Calibrated Probabilistic Quantitative Precipitation Forecasts Based on the MRF Ensemble

F. ANTHONY ECKEL AND MICHAEL K. WALTERS

Air Force Institute of Technology, Wright-Patterson AFB, Ohio

Manuscript received 30 March 1998, in final form 15 June 1998

ABSTRACT

Probabilistic quantitative precipitation forecasts (PQPFs) based on the National Centers for Environmental Prediction Medium-Range Forecast (MRF) ensemble currently perform below their full potential quality (i.e., accuracy and reliability). This unfulfilled potential is due to the MRF ensemble being adversely affected by systematic errors that arise from an imperfect forecast model and less than optimum ensemble initial perturbations. This research sought to construct a calibration to account for these systematic errors and thus produce higher quality PQPFs.

The main tool of the calibration was the verification rank histogram, which can be used to interpret and adjust an ensemble forecast. Using a large training dataset, many histograms were created, each characterized by a different forecast lead time and level of ensemble variability. These results were processed into probability surfaces, providing detailed information on performance of the ensemble as part of the calibration scheme.

Improvement of the calibrated PQPF over the current uncalibrated PQPF was examined using a separate, large forecasting dataset, with climatological PQPF as the baseline. While the calibration technique noticeably improved the quality of PQPF and extended predictability by about 1 day, its usefulness was bounded by the intrinsic predictability limits of cumulative precipitation. Predictability was found to be dependent upon the precipitation category. For significant levels of precipitation (high thresholds), the calibration designed in this research was found to be useful only for short-range PQPFs. For low precipitation thresholds, the calibrated PQPF did prove to be of value in the medium range.

1. Introduction

The capability of ensemble forecasting to provide additional information for weather prediction is now well established (Hamill and Colucci 1997, 1998; Tracton and Kalnay 1993; Toth and Kalnay 1993; Molteni et al. 1996). In order to make output from ensemble forecasts useful operationally, the large amount of information must be condensed. One such technique is to create a probabilistic forecast for a weather parameter of interest. This research focused on probabilistic quantitative precipitation forecasts (PQPFs) based on the ensemble run of the Medium-Range Forecast (MRF) model. A PQPF is an answer to the question, what is the chance of getting greater than some amount of precipitation over a particular time period at a particular location?

Because atmospheric models are imperfect and techniques for generating ensemble perturbations are less than optimum, leading to underdispersive ensembles, the predictive power of ensemble forecasts is reduced by systematic error, that is, errors that occur repeatedly. It is possible to apply a calibration to account for this error and thus improve upon the ensemble forecast (Hamill and Colucci 1997, 1998). The goal of this research was to design a calibration to improve the quality (both accuracy and reliability) of the current, MRF-ensemble-based PQPF. At the National Centers for Environmental Prediction (NCEP), research with this same goal is under way using a very different technique that involves fitting and adjusting three-parameter gamma distributions to ensemble forecast output (Toth et al. 1998).

The procedure followed in this research was as follows. 1) Construct a calibration using a large sample of archived ensemble data and corresponding observations following the method of Hamill and Colucci (1997, 1998). 2) Using a separate large sample of archived ensembles, produce calibrated PQPF with the new calibration, uncalibrated PQPF using the current method, and climatological PQPF. 3) Measure the improvement of the calibrated PQPF over the uncalibrated PQPF using climatological PQPF as the baseline.

Section 2 of this paper describes the datasets applied in this research and how they were processed. Section 3 details the design of the calibration. Section 4 discusses the statistical techniques used to measure the improvement in PQPF realized due to the calibration and discusses findings concerning the limits of pre-
Table 1. Division of forecast case days into training and forecasting datasets. Forecasts were initialized at 0000 UTC on the dates shown.

<table>
<thead>
<tr>
<th>Month</th>
<th>Training data</th>
<th>Forecasting data</th>
<th>Dates of ensemble forecast case days</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEP 96</td>
<td>X</td>
<td></td>
<td>20 21 22 23 24 25 26 27 28 29 30</td>
</tr>
<tr>
<td>OCT 96</td>
<td>X</td>
<td></td>
<td>2 3 4 5 6 7 8 9 10 11 12 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>NOV 96</td>
<td>X</td>
<td></td>
<td>2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>DEC 96</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>JAN 97</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 19 20 21 22 23 24 25 26 27 28 30 31</td>
</tr>
<tr>
<td>FEB 97</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 31</td>
</tr>
<tr>
<td>MAR 97</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 7 10 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>APR 97</td>
<td>X</td>
<td></td>
<td>1 2 3 5 6 7 8 9 10 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>MAY 97</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>JUN 97</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>JUL 97</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 13 14 15 16 18 19 21 22 23 24 25 26 27 28 29 30 31</td>
</tr>
<tr>
<td>AUG 97</td>
<td>X</td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 11 12 13 15 16 17 19 20 21 22 23 24 25 27 28 29 30 31</td>
</tr>
<tr>
<td>SEP 97</td>
<td>X</td>
<td></td>
<td>1 2 3 6 7 8 9 10 11 12 13 14 15 23 27 28 29 30 31</td>
</tr>
<tr>
<td>OCT 97</td>
<td>X</td>
<td></td>
<td>3 4 5 6 12 13 14 16 19 20 24 25 26 27 28 30 31</td>
</tr>
<tr>
<td>NOV 97</td>
<td>X</td>
<td></td>
<td>4 6 7 8 9 10 13 14 15 19 20 21 22 23 26 28</td>
</tr>
</tbody>
</table>

dictability of cumulative precipitation. Section 5 presents conclusions and recommendations.

2. Research data

a. Ensemble data

A total of 358 daily gridded MRF ensemble forecast cases from the period September 1996 through November 1997 were used (Table 1). Many forecast case days over this period were missing due to computer or archival problems. The archived forecasts, originally produced as an operational forecasting tool at NCEP, were provided by the Climate Diagnostics Center of the National Oceanic and Atmospheric Administration.

This research focused on PQPF for 24-h cumulative precipitation (pcp24). Each ensemble forecast, with an initial time of 0000 UTC and a 2.5° gridpoint resolution, consisted of 17 ensemble members valid every 12 h over a 16-day period. For a complete description of the NCEP ensemble system, see Toth and Kalnay (1993), Tracton and Kalnay (1993), and NCEP (1997). Although 12-h cumulative precipitation is available, pcp24 was used since it corresponded to the verification data (valid 1200–1200 UTC) described below. Therefore, for each ensemble member, the 1200–0000 and the 0000–1200

![Fig. 1. MRF 2.5° grid over the contiguous United States. Grid points (centers of grid boxes) marked with a dot had both pcp24 observations and climatological data and were used for both construction of the calibration and evaluation of PQPF quality. Grid points marked with a star had no climatological data so they were only used for calibration construction.](http://journals.ametsoc.org/doi/pdf/10.1175/1520-0434(1998)013<1132:CPQPFB>2.0.CO;2)
cumulative precipitation forecasts were combined over the forecast valid period to obtain $pcp_{24}$ forecasts valid at the 36- (1.5-), 60- (2.5-), . . . , and 372-h (15.5-day) forecast lead times. Each lead time refers to $pcp_{24}$ in the previous 24 h.

The full dataset of 358 forecast case days was divided by month into a training dataset and a forecasting dataset, as shown in Table 1. Here dates are the initialization date of each case day. The training dataset was used exclusively for construction of the calibration. The forecasting dataset was used in generating PQPF after development of the calibration, so forecasts were made with no prior knowledge of the verification value. This allowed for fair evaluation and comparison of PQPF quality.

b. Observational data

For both construction of the calibration and measurements of PQPF skill, observations of cumulative precipitation were needed that were consistently representative of the region within each box. This was a difficult problem since precipitation is highly variable in both space and time, and the 2.5° grid boxes are large (e.g., the area of a 2.5° grid box at 40°N is approximately 60 000 km²).

A reasonable method for generating representative observations is to average the observations from many rain gauges within each grid box (Baldwin 1997). NCEP provided $pcp_{24}$ observations valid daily at 1200 UTC for MRF 2.5° grid points within the conterminous United States (Fig. 1) for September 1996 through November 1997. The data were prepared by NCEP using reports from a network of approximately 10 000 rain gauges belonging to the National Weather Service’s River Forecast Centers. For a more detailed description of how these data were processed, see Baldwin (1997).

c. Climatological data

PQPF based on climatology (i.e., for a particular region on a particular day, what is the average chance of receiving greater than a threshold amount of $pcp_{24}$?) was required to determine the skill of the PQPF based on the MRF ensemble. More specifically, climatological PQPF was needed for the same grid boxes as the forecast and verification data. A good way to generate such PQPF is through a technique that uses the climatological probability distribution function of $pcp_{24}$ (Wilks 1995), as described in section 4 below. Unfortunately, the network of rain gauges that was used for the $pcp_{24}$ verification data has not been in service long enough to provide a climatological database from which to obtain representative distributions. The alternative approach used here was to use climatological data from meteorological stations.

The first step was to estimate the theoretical distribution of $pcp_{24}$ at each station, for each day of the year, using the daily $pcp_{24}$ observations from 120 U.S. meteorological stations from 1948 to 1991. The gamma distribution was chosen as the most appropriate because of its ability to represent the phenomenon of precipitation. The $pcp_{24}$ data contained many zero values, which makes either the method of moments or standard maximum likelihood estimation techniques unreliable methods of estimating the gamma parameters ($\alpha$ and $\beta$). This problem was avoided by applying Wilks’s (1990) method, which is specifically designed for data that contain many zeros.

To increase sample size per distribution and simplify
processing, the year was grouped into 73, 5-day periods. It was assumed that the climatological pcp24 does not change significantly at a given location over any consecutive 5-day period. The resulting parameters for each station had a notable seasonal shift, as was expected, but with considerable noise that had to be smoothed out (Fig. 2). The smoothing routine computed a centered, 11-period weighted average where weights were defined by inverse distance from the center (i.e., the data point four periods away from the center was weighted by \( \frac{1}{4} \)). Since the data were treated as periodic, late December was smoothed with early January.

The next step was to fit the distributions to the 2.5° grid. A method such as a Barnes (1964) interpolation was ruled out since it would produce poor results by incorporating data from stations unrepresentative of a grid box’s climatology. It was decided that the best method was to apply a slightly modified nearest neighbor technique. If a station within a grid box was subjectively determined to be representative for the entire grid box, then that station’s climatological gamma distributions were used to represent that grid point. For example, the distributions analyzed for Omaha, Nebraska, were used for the grid box centered at 42.5°N, 97.5°W even though this station is at the edge of the grid box. Climatic precipitation has a low spatial variability in this region so any station is representative of a large area. Grid boxes without any stations, or with stations considered unrepresentative, were excluded from climatological PQPF calculations (Fig. 1).

3. The calibration

The backbone of the calibration, modeled after work done by Hamill and Colucci (1997, 1998), was the verification rank histogram. A verification rank is the ranking of the verification value (observed pcp24 amount)
when it is pooled with the ordered members of the ensemble forecast at a grid point. Since there are 17 ensemble members, there are 18 possible verification ranks. The verification rank histogram tracks the performance of the ensemble by tallying the number of times the verification occurs in each verification rank. After many independent samples of forecast and observation pairs, systematic errors in the ensemble reveal themselves through nonuniformity of the histogram’s distribution (Anderson 1996).

Figure 3a is an example of a verification rank histogram for all 5.5-day forecasts. A probability at each verification rank is determined by dividing the total number of times the verification occurred in a rank by the total number of forecast/observation samples. The highest probability of about 0.15 in rank 18 means that over the sample space, the most common event was that all 17 ensemble members underforecast the pcp24 amount.

The idea introduced by Hamill and Colucci (1997, 1998) was that the information in the verification rank histogram can be used in interpreting and correcting the ensemble forecast. This in effect accounts for the sys-
a. Use of correlated data

For the construction of a verification rank histogram, it was briefly noted that samples should be independent; that is, no correlation should exist between samples (Anderson 1996). If correlated data were used, population in the ranks could become unrepresentative. This presented a dilemma for this research because precipitation data are highly correlated over small spatial scales and short timescales. Since data points in the sample space were on average 200 km and 1 day apart, a high degree of correlation was considered likely. To reduce the correlated errors and thus produces a calibrated probabilistic forecast. For a detailed description of this method, hereafter referred to as the weighted ranks method, see Hamill and Colucci (1997, 1998).

Fig. 8. Probability surfaces of 5.5-day forecasts constructed for the different weather regimes of the (a) NE United States in summer and (b) SW United States in winter. The differences in the two surfaces, (b)–(a), is shown in (c).

Fig. 9. Determination of climatologically based CAT1 PQPF. Probability of receiving pcp above the CAT1 threshold of 2.54 mm (marked with an asterisk) is 32%, the shaded area under the climatological PDF, also known as the 1 - p value.

relation to an acceptably low level, the sample space would have to be thinned out.

However, it was hypothesized that use of correlated data presented no such problem in the present application. If sampling were done over a very disperse geographical region, the unrepresentativeness of the ranks might be balanced out. To test the hypothesis, the verification ranking process was run twice, sampling once from the original, highly correlated data and again from sampled data with reduced correlation. If the resulting histograms were to show no statistical difference, it could be confidently concluded that use of correlated data is legitimate in making verification rank histograms for a sample space with disperse data.

This hypothesis was tested using ensemble forecasts at various lead times. One verification rank histogram was created using the complete sample space (all case days, all grid points, for one lead time), a total of 41 704 verifications (Fig. 3a). The sample space was then divided up to produce two more histograms from data with lower correlation. Forecast case days were divided by every other day giving two distinct sets of forecast cases (called A and B) with at least two days (more for missing days) between forecast initial times. The grid (Fig. 1) was divided up in checkerboard fashion giving two distinct sets of grid points (called a and b) with

Fig. 10. Ranked probability skill score (RPSS) results for all forecast lead times.
increased distance between points. One verification rank histogram used data from $A$ with $a$ (Fig. 3b) while the other used $B$ with $b$ (Fig. 3c). The size of the sample space for these histograms was 10,634 verifications, one-quarter of the complete sample space.

Figure 3 compares the resulting verification rank histograms for just 5.5-day forecasts. From a visual examination of Fig. 3, it is difficult to see any differences at all between the three histograms. To be thorough, $\chi^2$ tests were done to test the statistical significance of the differences in the histograms. All comparisons strongly passed with $p$ values of 1.0. Tests performed at other forecast lead times gave similar results.

Since statistically equivalent verification rank histograms were produced from datasets with probable different levels of correlation, it was concluded that use of correlated data in this research was legitimate. This was a critical finding, because the samples had to be finely divided up into groups of similar characteristics in order to produce a robust calibration. If these divisions were made from a thinned out sample space to achieve low correlation, there would not have been enough data to accomplish the desired calibration.

It may be argued that the results of this test are suspect since the level of correlation in the complete sample space and in the subsample spaces was not rigorously established. However, this is a moot point because, as will be shown in section 4, the calibration technique worked quite well regardless of whether or not use of correlated data is valid in constructing verification rank histograms.

\[ \text{Fig. 11. Brier skill score (BSS) results for all four PQPF categories and all forecast lead times.} \]

\[ \text{CAT1 BSS} \]

\[ \text{CAT2 BSS} \]

\[ \text{CAT3 BSS} \]

\[ \text{CAT4 BSS} \]

\[ \text{Forecast Lead Time (Days)} \]

\[ \text{1.5} \quad 2.5 \quad 3.5 \quad 4.5 \quad 5.5 \quad 6.5 \quad 7.5 \quad 8.5 \quad 9.5 \quad 10.5 \quad 11.5 \quad 12.5 \quad 13.5 \quad 14.5 \quad 15.5 \]

\[ \text{b. Construction of the calibration} \]

Although it is possible to make a verification rank histogram from the total sample space (all grid points and lead times of training dataset), the results would be too generic to produce a robust calibration. The key to this research was to construct many verification rank histograms, each able to account for a slightly different systematic error. A logical choice for dividing up histograms was by ensemble variability, as shown by Hamill and Colucci (1997, 1998), which is an important characteristic of an ensemble forecast. Ensemble variability was quantified by the sample standard deviation ($s$) of the 17 ensemble members at a grid point. The value of $s$ has large temporal and spatial variability. Generally, the wetter the forecast at a point, the higher the value of $s$.

Hamill and Colucci (1997, 1998) constructed only three verification rank histograms based on high, medium, and low $s$, since they were limited by a small number of samples. Since this research had a plethora of data, a more detailed stratification was possible. With roughly 20,000 sample verifications available in the training dataset at each lead time, it was possible to divide $s$ into many class intervals while at the same time maintaining large enough subsample space sizes. A unique verification rank histogram could then be constructed for each interval. Sixteen class intervals were chosen, based on the distribution of $s$, giving around 1000 samples for each histogram. The range of the class intervals increases with increasing $s$, since it was found that the sampling distribution of $s$ is skewed to the right with a peak near zero. Table 2 gives the class intervals used in constructing the 16 histograms for 2.5-day forecasts. Each lead time was handled in a similar manner and produced notably different results. Note that totally dry forecasts ($s = 0.0$) are a unique class, not included as part of the calibration.

The resulting 16 verification rank histograms gave a very interesting picture when viewed together (Fig. 4a). Discounting the noise due to the reduced sample sizes, the rank probability appeared to vary smoothly along the class intervals as well as across the ranks. This result can be anticipated if the systematic error does have some dependence on ensemble variability. The consequence of this result is that it was possible to fit functions to the behavior at each rank. Instead of a discrete number of verification rank histograms at each lead time, there could now be an infinite number.

Before fitting functions at each rank, each histogram...
was smoothed to remove the noise between the ranks. The probability values of the outside ranks (1 and 18) were not changed since smoothing would alter them too much. The histograms were then normalized so the probability summed over all the ranks would equal one for each histogram. Figure 5 shows the raw and processed data for class interval 3 of 2.5-day forecasts. Figure 4b shows all 16 smoothed verification rank histograms.

To fit a function at each rank, the median of each class interval (Table 2) was chosen to represent the value of the independent variable (standard deviation) for the corresponding values of the dependent variable (rank probability) for each of the 18 separate functions. It was found that third-order polynomials made an excellent fit at any rank once the independent variable was transformed to a natural log scale. Figure 6 is an example of this process for the function fit to rank 17 for 2.5-day forecasts.

For comparison, Fig. 4c displays the resulting histograms for all the original class intervals of the 2.5-day forecasts. The information now available should no longer be displayed in this fashion since any one of the multitude of possible histograms does not represent a discrete range of standard deviation anymore. For the 2.5-day forecasts example, there is now a continuum (Fig. 7b) where any particular value of standard deviation determines a histogram with a unique set of 18 rank probabilities.

The name given by the authors to this continuum of verification rank histograms is a probability surface. Technically, it is not a surface because while it is continuous along the ln(s) axis, the rank numbers are still discrete integer quantities. The surface is really just a series of 18 separate functions. However, viewing the functions as a surface is a useful tool for interpreting and understanding how they work in the calibration scheme.

Notice that the ln(s) scale extends beyond the range of the transformed median values in both directions. For these values, the functions are extrapolated to ln(s) values outside of the original data values so the rank distribution becomes less realistic. The lowest possible value of s is 0.024 mm, which transforms to −3.73 on the log scale. The high value plotted of 4.0 is for s = 55 mm, an extremely high value that rarely occurs. While the extrapolated regions of the probability surface may be less realistic, this does not have a serious impact on the calibration since the s values rarely occur there.

On some surfaces, the extrapolated curves fall below 0.0 probability. For these cases, the rank probabilities are frozen to the last s values that did not have a probability value less than 0.0. All s values beyond this point then repeat the same histogram creating a leveling effect of the surface (Figs. 7d–f). This also has little affect on the calibration for the same reason noted above.

Visual inspection of Fig. 7 reveals similarities of the probability surfaces created for different lead times but significant differences as well. The most noticeable changes occur within the first 6 days of the forecast period (Figs. 7a–d). For forecasts beyond 6 days, the probability surfaces still steadily change with increasing lead time but not as drastically.

As previously mentioned, these probability surfaces are track records of how the ensemble performed over the training dataset. To interpret the surfaces, consider the probability surface for 3.5-day forecasts (Fig. 7c). For forecasts with negative ln(s) values (very little precipitation), the high probability in the upper ranks means that the verification value usually was greater than all or most of the ensemble members’ forecast value of pc p 24 . In other words, the ensemble frequently underforecasted pc p 24 in these conditions. For ensemble forecasts with ln(s) around 0.0, the highest probabilities are in both the lower and higher ranks indicate that the ensemble often underforecasted or overforecasted pc p 24 amounts. Last, for ensemble forecasts with positive ln(s) values (high precipitation), the probability shifts mostly into the lower ranks so the ensemble usually overforecasted pc p 24 .

This common feature among the probability surfaces of high probability in the extreme ranks, similar to the findings of Hamill and Colucci (1997, 1998), indicates that the ensemble often fails to encompass the verification. This is most likely due to limited dispersion among ensemble members. For low dispersive ensemble forecasts, the verification often occurs outside of the range of forecast values. This equates to either an overforecasting or an underforecasting event. This subject is discussed further in section 5.

c. Regime dependence

The verification rank distribution was shown to have a strong dependence on both ensemble standard deviation and forecast lead time. This indicates that the strategy of using these probability surfaces should produce a good calibration. Is there a further stratification that may provide a better calibration? To examine this possibility, the regime dependence of the probability surfaces was briefly explored.

A weather regime is described by a defined set of conditions over an area in which some similar relationships and influences on atmospheric dynamics are shared (i.e., low atmospheric stability, warm ocean temperatures, continental polar air, etc.). Because the ensemble’s systematic errors will likely differ between regimes due to differences in dynamical forcing between regimes, a more specific calibration for each regime would likely result in a more robust calibration scheme. Regimes defined by geography and season were explored to determine the possible significance of regime dependence for the calibration.

Separate probability surfaces for 5.5-day forecasts were created using data from two completely different regimes. If these surfaces proved to be different, then
a regime dependence for the surfaces could be confirmed. Two rough geographical regimes were created by dividing the grid (Fig. 1) diagonally from Washington to Florida, giving a northeast U.S. regime and a southwest U.S. regime. Two seasonal regimes were defined as summer (May through September) and winter (November through March). Since this exploratory study of regime dependence was not intended to be part of the calibration, the entire archive of ensemble cases was used (except for the transition months of October and April).

Various combinations of these regimes were used to produce 5.5-day forecast probability surfaces. The resulting surfaces were similar but also had noticeable differences. Figure 8 shows the probability surfaces for the northeast U.S. regime in summer and the southwest U.S. regime in winter. The same general pattern is evident in both, but they do produce different calibrations. For example, the larger probabilities in rank 18 for the southwest United States in the winter regime (Fig. 8b) means that the ensemble is much more likely to underforecast precipitation events in that regime. There is a regime dependence.

This finding has two important conclusions: 1) A more specific calibration than the one used in this research is possible, and 2) the calibration developed here is only valid in the regime over which it was developed (the United States for all seasons). Because the differences in the probability surfaces between regimes are small when compared with the probability magnitude of the more general 5.5-day surface (Fig. 7d), the benefit of further stratification of probability surfaces may be small compared to the amount of effort it would take in defining appropriate regimes. For this reason, regime-dependent probability surfaces were not pursued further in this research.

4. The calibration’s improvement of PQPF

The last step of this research was to measure improvement of the calibrated PQPF over the current uncalibrated PQPF using climatological PQPF as the base-
line. This section begins by describing how the three distinct PQPFs (uncalibrated, climatological, and calibrated) were generated. It then presents the results of various statistical techniques and discusses the findings.

Currently, PQPF is produced operationally at NCEP from the MRF ensemble with a method defined here as the *democratic voting* method. As the name implies, each ensemble member gets an equal vote on the probability of precipitation occurring above some threshold. Mathematically, the number of ensemble members that forecast above the threshold are first tallied. Dividing the tally total by the total number of members yields the probability. This method produces an approximate PQPF that does not account for systematic error.

This research applied the calibration to the four PQPF categories currently used by NCEP over the contiguous United States. These categories are CAT1: pcp$_{24}$ ≥ 0.1 in. (2.54 mm), CAT2: pcp$_{24}$ ≥ 0.25 in. (6.35 mm), CAT3: pcp$_{24}$ ≥ 0.5 in. (12.70 mm), and CAT4: pcp$_{24}$ ≥ 1.0 in. (25.40 mm).

Climatological PQPF was produced using the climatological gamma distributions described above (Wilks 1995). For example, for the grid box centered at 47.5°N, 122.5°W in western Washington, PQPF for CAT1 for a day in mid-January was computed using the $1 - p$ value of the climatological pcp$_{24}$ distribution (Fig. 9). The result of 32% means that cumulative precipitation of 2.54 mm or higher is normally observed in about one out of three days in this region in mid-January. Higher thresholds of course result in lower PQPF and vice versa.

The technique for generating calibrated PQPF from a verification rank histogram, termed the weighted ranks method by the authors, is well described in Hamill and Colucci (1997, 1998). Basically, the method uses the information on past performance of the ensemble, contained in the verification rank histograms, as a systematic error filter for producing calibrated probability forecasts from an ensemble forecast. The only change applied here was that the process for computing PQPF for a category had to be adapted for NCEP’s categories of overlapping pcp$_{24}$ ranges, which, unlike model output statistics categories, are not mutually exclusive and collectively exhaustive.

Quality of PQPF was measured and compared with three different tools, namely, the Brier skill score,
ranked probability skill score, and the reliability diagram. The calibrated PQPF produced by the weighted ranks method proved to produce a significant improvement over the uncalibrated democratic voting method.

a. Ranked probability skill score

The ranked probability skill score (RPSS) is a mean-square error (mse) measure for multicategory forecasts with respect to climatological forecasts (Wilks 1995). It can be used to compare the relative accuracy of two forecasts over a large sample as well as establish the limits of predictability of these forecasts. The skill of climatologically based forecasts is 0.0, so positively scored forecasts beat climatology while negative ones do not.

These RPSS results (Fig. 10) may be considered robust because of the large sample size applied. Sample sizes for each lead time varied within a few hundred of 15 000 due to a few missing days of observations. The critical finding is that the calibration was successful; that is, PQPF derived using the weighted ranks method shows a significant improvement in accuracy over the democratic voting method. The observed decrease in skill for both methods with increasing forecast lead time is expected. The fact that the skill scores converge means that the effectiveness of the calibration also decreases with forecast lead time.

Lorenz (1963) demonstrated that atmospheric modeling has predictability out to only a finite amount of time into the future. Therefore, it is expected that a climatologically based forecast will become the better forecast at some lead time. In other words, the RPSS should decrease below zero at some point and remain negative thereafter, as displayed in Fig. 10. Since skill scores are relative to climatologically based PQPF, exactly where the score becomes negative is dependent on the mse of the climatological PQPF. The shape of the RPSS plot over the valid period is set, but it can shift upward or downward depending upon the accuracy of the climatological PQPF. This possibility is explored in more detail below.

As shown in Fig. 10, the calibration extended predictability by about 1 day up to a lead time of 6 days. Predictability out to only 6 days is not a surprising finding considering the variability of precipitation. However, since the RPSS is an overall score for all categories, this estimate of predictability is dominated by the lower threshold categories. Predictability at higher threshold categories is actually much worse than is indicated here, as discussed below.

b. Brier skill score

The Brier skill score (BSS) is a breakout of RPSS by category. Sample sizes are the same as with the RPSS, but now an mse measure is found for each category separately, and then referenced to the mse of climatologically based forecasts.

The general conclusions of the BSS results (Fig. 11) are the same as with the RPSS—the calibration improves PQPF accuracy and extends predictability. The feature that is now evident is that both skill and predictability steadily decrease with increasing category. In other words, higher thresholds of cumulative precipitation are more difficult to predict. For example, Fig. 11 shows that the CAT4 forecasts produced by the democratic voting method have negative skill at all time periods. This finding agrees with intuition in that the more extreme an event, the harder it is to forecast.

There is an odd feature of these graphs that is most pronounced for CAT4. Skill steadily decreases to a point and then appears to recover somewhat before leveling off. This affect, explained in the case study below, allows the weighted ranks CAT4 BSS to stay barely positive up to almost 5 days, which is better than CAT3. This inconsistency leads to the conclusion that the climatologically based PQPF is somehow deficient, making the MRF-based PQPF appear to have more predictability. The plots of BSS may be unrealistically shifted too high.

The deficiency in the climatologically based PQPF is likely the result of the method used to fit the distribution...
from one station to a grid point. Cumulative precipitation at one station can contain many extreme events that may make the gamma distribution unrepresentative of the precipitation over the entire grid box. An alternative explanation is that precipitation during the period of the forecast dataset may have departed significantly from climatological norms. This would cause larger mse in the climatological PQPF and shift the MRF-based PQPF skill upward. An extremely intense El Niño did begin halfway through the forecasting dataset, which may have caused significant departures from climatology in the precipitation data used.

c. Reliability diagram

A reliability diagram is a graphic display of the performance of a set of probabilistic forecasts for a certain category. Observed relative frequency of occurrence in the category is plotted as a function of forecast probability for the category (Wilks 1995). For simplification and to increase the population at each probability bin, forecast probabilities are rounded to the nearest 10%.

Reliability diagrams for all four PQPF categories for 1.5-day forecasts are displayed in Fig. 12. These diagrams show that the calibrated PQPF of the weighted ranks method consistently outperformed the uncalibrated PQPF of the democratic voting method; that is, weighted ranks plots of observed relative frequency are closer to the perfect forecast diagonal. However, it is evident that the calibrated forecasts exhibit the same tendency toward overforecasting in the mid- and high forecast probabilities and underforecasting the very low forecast probabilities. It is also evident that reliability decreases for the higher categories (thresholds), which agrees with the BSS analysis.

A skill score (SS) can be computed from a reliability diagram from (resolution — reliability)/uncertainty (Wilks 1995). The SS is similar to the BSS with two differences: 1) The SS uses the forecast probabilities that were rounded to the nearest 10%, whereas the BSS uses the actual forecast probabilities; and 2) the SS is measured with respect to the sample climatology, whereas the BSS is measured with respect to the long-term climatology. The sample climatology is the overall frequency of occurrence in a category, within the sample space. It can be thought of as the short-term climatology for just the forecast dataset.

The SS results (Fig. 13) are very similar to the BSS results (Fig. 11) with two major exceptions: 1) The SS curves are shifted lower, and 2) the two PQPF methods converge more tightly. While both scores have their limitations, the SS is probably more representative. The SS does not suffer from the potential problems of the climatological analysis mentioned above. This is evident in the CAT4 SS, which does not hover above zero for an extended period as with the BSS.

d. Case study

This brief PQPF case study demonstrates the calibration’s effects and explains the source of the peculiarities noted in the skill and reliability analysis above. From 1200 UTC 7 November 1996 to 1200 UTC 8 November 1996, an intense extratropical cyclone caused greater than 25.4 mm of rain (CAT4 threshold) to fall over a large area from the Gulf of Mexico into the Ohio Valley (Fig. 14a). This case study compares the uncalibrated (democratic voting) and the calibrated (weighted ranks) PQPF from the days leading up to the precipitation event (Figs. 14b–e). With just this one study, it is difficult to effectively demonstrate all the aspects of the calibration’s improvements. This case was selected because it shows the skill score recovery phenomenon so well.

Comparing the two PQPFs, the most obvious effect of the calibration is to decrease higher forecast probabilities. This results in the weighted ranks PQPF being more reliable than the democratic voting PQPF, which is so often an overforecast for higher forecast probabilities as seen in the CAT4 reliability diagrams (Figs. 12f and 12h). This effect is especially apparent in the 2.5-day PQPF in northern Ohio where pcp24 failed to occur in CAT4.

The other effect of the calibration is to increase the lower forecast probabilities. This is most evident by the extension of the 0% isopleth in the weighted ranks PQPF charts. This also improves reliability because the democratic voting PQPF has a slight underforecasting problem in the lower forecast probabilities. There is a small region in South Carolina and Georgia where democratic voting PQPF was zero but pcp24 was observed in CAT4.

While the calibration adjusted for both of these problems (i.e., overforecasting higher forecast probabilities and underforecasting lower probabilities), it still suffered from these problems to a lesser degree. Although the calibration mostly decreased the probability in the region where pcp24 was observed in CAT4, it resulted in improved accuracy by dramatically decreasing forecasts of higher probabilities in the larger regions where pcp24 was not observed in CAT4. For example, compare the size of the area of 50% probability produced by each of the two methods shown in Fig. 14b.

Now consider the phenomenon of the peculiar recovery of the skill scores after first dropping below zero (Figs. 11 and 13). The phenomenon is a result of the different rates at which the accuracy and the precision of the forecasts decrease. Accuracy can be visually assessed in these figures as the ability of the highest forecast probabilities to be in the region of occurrence. Precision can be seen as the gradient of forecast probability (i.e., tight gradient equates to high precision).

For the 1.5-day PQPF, there is fair accuracy and high precision, resulting in low mse and good skill. For 2.5-day PQPF, accuracy drops dramatically but precision does not decrease as rapidly, causing large mse and very
Fig. 14. Representative case study for CAT4 PQPF: (a) Observed isohyet of $pcp_{24} > 25.4$ mm (CAT4 threshold) for 1200 UTC 7 Nov 1996 to 1200 UTC 8 Nov 1996. Uncalibrated (democratic voting) and calibrated (weighted ranks) PQPF for the valid period produced from...
poor skill. This is reflected in the dip in the BSS or SS curve for CAT4. For the 3.5-day PQPF, accuracy changes little but precision continues to decrease, which lowers the mse and thus increases the skill. In the lower categories (thresholds), this rebound of the skill score is less pronounced and happens later in the valid period because areas of occurrence for lower thresholds are much larger.

e. Limits of predictability

Referring back to Fig. 13, the limits for skillfully predicting cumulative precipitation with the MRF ensemble were estimated as approximately 6.0 days for CAT1, 4.4 days for CAT2, 3.3 days for CAT3, and 2.1 days for CAT4. To estimate the general relationship between predictability and pcp24 threshold, a third-order polynomial function was fit to the four limits of predictability after a natural log transformation of the pcp24 threshold values (Fig. 15). The curve suggests that predictability falls off sharply at first with an increasing pcp24 threshold then decreases more gradually at higher levels of pcp24, threshold.

This result is consistent with the ideas presented by Lorenz (1969), who proposed that smaller scales of motion have shorter ranges of predictability. Low levels of pcp24 are generally associated with widespread precipitation events occurring on the synoptic scale. An event with a large pcp24 amount occurs on a much smaller scale, more likely due to convective activity than a synoptic-scale storm. Therefore, predictability should decrease with increasing pcp24 threshold (i.e., decreasing scale).

The limits of predictability displayed in Fig. 13 are specific to the data and design of this research. The major factors that influenced these results are 1) the model error of the MRF, 2) the nonoptimum ensemble perturbation scheme, and 3) the spatial and temporal
scales of the forecast verification. The first two factors were compensated for to some degree by the calibration. However, the calibration can only nudge the probability in the right direction, not correct significant deficiencies. The third factor, influence of scale on predictability, is well documented in Islam et al. (1993). Basically, the larger the scale (both spatial and temporal) over which predictions are made, the better the predictability, and vice versa. Since this research used such a large scale (2.5° grid and a 24-h period), it is likely that the resulting limits of predictability are optimistic. Cumulative precipitation predictability for smaller scales (i.e., finer grid spacing and shorter period) may be much less.

5. Conclusions and recommendations

Calibrated PQPF produced by the weighted ranks method does dramatically improve the quality of PQPF, as was clearly shown by the findings of the various measurement tools discussed in section 4. Unfortunately, the calibration can only increase skill for a relatively short period. In comparing the calibrated and uncalibrated PQPFs, it was discovered that the limit of predictability for significant amounts of cumulative precipitation is rather short.

Referring back to Fig. 15, it was found that for higher levels of cumulative precipitation, of more concern to meteorologists and their customers, PQPF based on the MRF ensemble lacks skill in the medium range. The spatial and temporal variability of high amounts of precipitation may be too great to be forecast with skill beyond the short range. However, the MRF-ensemble-based PQPF is of value in the medium range for very low thresholds of cumulative precipitation. In other words, in the medium range it may be possible to forecast the probability of precipitation versus dry conditions with skill but not possible to predict precipitation quantity. These conclusions may apply to all forecasts of cumulative precipitation, not just to forecasts based on the MRF ensemble.

The calibration designed by this research should be considered for implementation by NCEP to improve the skill of PQPF over the United States. This is of course provided that this calibration is more robust than the technique currently under development. A comparison of the skill found in this research to the skill of NCEP’s calibrated PQPF needs to be accomplished to answer this question. Whichever technique proves better, that calibration process should then be applied to other regions of the world where improved PQPF would be beneficial. Additionally, in using PQPF products, forecasters should be aware of the limits of predictability. For example, PQPF for CAT4 is available for out to 9 days. However, this research showed that even after calibration, such a forecast is only of value out to about 2 days.

The lack of dispersion in the MRF ensemble indicated by the probability surfaces needs to be further investigated. It is a sign that there may be a serious deficiency in the MRF ensemble that may be from many sources. 1) It could be that the perturbation method, breeding of growing modes, does not produce initial conditions close enough to the maximum growing mode to make ensemble members diverge fast enough; or 2) the MRF ensemble may not have enough members to consistently capture all the variability; or 3) the MRF model may not contain adequate resolution to represent the small-scale variability of precipitation; or 4) the MRF model may include significant biases. Whatever the case, there is definitely room for improvement in the MRF ensemble. Until this deficiency is largely corrected, a post-processing calibration like the one designed in this research must be applied.

Acknowledgments. This work represents a portion of the first author’s (FAE’s) M.S. research under the supervision of the second author (MKW) at the Air Force Institute of Technology. The authors would like to thank Tom Hamill at NCAR and Zoltan Toth at NCEP for their assistance in this research, Mike Baldwin and Russell Martin at NCEP for their inputs on the observed precipitation data, Andrew Loughe at CDC for the archived ensemble data, Dan Wilks for providing his code for MLE of gamma parameters, and two anonymous reviewers for comments and suggestions that improved the manuscript.

REFERENCES