Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application

Thomas M. Hamill and Jeffrey S. Whitaker

Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 17 October 2005, in final form 7 February 2006)

ABSTRACT

A general theory is proposed for the statistical correction of weather forecasts based on observed analogs. An estimate is sought for the probability density function (pdf) of the observed state, given today's numerical forecast. Assume that an infinite set of reforecasts (hindcasts) and associated observations are available and that the climate is stable. Assume that it is possible to find a set of past model forecast states that are nearly identical to the current forecast state. With the dates of these past forecasts, the asymptotically correct probabilistic forecast can be formed from the distribution of observed states on those dates.

Unfortunately, this general theory of analogs is not useful for estimating the global pdf with a limited set of reforecasts, for the chance of finding even one effectively identical forecast analog in that limited set is vanishingly small, and the climate is not stable. Nonetheless, approximations can be made to this theory to make it useful for statistically correcting weather forecasts. For instance, when estimating the state in a local region, choose the dates of analogs based on a pattern match of the local weather forecast; with a few decades of reforecasts, there are usually many close analogs.

Several approximate analog techniques are then tested for their ability to skillfully calibrate probabilistic forecasts of 24-h precipitation amount. A 25-yr set of reforecasts from a reduced-resolution global forecast model is used. The analog techniques find past ensemble-mean forecasts in a local region that are similar to today's ensemble-mean forecasts in that region. Probabilistic forecasts are formed from the analyzed weather on the dates of the past analogs. All of the analog techniques provide dramatic improvements in the Brier skill score relative to basing probabilities on the raw ensemble counts or the counts corrected for bias. However, the analog techniques did not produce guidance that was much more skillful than that produced by a logistic regression technique. Among the analog techniques tested, it was determined that small improvements to the baseline analog technique that matches ensemble-mean precipitation forecasts are possible. Forecast skill can be improved slightly by matching the ranks of the mean forecasts rather than the raw mean forecasts by using highly localized search regions for shorter-term forecasts and larger search regions for longer forecasts, by matching precipitable water in addition to precipitation amount, and by spatially smoothing the probabilities.

1. Introduction

Despite much recent progress in numerical weather prediction, weather forecasts are still subject to error, both as a result of the growth of initial-condition errors and model errors. Near-surface forecasts and forecasts of hydrologic variables such as precipitation or cloud properties are particularly error prone, in part because these physical processes often occur at scales below those resolved by the model. These effects must be parameterized, and developing accurate parameterizations is a difficult endeavor. As computational power increases, forecast models have been updated and increased in resolution to address these problems.

A complementary pathway to improved forecasts for users is to utilize a known weather forecast model consistently so that a long time series of past weather forecasts is available. If the climate is relatively stable, then the errors in past similar weather scenarios can be used to statistically correct the current numerical forecast. This approach is, of course, well established, being the essence of model output statistics (MOS) techniques (Glahn and Lowry 1972; Carter et al. 1989). If today's numerical forecast indicates relatively ordinary conditions, then perhaps the past few months or year will have exhibited enough other similar scenarios that the

Corresponding author address: Dr. Thomas M. Hamill, NOAA/ESRL/PSD, R/PSD 1, 325 Broadway, Boulder, CO 80305-3328.
E-mail: tom.hamill@noaa.gov

current forecast can be properly adjusted. But what if the weather is relatively unusual? Suppose high rain amounts are forecast for a desert location; it is likely that there will have been few similar forecast events at that location that can be used to determine how to correct the forecast. If a model's systematic errors are similar throughout a region, then the effective sample size is increased by pooling the training data over many geographic locations. However, if the forecast errors are regionally dependent, there may be no effective substitute for a training database that spans many years or decades.

The presumed benefit of large training datasets motivated our foray into "reforecasting," the production of a large dataset of retrospective forecasts using the same model that is run operationally. Recently, we produced a sample reforecast dataset (Hamill et al. 2004; Hamill et al. 2006, hereafter HWM06). The novel feature of this prototype dataset was the extraordinary length and volume of the reforecast training dataset, 25+ years of 2-week ensemble forecasts initially centered on a reanalysis state. Such a large training dataset may permit accurate statistical adjustments even for some relatively rare events. A disadvantage of relying on reforecasts is the computational expense of generating them. To reduce this expense, we used a 1998 version of the National Centers for Environmental Prediction's (NCEP's) Global Forecast System (GFS) at a reduced, T62 resolution; certainly, it would be preferable to use a newer, higher-resolution model.

In the HWM06 article, a simple, skillful, two-step analog statistical correction technique was introduced as a way of making probabilistic forecasts, and we return to consider this analog technique more closely here. The first step in the analog technique was to compare the current forecast to all "past"[1] forecasts at a similar time of the year in a local region. Second, the dates of the closest matches were determined, and an ensemble was formed from the higher-resolution analyzed weather on those dates from which probabilities could be calculated from the event frequency. A gridded field of probabilities was produced, tiling together the probabilities computed for each independent region. Probabilistic forecasts from this ensemble were both reliable and specific when compared to forecasts generated from a more recent, higher-resolution version of the NCEP's GFS ensemble. The analog technique was able to correct the forecast bias and ensemble spread deficiencies and downscale the output to the scale of the higher-resolution, 32-km precipitation analysis (the North American Regional Reanalysis; Mesinger et al. 2006).

As HWM06 was an overview article, much detail and context were missing, and this article is intended to provide this context. Specifically, the purpose of this article is 1) to provide an underlying theoretical basis for use of analog techniques and explain the practical approximations that must be made in order to apply it; 2) to compare the analog technique against a few logical alternatives, such as logistic regression; and 3) to explore whether the simple analog technique of HWM06 can be enhanced further through slight algorithmic variations. The intent of this article, however, is not to provide an exhaustive comparison of the myriad of possible calibration techniques that exist in the literature. The few nonanalog methods that we test are included primarily to help understand the reasons why the analog methods provide such an improvement over probabilities estimated from the raw ensemble forecasts. A rigorous comparison against other calibration techniques would indeed be interesting, but many of them were designed with small training datasets in mind, applying approximations such as compositing training data over many locations.

Below, we first provide a theoretical underpinning for this two-step analog technique (section 2). The datasets and a variety of specific statistical correction techniques are then described (section 3), and a comparison of these techniques are provided (section 4), with conclusions (section 5).

The data used in this study are also freely available (see appendix A), and we encourage others who would be interested in testing their methods to use this dataset and compare their results against the benchmarks set here.

2. Theoretical basis of the analog technique and simplifying assumptions

Let us suppose that we have an ensemble of gridded-forecast model states for a particular time. Assume that there are n components to the state vector, and m ensemble members. We thus have a mn component forecast vector $\mathbf{x}^f$ composited from the ensemble members' forecasts,

$$\begin{aligned}\mathbf{x}^f &= [x_1^f(1), \ldots, x_1^f(m), \ldots, x_n^f(1), \ldots, x_n^f(m)] \\ &= (\mathbf{x}_1^f, \ldots, \mathbf{x}_n^f). \qquad (1)\end{aligned}$$

Suppose we are interested in the p-dimensional observed state of the atmosphere

[1] "Past" is used somewhat euphemistically here; in fact a cross-validation technique (Wilks 1995, p. 194) was used, so when processing early years in the 25-yr record, the latter years were used for training data.

$$\mathbf{x}^t = (x_1^t, \ldots, x_p^t) \tag{2}$$

at the same time; this could be the state at grid points or specific locations. The probabilistic weather forecast problem is then conceptually simple; we seek

$$f(\mathbf{x}^t|\mathbf{x}^f), \tag{3}$$

where $f(\cdot)$ denotes the probability density function; that is, we want to accurately quantify the probability distribution of the observed state of the atmosphere, given the ensemble forecast. Were the observed state comprised of the same variables at the same locations as the forecast state and the forecast model perfect (i.e., chaos was the only source of error, and the ensemble perfectly represented this), then the relative frequency from the ensemble would provide an adequate definition of any univariate event probability, accurate within sampling error,

$$P(x_i^t > T) = \frac{1}{m}\sum_{j=1}^{m} I[x_i^f(j), T], \tag{4}$$

where T is the threshold for some chosen event, $I[x_i^f(j), T] = 1$ when $x_i^f(j) > T$, and 0, otherwise. Unfortunately, ensemble forecasts are typically quite imperfect due to model errors and deficiencies in the method of constructing the ensemble.

If the climate was stable and it was possible to compute a nearly infinite set of reforecasts with associated verification data, then it would be possible to compute Eq. (3) directly, even in the presence of model error. With this nearly infinite ensemble, we could simply find past forecast states that were almost identical to the current forecast state and then determine Eq. (3) from the distribution of the observed states on those dates. Suppose there are s reforecasts of the same forecast lead time that are practically identical to the current forecast at that lead. Let $\mathbf{x}^{t|r} = [\mathbf{x}^{t|r}(1), \ldots, \mathbf{x}^{t|r}(s)]$ denote the collection of the s associated past observed states on the dates of the nearly identical reforecast analogs. Here "t" is shorthand for "truth" and "r" for reforecast. Then to find the event probability at a given location,

$$P(x_i^t > T) = \frac{1}{s}\sum_{k=1}^{s} I[x_i^{t|r}(k), T], \tag{5}$$

where $I[x_i^{t|r}(k), T] = 1$, when $x_i^{t|r}(k) > T$, and $I[x_i^{t|r}(k), T] = 0$, otherwise. All that is being done here is to determine the fraction of time when the threshold is exceeded by using the observed data associated with the chosen analogs. If the observed state is actually describing the atmospheric state at much smaller scales than the original forecast, then this procedure amounts to a statistical downscaling (Zorita and von Storch 1999). Note that it is possible also to write a more general form of Eq. (5) accounting for the possibility that the observed and model state components are different; for example, the observed state is virtual temperature and the model state includes temperature and humidity.

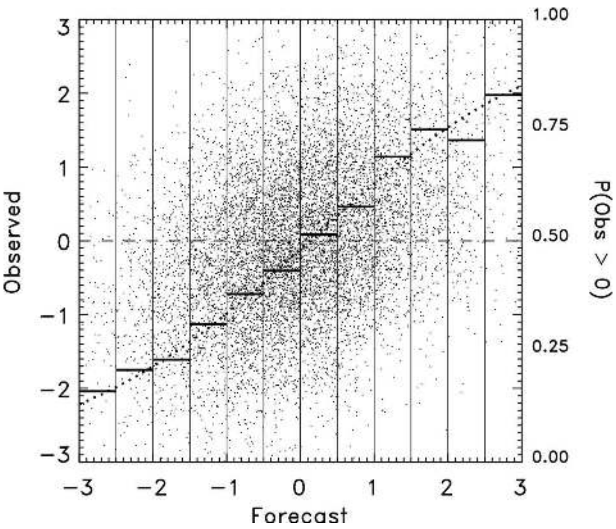


FIG. 1. Illustration of two methods for setting probabilities using synthetic reforecast data. Dots are the reforecast data (abscissa) and the associated observed value (ordinate). Vertical bars denote bins for considering "nearby" analogs. Heavy horizontal solid lines are the probabilities set by relative frequency in these bins (axis labels on right). Dashed line is the probability set by logistic regression.

This process of Eq. (5) is conceptually illustrated in Fig. 1 with a synthetic 10 000-day reforecast dataset. Here we have created a time series of reforecast and associated observed data; the state is a scalar and the forecast is deterministic, so the problem can be visualized two-dimensionally. Consider the event that the true state is >0.0. Suppose our criteria for closeness of a reforecast to the current day's forecast was to be within a window of 0.5 units. To apply Eq. (5), we find the forecast points in vertical columns of this width and then count the fraction with observed data >0.0; the horizontal bars in Fig. 1 provide the probability based on this simple count. Also plotted is a fitted logistic regression curve (Wilks 1995) to the data. In comparison to the analog process, the logistic regression parameters are fit using all of the scatterplot data at once, rather than just the data from close forecast analogs. Depending on the data, the smooth S-shaped logistic-regression curve may provide a better or worse fit than the analogs when sample size is finite.

It is worth considering the asymptotic error characteristics of such a forecast approach as skill increases or decreases. If the forecast is totally uncorrelated with the observed data, then using (5) will reproduce the climatological distribution within sampling error. If the forecast system's fidelity improves so that the correlation of forecast and observed approaches 1.0, the probabilistic forecasts will become increasingly sharp without losing reliability. In the asymptotic limit that the forecast error approaches zero, the probabilistic forecast will approach a perfect deterministic forecast. In this case, of course, a reforecast would be unnecessary, but, as this asymptotic limit is only of theoretical concern (Lorenz 1963), it is at least comforting to know that the performance of the statistical analog approach in Eq. (5) will improve as the forecast model improves and score no worse than climatology.

The analog process is quite simple, as illustrated in Fig. 1, but suppose the model state is a 100-member ensemble forecast of winds, temperatures, humidity, and geopotential at millions of grid points covering the globe. Even with billions of years of reforecasts, it may prove difficult to find many close global analogs (Lorenz 1993, p. 86; Van den Dool 1994). And were a reforecast available over such a long period of time, the climate, and indeed the continents themselves, would not be very stable. Hence, simplifying assumptions are required. Some possible assumptions may include the following:

- If we are concerned specifically with assessing the probabilities at a particular location, only the forecast model state around that location may be needed; for example, to estimate probabilities for Washington, D.C., it is only necessary to find the dates of past forecasts matching today's D.C.-area forecast; matching or not matching at other distant locations is irrelevant (Van den Dool 1989). In the terminology of linear regression, the model forecast state at the distant locations would not make useful predictors.
- If provided with a forecast ensemble, it may be unnecessary to match all the aspects of the ensemble; matching the mean state may be sufficient or perhaps the mean and the spread, rather than requiring that each member match.
- If considering an event like surface temperature, it may be sufficient to match reforecasts of surface temperature alone, ignoring other forecast aspects such as upper-level winds or temperatures.
- A shorter reforecast period may be necessary so that the climate and analysis quality is approximately stationary. If the climate is changing extremely rapidly or the skill of recent forecasts is much larger than older forecasts due to changes in the observation network, a skill increase from improved sampling from a longer reforecast archive may not be realized.

3. Datasets and methods

Below, we provide a brief description of the reforecast and verification datasets, and then more detail on various statistical correction techniques and the metrics for evaluating forecast skill. Our focus is on the calibration of probabilistic forecasts of 24-h precipitation amount on a Lambert-conformal grid covering the conterminous United States at the scale of the analyzed verification data, ~32 km.

a. Reforecast and verification datasets

HWM06 provide a more complete description of the reforecast dataset. The forecast model is a 28-level, T62 resolution version of NCEP's Global Forecasting System model using physics that were operational in the 1998 version of the model. Precipitation forecasts used here were archived on a 2.5° latitude–longitude grid, though in this study they have been interpolated to an ~250 km grid on a Lambert-conformal projection corresponding to every eighth grid point in the North American Regional Reanalysis (NARR) (Mesinger et al. 2006). A 15-member ensemble was produced every day from 1979 to current, starting from 0000 UTC initial conditions and integrated to 15-days lead. The ensemble initial conditions consisted of a control initialized with the NCEP–National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996) and a set of seven bred pairs of initial conditions (Toth and Kalnay 1993, 1997) re-centered each day on the reanalysis initial condition. The breeding method was the same as that used operationally in January 1998. The forecasts extended to 15-days lead, with data archived every 12 hours. Winds, temperature, and geopotential height are available at the 850-, 700-, 500-, 250-, and 150-hPa levels. Ten-meter wind components, 2-m temperature, mean sea level pressure, accumulated precipitation, convective heating, precipitable water, and 700-hPa relative humidity were also archived.

For verification, the NARR 24-h precipitation analyses were used. This data and the observations used to generate it are described in Mesinger et al. (2006). The data was on a 32-km Lambert-conformal grid. Only grid points over the conterminous United States were used.

For all subsequent experiments, the dataset will consist of reforecasts from 1 January 1979 to 31 December 2003, 25 years of forecasts. We focus on daily precipitation forecasts for the first six days.

For access to this reforecast and precipitation analysis data, please see appendix A.

b. Probabilistic estimation techniques

We now briefly review 10 different techniques for estimating precipitation event probabilities. The first does not use the reforecast dataset; the rest do.

1) Ensemble relative frequency

The simplest approach uses no statistical calibration. The relative frequency of event occurrence is estimated directly from the 15-member ensemble, interpolated to the 32-km NARR grid. For example, if 3 of the 15 members at a point indicate greater than 25-mm rainfall, the probability of that event is set to 20%.

2) Bias-corrected relative frequency

In this procedure, probabilistic forecasts are generated from an ensemble of forecasts, where each member has been bias corrected according to the long-term bias statistics for that grid point and time of year. This follows a technique proposed by Y. Zhu (NCEP, 2005, personal communication). Let $F_Y^C(y)$ denote the cumulative distribution function (CDF) of the analyzed 24-h precipitation amount, defined by

$$F_Y^C(y) = P_Y^C(Y \le y). \tag{6}$$

Here Y is the random variable, y the specific amount being considered, and $P(\cdot)$ indicates the probability, which will be determined by frequency from a large sample. Similarly, define a CDF for the 24-h ensemble forecast amount, $F_X^E(x)$ defined by

$$F_X^E(x) = P_X^E(X \le x). \tag{7}$$

The technique is then rather simple. For a given day of the year, we compute $F_Y^C(y)$ and $F_X^E(x)$ using the 25 years × 91 days (centered on the day of interest) of analyzed precipitation and interpolated member forecasts. Then, for a given ensemble forecast on that day with the value x, we determine a value y such that $F_Y^C(y) = F_X^E(x)$. The ensemble member forecast x is then replaced with the value y. This is illustrated in Fig. 2 for today's hypothetical forecast and the forecast and analyzed CDFs. As implemented at NCEP in 2004, the technique is implemented differently; only a short training sample is used, such as the past 30 days, and the technique generates CDFs that are not location specific, instead representing an average over the domain.

3) Basic analog technique

This procedure was described in HWM06, and a simple pictorial representation of the method is provided in Fig. 3. As suggested in section 2, application of the full analog theory assumes a nearly infinite training sample. Without this, we adopt several of the simplifying assumptions; namely, we search only for local analogs, match the ensemble-mean fields, and consider only the model forecast of precipitation (no winds, temperature, geopotential, etc.) in selecting analogs.

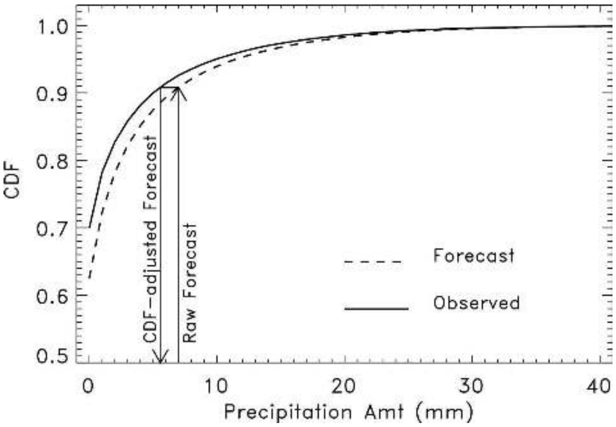


FIG. 2. Illustration of the bias-correction technique described in section 3b(2). Dashed line denotes the forecast CDF; solid line the observed CDF. A raw forecast of 7 mm is at approximately the 91st percentile of the forecast CDF; the 91st percentile of the observed CDF is approximately 5.6 mm. Thus, the precipitation forecast is changed from 7 to 5.6 mm.

The first step of the procedure is to find the closest local reforecast analogs to the current numerical forecast. Within a limited-size region, the forecast for the day under consideration (the map in the top row in Fig. 3) is compared against past forecasts in that same region and at the same forecast lead. Specifically, the ensemble-mean precipitation forecast pattern is computed at a subset of 16 coarse-mesh reforecast grid points, the 16 dots in each panel of Fig. 3, surrounding the region where probabilities are to be defined (the region enclosed by the dashed line). These coarse-mesh reforecast grid points are separated by ~32 km × 8 = 256 km, eight times coarser than the 32-km analyzed data. The ensemble-mean forecast pattern at these 16 points are compared to ensemble-mean reforecast patterns at these 16 points in all the other years.[2] However, only those reforecasts within a window of 91 days are

[2] If there is a large change in forecast skill or in the climate, then the cross-validation techniques used here may overestimate the skill that will be achieved in operational implementation, when only past training data are available. There are indications that the frequency of intense precipitation has been increasing slightly over the past few decades (Groisman et al. 2005, Table 1), but not enough, our tests show, to invalidate the use of cross validation.

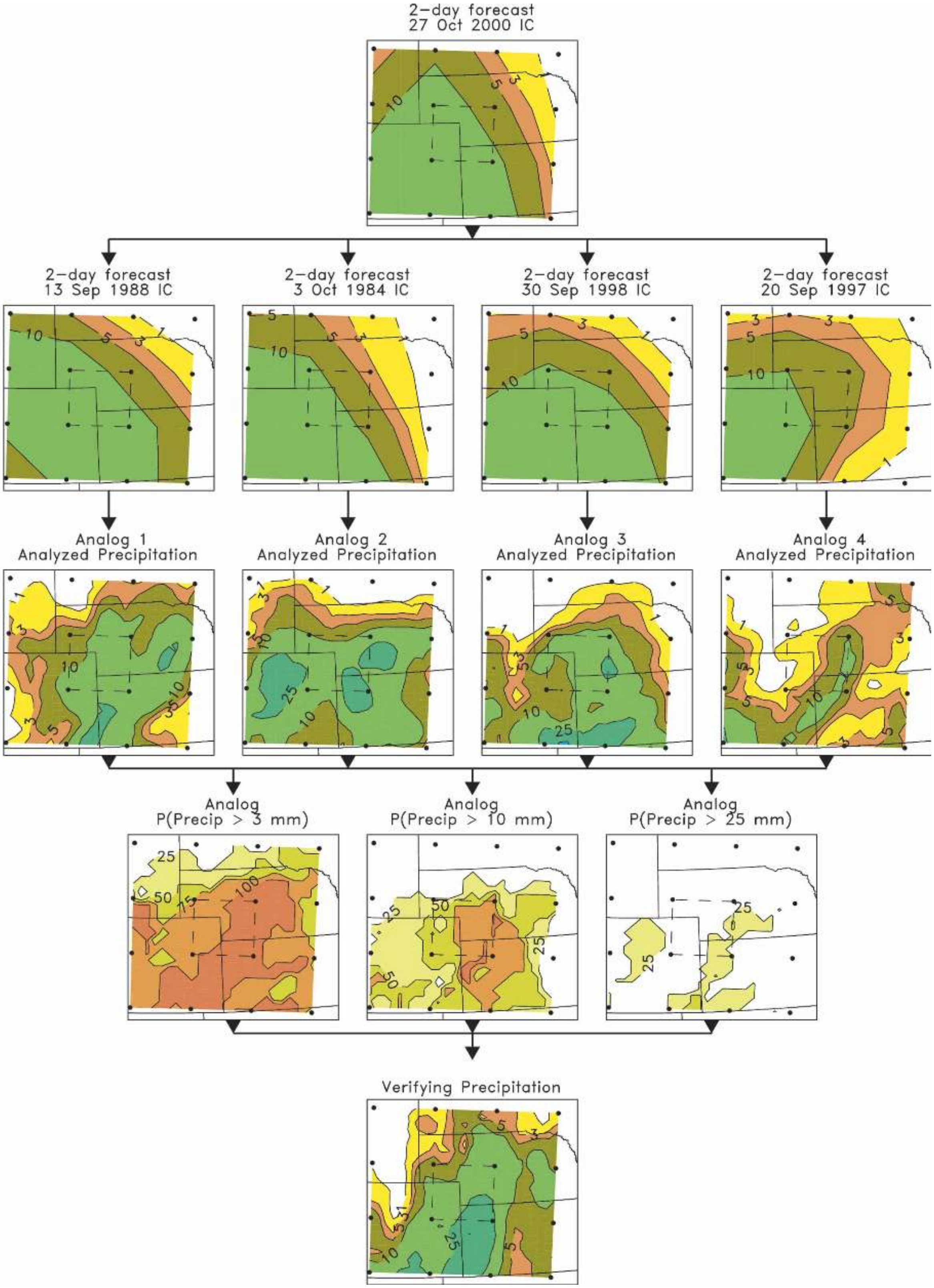


FIG. 3. Illustration of the basic analog technique for a 2-day forecast. The coarse-mesh ensemble-mean precipitation forecast is shown in the first row, defined at the 16 dots and contoured here. Analogs and probability forecasts are desired for the dashed box in the middle. The four closest matching 2-day ensemble-mean forecasts are shown in the second row, and the higher-resolution analyzed weather on those dates are shown in the third row. Probabilistic forecasts formed from the analyzed analogs are shown in the fourth row for 3-mm, 10-mm, and 25-mm thresholds, and the analyzed data are shown in the bottom row.

compared, that is, a ±45 day window around the date of the forecast; this window is used under the presumption that model biases may change substantially with the time of year. The root-mean-square (rms) difference between the current forecast and each reforecast is then computed, averaged over the 16 grid points. The *n* historical dates with the smallest rms difference are chosen as the dates of the analogs. The four closest-matching forecast patterns are shown in the second row of Fig. 3. The next step is the formation of an ensemble of 32-km NARR-analyzed precipitation on the dates of the closest *n* forecast analogs. This ensemble is the third row in Fig. 3. Probabilistic quantitative precipitation forecasts (PQPFs) are then generated by using the relative frequency of the event in the analyzed ensemble; for example, if three of the four analyzed members at a grid point had greater than 10 mm of accumulated rain, the probability of exceeding 10 mm at that grid point was set to 75%. Sample probability forecasts from the four-member ensemble are shown in the fourth row in Fig. 3. Note, however, that the probabilities are retained only in the region enclosed by the dashed box. Probabilities in adjacent regions are computed by shifting the search region one coarse-mesh grid point. A national-scale 32-km PQPF is generated by tiling together the local PQPFs; further discussion of this is supplied in section 3b(10) below (see Fig. 4). The final step of the process is to compare the probability forecasts to the analyzed precipitation, which is shown in the bottom row of Fig. 3.

Commonly, many more than the four members shown in this figure would be used. In actuality, ensembles of size 10, 25, 50, and 75 were computed. In subsequent figures, the skill scores will be plotted only for the optimal size. In general, the optimal size was smaller for heavier precipitation events and shorter forecast leads. For more information on the optimal ensemble size; see HWM06 Fig. 7.

4) Logistic regression

Logistic regression estimates event probabilities at a particular location through an equation of the form

$$P(x_i^t > T) = \frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\phi_1 + \cdots + \hat{\beta}_n\phi_n)}, \qquad (8)$$

where $\hat{\beta}_0, \ldots, \hat{\beta}_n$ are fitted regression coefficients and $\phi_1, \ldots, \phi_n$ are model predictors based on the forecast data. Here $n = 2$, ϕ_1 is the square root of the ensemble-mean precipitation amount interpolated to the observation location (the square root transformation made the data less positively skewed), and ϕ_2 is the column ensemble-mean precipitable water interpolated to the observation location, measured in millimeters. The regression coefficients are determined in a cross-validated manner; when coefficients are developed for grid points for a particular year, that year's forecasts are excluded from the training data. As with the basic analog technique, a 91-day window of forecasts is used. Multiplied by the 24 years of training data, 2184 training samples are produced. Logistic regression techniques were tried without precipitable water and without the power transformation; both were somewhat less skillful, and results for these will not be presented.

Unlike the analog technique, the logistic regression technique uses all of the data to determine the regression curve, not just a subset of close analogs. This has the advantage of increasing the sample size but the potential disadvantage that cases very dissimilar to the forecast of interest on that day are also being used to estimate the probabilities. Another disadvantage of the logistic regression technique is it provides only a probability for the event threshold under consideration; if probabilities are desired for a different threshold, the regression analysis must be repeated. In comparison, when using analog techniques, once the analog members are chosen, probabilities can be defined quickly for any event threshold.

5) Basic technique using individual members

This technique is similar to the basic technique [section 3b(3)], but instead of searching for analogs of the ensemble mean, the member 1 forecast is compared against past member 1 reforecasts, the dates of the five closest matches are noted, and the process is repeated for the rest of the 15 members, producing 75 (possibly nonunique) dates. If an ensemble of size smaller than 75 is to be formed, the dates of the closest pattern matches from the set of 75 dates are used. Comparing the skill of these forecasts to that of the basic technique will indicate whether the information content can be distilled down to the ensemble mean or whether there was extra information in each member.

Note that a slight algorithmic variant was also tried, whereby forecast matches were permitted between members; for example, the member 1 forecasts are compared against members 1–15; this variant produced forecasts that were no more skillful (not shown).

6) Basic technique including precipitable water

This technique repeats the basic analog technique from section 3b(3), but instead of only matching the ensemble-mean precipitation amount, column precipitable water is included as well. When measuring the

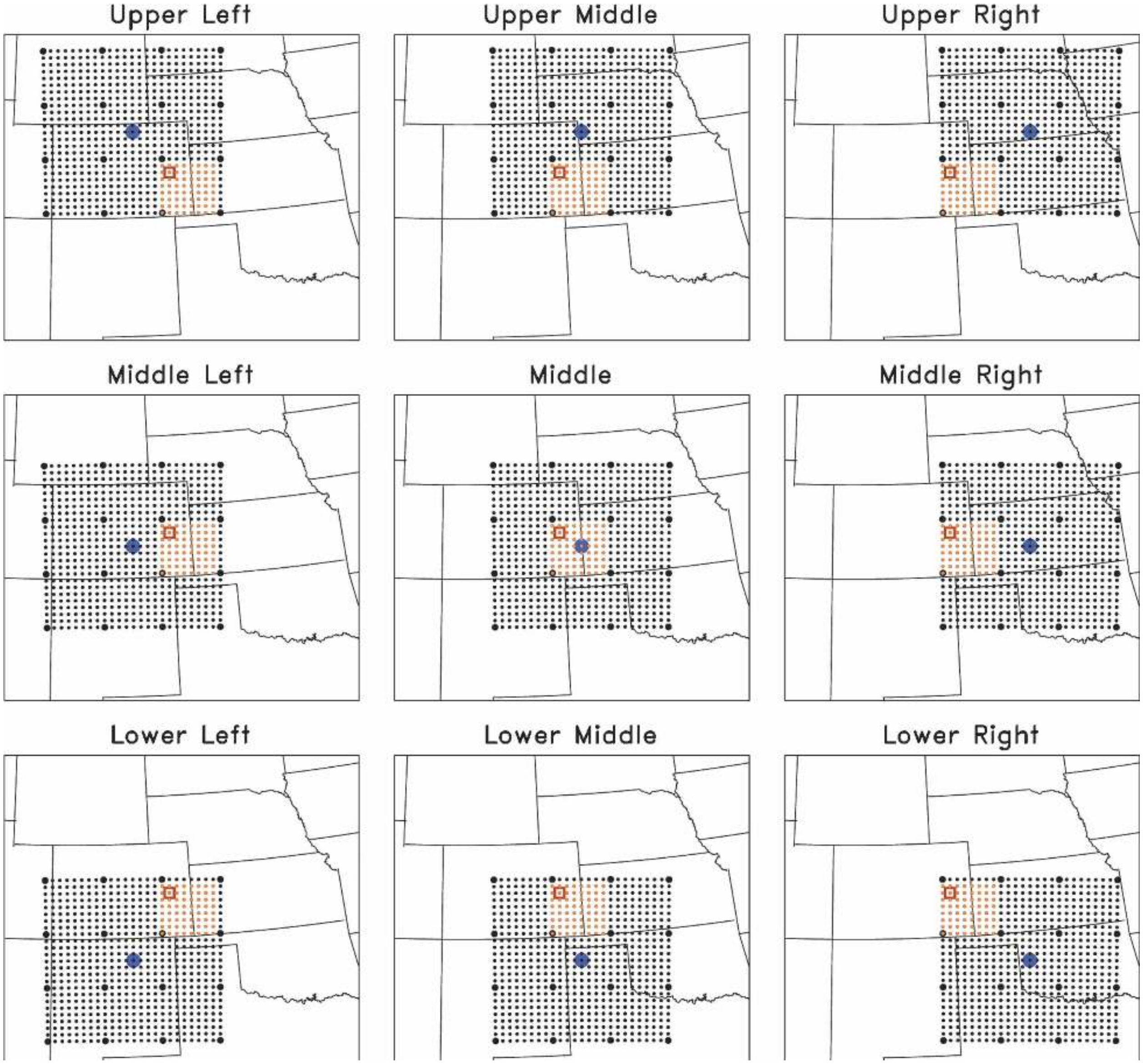


FIG. 4. Illustration of the smoothing algorithm. Probabilities are sought in this case for the NARR grid points colored orange. The nine panels are the nine regions where analog dates have been calculated that overlap the orange grid points. Other NARR grid points are denoted by small black dots. Analog matches are calculated by forecast similarity at the large black dots; the center of each analog search region is denoted by the blue dot. For the orange grid point highlighted by the red box, the final probability is a weighted sum of the probabilities of the nine estimates, the weight being determined by the relative distance of the blue dot in that panel from the red box.

closeness of a past reforecast, the ensemble-mean precipitation is weighted by 70% and the precipitable water by 30%.

7) Basic technique including 2-m temperature and 10-m winds

This technique repeats the basic analog technique from section 3b(3) but, instead of only matching the ensemble-mean precipitation amount (millimeters), 2-m temperature (kelvin) and 10-m u and v components of the wind (meters per second) are included as well. When measuring the closeness of a past reforecast, the differences in precipitation, temperature, and wind speed between today's forecast and the past reforecasts at the 16 grid points are squared and then summed, and the analog dates are those with the smallest sums.

8) Rank analog technique

This technique is generally the same as the basic analog technique described in section 3b(3) with one exception. When determining the closest matches, at each of the 16 grid points, the rank is computed for today's precipitation forecast amount when pooled with the reforecasts. Similarly, the rank of the precipitation amount is determined at each grid point for each of the reforecasts in the temporal window. The analog dates are those with the lowest sum of the absolute value of rank differences over the 16 points.

The rank analog technique is included here because results will show (section 4, Fig. 13) that at early leads the basic analog technique produced somewhat unreliable forecasts, underforecasting precipitation probabilities. Upon closer examination, it was determined that this was primarily because the distribution of precipitation forecast amounts was skewed with lighter amounts more common than heavier amounts. Consequently, the basic technique's closest forecast analogs more commonly had slightly less precipitation than today's forecast more often than they had slightly more precipitation. Using a rank-based approach was proposed as a way of ensuring that more equal numbers of heavier and lighter forecast events were used as analogs.

9) Rank analog with smaller search region

This technique repeats the rank analog technique from section 3b(8), but instead of finding a match over a set of 16 grid points (see Fig. 3), only the four center grid points are used in determining the dates of the reforecast analogs. A comparison of this against the rank analog will indicate whether the size of the search region is an important determinant of forecast skill.

10) Smoothed rank analog technique

Most of the prior analog approaches discussed in this article produce probability estimates for an 8 × 8 box of 32-km grid points, finding analogs using the surrounding large-scale forecast fields (see Fig. 3). To produce a national map of the probabilities, the process is repeated for other regions and the final map is a tiled composite of the 8 × 8 patches. Unfortunately, sometimes the dates of the analogs can be quite different for one set of 8 × 8 boxes when compared to its adjacent sets. This may result in a discontinuity of the probabilities at the boundaries between patches. Accordingly, we test a simple smoothing algorithm to eliminate this effect. Aside from the smoothing applied at the end, this method will be identical to the rank analog technique, described previously.

To understand the smoothing, consider Fig. 4. Say we seek to estimate the probabilities on the 32-km grid at the orange grid points. The analogs are determined by matching today's forecast at the large-scale grid points (large black dots) to past forecasts at these same dots. For the orange dots, there are actually nine separate regions where analogs and the subsequent probabilities are calculated that overlapped the orange dots, shown in the nine panels of Fig. 4. In all previously described analog algorithms, estimates at eight of these are thrown away, and only the probabilities from analogs from the middle search region are used.

Here the smoothing algorithm uses several of the overlapping regions' estimated probabilities. Consider the orange grid point surrounded by the red box in Fig. 4.

Let w_{ul}, w_{um}, w_{ur}, w_{ml}, w_m, w_{mr}, w_{ll}, w_{lm}, and w_{lr}, denote the weights applied to the probability estimates using the upper left box, the upper middle box, and so on. Let d_{ul}, d_{um}, d_{ur}, d_{ml}, d_m, d_{mr}, d_{ll}, d_{lm}, and d_{lr} indicate the distance between the center point of each calculation region (the blue dot) and the red box, and let w'_{ul}, w'_{um}, w'_{ur}, w'_{ml}, w'_m, w'_{mr}, w'_{ll}, w'_{lm}, and w'_{lr} represent a nonnormalized weight. Here, define the threshold distance D to be $\sqrt{128}$, the distance in NARR grid points between the upper-left and middle blue dots. First, a nonnormalized weight is calculated according to

$$w'_{xx} = \begin{cases} \dfrac{D - d_{xx}}{D + d_{xx}}, & d_{xx} < D \\ 0, & d_{xx} \geq D, \end{cases} \tag{9}$$

where $_{xx}$ is one of the nine regions, for example, $_{ul}$. After all nine nonnormalized weights are calculated, then weights are normalized so that the weights sum to 1.0. For example,

$$w_{ul} = \frac{w'_{ul}}{w'_{ul} + w'_{um} + w'_{ur} + w'_{ml} + w'_m + w'_{mr} + w'_{ll} + w'_{lm} + w'_{lr}}. \tag{10}$$

For example, the nine weights for the red box from upper left to the lower right are 0.134, 0.188, 0.000, 0.261, 0.379, 0.004, 0.004, 0.029, and 0.000.

c. Performance measures

A primary metric of forecast performance will be the Brier skill score (BSS; Wilks 1995). The precise definitions of BS_f and BS_c and the climatological reference are described in appendix B, along with a cautionary warning on how the BSS may overestimate geographically averaged skill.

Because of the extremely large sample size of the forecasts, even small differences in the BSS tended to be statistically significant. Tests of significance were

TABLE 1. Brier skill score for various forecast techniques at 2.5 mm, averaged over the 25 years. The last row provides the amount of difference between two forecasts that is considered statistically significant according to a two-sided test with $\alpha = 0.05$ (cf. Wilks 1995, p. 117). Highest score for a particular day is in boldface type.

	Day					
Technique	1	2	3	4	5	6
1) Ensemble relative frequency	0.0840	−0.0486	−0.1098	−0.1624	−0.2117	−0.2552
2) Bias-corrected relative frequency	0.2642	0.1753	0.0597	−0.0424	−0.1318	−0.2033
3) Basic analog	0.4026	0.3443	0.2648	0.1923	0.1335	0.0853
4) Logistic regression	0.4108	0.3395	0.2564	0.1842	0.1266	0.0815
5) Basic using individual members	0.4061	0.3414	0.2555	0.1774	0.1155	0.0692
6) Basic including precipitable water	0.4080	0.3486	0.2687	0.1969	0.1378	0.0898
7) Basic including 2-m temperature and 10-m winds	0.3803	0.3312	0.2565	0.1881	0.1319	0.0875
8) Rank analog	0.4195	0.3555	0.2726	0.1965	0.1360	0.0865
9) Rank analog with smaller search region	0.4194	0.3496	0.2635	0.1871	0.1272	0.0791
10) Smoothed rank analog	**0.4260**	**0.3613**	**0.2779**	**0.2020**	**0.1415**	**0.0925**
Difference that is statistically significant, two-sided test, $\alpha = 0.05$.	0.0010	0.0009	0.0008	0.0007	0.0006	0.0006

evaluated with the block bootstrap technique described in Hamill (1999).

Reliability diagrams (Wilks 1995) will also be used to illustrate the degree of correspondence between forecast probabilities and analyzed relative frequencies.

4. Results

Tables 1 and 2 provide the BSS for each of the 10 techniques discussed in the previous section. The bias-corrected relative frequency technique improved the forecast of 2.5-mm forecasts somewhat compared to the ensemble relative frequency technique, but it tended to lower the skill of the 25-mm forecasts. All remaining methods provided a consistent, very large improvement over the ensemble relative frequency technique. With the exception of the bias-corrected relative frequency, the skill differences between the various reforecast-based calibration techniques were much smaller; having achieved most of the skill with the basic analog technique or the logistic regression technique, other methods had only slightly larger or smaller skill scores. This is not to suggest that the daily PQPFs were necessarily very similar; for example, the logistic regression technique tended to overforecast high probabilities at 25 mm, while the analog techniques underforecast them (not shown). But in practice, their overall skills were similar.

Let us examine these results in more detail. First, consider forecasts from the ensemble relative frequency technique. Figure 5 provides a plot of the BSS of this technique as a function of time of the year and the forecast lead time. Unsurprisingly, skill was larger at short leads and larger in the cool season. Many of the forecasts were less skillful than the reference seasonal climatological distribution. Oddly, shorter warm-season forecasts had more negative skill than longer warm-season forecasts. Light precipitation forecasts in April were particularly unskillful; a subsequent examination of reliability diagrams (not shown) showed that light

TABLE 2. As in Table 1 but for 25 mm.

	Day					
Technique	1	2	3	4	5	6
1) Ensemble relative frequency	0.0534	−0.0668	−0.0624	−0.0473	−0.0535	−0.0559
2) Bias-corrected relative frequency	0.0105	−0.0503	−0.0684	−0.0731	−0.0860	−0.0894
3) Basic analog	0.1816	0.1298	0.0887	0.0597	0.0357	0.0201
4) Logistic regression	**0.1895**	0.1205	0.0831	0.0572	0.0350	0.0219
5) Basic using individual members	0.1856	0.1267	0.0815	0.0504	0.0278	0.0131
6) Basic including precipitable water	0.1841	**0.1319**	0.0903	0.0607	0.0370	0.0212
7) Basic including 2-m temperature and 10-m winds	0.1715	0.1245	0.0854	0.0587	0.0363	0.0217
8) Rank analog	0.1727	0.1260	0.0878	0.0588	0.0350	0.0193
9) Rank analog with smaller search region	0.1860	0.1280	0.0865	0.0568	0.0326	0.0176
10) Smoothed rank analog	0.1832	0.1318	**0.0912**	**0.0621**	**0.0378**	**0.0222**
Difference that is statistically significant, two-sided test, $\alpha = 0.05$.	0.0015	0.0012	0.0010	0.0009	0.0007	0.0007

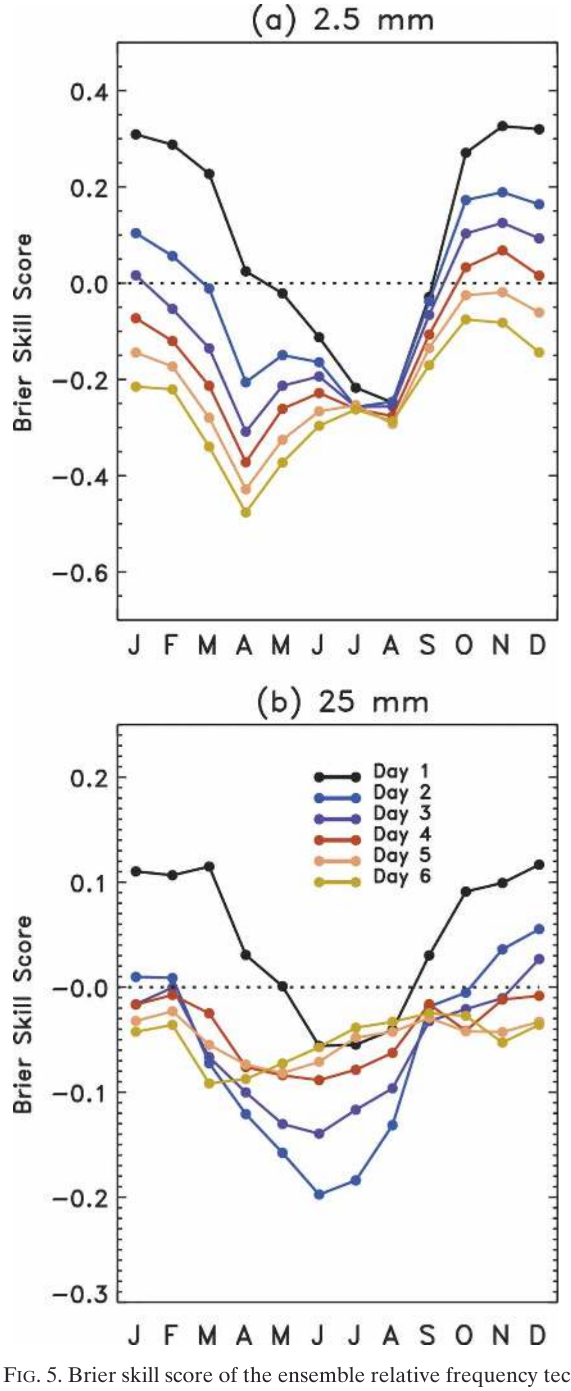


FIG. 5. Brier skill score of the ensemble relative frequency technique as a function of the time of the year and the lead time of the forecast: (a) Skill at 2.5 mm and (b) skill at 25 mm.

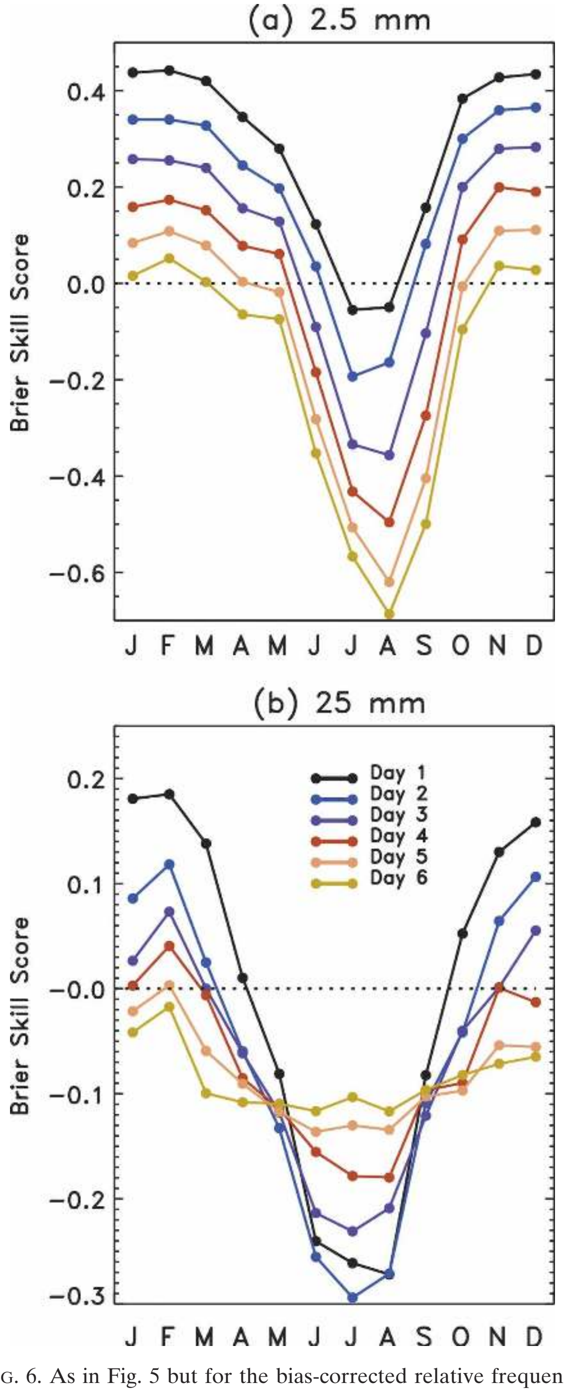
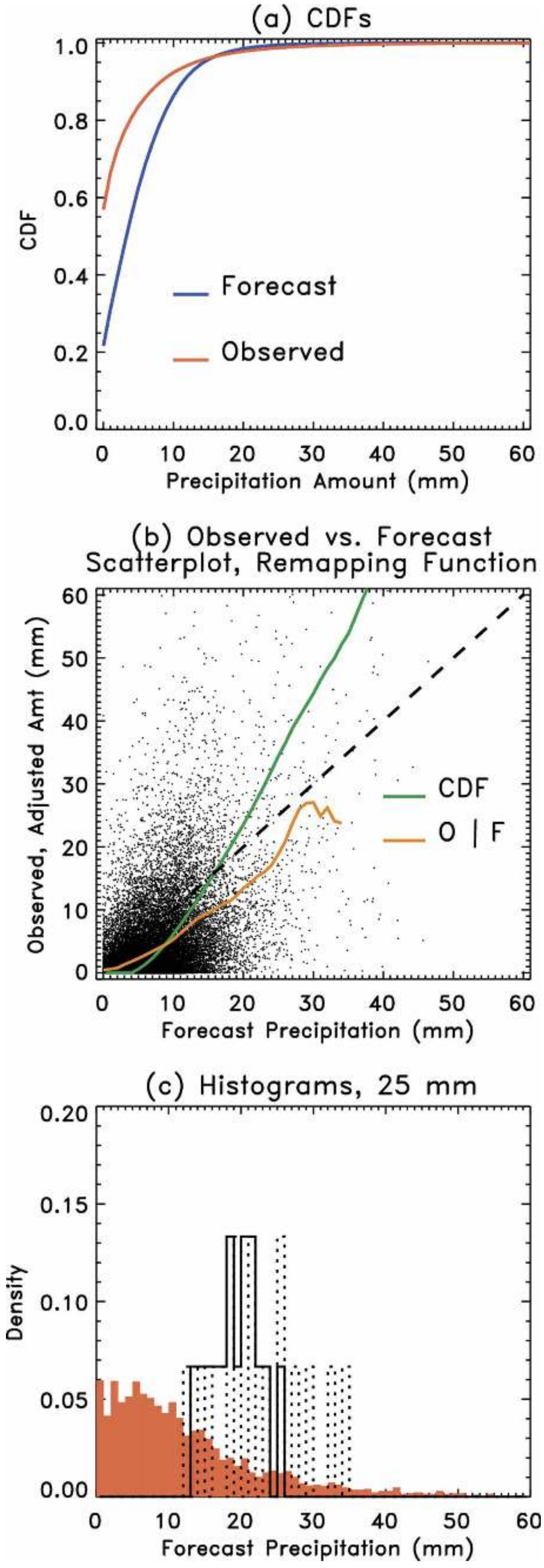


FIG. 6. As in Fig. 5 but for the bias-corrected relative frequency technique.

amounts were overforecast in April much more commonly than at other times of the year.

The bias-corrected relative frequency technique improved the cool-season precipitation forecasts but tended to make the warm-season forecasts even less skillful (Fig. 6). How could this bias-correction technique worsen the forecast? To understand this, we chose an 8 × 8 set of 32-km NARR grid points centered in northern Mississippi and examined the 1-day forecasts in mid-August, a date and location where the skill of 25-mm forecasts decreased. Figure 7a shows the average CDFs for the forecast and analyzed over these grid points. Light precipitation events were forecast



much too frequently; for example, a 2-mm forecast was approximately at the 38th percentile of the forecast's cumulative distribution, while the 38th percentile of the analyzed distribution was 0 mm. Conversely, precipitation events above 16 mm were forecast less commonly than were analyzed. Figure 7b provides a scatterplot of 1-day forecasts of a single member from the ensemble, plotted against the analyzed precipitation. Plotted over top, the green line illustrates the function that relates adjusted amount to forecast precipitation based on the CDF differences. The orange line indicates the mean of the conditional distribution of the analyzed amounts given the forecast, plotted using a running-line smoother with a window width of 4 mm (Hastie and Tibshirani 1990). Using the CDF adjustment, all forecasts below ~5 mm were adjusted to zero precipitation, while a forecast precipitation amount of 30 mm was adjusted to ~45 mm. Figure 7c illustrates the pdf adjustment for an ensemble forecast with a mean of ~25 mm. The dashed histogram indicates the frequency distribution of the adjusted ensemble forecast, while the red histogram indicates the conditional distribution of analyses, given an ensemble-mean forecast of between 23 and 27 mm. As can be seen, the adjustment shifted the distribution further away from the conditional distribution of analyses, so, averaged over many similar cases, these forecasts should have scored worse than the uncorrected forecast.

At first glance, the difference between the CDF adjustment and the adjustment implied by the conditional distribution of analyses appears contradictory: if there were truly more high-precipitation events in the analyzed CDF than in the forecast CDF, then why would the conditional distribution of analyzed events, given the 25-mm forecast, have a mean analyzed value lower than the mean forecast? The discrepancy was due to the lack of a strong relationship between forecast and analyzed data; that is, the largest analyzed amount in the sample pool did not occur when the forecast was the largest. Had the forecasts and analyses been very highly

←

FIG. 7. (a) Illustration of forecast and analyzed CDFs for 1-day forecasts in northern Mississippi during August. (b) Scatterplot of one ensemble member's forecast vs analyzed forecast. Red curve illustrates the remapping that will occur between a forecast precipitation amount and the corrected amount, based on the CDF correction technique. Orange curve denotes remapping between forecast and mean analyzed given the forecast. (c) For 25 mm, a typical ensemble forecast distribution with a mean of between 23 and 27 mm (solid line), the adjusted distribution (dashed line), and the conditional distribution of the analyzed values given an ensemble mean of between 23 and 27 mm (in red).

related, then the conditional distribution of analyzed events given a 25-mm ensemble-mean forecast would, indeed, be larger than 25 mm—far different than the unconditional climatology. In fact, as shown by the difference between the green and orange lines, the CDF bias correction was in the wrong direction for forecasts above ~17 mm, where the CDF correction suggested a mapping of the forecast to higher amounts while the mean analysis given the forecasts indicated the preferred mapping was toward lower amounts. Asymptotically, then, the performance of this bias correction based upon differences in the CDFs was likely to make already bad forecasts (i.e., deficient in spread and poorly correlated with analyses) worse when the analyzed and forecast CDFs differed. The stronger the forecast–analysis relation, the more one can expect the CDF-based bias correction to improve forecast skill. Still, the CDF-adjustment method by construction does not correct for spread deficiencies in the ensemble. Hence, even with a perfect forecast–analysis relationship, this calibration technique may not result in as skillful probabilistic forecasts as other methods that include spread corrections.

Suppose a bias correction was based on some type of regression rather than the CDF technique. Then, if forecasts and analyses were uncorrelated, all member forecasts would be adjusted to the climatological mean, regardless of their initial value, converting the ensemble into a deterministic forecast. And were the mean of the forecast distribution shifted but the spread of the ensemble preserved, one could envision situations where the shift could create members with unmeteorological, negative-valued precipitation amounts. The overall lesson seems to be that it will be difficult to improve probabilistic forecasts through some simple bias adjustments; errors in the mean and in the spread should both be addressed.

We return to examining the rest of the correction methods, all of which did improve upon the BSS compared to the ensemble relative frequency. Figure 8 shows the BSS of the basic analog approach for each month. The forecasts were almost universally skillful relative to climatology, with more skill in the cool season and more skill at short leads and lesser amount thresholds.

Figure 9 shows that the logistic regression technique performed quite similarly. The forecasts were slightly more skillful than the basic analog forecasts at day 1 and generally similar or less skillful than the basic analog at longer leads. In Fig. 9, this skill comparison is visualized through the use of the gray shading. The height of the gray-shaded line indicated the skill of the reference analog forecasts. Hence, when the shading was above the plotted line for the logistic regression technique, this indicated that the basic analog technique had a higher BSS in proportion to the thickness of the gray shading. When the shading was below the logistic regression line, the basic analog technique had a proportionally lower BSS. Because of the large sample size, even small differences tended to be statistically significant. Applying a block bootstrap technique (Hamill 1999), the magnitude of yearly differences that were statistically significant at the 95% confidence level are presented in the last row of Tables 1 and 2; monthly differences that were significant were typically two or three times larger.

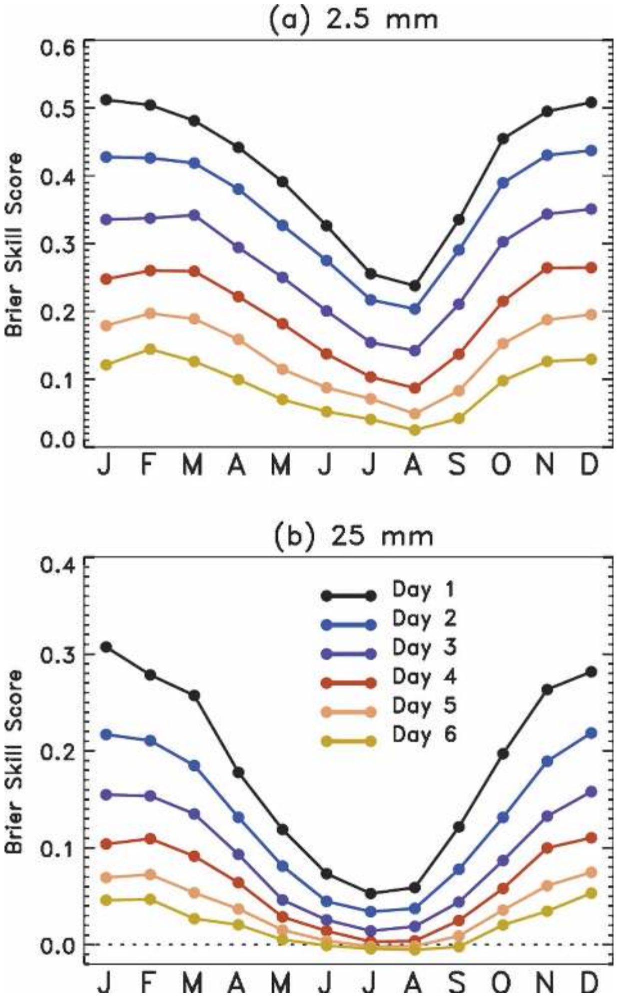


FIG. 8. BSS of the basic analog technique as a function of the month and the lead time of the forecast: (a) Skill at 2.5 mm and (b) skill at 25 mm.

Was there an advantage to fitting individual ensemble members rather than the ensemble mean? Figure 10 presents the BSS for the basic technique using individual members, along with a comparison of skill

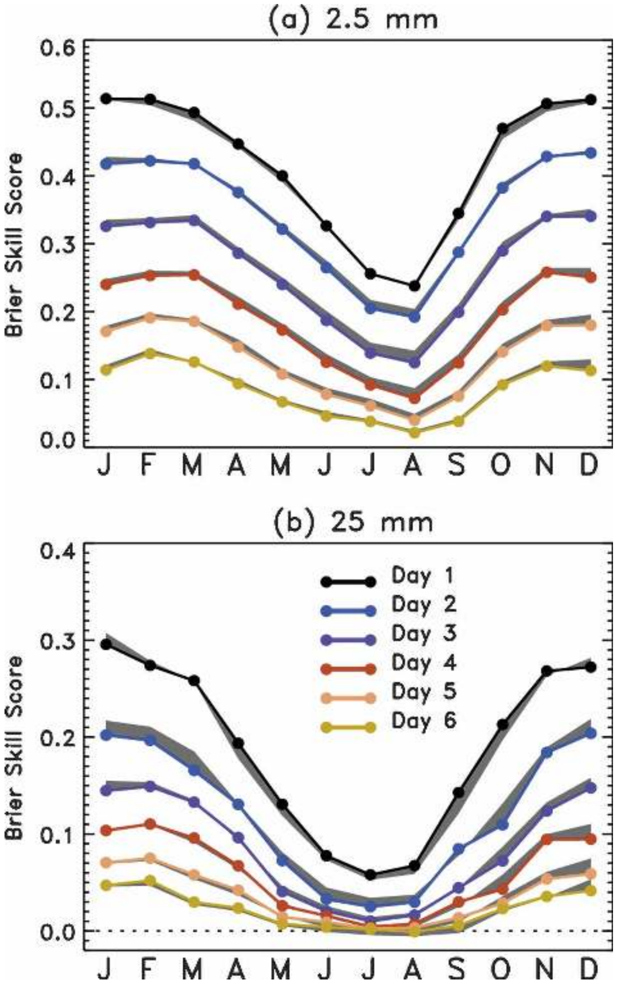


FIG. 9. Monthly BSS of the logistic regression technique, as in Fig. 8. Gray shading on forecasts indicates skill difference relative to basic analog approach in Fig. 8; gray shading below the reference line indicates more skill than the basic analog approach, and shading above the reference line indicates less skill.

relative to the basic analog technique. Fitting individual members provided forecasts of approximately equal skill for short leads but, for longer-range forecasts in the cool season, the skill was considerably worse when fitting individual members. We hypothesize that, at longer leads in the cool season, the filtering properties of the ensemble mean were helpful in extracting the predictable signal obscured by the chaotic error growth; when fitting individual members, one was fitting more noise than signal at the longer leads. In the summer, we hypothesize that model systematic errors played a more dominant role in limiting forecast skill and that the role of chaotic error growth was secondary.

The logistic regression technique included an extra predictor for precipitable water. If this extra predictor were incorporated into the analog technique, would the skill improve as well? Figure 11 presents the skill of the basic technique including precipitable water. While warm-season forecasts of light precipitation amounts were improved somewhat, the skill of the two methods was otherwise very comparable. Perhaps in the warm season, the forecast precipitation amount is very sensitive to the vagaries of the convective parameterization and its triggering scheme; if the parameterization is not uniformly accurate, then the extra predictor, precipitable water, can provide useful information.

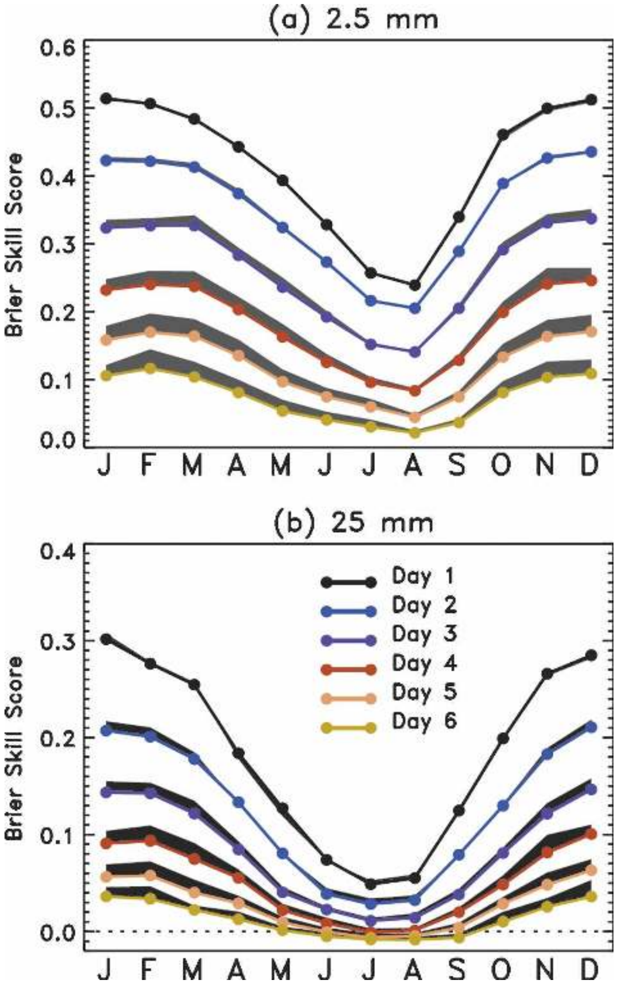


FIG. 10. Monthly BSS of the basic technique using individual members. Skill differences (gray shading, as in Fig. 9) are relative to the basic analog technique in Fig. 8.

When 2-m temperatures and 10-m winds were included as predictors into the basic analog technique, the skill was uniformly poorer than the basic analog technique (Fig. 12). While it is possible that temperatures and winds may have some predictive value in some circumstances, in this case the prespecified equal weighting of precipitation, temperature, and wind components was not a good choice; the de-emphasis of fore-

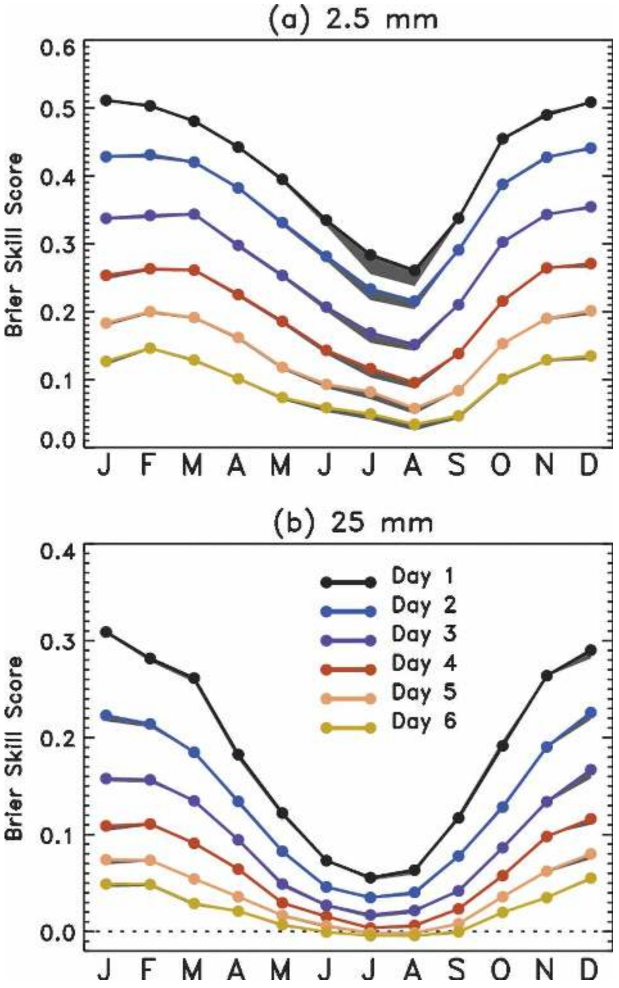


FIG. 11. Monthly BSS of the basic technique including precipitable water, with skill again compared via shading relative to the basic analog technique in Fig. 8.

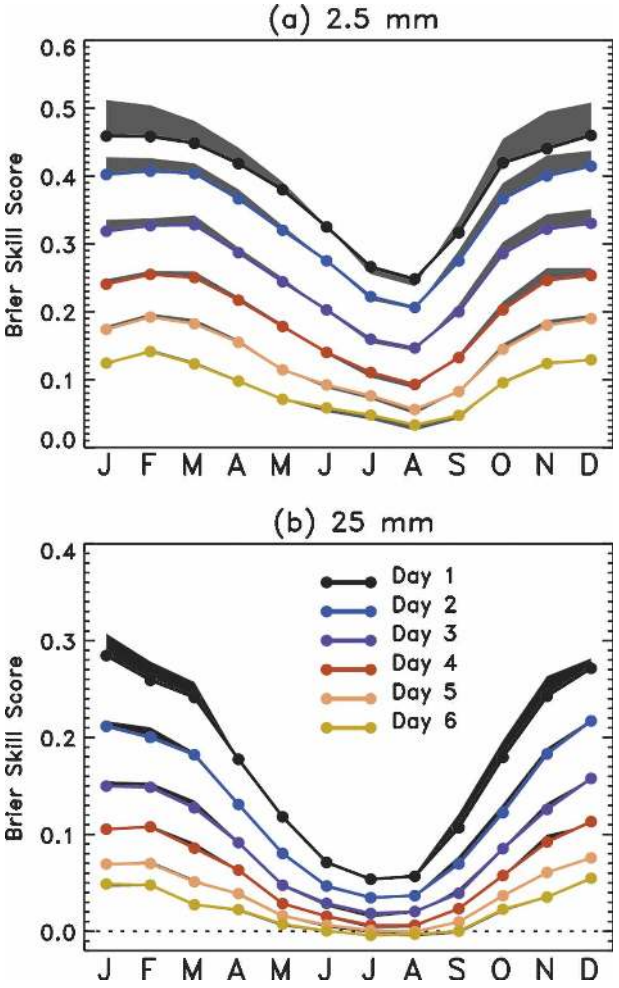


FIG. 12. Monthly BSS of the basic technique including 2-m temperatures and 10-m winds, with skill again compared via shading relative to the basic analog technique in Fig. 8.

cast precipitation as a predictor lessened the skill. However, a potential advantage of choosing analogs by a multivariate fit of precipitation, temperature, and winds is that, if some users desired information on joint probabilities (how likely is it to be cold, windy, *and* wet?), the joint probability distribution could have been estimated directly from this set of analogs, while analogs chosen by a closeness of fit of precipitation forecasts would likely not be of much use for estimating winds, temperatures, or joint distributions.

As indicated in section 3b(8), one deficiency of the basic analog technique that would be desirable to correct was a tendency for underforecasting precipitation probabilities, especially at short leads (Fig. 13a). This was due to the skewed, often exponentially shaped, climatological pdf of forecast precipitation causing a bias in the selection of closest analogs toward those with less forecast amounts. When the rank analog technique was used, the reliability was markedly improved (Fig. 13b): the BSS was also substantially higher for the 2.5-mm forecasts, especially at the short forecast leads, and the skill improvement was consistent across seasons (Fig. 14). However, the 25-mm rank-analog forecasts were slightly less skillful than the basic analog. The rank analog forecasts were more reliable (not shown) but they were slightly less sharp.

Would forecasts be improved if the rank analog technique used a smaller search region than the 16 points in Fig. 3? When using the inner 2 × 2 grid points, at short leads the skill of the 25-mm forecasts was improved substantially relative to the rank analog technique (Fig. 15). However, for the lighter precipitation amounts, the longer-lead forecasts in the warm season were slightly worse when using the smaller search region. The improvement at short leads for the high-precipitation threshold indicated that the important predictor was

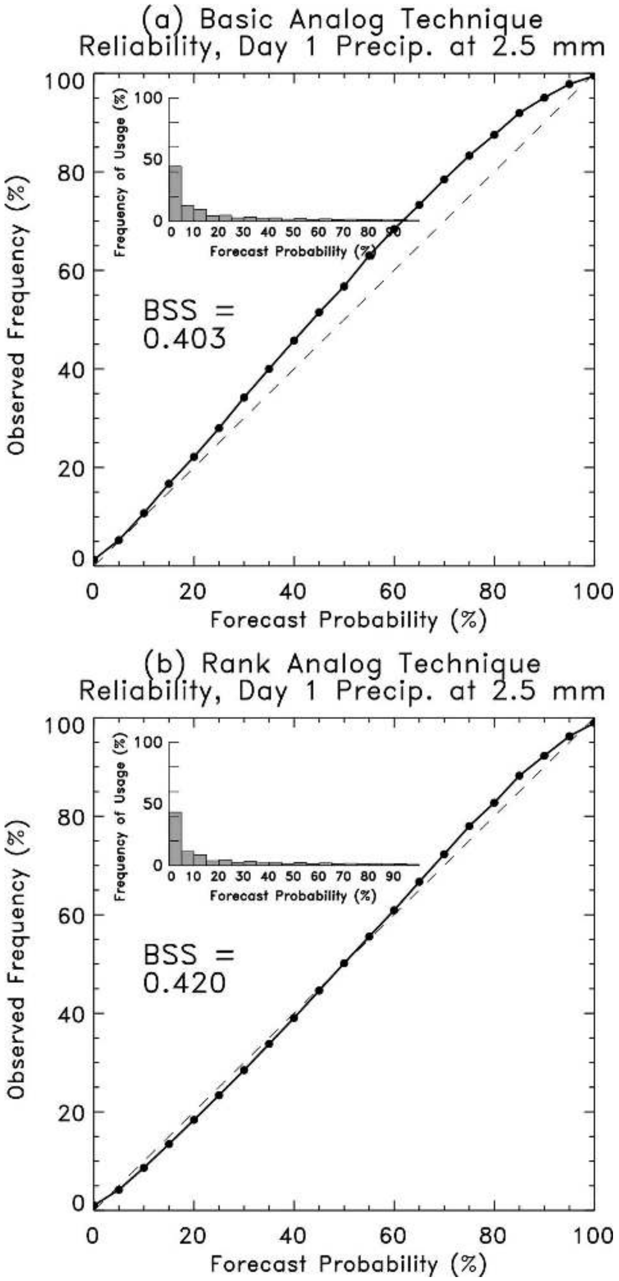


FIG. 13. Reliability diagrams for 2.5-mm 1-day forecasts from (a) 50-member basic analog technique and (b) 50-member rank analog technique.

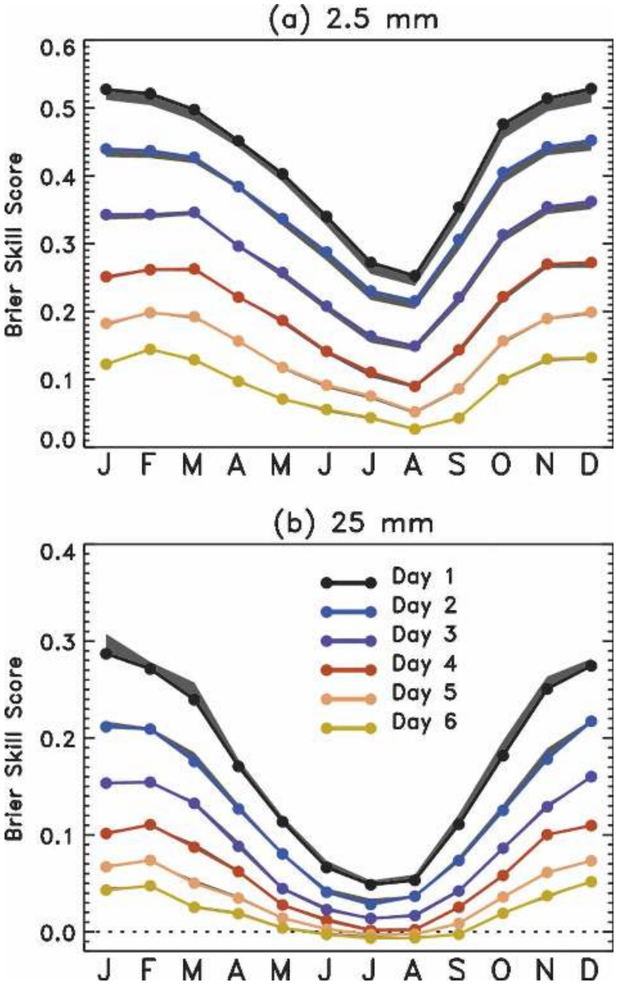


FIG. 14. Monthly BSS for the rank analog technique, with skill again compared via shading relative to the basic analog technique in Fig. 8.

the local precipitation forecast, and matching the pattern in the larger surrounding region was a less important consideration. The reason for the shorter-term forecasts being improved with the smaller search region was that systematic errors of the position bias were much smaller for the shorter-range forecasts. A cross-correlation analysis was performed that determined which NARR grid point had the highest rank-correlated precipitation analysis with a given forecast grid point's ensemble-mean precipitation during the summer. Averaged over the conterminous United States, 4.85 grid points separated the highest-correlated analyzed location from the original grid point for a 1-day forecast and 8.64 grid points for a 5-day forecast.

Finally, consider the effect of the smoothing algorithm discussed in section 3b(10). This technique was the same as the rank analog technique, but now the probability forecasts were smoothed to eliminate discontinuities in the probabilities along box boundaries. The smoothing produced a very slight improvement in the 2.5-mm forecast skill but improved the 25-mm forecast skill more substantially, especially at short leads (Fig. 16). An example of the subtle of the smoothing is shown in Fig. 17. Notice that the probability discontinuities in central Tennessee and western Georgia were smoothed between Figs. 17a and 17b.

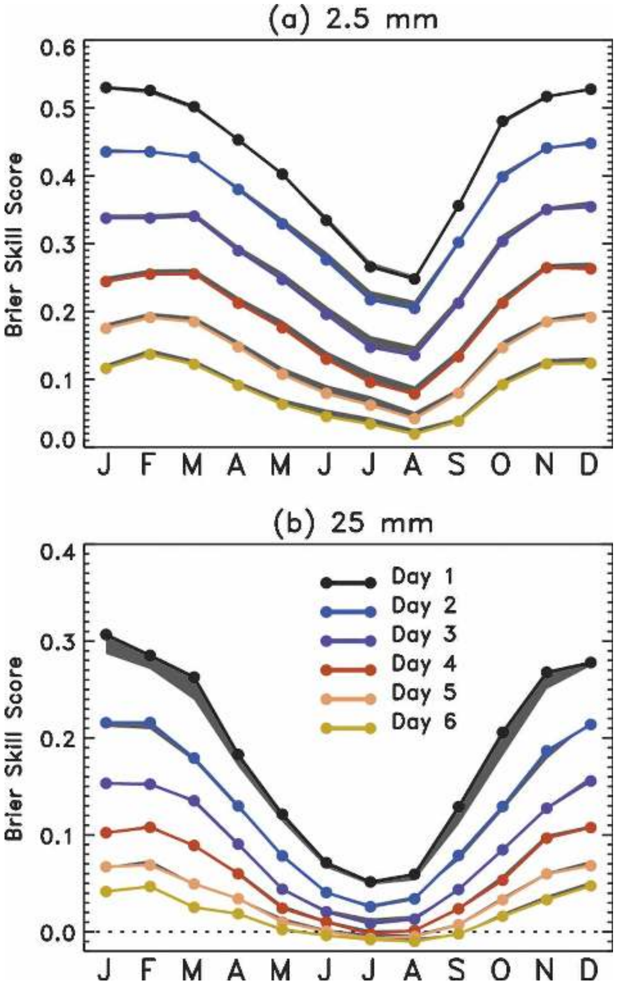


FIG. 15. Monthly BSS of the rank analog with smaller search region technique. Skill differences here are with respect to the rank analog technique in Fig. 14.

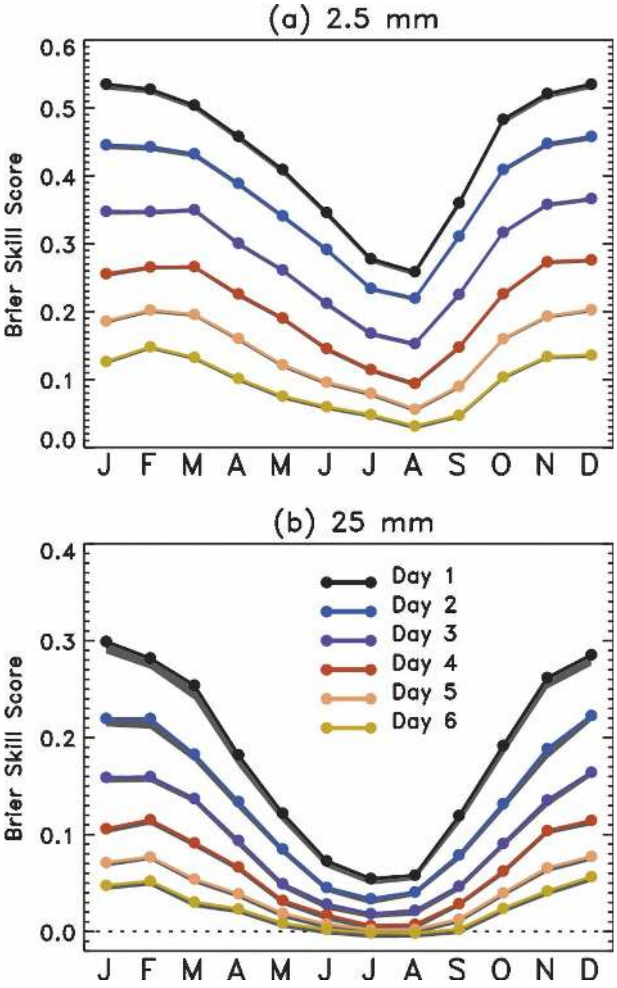


FIG. 16. Monthly BSS of the smoothed rank analog technique. Skill differences here are with respect to the rank analog technique in Fig. 14.

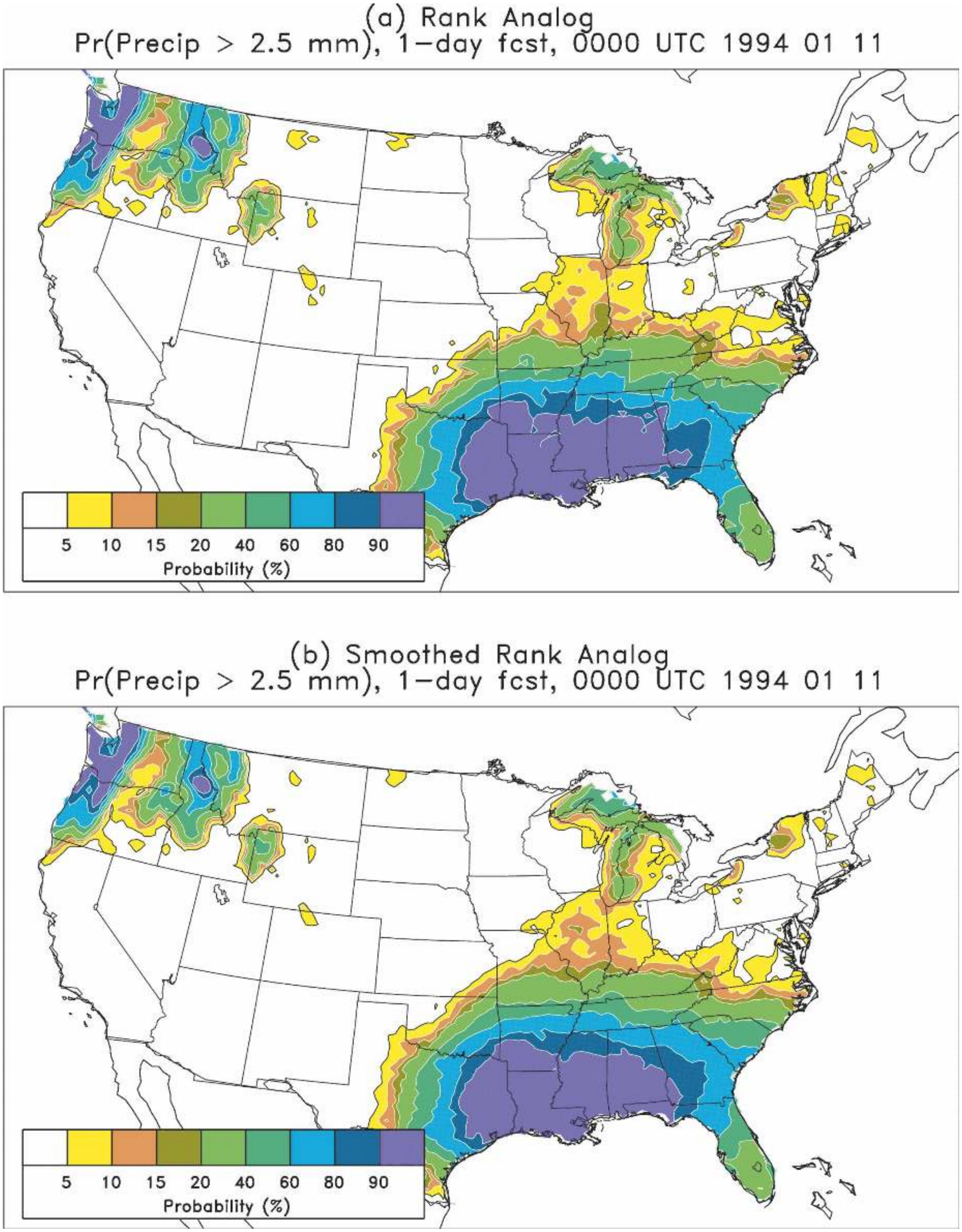
5. Conclusions and discussion

In this article we have examined the skill of probabilistic forecasts of 24-h precipitation amount from a variety of analog techniques that utilized a new, 25-yr global reforecast dataset produced by NOAA.

A general theory for probabilistic weather forecasting based on analogs was first proposed. Suppose an estimate is sought for the analyzed state's pdf given today's numerical forecast. Suppose also that we were provided with a nearly infinite set of reforecasts (hindcasts) and associated observations and that the climate was stable. Then, past model forecast states could be identified that are nearly identical to the current forecast state. Given the dates of these past analog forecasts, the asymptotically correct probabilistic forecast can be formed from the distribution of analyzed states on those dates.

This general theory could not be applied to global weather prediction, given a limited set of reforecasts, for the chance of finding even one similar forecast analog in that limited set is highly improbable, and the climate and forecast quality are not stable over these extremely long periods. However, approximations can be made to this theory to make it useful for statistically correcting weather forecasts. For instance, when estimating the local pdf of the analyzed state given the forecast, it was possible to choose the forecast analogs only based on the local weather. There commonly is an ample supply of highly similar local forecasts given a modest-length reforecast to compare. However, the rarer the event, the more difficult it is to find close forecast analogs.

We then examined several approximate PQPF analog forecast techniques using a 25-yr set of ensemble reforecasts. This period is presumably short enough for



(a) Rank Analog
Pr(Precip > 2.5 mm), 1-day fcst, 0000 UTC 1994 01 11

(b) Smoothed Rank Analog
Pr(Precip > 2.5 mm), 1-day fcst, 0000 UTC 1994 01 11

FIG. 17. Probability of greater than 2.5-mm precipitation for the 24-h period starting 0000 UTC 11 January 1994 from (a) rank analog technique and (b) smoothed rank analog technique.

the precipitation climate to be considered stable while long enough to provide an adequate training dataset, even for the calibration of relatively rare events. The analog techniques found past ensemble-mean forecasts in a local region that were similar to today's ensemble-mean forecasts in that region and formed probabilistic forecasts from the analyzed weather on the dates of the past analogs. All of the analog techniques provided dramatic improvements in the Brier skill score relative to basing probabilities on the raw ensemble counts or the counts corrected for bias. However, the analog techniques were generally similar in skill to those from a logistic regression technique. The analog techniques, however, are computationally less expensive and, once the analogs have been chosen, probabilities can quickly be estimated for any threshold. In comparison, the logistic regression techniques are more expensive, and the regression coefficients must be computed independently for each precipitation threshold.

Comparing the skill of various analog techniques, we found the following: 1) Selecting analogs for each member rather than for the ensemble mean generally decreased the forecast skill. 2) Finding analogs by matching not only mean forecast precipitation but also mean forecast precipitable water improved short-range, warm-season forecasts. 3) Finding analogs by matching surface winds and temperatures in addition to precipitation decreased the precipitation forecast skill. 4) Finding analogs based on the closeness of the relative rank of the mean forecast rather than its magnitude improved reliability at the short forecast leads. 5) A smaller search region was preferable when finding analogs for short-range forecasts, and a larger search region was preferable for longer-range forecasts. 6) Smoothing increased the skill of the forecasts slightly.

We also considered the effectiveness of a proposed bias-correction technique that adjusted precipitation amounts so that over many cases the forecast cumulative density function would match the analyzed cumulative density function. This procedure has been used in a slightly modified form at NCEP since 2004. This technique tended to improve forecast skill relative to the raw ensemble in the wintertime but worsen it in the summer. We determined that a CDF correction was generally unwise when the forecast and analyzed data are not highly correlated.

Despite the demonstrated skill and reliability of the reforecast-based techniques, many may believe it unwise to utilize forecast products from a T62, 1998 version of the NCEP GFS. Would it be wiser to base a PQPF upon raw output from more recent, higher-resolution model forecasts? While numerical precipitation forecast skill undoubtedly has improved in the past 10 years, there is reason to believe that the analog reforecast products demonstrated here are still competitive with these much newer, higher-resolution forecast models (HWM06). Calibrated products based on future, higher-resolution reforecasts should be even more competitive for, even with a better model, calibration with reforecasts still has been shown to provide substantial benefit (Whitaker et al. 2006).

For reforecasts to be of most benefit, the current numerical forecast should be conducted with the same model and data assimilation methods used in the production of the reforecast. This of course requires freezing the forecast model. Considered in isolation, this approach would be unattractive to weather prediction facilities, which would prefer to implement forecast model improvements quickly. Perhaps a dual-track system can be used, whereby a comparatively inexpensive fixed, reduced-resolution version of the forecast model is run alongside the frequently upgraded, operational higher-resolution version. The reduced-resolution version would have access to a reforecast dataset, computed offline. Users can choose for themselves whether they prefer guidance from the statistically adjusted reforecast model, raw guidance from the newer model, or some blend. Perhaps every few years, a new reforecast dataset would be produced with a more recent, higher-resolution version of the model so that the calibrated probabilistic guidance could leverage the improvement in the forecast models.

The literature has yet to demonstrate that quantum jumps in skill can be achieved with small training datasets. The skill increases in precipitation forecasts that we have demonstrated here are equivalent to the skill increases afforded by many years of sustained model development by a large staff of scientists. While we have concentrated here on demonstrating a calibration technique for precipitation, the statistical problems with precipitation are likely to be much more difficult than, say, for other commonly desired weather elements such as surface temperature. We expect that with a long reforecast dataset (saving more forecast variables than we did for this pilot project), it should be possible to produce calibrated probabilistic forecasts for even the thorniest of problems, such as precipitation type or severe-weather probability.

The U.S. National Weather Service is currently considering how to make skillful, reliable probabilistic weather forecasts a part of its National Digital Forecast Database (Glahn and Ruth 2003; Mass 2003a,b; Glahn 2003, 2005; Abrams 2004). Perhaps reforecast-based techniques are the most straightforward and promising way to achieve this goal.

Acknowledgments. The first author's participation was partially supported by NSF Grant ATM-0130154. We thank two anonymous reviewers for their meticulous reviews that added substantially to the quality of this manuscript.

APPENDIX A

Accessing Reforecast Data

A web form for downloading general reforecast data is available online at http://www.cdc.noaa.gov/reforecast. Using this form, precipitation and other forecast fields can be readily downloaded.

For those who wish to test their own precipitation forecast methods against those described here, the specific reforecast and training data used in this experiment are available at http://www.cdc.noaa.gov/reforecast/testdata.html. This includes forecast and analyzed precipitation data as well as selected FORTRAN90 codes.

APPENDIX B

Calculating Brier Scores and the Brier Skill Score

We indicate the specific method for calculating the Brier skill score (BSS). The method of calculation can have a large impact on the resulting scores. We use a conventional method for calculating BSS throughout this paper, but we shall demonstrate that this conventional method tends to overestimate skill.

The BSS is commonly calculated as

$$\mathrm{BSS} = 1.0 - \frac{\mathrm{BS}_f}{\mathrm{BS}_c}, \tag{B1}$$

where BS_f denotes the Brier score of the forecast and BS_c the Brier score of climatology. To calculate BS_f and BS_c, assume that we desire scores for an event threshold T. Let $p^f_{i,j}$ denote the probability of a forecast at grid box i and time j, where $i = 1, \ldots, n_x$, and $j = 1, n_d$. Here, n_x indicates the number of model grid points and n_d indicates the number of sample days; $p^c_{i,j}$ is similarly defined and denotes the climatological probability, which is allowed to vary with grid location and time of the year. Here, a running 60-day window was used to define the climatology; for example, the climatology for 1 December uses data from 1 November to 31 December for the period 1979–2003. Let $O_{i,j}$ denote the observed state at that grid point and date and let $I^o_{i,j}$ be an indicator variable, where $I^o_{i,j} = 1$, if $O_{i,j} > T$, and $I^o_{i,j} = 0$, if $O_{i,j} \leq T$. Then BS_f is calculated according to

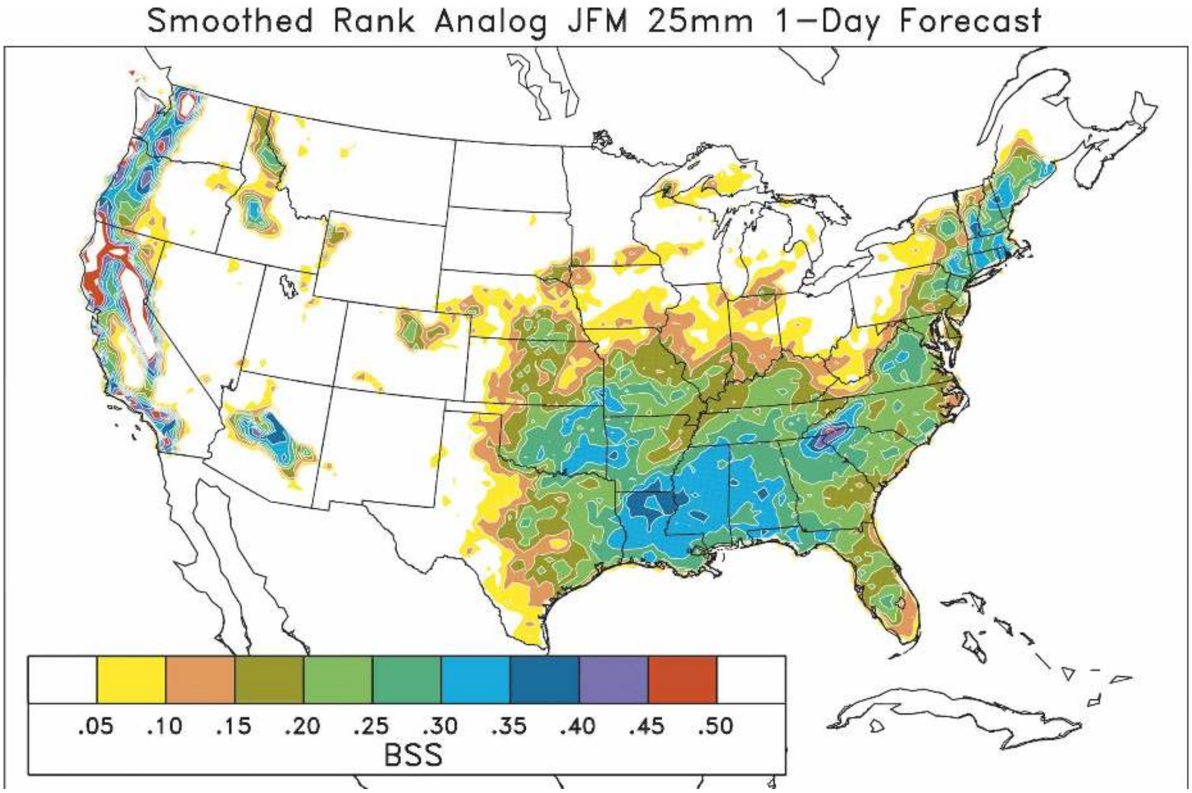


FIG. B1. Map of January–March (JFM) Brier skill score of 1-day 25-mm forecasts over the conterminous United States for the smoothed rank analog technique.

$$\mathrm{BS}_f = \sum_{i=1}^{n_x} \sum_{j=1}^{n_d} (p^f_{i,j} - I^o_{i,j})^2 \qquad \text{(B2)}$$

and

$$\mathrm{BS}_c = \sum_{i=1}^{n_x} \sum_{j=1}^{n_d} (p^c_{i,j} - I^o_{i,j})^2. \qquad \text{(B3)}$$

The BSS is then calculated using Eq. (B1).

While this method of calculation is quite straightforward and is commonly used, the resulting BSS can be quite different than the arithmetic average of BSS over all grid points. Figure B1 shows a geographic map of the day-1 25-mm wintertime BSS across the conterminous United States for the logistic regression method. The BSS reported for this day and threshold, according to Fig. 9, is approximately 0.27. Far less than half the map in Fig. B1 is covered by grid points with BSS $>$ 0.27. The underlying issue is that two grid points with different climatologies will receive different weights; in the calculation of Eq. (B3), wet grid points will receive more weight since the term $(p^c_{i,j} - I^o_{i,j})^2$ tends to be very small on average at dry grid points. This problem is accentuated at the high-precipitation thresholds. A similar problem is discussed in Juras (2000) and possible alternative methods of calculation are discussed in Hamill and Juras (2006).

REFERENCES

Abrams, E., 2004: Implementation and refinement of digital forecasting databases. *Bull. Amer. Meteor. Soc.,* **85,** 1667–1672.

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

Glahn, H. R., 2003: Comments on "IFPS and the future of the National Weather Service." *Wea. Forecasting,* **18,** 1299–1304.

——, 2005: Comments on "Implementation and refinement of digital forecasting databases." *Bull. Amer. Meteor. Soc.,* **86,** 1315–1318.

——, and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

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

Groisman, P. Ya., R. W. Knight, D. R. Easterling, T. R. Karl, G. C. Hegerl, and V. N. Razuvaev, 2005: Trends in intense precipitation in the climate record. *J. Climate,* **18,** 1326–1350.

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.,* in press.

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

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

Hastie, T. J., and R. J. Tibshirani, 1990: *Generalized Additive Models*. Chapman and Hall, 335 pp.

Juras, J., 2000: Comments on "Probabilistic predictions of precipitation using the ECMWF ensemble prediction system." *Wea. Forecasting,* **15,** 365–366.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1993: *The Essence of Chaos*. University of Washington Press, 227 pp.

Mass, C. F., 2003a: IFPS and the future of the National Weather Service. *Wea. Forecasting,* **18,** 75–79.

——, 2003b: Reply. *Wea. Forecasting,* **18,** 1305–1306.

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

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.,* **117,** 2230–2247.

——, 1994: Searching for analogues, how long must we wait? *Tellus,* **46A,** 314–324.

Whitaker, J. S., X. Wei, and F. Vitart, 2006: Improving week-two forecasts with multimodel reforecast ensembles. *Mon. Wea. Rev.,* **134,** 2474–2489.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate,* **12,** 2474–2489.