Reliable Probabilistic Quantitative Precipitation Forecasts from a Short-Range Ensemble Forecasting System

DAVID J. STENSRUD
NOAA/National Severe Storms Laboratory, Norman, Oklahoma

NUSRAT YUSSOUF
Cooperative Institute of Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 5 October 2005, in final form 13 March 2006)

ABSTRACT

A simple binning technique is developed to produce reliable 3-h probabilistic quantitative precipitation forecasts (PQPFs) from the National Centers for Environmental Prediction (NCEP) multimodel short-range ensemble forecasting system obtained during the summer of 2004. The past 12 days’ worth of forecast 3-h accumulated precipitation amounts and observed 3-h accumulated precipitation amounts from the NCEP stage-II multisensor analyses are used to adjust today’s 3-h precipitation forecasts. These adjustments are done individually to each of ensemble members for the 95 days studied. Performance of the adjusted ensemble precipitation forecasts is compared with the raw (original) ensemble predictions. Results show that the simple binning technique provides significantly more skillful and reliable PQPFs of rainfall events than the raw forecast probabilities. This is true for the base 3-h accumulation period as well as for accumulation periods up to 48 h. Brier skill scores and the area under the relative operating characteristics curve also indicate that this technique yields skillful probabilistic forecasts. The performance of the adjusted forecasts also progressively improves with the increased accumulation period. In addition, the adjusted ensemble mean PQPFs are very similar to the raw ensemble mean PQPFs, suggesting that the method does not significantly alter the ensemble mean forecast. Therefore, this simple postprocessing scheme is very promising as a method to provide reliable PQPFs for rainfall events without degrading the ensemble mean forecast.

1. Introduction

Even though remarkable improvements have been made over the past decades in the deterministic predictions of temperature, humidity, winds, and other forecast variables using numerical weather prediction models, the improvements in quantitative precipitation forecasts (QPFs) are still relatively slow (Sanders 1986; Applequist et al. 2002). The considerable difficulties surrounding the production of accurate QPFs beyond a few hours, in tandem with the large societal impacts of precipitation, has led to the call for a focus on probabilistic QPFs (PQPFs) as the overarching strategy to provide user guidance (Fritsch and Carbone 2004). The reasons given for this focus on PQPFs is that risk guidance is facilitated by probabilistic guidance, skillful deterministic prediction of precipitation will be limited to a few hours at best for the foreseeable future, the parameterized convection used in present operational models limits the improvements to QPFs, and postprocessing techniques can likely be developed to generate reliable PQPFs (Fritsch and Carbone 2004). While improvements to model QPFs are strongly desired, this strategy implies that improvements to deterministic forecasts are needed principally to improve PQPFs. Precipitation forecasts after only a few hours should be viewed only from a probabilistic perspective because of our limited understanding of precipitation processes, difficulties in providing accurate small-scale initial conditions, and the chaotic nature of the atmosphere. It is important to further emphasize that more accurate probabilistic rainfall prediction may benefit various economic sectors, such as agriculture, aviation, electric

Corresponding author address: Dr. David J. Stensrud, NSSL/FRDD, National Weather Center, 120 David L. Boren Blvd., Norman, OK 73072.
E-mail: david.stensrud@noaa.gov

DOI: 10.1175/WAF968.1

© 2007 American Meteorological Society
power generation, and road transportation (Fritsch et al. 1998; Stensrud 2006).

It is well known that using the raw output from forecast ensembles leads to an overprediction of the likelihood of precipitation and PQPFs that are less reliable and less skillful as the precipitation thresholds increase (Mullen and Buizza 2001). However, it is clear that the raw ensemble data still provide useful information. Du et al. (1997) show that raw 80-km forecasts from a short-range ensemble forecasting (SREF) system provide skillful 6-h PQPFs relative to probabilistic forecasts from several operational centers that use different models, as Ebert (2001) further finds that the ECMWF EPS produces skillful predictions for a precipitation threshold of 1 mm day$^{-1}$ past one week. The accuracy of the EPS decreases as the precipitation threshold increases, such that forecasts of 50 mm day$^{-1}$ are not skillful even at day 1. It appears that raw ensemble PQPFs also benefit from ensembles that use different models, as Ebert (2001) finds that a poor man’s ensemble of seven independent model forecasts from several operational centers provides skillful forecasts of 50 mm day$^{-1}$ rainfall amounts for day 1, an improvement over the results from the much larger ECMWF EPS.

While the raw data from ensemble forecasting systems provide moderately skillful probabilistic information for lower precipitation thresholds, postprocessing these ensemble data is expected to produce significant improvements in the reliability and skill of the PQPFs and may also yield skillful probabilistic forecasts for higher precipitation thresholds (Fritsch and Carbone 2004). Hamill and Colucci (1997, 1998) and Eckel and Walters (1998) find that large improvements accrue when postprocessing is performed on ensemble forecasts of 24-h rainfall totals. In particular, using the ensemble forecast data in conjunction with the probability information contained implicitly within rank histograms produces more highly calibrated PQPFs and improved forecast skill for 24-h precipitation forecasts (Hamill and Colucci 1997; Eckel and Walters 1998). Eckel and Walters (1998) conclude that this approach dramatically improves PQPFs from the medium-range forecast model over the United States.

While these results illustrate the potential for reliable PQPFs to be produced from a variety of different ensemble systems, especially when postprocessing is used to reduce the tendency to overpredict the probability of precipitation for all thresholds (Mullen and Buizza 2001), these postprocessing approaches have been applied only to 24-h accumulation periods and on verification grids of 80 km or larger. Ensemble postprocessing methods to obtain reliable PQPFs for periods less than 24 h on verification grids more consistent with the forecast output grids available today in short-range numerical weather prediction have not been examined. One reason for the lack of such studies may be that less predictable scales are better sampled on a finer grid over shorter accumulation periods (Islam et al. 1993), suggesting that applying postprocessing techniques to these shorter time periods and over smaller regions will likely be less skillful than those reported for the longer 24-h accumulation periods. Regardless of these expectations, a postprocessing technique designed specifically for producing reliable 3-h PQPFs is developed. Results shown later indicate that this technique provides reliable PQPFs for accumulation periods as short as 3 h and that these 3-h values can be summed over multiple time periods to provide reliable PQPFs for longer accumulation periods as well.

The forecast data used in this study were collected as part of the New England High-Resolution Temperature Program (Stensrud et al. 2006), designed to improve predictions of near-surface variables over New England. A multimodel SREF system is chosen as the dataset used to test the postprocessing technique and is composed of the 15-member National Centers for Environmental Prediction (NCEP) SREF system (McQueen et al. 2005) and the NCEP 12-km operational Eta Model (Black 1994). Another unique aspect of this study is the use of the NCEP stage-II precipitation analyses for the precipitation observations used to both adjust and verify the ensemble forecasts.

The ensemble and the verification data are described in section 2. Section 3 contains a description of the postprocessing technique to adjust the precipitation forecasts. The forecast accuracy and skill of the new technique are examined in section 4, followed by a final discussion in section 5.

2. Data

a. Ensemble data

The data collection for this experiment started on 1 June and ended on 15 September 2004 for a total of 107 forecast days. The multimodel SREF system consists of 16 members provided by NCEP. Out of these 16 ensemble members, 15 members are from the 32-km operational SREF system and the other member is from the 12-km operational Eta Model. The 12-km Eta Model starts at 1200 UTC each day, whereas the 15-
member SREF ensemble forecasts start at 0900 UTC each day. Ten of the SREF members are from the Eta Model (Black 1994), with the remaining five members from the regional spectral model (RSM; Juang and Kanamitsu 1994). The Eta Model SREF system forecasts contain two runs from the control initial condition and eight runs using perturbations from the breeding of growing modes technique (Toth and Kalnay 1997). These runs use either the Betts–Miller–Janjic, or the Kain–Fritsch convective parameterization schemes. The RSM runs contain both initial condition (one run) and model convective scheme perturbations in which either the simple Arakawa–Schubert or the relaxed Arakawa–Schubert schemes (two runs per pair) are used. The RSM runs also include two perturbation pairs from the breeding of the growing mode technique.

Because of hardware problems, there may be one or more model forecasts missing on a given day, in which case the remaining models are used in the calculations. The accumulated precipitation forecasts for all these models are available on a common 40-km grid every 3 h starting from 1200 UTC and extending out to 48 h.

b. Precipitation data

The national stage-II precipitation analysis (Baldwin and Mitchell 1997) developed at NCEP is used as the observed precipitation dataset for this experiment. It is based on a multisensor precipitation algorithm developed in the Office of Hydrological Development (Seo 1998). The stage-II precipitation analysis is a blend of approximately 3000 automated, hourly rain gauge observations with hourly rainfall estimations from approximately 140 Weather Surveillance Radar-1988 Doppler radars over the contiguous United States. The data are available on a Hydrologic Rainfall Analysis Project map, which uses a polar stereographic map projection and has a spatial resolution of approximately 4 km × 4 km (Schaeke 1989). The stage-II analysis contains a high spatial coverage, but it does not have any manual quality control steps. The gauge data undergo a few initial quality control steps, however, that include a gross error check on the gauge data and subjective examination of any consistently bad rain gauges. The mean biases of radar estimates also are removed prior to the multisensor analysis (Smith and Krajewski 1991), although no attempts are made to remove range-dependent biases.

To compare the ensemble forecasts against the stage-II analyses, it is necessary to place the observed precipitation analyses onto the same grid and to use the same accumulation period (Hamill and Colucci 1997; Eckel and Walters 1998; Mullen and Buizza 2001). Therefore, the hourly precipitation data are summed to produce 3-h accumulated quantitative precipitation estimates and then averaged to the same 40-km grid as the ensemble members. The averaging is a simple areal mean (box average) of all the precipitation values within each of the 40-km model grid boxes (Baldwin 1997; Mullen and Buizza 2001). In general, there are around 6500 observed 40 km × 40 km grid points (Fig. 1) available from this analysis, with the total number of grid points on any given day varying because of radar data availability. The ensemble forecasts are evaluated only for grid points at which observations are available.

3. Simple binning technique for PQPFs

A simple binning technique is developed to postprocess precipitation forecasts from the ensemble system by using information from previous forecasts and observations. Stensrud and Yussouf (2005) show that using a 12-day window length to bias correct forecasts of near-surface variables such as 2-m temperature, 2-m dewpoint temperature, and 10-m wind speed works well, producing ensemble mean forecasts that are equal to or more accurate than those produced by the operational model output statistics packages. The bias-corrected forecasts also provide reliable probabilistic forecasts for these variables. Motivated by these results, a 12-day window is also chosen as the adjustment window length for this study. Out of the 107 forecast days, the first 12 days of this experiment are used only to provide the first adjusted forecasts, leaving a total of 95 case days for evaluation.

It is assumed that the location of the model grid point is unimportant and that only the precipitation amount and forecast time matters in adjusting the forecast precipitation amounts. This is equivalent to assuming that a forecast 0.5-cm rainfall event in the northwestern United States associated with upslope flow has the

![Fig. 1. Map of the United States indicating the locations of the stage-II analysis (dots) used as verification data in this study.](image-url)
same characteristics as a forecast 0.5-cm rainfall event over central United States associated with deep convection. While this likely is not true, Gallus and Segal (2004) show that model forecasts of precipitation amounts become more likely as the precipitation amounts increase, a fact well known to operational forecasters for many years. Thus, basing the adjustment procedure on the forecast precipitation amount and not the forecast precipitation location appears to be a reasonable initial assumption. This type of approach can help adjust systematic model errors, but not random errors.

Because only the forecast precipitation amounts are important, the postprocessing approach uses 22 preselected bins that represent a mutually exclusive and collectively exhaustive separation of the 3-h forecast precipitation totals. Only nonzero forecast precipitation amounts are examined, such that grid points with zero 3-h forecast precipitation amounts are ignored and remain unchanged after postprocessing. The bins are selected a priori such that the number of 3-h forecast precipitation amounts associated with each bin are roughly similar for all 22 bins when examined over the summer season. This requires the width of the precipitation bins to be narrower for the smaller and more common 3-h forecast precipitation amounts and wider for the larger and less common 3-h forecast precipitation amounts. The same bins are used for all models and for all forecast times. It may be that the application of this approach to different seasons necessitates that the bin widths be changed in order to maintain the rough consistency in the number of 3-h forecast precipitation amounts associated with each bin.

The postprocessing begins by identifying which of the 22 preselected bins the 3-h accumulated forecast nonzero precipitation amount falls into for each grid point from each ensemble member over the past 12 days (Table 1), with each forecast time evaluated separately. The bins are numbered from 1 to 22, with forecast precipitation amounts larger than 0 in. or less than or equal to 0.005 in. (0.0127 cm) associated with bin 1, forecast amounts larger than 0.005 in. or less than or equal to 0.01 in. (0.0254 cm) associated with bin 2, and so on as indicated in Table 1. Again, grid points with zero fore-

<table>
<thead>
<tr>
<th>Bin</th>
<th>Precipitation amounts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Greater than (in.)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
</tr>
<tr>
<td>6</td>
<td>0.04</td>
</tr>
<tr>
<td>7</td>
<td>0.05</td>
</tr>
<tr>
<td>8</td>
<td>0.06</td>
</tr>
<tr>
<td>9</td>
<td>0.07</td>
</tr>
<tr>
<td>10</td>
<td>0.08</td>
</tr>
<tr>
<td>11</td>
<td>0.09</td>
</tr>
<tr>
<td>12</td>
<td>0.10</td>
</tr>
<tr>
<td>13</td>
<td>0.20</td>
</tr>
<tr>
<td>14</td>
<td>0.30</td>
</tr>
<tr>
<td>15</td>
<td>0.40</td>
</tr>
<tr>
<td>16</td>
<td>0.50</td>
</tr>
<tr>
<td>17</td>
<td>0.60</td>
</tr>
<tr>
<td>18</td>
<td>0.70</td>
</tr>
<tr>
<td>19</td>
<td>0.80</td>
</tr>
<tr>
<td>20</td>
<td>0.90</td>
</tr>
<tr>
<td>21</td>
<td>1.0</td>
</tr>
<tr>
<td>22</td>
<td>2.0</td>
</tr>
</tbody>
</table>

**Fig. 2.** Schematic illustration of the simple binning technique used to adjust the ensemble precipitation forecasts. The model 3-h forecast precipitation totals are evaluated at each grid point (small black squares) in order to determine the bin into which the observed 3-h precipitation totals are collected. Both zero (white) and nonzero (black) precipitation values are stored in the bins. The past 12-days’ worth of model forecast and observed precipitation values are used to populate the bins, with different models and different forecast times evaluated separately. For today’s forecast, the model forecast precipitation amount is again used to determine the bin, and then a random selection from all the collected observed precipitation amounts is made and replaces the model forecast amount.
cast precipitation are ignored. Once the forecast precipitation bin is identified for a given grid point, the observed 3-h precipitation amount from the NCEP stage-II analyses associated with this grid point is saved in the bin (Fig. 2). Both zero and nonzero observed 3-h accumulated precipitation amounts are collected into the bins. Thus, the model 3-h forecast accumulated precipitation is used to select the bin, while the observed 3-h accumulated precipitation from the stage-II analysis is stored in the bin. All grid points within the model domain that overlap with the NCEP stage-II analyses are evaluated. Using this approach a total of 5984 distinct bins (22 bins \( \times \) 16 models \( \times \) 17 forecast times) are populated from roughly 624 000 individual observed 3-h accumulated precipitation amounts (6500 observations \( \times \) 8 observation times per day \( \times \) 12 days). The number of observed 3-h precipitation amounts collected within each of the 22 preselected bin varies depending upon the weather patterns during the past 12-day period. However, because the bins use observations from across the United States the likelihood that a given bin contains no observations is small.

The observed values of 3-h precipitation accumulations stored within each bin from the past 12 days are used to adjust today’s forecast. First, the bin for today’s 3-h forecast precipitation amount for each grid point, each ensemble member, and each forecast time is determined. Then a random selection of an observed 3-h precipitation amount from the pool of past observed values associated with this forecast bin, forecast model, and forecast time is made. This randomly selected observed amount replaces the present day’s model forecast precipitation amount at that grid point and forecast time. This is done for all the model grid points with nonzero precipitation totals. All the nonzero forecast precipitation amounts are replaced by a randomly selected observed precipitation amount. Forecast precipitation amounts of zero are not changed. The end result is an adjusted precipitation field that shows more small-scale structure than the raw forecast in the regions with smaller precipitation totals, yet largely maintains the forecast region of larger precipitation amounts (Fig. 3).

One of the outcomes of this postprocessing approach is that the adjusted precipitation field shows more small-scale structure than the raw forecast in the regions with smaller precipitation totals, yet largely maintains the forecast region of larger precipitation amounts (Fig. 3).

The observed values of 3-h precipitation accumulations stored within each bin from the past 12 days are used to adjust today’s forecast. First, the bin for today’s 3-h forecast precipitation amount for each grid point, each ensemble member, and each forecast time is determined. Then a random selection of an observed 3-h precipitation amount from the pool of past observed values associated with this forecast bin, forecast model, and forecast time is made. This randomly selected observed amount replaces the present day’s model forecast precipitation amount at that grid point and forecast time. This is done for all the model grid points with nonzero precipitation totals. All the nonzero forecast precipitation amounts are replaced by a randomly selected observed precipitation amount. Forecast precipitation amounts of zero are not changed. The end result is an adjusted precipitation field that shows more small-scale structure than the raw forecast in the regions with smaller precipitation totals, yet largely maintains the forecast region of larger precipitation amounts (Fig. 3).

One of the outcomes of this postprocessing approach is that the adjusted precipitation field shows more small-scale structure than the raw forecast in the regions with smaller precipitation totals, yet largely maintains the forecast region of larger precipitation amounts (Fig. 3).
is that it naturally helps to correct the individual model systematic biases. Bias is simply the number of forecast events divided by the number of observed events, so an unbiased forecast has a bias of 1 (Wilks 1995). Numerical models tend to overpredict the precipitation area, yielding a bias greater than 1 for smaller precipitation thresholds (e.g., Black 1994). The tendency of model forecasts to overpredict the precipitation area also is clearly seen in this study (Fig. 2). By randomly selecting from the observed precipitation amounts (including zeroes) associated with a given narrow precipitation range, the technique naturally ends up yielding fewer points with rainfall (cf. Figs. 2 and 3) and thereby lowers the bias toward 1. Numerical models also tend to underforecast convective rainfall amounts (sometimes by a factor of 2 or more) over the eastern two-thirds of the United States (Funk 1991). Again, by randomly selecting from the observed precipitation amounts associated with a given narrow forecast precipitation range, a systematic model tendency to underforecast the larger precipitation amounts largely can be corrected. When these adjustments are viewed from the probabilistic perspective of the entire ensemble, the adjusted frequencies at which forecast precipitation amounts are exceeded should match much closer with the observations.

**Fig. 6.** Attribute diagrams for the 3-h adjusted (solid line) and raw (dashed line) ensemble PQPFs equal to or exceeding (a) 0.005 in. (0.127 mm), (b) 0.01 in. (0.254 mm), (c) 0.025 in. (0.635 mm), (d) 0.05 in. (1.27 mm), and (e) 0.10 in. (2.54 mm). Error bars indicate 90% confidence interval for adjusted (black bars) and raw (gray bars) PQPFs. Inset histogram indicates the frequency of usage of each forecast probability category for the raw ensemble probabilities (white bar) and adjusted ensemble probabilities (black bar). All 95 forecast days and all grid points are used in this analysis.
Once the postprocessing is complete at all grid points for all ensemble members and forecast times, the adjusted forecasts are used to create PQPFs by assuming that each ensemble member is equally likely. Recall that no adjustments are done for precipitation forecasts of 0 and it also is assumed that 3-h accumulated precipitation forecasts greater than 7.62 cm (3 in.) should not be altered. Comparisons of ensemble PQPFs before and after postprocessing indicate that the technique typically lowers the probabilities for precipitation amounts exceeding the smaller threshold values (Fig. 4). Because the raw ensemble forecasts, as shown later, overpredict the probability of precipitation for all thresholds, these lower adjusted probabilities are greatly desired. The general shapes of the region of highest probabilities are the same both before and after adjustment, although there is more small-scale structure in the adjusted probabilities. Verification of the 3-h probability of precipitation greater than 0.05 in. (1.27 mm) from this same forecast time indicates that the raw ensemble has some resolution but no skill, whereas the adjusted ensemble has skill and resolution (Fig. 5). In addition, an ensemble mean precipitation forecast can be produced through simple averaging of all the ensemble members.

By using this method, adjustments may be made at...
each of the ∼6500 grid points, each of the 16 models, and each 3-h forecast interval for all the 95 forecast days. Forecasts resulting from the application of this simple binning adjustment technique are referred to as the adjusted ensemble. In addition to 3-h precipitation, it is of interest to evaluate the performance of this technique for longer accumulation periods. Therefore, to obtain 6-, 12-, 24-, and 48-h adjusted accumulated precipitation forecasts, the adjusted 3-h forecast precipitation amounts simply are summed over 6-, 12-, 24-, and 48-h periods at each grid point and ensemble member. These adjusted forecasts again are used to create PQPFs by assuming that each ensemble member is equally likely.

4. Results

To investigate the behavior of the adjusted ensemble to provide reliable PQPFs, attribute diagrams (Wilks 1995) for 3-, 6-, 12-, 24-, and 48-h precipitation totals with thresholds varying from 0.0127 to 5.08 cm (0.005–2.0 in.) are generated along with the estimation of 90% confidence bounds. Results shown are from all 95 forecast days and all grid points at the forecast times indicated. Statistical significance is determined using a bootstrap technique (Mullen and Buizza 2001), where resamples are randomly selected from the pool of 95 days for each forecast hour and probability statistics for each of those resamples are generated. This resampling
The procedure is repeated 10,000 times to estimate the 90% confidence bounds of the probability statistics. If the confidence intervals associated with the raw and adjusted ensembles do not overlap, then assuming a normal distribution the differences are significant at more than the 95% level. These diagrams (Figs. 6–10) show that the PQPFs from the adjusted ensemble system consistently outperform the PQPFs of the raw ensemble system for a variety of forecast times and precipitation thresholds. The differences between the raw and adjusted ensemble results are significant at the 95% confidence level when the confidence bounds do not overlap. The raw ensemble system has no skill for the higher precipitation amounts (≥0.10 in.) for 3-h accumulation period (Figs. 6), while the adjusted ensemble shows skill and generally reliable PQPFs for the same accumulation period. For longer accumulation periods, the raw ensemble often is skillful for the smaller precipitation thresholds, but tends to lose skill as threshold amounts increase (Figs. 7–10). In contrast, the adjusted ensemble is skillful even for higher precipitation thresholds (Figs. 7–10). These results highlight the benefits gained through the simple postprocessing technique. Note that the wide confidence intervals seen occasionally in both the raw and adjusted PQPFs for some of the higher probabilities are due to
having very few occurrences of these events in the dataset.

While the adjusted ensemble has a tendency to underestimate the probability of occurrences for the smaller precipitation totals (Figs. 6–10), the adjusted ensemble forecasts exhibit good reliability for the moderate and larger precipitation totals. The erratic behavior of some of the probabilities for the larger accumulations indicates the insufficiency of the sample size for these larger precipitation amounts. Results from the 24-h accumulation period (Fig. 9) show reliabilities similar to those reported by Hamill and Colucci (1997) and Eckel and Walters (1998), but with improved reliability for the higher precipitation amounts. The unique contribution of the simple binning adjustment technique is that it provides reliable forecasts even for periods as short as 3 h. Note that precipitation thresholds higher than 5.08 cm (2.0 in.) are not examined because of their very infrequent occurrence during the period of this experiment.

The Brier score (BS; Brier 1950), a mean square error of the probability forecasts, along with the 90% confidence interval, is calculated from the adjusted and raw forecast probabilities for 3-, 6-, 12-, 24-, and 48-h precipitation amount forecasts with threshold values ranging from 1 to 50 mm. In illustrating the results, a skill score is used (Wilks 1995) in which the climatological frequency of the event is determined from the full 95-day observational dataset for each threshold amount. Results from this Brier skill score (BSS) indicate that the adjusted ensemble generally is more skillful than the raw ensemble for accumulation periods of 24 h or less, with improvements often seen in the largest precipitation amounts for accumulation periods of 12 h or longer (Figs. 11c,d,e). In addition, for the 3- and 6-h accumulation periods the adjusted ensemble has significantly better BSS values than the raw ensemble for the smaller precipitation amounts (Figs. 11a,b). A decomposition of the BS (Wilks 1995) indicates that the adjusted ensemble has improved reliability, but also has smaller resolution, often yielding little change to the BS.

The skill of ensemble forecasts is expected to change with precipitation amount, lead time, and accumulation period. Results indicate that the adjusted ensemble skill decreases both with increasing precipitation amounts and with increasing forecast lead time (Fig. 12). Not surprisingly, these same tendencies also are seen in
other ensemble studies (Eckel and Walters 1998; Ebert 2001; Mullen and Buizza 2001). However, the adjusted ensemble skill increases with the length of the accumulation period (Fig. 13) out to the 48-h time period, the longest accumulation period that can be examined with these data.

Finally, the area under the relative operating characteristic (ROC) curve provides a useful measure of the ability of probabilistic forecasts to discriminate dichotomous events (Mason 1982). A value of 1.0 indicates a perfect probabilistic forecast system, an area of greater than 0.8 is considered good, an area of 0.7 is the lower limit of a useful prediction system, and an area of 0.5 or less indicates a useless forecast (Buizza et al. 1999). The area under the ROC curve is larger in general for the 48-h accumulations indicating the greater skill of the adjusted ensemble system for longer accumulation periods (Fig. 14). As the precipitation amount increases the ensemble forecast skill decreases. The area under the ROC curve for the 48-h accumulation period is 0.85 for 1 mm forecasts and stays above 0.7 for forecasts up to 25 mm. The ROC area of the 24-h forecast is very close to the value from the 48-h forecasts for lower precipitation thresholds and is smaller than the values from the 48-h forecasts for higher precipitation thresholds. The ROC areas calculated from the 24-h adjusted ensemble accumulations valid at 24 and 48 h are similar to the results from the ECMWF EPS as shown by Mullen and Buizza (2001).

To further explore the ability of the multimodel ensemble system to provide probability information, rank histograms (Hamill 2001) are generated for the adjusted and raw ensemble systems. Only forecast days with all 16 ensemble members are considered in the rank histogram. The verification plus the ensemble values are sorted from lowest to highest and the rank of the verification is determined from the sorted joint distribution. If ties occur among the verification and en-

![Fig. 11. BSS as a function of rain threshold (mm) for (a) 3-, (b) 6-, (c) 12-, (d) 24-, and (e) 48-h adjusted (black solid line) and raw (gray dashed line) PQPFs. All 95 forecast days and all grid points are used in this analysis.](image)
semble values, random deviates are added before determining the rank (Hamill 2001). This is done for the 3-, 6-, 12-, 24-, and 48-h precipitation totals at all forecast times and locations during the 95-day period. Results are similar for both the adjusted and raw ensemble system (Fig. 15). Some of the local peaks in the rank histograms may be due to the use of multiple models in the ensemble with distinctly different precipitation distributions. The verification falls frequently at the higher end of the distribution, indicating the tendency of the ensemble system to underpredict precipitation amounts. This tendency is also seen in the underprediction of the PQPFs for lighter precipitation accumulations shown earlier (Figs. 6–10), which are the most common and dominate the rank histogram calculations.

While this study is focused on the ability to produce reliable PQPFs, the deterministic ensemble mean precipitation forecast may also be of interest to some end users of the ensemble data. To explore whether the simple binning technique used to adjust the ensemble forecasts produces any significant changes to the ensemble mean precipitation forecast, the mean error (bias), mean absolute error (MAE), and the root-mean-square error (rmse; Wilks 1995) at each forecast hour are calculated for both the adjusted and raw ensemble mean QPFs. Results indicate that the MAE and bias for the 3-h adjusted mean QPFs are smaller than the raw values at all forecast times whereas the rmse is larger at most of the times (Fig. 16). This reflects slightly higher error variances in the adjusted mean QPFs compared with the raw QPFs. Results also indicate that the dif-
ferences in bias and MAE are often significant at the 95% level, whereas the differences in rmse are significant at this level only for several of the forecast times. Results are more mixed for the 12-h accumulation times, with the adjusted ensemble sometimes less accurate and sometimes more accurate than the raw ensemble (Fig. 16). These results suggest that the simple binning technique is not producing significantly larger errors in the ensemble mean precipitation forecasts, and may even be improving these ensemble mean forecasts for shorter accumulation periods.

5. Discussion

A simple postprocessing technique is developed to adjust short-range ensemble forecasts to yield reliable PQPFs for forecast periods as short as 3 h. This approach is evaluated using short-range ensemble forecast data and NCEP stage-II precipitation analyses during the summer of 2004. The ensemble system comprises the NCEP 15-member 32-km SREF system and the 12-km operational Eta Model, with all model and observation data available on a common 40-km grid. The PQPFs for 3-, 6-, 12-, 24-, and 48-h accumulation periods that exceed various selected threshold values reveal that the adjusted ensemble probabilities are significantly more skillful and reliable than the raw ensemble probabilities. The BSSs are also improved when using the adjusted ensemble for the lower precipitation thresholds over 3- and 6-h accumulation periods, and for the higher precipitation thresholds over 12- and 24-h accumulation periods, in comparison with the raw ensemble. Results further indicate that the accuracy of the ensemble mean forecast from the adjusted ensemble is comparable to that from the raw ensemble system. This indicates that the simple binning technique is not altering the accuracy of the ensemble mean as it provides a much more skillful PQPF. For a given forecast, the simple binning adjustment technique produces more small-scale structures in the forecast ensemble mean precipitation fields than found in the original ensemble mean.

Fig. 15. Rank histogram of adjusted (black bars) and raw (gray bars) ensemble precipitation forecasts for (a) 3, (b) 6, (c) 12, (d) 24, and (e) 48 h at all forecast times over all the observing locations for the period of study.
Because the ensemble forecasts are only available out to 48 h, the performance of the adjusted ensemble system beyond this 48-h period cannot be evaluated. In addition, this study also evaluates the ensemble forecasts for a single summer season and so the skill of the adjustment technique during winter is unknown. Many researchers now believe that probability forecasts are more useful than the traditional deterministic forecasts (Fritsch et al. 1998). Indeed, Murphy and Winkler (1979) argue that forecasts must include probabilistic information. Thus, this simple postprocessing scheme can be used to provide reliable and skillful PQPFs from any ensemble forecasting system.

Acknowledgments. The authors are thankful to Jun Du and Jeff McQueen for providing the output from the forecast models used in this experiment. Special thanks to Mike Baldwin for providing assistance in obtaining verification data and codes to convert the data, and to Harold Brooks for helpful discussions. Two anonymous reviewers are thanked for their helpful and constructive comments that improved this presentation. Provision made by NCEP in order for us to obtain the stage-II data via FTP is gratefully acknowledged. Local computer assistance provided by Doug Kennedy, Steven Fletcher, Brett Morrow, and Karen Cooper is greatly appreciated. Partial funding for this research was provided under NOAA-OU Cooperative Agreement NA17RJ1227.

REFERENCES


