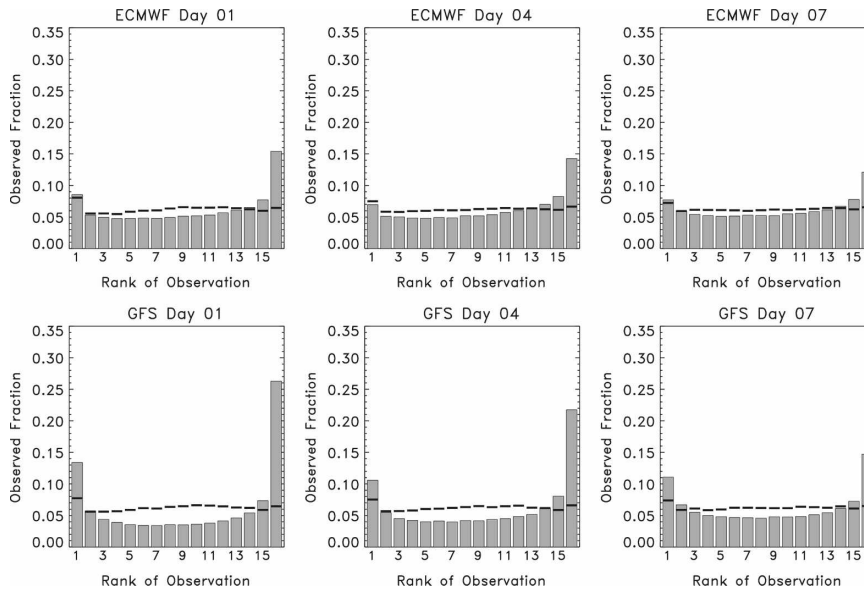

FIG. 2. Rank histograms for 2-m temperatures from (top) ECWMF and (bottom) GFS ensembles at (left) 1-, (middle) 4-, and (right) 7-day leads. Histograms denote the raw ensemble and solid lines the calibrated ensembles.

in Hamill (1999); in this case, 4000 iterations of a resampling procedure were used, shuffling the data in blocks of case days. The CRPSS was computed using Eq. (8) for the two resampled sets, and the difference in CRPSS was used to build up the distribution for the null hypothesis. Confidence interval data are not plotted here; for the 20-yr ECMWF reforecast experiments, the 95% confidence intervals for calibrated versus raw ensembles were small, from ± 0.033 at the half-day lead to ± 0.02 at the 10-day lead.

4. Results

a. Twenty-year weekly training data

Figure 2 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 for ECMWF forecasts. After calibration with NGR, the rank histograms were much flatter, although there still was some slight excess of population of the lowest rank. Probably the assumption of Gaussianity underlying the NGR was not strictly appropriate; although forecast PDFs may have somewhat more Gaussian distributions than climatology, notably 416 of the 439 stations exhibited a negative skew of their observed 2-m temperature distributions.

The general similarity of the rank histogram shapes from the ECWMF and GFS ensembles may be somewhat misleading as to the characteristics of these ensembles. Figure 3 provides a plot of average spreads [the standard deviations of the ensemble perturbations about their means plus observation noise with variance R; Eq. (6)] and the rmses [Eq. (7)] from the raw ensembles. In a perfect ensemble forecast where ensemble spread is due solely to chaotic growth of initial condition errors, these two curves should lie on top of each other. Neither the ECMWF nor the GFS ensembles had a spread nearly as large as the rmse, indicating that model biases were large. However, the rmse of the ECMWF ensemble was substantially smaller than that of the GFS, indicating that its forecasts should have higher skill.

We now consider the overall CRPSS of the calibrated and uncalibrated forecasts in Fig. 4. Several main points can be made. First, as suggested by Fig. 3, the raw ECMWF forecasts were indeed more skillful than the GFS forecasts. Second, although 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 benefitted greatly from the statistical correction. Although 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, the multimodel NGR forecast consistently outperformed the calibrated ECMWF forecast by a

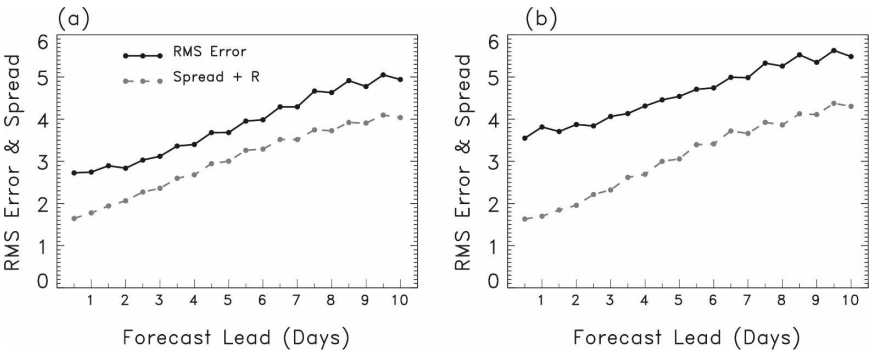

FIG. 3. Average ensemble spread (with members additionally perturbed with random sample of observation error drawn from R) and rmse of 2-m temperature forecasts from the (a) ECMWF and (b) GFS ensembles.

small amount, indicating that there was some independent information provided by the older, less sophisticated GFS. This is consistent with many previous results from the combination of information from multiple models using smaller training datasets (e.g., Vislocky and Fritsch 1995, 1997; Krishnamurti et al. 1999). Fifth, the forecast skills have a slight stair-step appearance—primarily because the reference climatological CRPS are larger for the 0000 UTC forecasts (days 1, 2, etc.) than for the 1200 UTC forecasts (days 0.5, 1.5, etc.)—which, following Eq. (9), will result in higher skills, assuming a smaller (or no) diurnal variation in the forecast CRPS. Finally, note that even at day 10 there is still some skill in the calibrated ECMWF and multimodel 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 5 demonstrates 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% of the improvement in skill in the ECMWF forecasts can be attributed to this simple bias correction; the NGR added the remaining 20%–40% 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.

Figures 6a–c show the geographic distributions of day 2 skill for the raw, NGR, and bias-corrected forecasts, respectively. The raw forecasts were commonly deficient in skill in the complex terrain of the western United States and Canada, presumably because the simplified terrain heights of the forecast model differed from that of the actual stations, with concomitant errors in the estimation of surface temperatures. It appeared that a simple bias correction achieved most of the impact for the stations with particularly unskillful raw forecasts. This is demonstrated in Fig. 6d. Here, the fractional improvement of the bias correction is plotted as a function of the raw and calibrated forecasts. Letting C_{RAW}, C_{NGR}, and C_{BC} denote the CRPSS of the raw, calibrated, and bias-corrected forecasts, respectively, the fractional improvement Fr is computed as $(C_{\mathrm{BC}} - C_{\mathrm{RAW}})/(C_{\mathrm{NGR}} - C_{\mathrm{RAW}})$. Figure 6d shows several interesting characteristics. First, note that the effect of the NGR calibration was primarily to improve forecasts that started off as particularly unskillful by homogenizing the resultant skill relative to the highly varying skills seen in the raw forecasts. Second, in general the locations that had relatively large improvements through the NGR calibration achieved a greater fraction of this from the bias correction than did the locations that had smaller improvements. Overall, the

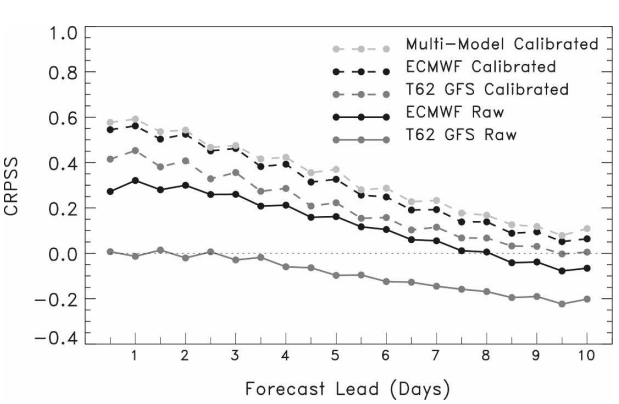

FIG. 4. CRPSS of surface temperature forecasts with and without calibration.

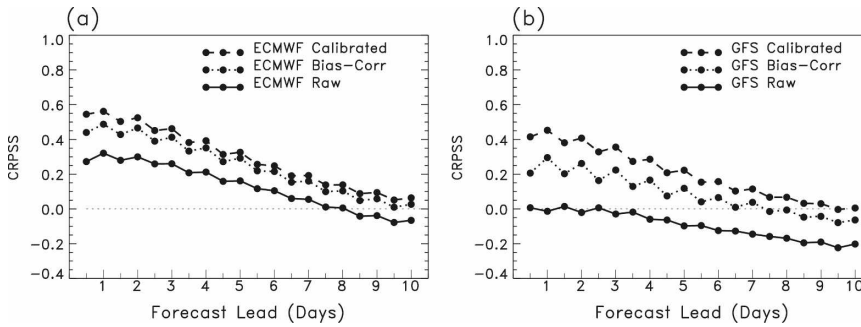

FIG. 5. CRPSS including bias-corrected ensemble forecasts for (a) ECMWF and (b) GFS forecasts.

large improvements from bias corrections may indicate that additional resolution may be helpful, leading to smaller mismatches between model terrain height and station elevation (see also Buizza et al. 2007).

b. *Differences between 20-yr weekly and 30-day daily training datasets*

To facilitate a comparison of long and short training datasets, 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 dataset. 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 dataset. Forecasts were compared for the period of 1 September–1 December 2005. Nonhomogeneous Gaussian regression was again used for the calibration. Figure 7 shows that at short forecast leads, the 30-day training dataset provided approximately equal skill improvements relative to the 20-yr training dataset for the ECMWF model, and marginally less for the GFS. However, as the forecast lead increased, then the benefit of the longer training dataset became 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 dataset 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). If errors were synoptically dependent and a regime change took place in the intervening 9 days, the training set at 1-day lead will have had samples from the new regime but the training set at 10-days lead will not. Second, determining the bias to a prespecified tolerance will require more samples at long leads than at short leads. At these long leads, the proportion of the error attributable to bias shrinks because of the rapid increase of errors due to chaotic error growth. This is shown in Fig. 8; for the ECMWF model, this decreased from ~0.54 at the half-day lead to ~0.28 at the 10-day lead. Consequently, because the overall error grows as the forecast lead increases and a larger proportion of it is attributable to random errors, determining the bias to a prespecified tolerance requires more samples. The third reason was that the short-lead forecast training datasets were composed of samples that tended to have more independent errors than the longer-lead training datasets. The ECMWF 1-day lagged correlation of forecast minus observed values 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),

$$n' = n\frac{1-\rho_1}{1+\rho_1}, \tag{10}$$

with $n = 30$, this indicates that the effective sample size was approximately 20 at the short leads and 10 at the longer leads. The once weekly, 20-yr reforecast dataset should, in comparison, be composed of samples that are truly independent of each other.

Considering again the puzzling result of similar skill at short leads, we hypothesize that the two factors here may have contributed to underestimating the potential skill that can be obtained with a properly constructed long training dataset. First, one limitation of the ECMWF datasets was that for the 2005 data, all forecasts were initialized with 4DVAR, but the

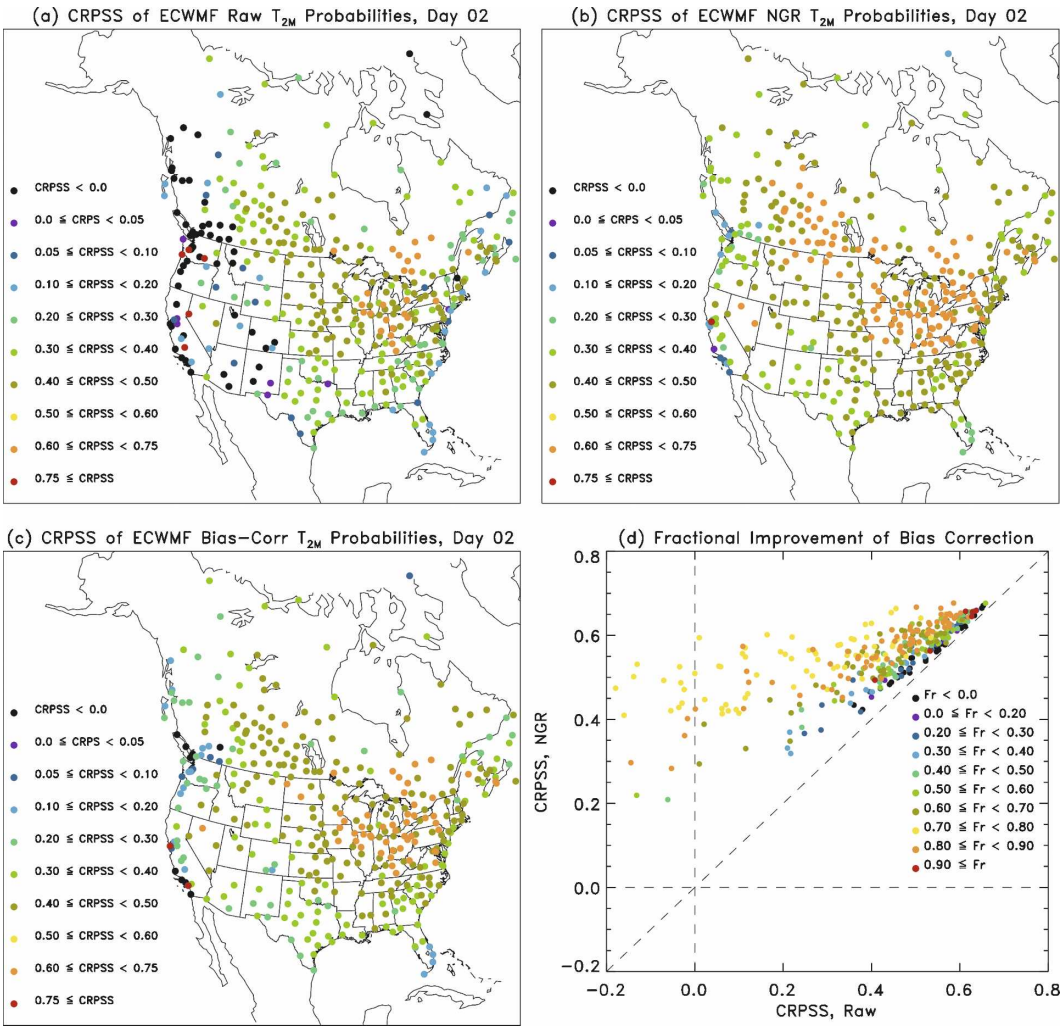

FIG. 6. CRPSS of raw 2-day forecasts from (a) the ECMWF model, (b) as in (a), but for calibrated NGR forecasts, and (c) as in (a), but for bias-corrected forecasts. (d) Fractional improvement Fr gained from bias correction as a function of the CRPSS from raw and NGR forecasts.

1982–2001 reforecast data were initialized with 3DVAR. It is thus possible that the ECMWF short-term reforecasts may have subtly different biases than the 2005 real-time forecasts, differences that may diminish with the forecast lead. This would affect the calibration of the short-term forecasts. Notice that Fig. 7b shows a somewhat larger benefit from long training datasets with the GFS, where a consistent data assimilation system was used. Second, the calibration with the full reforecast training dataset here used only the model forecast temperature as a predictor. Perhaps the short training dataset benefits from having samples with a more similar set of land surface conditions. If this is the case, then perhaps a multipredictor regression analysis including, say, soil moisture content as an additional predictor would improve the reforecast calibrations.

c. *Differences between 20-yr weekly and 25-yr daily training datasets in the GFS*

Figure 9 shows the CRPSS of GFS forecasts from the raw ensemble after a calibration with the 20-yr weekly training dataset and with the full 25-yr daily training dataset. When training with the 25-yr daily data, the training data were used in a window of ±15 days around the date being forecast; for example, forecasts for 16 September used 1 September–1 October reforecasts for training data. Data for the year being forecast were excluded (cross validation), and all days between 1 September and 1 December were validated, as opposed to the weekly samples used in Fig. 4. As Fig. 9 shows, the 25-yr, daily training dataset provided a small but consistent improvement over the 20-yr, weekly

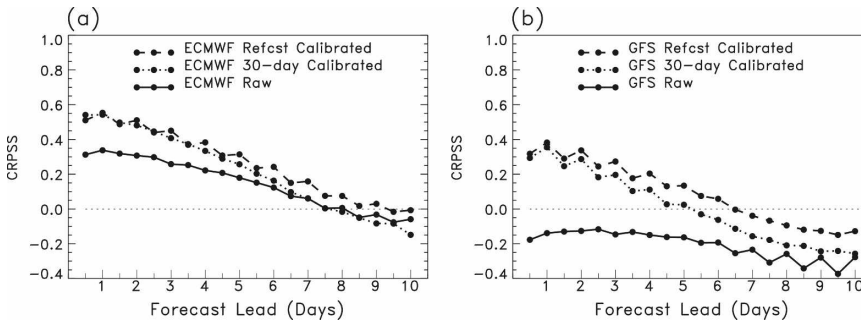

FIG. 7. Comparison of CRPSS using 30-day and 20-yr training datasets for the period 1 Oct–1 Dec 2005: (a) ECMWF and (b) GFS data.

training dataset. Why was the improvement not larger? First, of course, the baseline for the comparison used 20 yr of weekly forecasts × 5 weeks of centered data = 100 samples, a respectably large number for the estimation of the four NGR parameters. Further, as noted in Hamill et al. (2004), forecast errors may be correlated from one day to the next, so using daily versus weekly samples does not necessarily mean that the effective sample size (Wilks 2006, p. 144) will be 7 times larger with daily samples.

5. Conclusions

A prior series of articles (Hamill et al. 2004, 2006; Hamill and Whitaker 2006, 2007; Whitaker et al. 2006; Wilks and Hamill 2007) have discussed the benefit of calibrating probabilistic forecasts using the large training datasets from an ensemble reforecast dataset from a 1998 version of the NCEP GFS. This dataset is now 10 yr old, and it is not clear whether the large positive benefits from the large training dataset would still occur with a newer, higher-resolution model with its reduced systematic errors. Recently, ECMWF developed a limited reforecast dataset consisting of a once weekly, 15-member reforecast for the period 1 September–1 December 1982–2001. These forecasts were conducted using the model version operational in the second half of 2005, a T255-resolution version of the forecast model. Although the once weekly reforecasts were sparser than the daily reforecasts from the GFS, the ECMWF reforecast dataset still spanned 2 decades of diverse climatological regimes. Accordingly, we performed an analysis of the skill that can be gained from calibration of surface temperatures using these training datasets.

Both the ECMWF and GFS raw ensemble surface temperature forecasts were found to be biased and/or underdispersive, noted from the excess populations of the extreme ranks in their rank histograms. This tendency was more pronounced at the short forecast leads. However, after calibration with nonhomogeneous Gaussian regression (NGR), the rank histograms were flatter, although the lowest rank was still populated slightly more than was appropriate with a perfectly calibrated ensemble.

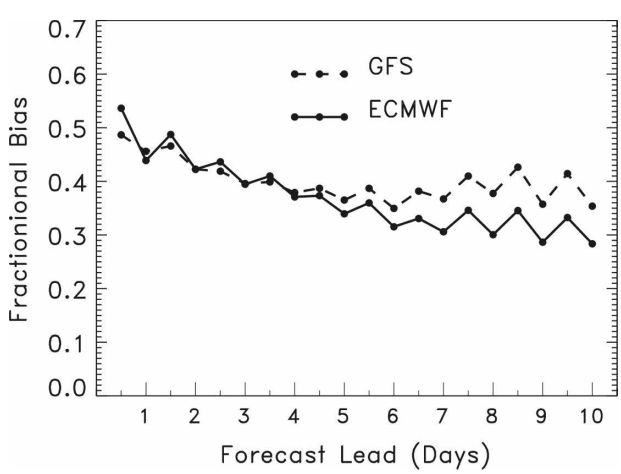

FIG. 8. Fractional bias, the fraction of the total rmse that can be attributed to systematic error, as a function of forecast lead.

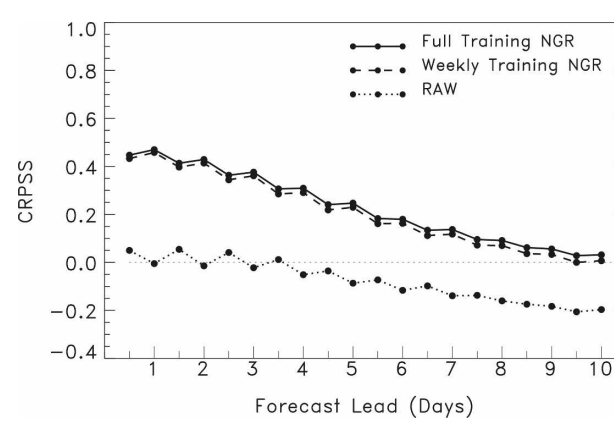

FIG. 9. CRPSS of GFS forecasts from raw ensemble, with 20-yr weekly training dataset and 28-yr daily training dataset.

The skill of these forecasts was measured with a modified version of the continuous ranked probability skill score (CRPSS), with the computation adjusted to remove the tendency to award fictitious skill due to variations in the forecast climatology (Hamill and Juras 2006). Climatology provided the no-skill reference. Using this skill measure, the raw GFS ensemble forecasts had near zero to negative skill at all leads due to the presence of large forecast biases. The ECMWF raw forecasts retained positive skill to approximately 8 days.

After calibration with NGR, the postprocessed GFS forecasts exceeded the skill of the uncalibrated ECMWF forecasts at all leads. Here, the GFS training data were subsampled to the same weekly, 20-yr set of dates as in the ECMWF reforecast. However, the reforecast-based, calibrated ECMWF forecasts were much more skillful than both the GFS calibrated forecasts and the ECMWF uncalibrated forecasts, although the absolute amount of skill increase from calibration was smaller for ECMWF than for the GFS. Nonetheless, the ECMWF skill improvement was substantial; for example, the skill of a calibrated, 4–5-day ECMWF forecast was comparable to the skill of an uncalibrated 1-day forecast. Approximately 70% of the improvement of the ECMWF could be attributed to a simple correction of mean bias in the forecasts, with a slightly smaller percentage in the GFS. The ECMWF raw forecasts were observed to have particularly low skill at stations in the intermountain western United States, perhaps due to larger discrepancies between the model terrain and the station locations. Calibration was particularly successful in increasing the skill at these stations. Finally, a multimodel calibrated forecast was more skillful than either individual calibrated forecast.

The computation of an extensive reforecast dataset is expensive, and a new reforecast dataset may be needed each time a model change affects its systematic error characteristics. If the same benefit could be achieved with a much smaller set of recent forecasts, this would make operational calibration much easier. Accordingly, using 2005 data, we compared the calibration using the 1982–2001 reforecasts to calibration using the most recent 30 samples of forecasts from 2005. For the shorter forecast leads, the skill after calibration using this shorter training dataset was very similar to that achieved with large reforecast dataset. We hypothesize that this benefit may be attributable to the more recent samples being more similar in their error characteristics than those from the reforecast dataset, which samples other years of data. However, at longer leads, the reforecast dataset produced more skillful calibrated forecasts than the 30-day training dataset. This was likely due to at least three reasons: first, 30 days of training data for the longer-lead forecasts were more separated from the actual forecast day of interest (e.g., when calibrating a 10-day forecast, the most recent training sample is 10 days old because verification is not yet available for the more recent forecasts). Second, the number of samples necessary to estimate the bias to a prespecified tolerance generally increased with increasing forecast lead. And third, for forecasts at the longer leads, the samples on adjacent days tended to have correlated forecast errors, thereby reducing the effective sample size.

Although a daily reforecast dataset was yet not available for the ECMWF model, the impact of daily versus weekly samples could be evaluated with the GFS reforecast dataset. Using a 25-yr daily reforecast versus a 20-yr weekly forecast produced a small but noticeable improvement.

It is also possible that the calibration could be improved by including other predictors. Here we considered only 2-m temperature as a predictor. Perhaps the reason the 30-day training dataset shows such good results is that the training samples are from a regime with similar surface characteristics, such as soil moisture. If so, then the performance of a multiyear reforecast could be enhanced by including soil moisture as an additional predictor. An examination of the potential value of several other predictors may be useful before any operational implementation of a temperature-calibration scheme.

This article considered only the calibration of 2-m temperature forecasts. Our experience with precipitation calibration using the GFS reforecasts suggests that the benefit from calibration using short training datasets will be smaller than for temperature. The companion article to this paper (Hamill et al. 2008) examines the calibration of ECMWF and GFS precipitation forecasts in more depth and provides substantial further evidence for the value of large training datasets, even with a state-of-the-art model. Nonetheless, the value of large training datasets for temperature calibration was confirmed here, even for a current, state-of-the-art forecast model. Short training datasets were adequate for the short-lead forecasts, but to achieve benefits at all forecast leads, the longer training dataset proved useful.

Combined with the evidence in our companion paper and previous studies, there is now a growing body of literature indicating the potential utility of reforecast methodology for improving operational ensemble predictions.

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

REFERENCES

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

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

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

——, J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Quart. J. Roy. Meteor. Soc.,* **133,** 681–695.

Casella, G., and R. L. Berger, 1990: *Statistical Inference*. Duxbury, 650 pp.

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., Z. Toth, Y. Zhu, D. Hou, D. Unger, and S. Beauregard, 2006: The trade-off in bias correction between using the latest analysis/modeling system with a short, versus an older system with a long archive. *Proc. First THORPEX Int. Science Symp.,* Montréal, QC, Canada, World Meteorological Organization, 281–284.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

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

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.

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.

——, and S. J. Colucci, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, 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.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.,* **136,** 2620–2632.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multi-model superensemble. *Science,* **285,** 1548–1550.

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.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus,* **55A,** 16–30.

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. Roy. Meteor. Soc.,* **131,** 2961–3012.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.

——, and ——, 1997: Performance of an advanced MOS system in the 1996-97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.,* **78,** 2851–2857.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.,* **131,** 965–986.

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-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.