# 5.4 Postprocessing multimodel ensemble data for improved short-range forecasting

David J. Stensrud[1] and Nusrat Yussouf[2]
[1]NOAA/National Severe Storms Laboratory, Norman, Oklahoma 73069
[2]Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma 73069

## 1. INTRODUCTION*

A short-range multimodel ensemble forecasting system was developed and run during the summers of 2002-2004 in support of a National Oceanic and Atmospheric Administration program to improve near surface forecasts over the New England region. One of the goals of this project is to assess the potential for a short-range ensemble forecasting system to provide improved near surface predictions when compared against the statistical postprocessing routinely available from present operational models. In the United States, a multiple linear regression approach, called model output statistics (MOS; Glahn and Lowry 1972; Jacks et al. 1990) has been in use for three decades, and presently is used to postprocess forecasts of the nested grid model (NGM), Eta Model (ETA), and the aviation (AVN) run of the Global Spectral Model for hundreds of individual station locations. The MOS forecasts of near surface variables, such as 2-m temperature and dewpoint temperature, 10-m winds, and rainfall are more accurate than the raw model forecasts (Jacks et al. 1990) and are considered the standard by which to compare other techniques for predicting these variables in the United States.

One of the difficulties in implementing a MOS approach is that the data archive required to develop the regression equations must cover several seasons during which the numerical model remains unchanged. When models are updated frequently, as is typically the case at present operational centers, the requirement of a lengthy data archive makes implementing a MOS approach difficult and has led to approaches that neglect the model changes (Hart et al. 2004), the development of techniques that account for model changes (Ross 1989; Wilson and Vallée 2002), and the exploration of alternative approaches that do not require a long data archive (Homleid 1995; Mao et al. 1999). One approach that has shown success recently is a bias-corrected multimodel ensemble system in which a simple running mean bias correction is applied individually to each ensemble member (Stensrud and Yussouf 2003, 2005; Eckel and Mass 2005). This bias correction approach is easy to implement, assuming that output from a short-range ensemble system is available, and delivers postprocessed forecasts within a few weeks of first receiving the operational ensemble forecasts.

To illustrate the value of bias-corrected forecasts of 2-m temperature and dewpoint temperature and 10-m winds, bias-corrected forecasts are produced from the available ensemble members during the summer of 2003. These data also allow us to explore the value of mutimodel ensembles. Ensemble data from 2004 are used to investigate a new method for developing reliable probabilistic quantitative precipitation forecasts (PQPFs) from a short-range ensemble system. Results of the adjusted forecasts for both near surface variables and rainfall totals illustrate the great value that can be added to ensemble forecasts by simple post-processing techniques.

## 2. DATA

The data used in this study cover much of the summer seasons of 2003 and 2004. The data from 2003 are used to develop a postprocessing approach for near surface variables, such as 2-m temperature and dewpoint temperature and 10-m winds. The data from 2004 are used to develop a postprocessing approach for accumulated rainfall for periods as short as 3 h and as long as 48 h. The reason for the two different summers is that we developed the near surface postprocessing technique during the fall of 2003 and only after the successful completion of this project did we start work on the postprocessing technique for rainfall during the fall of 2004. Thus, we simply used the most

* Corresponding author: Dr. David Stensrud, NOAA/NSSL, 1313 Halley Circle, Norman, OK 73069. David.Stensrud@noaa.gov

recent ensemble forecast data to develop the respective techniques.

***a. 2003 ensemble system***

The models used in the 2003 ensemble system are the National Centers for Environmental Prediction (NCEP) Eta model (Black 1994), the regional spectral model (RSM: Juang and Kanamitsu 1994), the Rapid Update Cycle model (RUC: Benjamin et al. 1994, 2001), and the Weather Research and Forecast model (WRF: Klemp 2004). Up to 31 different forecasts are available at 3 h intervals starting from 1200 UTC and out to 48 h, depending upon the availability of the model forecasts in real time. The inclusion of the RUC and WRF model forecasts allows us to more clearly explore the importance of model diversity to forecast skill. The forecast data are available from 23 July through 15 September 2003, for a total of 55 forecast cases.

Sixteen of the model forecasts are from the Eta Model, with 15 forecasts from a 32-km version used in an experimental short-range ensemble system and one forecast from the 12-km operational version. The 15 ensemble forecasts are started at 0600 UTC and use both the breeding of growing modes technique (Toth and Kalnay 1993, 1997) and perturbations to the model convective parameterization and microphysics schemes (Du et al. 2004). While the details are found in Du et al. (2004), the 15 ensemble members use either the control initial condition (3 runs) or perturbations from two breeding of growing mode pairs (6 runs per pair), and use either the Betts-Miller-Janjic, relaxed Arakawa Schubert, or Kain-Fritsch convective parameterization schemes with a version of the Ferrier microphysics (Ferrier 2004). The 12-km operational Eta Model forecast is started at 1200 UTC. In addition, seven forecasts are from the 32-km RSM that starts at 0600 UTC and again contain both initial condition (bred modes) and model convective scheme perturbations in which either the simple Arakawa-Schubert or the relaxed Arakawa-Schubert schemes are used (Du et al. 2004). These 23 ensemble members are later grouped together as a subset of the full ensemble and designated the NCEP ensemble.

The full 31-member ensemble contains another 8 model forecasts from model versions that are not operational. Four additional forecasts are from two 22-km versions of the Eta Model that use either the Betts-Miller-Janjic (Eta) or the Kain-Fritsch (EtaKF) convective parameterization schemes. These two versions of the 22-km Eta Model are started from both the 0000 UTC Eta Model and Global Forecast System (GFS) initial conditions, and use a smaller horizontal domain than the operational Eta Model (Kain et al. 2001). Two forecasts are from the RUC started at 1200 UTC, one using a 10-km grid over just the northeastern United States and the other using a 20-km grid over the contiguous 48 states. The 20-km RUC forecasts use an initial condition created by an optimal interpolation scheme (Benjamin 1989), while boundary conditions are provided by the Eta Model forecasts. The 10-km RUC uses the 20-km RUC data for initial and boundary conditions. The final two forecasts are from the WRF model started at 1200 UTC, with one version again using a 10 km grid over just the northeastern United States and the other using a 20 km grid over the contiguous 48 states. The initial and boundary condition data are the same as for the RUC forecasts. All the models, except for the 10 km RUC and WRF forecasts, have domains that cover the contiguous 48 states.

***b. 2004 ensemble system***

The models used in the short-range ensemble for 2004 are only the NCEP Eta Model and the NCEP RSM. This ensemble this experiment consists of 16 members, with 15 members are from the operational SREF system (McQueen et al. 2005) and the other member is 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 32-km Eta Model (Black 1994), with the remaining 5 members from the 40-km regional spectral model (RSM: Juang and Kamanitsu 1994). The Eta Model SREF system forecasts contain two runs from the control initial condition and eight runs using perturbations from two breeding of growing mode pairs (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 (2 runs per pair) are used. The RSM runs also include two perturbation pairs from the breeding of growing mode technique. The data collection for this summer started on 1 June and ended on 15 September 2004 for a total of 107 forecast days.

### *c. Temperature and wind observations*

To compare the model results against surface observations of temperature, dewpoint temperature, and wind speed, the model data are bilinearly interpolated to the NWS observing station locations. A total of 1892 surface stations are available across the United States, Canada, and Mexico (Fig. 1) to use in determining the bias correction and for verification. Similarity theory (Stull 1988) is used to interpolate from the lowest model level to a 2-m height for temperature and dewpoint temperature and to a 10-m height for wind in a manner that is consistent with the model planetary boundary layer scheme. Owing to computer problems, one or more model forecasts may be unavailable on a given day, in which case the ensemble is created from the remaining members.

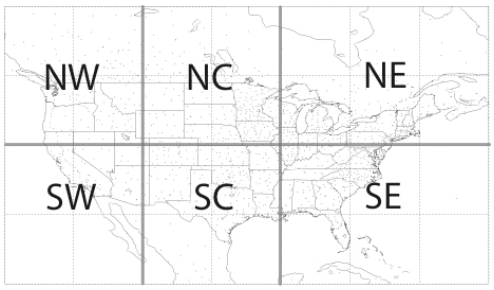


*Fig. 1. Map of the United States indicating the locations of the stations used in the bias correction approach (dots), and the six regions into which the model data are divided to evaluate the spread-skill relationship. These regions are northwest (NW), northcentral (NC), northeast (NE), southwest (SW), southcentral (SC), and southeast (SE).*

The surface observations used in determining the bias correction and in the verification of the forecasts are not quality controlled in any manner beyond that done by the National Weather Service. However, observational errors should be more detrimental to the verification of the bias correction approach than to MOS, since the observational error influences not only the forecast verification on the day of the error, but also influences the magnitude of the bias correction applied to the forecasts over subsequent days. Thus, improved quality control of the observations should act to further improve the bias correction approach in comparison with MOS.

### *b. Precipitation data*

The national Stage II precipitation analysis (Baldwin and Mitchell 1997) developed at NCEP is used as the observed precipitation data set for this experiment. It is based on a multi-sensor 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 WSR-88D radars over contiguous United States (CONUS). The data are available on a Hydrologic Rainfall Analysis Project (HRAP) map, which uses a polar stereographic map projection and has a spatial resolution of approximately 4 km × 4 km (Schaake 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 raingages. The mean biases of radar estimates also are removed prior to the multi-sensor analysis (Smith and Krajewski 1991), although no attempts are made to remove range-dependent biases.

In order to produce a valid comparison of the ensemble forecasts against the Stage II analyses, it is necessary to place the observed precipitation analyses onto the same grid and using the same accumulation period. Therefore, the hourly precipitation data are summed to produce 3-h accumulated quantitative precipitation estimates (QPEs) 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. In general, there are around 6500 observed grid points (Fig. 2) available from this analysis, with the total number of points on any given day varying due to radar data availability. The ensemble forecasts are evaluated only for points at which observations are available.

## 3. METHODOLOGY

### *a. Temperatures and winds*

The bias correction method for temperature, dewpoint temperature, and winds uses the past complete 12 days of data to calculate the bias of each ensemble member at each observation station and each forecast hour from 3 h to 48 h at 3-h intervals. A 12-day window is chosen after quantitative evaluation of the data with window

lengths varying from 2 to 25 days indicating that 12-day is a reasonable choice of bias correction window length (Stensrud and Yussouf 2005). The calculated 12-day bias values are then applied to today's forecast at each station and at each forecast hour. The ensemble mean then represents the average value of all the bias corrected forecasts. Since there are over 1800 station locations used, 17 forecast times, and 31 models, a total of nearly 1 million bias corrections are determined for each 48-h forecast cycle.

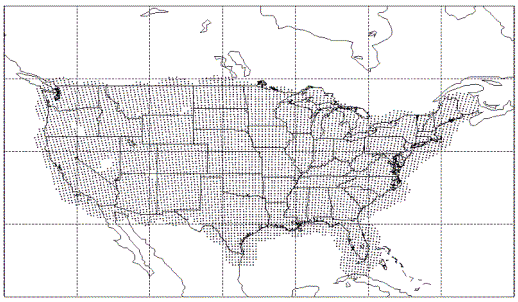

*Fig. 2. Map of United States indicating the locations of the stage II analysis (dots) used as verification data in this study.*

***b. Rainfall***

For this adjustment technique, 3-h forecast precipitation amounts from each ensemble member and forecast time over the past 12 days are separated into 22 preselected[1] bins and the observed 3-h precipitation values associated with each of these bins are saved. The location of the model grid point is unimportant - only the precipitation amount and forecast time matters. The number of observed 3-h precipitation amounts associated with each preselected bin varies depending upon the weather patterns during the past 12-day period. These stored values of observed precipitation amounts within each bin are then used to adjust today's forecasts. First, the bin for today's forecast 3-h precipitation amount is determined. Then a random selection of an observed precipitation amount from the pool of observed values associated with this forecast bin is made. This randomly selected observed amount replaces today's model forecast amount at that time and grid point. This is done for all the model grid points and individually for each ensemble member. Finally, all the adjusted forecasts are averaged to obtain the ensemble mean forecast, or used to obtain forecast probabilities. No adjustments are done for precipitation forecasts of 0 or of greater than 7.5 cm (3 in) of accumulated precipitation.

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.

## 4. RESULTS FOR TEMPERATURE AND WINDS

Results indicate that the mean bias-corrected ensemble (BCE) forecasts of 2-m temperature and dewpoint temperature generally have smaller MAE and rmse than either the ETA or AVN MOS (Figs. 3, 4). This is true for all forecast times for dewpoint temperature, whereas the MOS temperature forecasts can have smaller MAE values than the mean BCE during the nighttime hours (18-24 h, and 42-48 h). Results for wind speed are not quite as good, with the mean BCE wind speed forecasts generally as accurate or better than MOS forecasts during the daytime, but can be less accurate at night (not shown). Generally, if the differences in the MAE or rmse values of two post processing systems at a given forecast time are greater than 0.1 K or 0.1 m $s^{-1}$, then a Wilcoxon signed rank test (Wilks 1995) using the daily averages of the error measures as paired samples indicates that the differences are significant at the 95% level. Thus, the mean BCE forecasts of 2-m temperature and dewpoint temperature are more accurate than all of the MOS forecasts during the daytime hours at a 95% significance level. The

[1] The bins are $0<p\le0.0125$ cm, $0.0125<p\le0.025$ cm, $0.025<p\le0.050$ cm, $0.050<p\le0.075$ cm, $0.075<p\le0.10$ cm, $0.10<p\le0.125$ cm, $0.125<p\le0.150$ cm, $0.150<p\le0.175$ cm, $0.175<p\le0.20$ cm, $0.20<p\le0.225$ cm, $0.225<p\le0.25$ cm, $0.25<p\le0.50$ cm, $0.50<p\le0.75$ cm, $0.75<p\le1.0$ cm, $1.0<p\le1.25$ cm, $1.25<p\le1.50$ cm, $1.50<p\le1.75$ cm, $1.75<p\le2.00$ cm, $2.00<p\le2.25$ cm, $2.25<p\le2.50$ cm, $2.50<p\le5.00$ cm and $5.00<p\le7.50$ cm (0.0125, 0.025, 0.050, 0.075, 0.10, 0.125, 0.150, 0.175, 0.20, 0.225, 0.25, 0.50, 0.75, 1.0, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 5.00 and 7.50 cm equal 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.0, 2.0 and 3.0 in, respectively)

day-two daytime wind speed forecasts from the mean BCE also are more accurate than both the MOS forecasts.

The mean BCE values for the NCEP ensemble, which is comprised of only the 23 model forecasts from versions of the operational ensemble models, are not as accurate as the full BCE (Figs. 3, 4). Indeed, while the mean NCEP ensemble forecasts are often better than the MOS forecasts during some of the daytime hours, and are comparable in magnitude to the MOS forecasts during the nighttime, the improvement is less than that found with the full BCE. Thus, the additional 8 model forecasts from the other 3 numerical models (RUC, WRF, and 22-km Eta) are providing additional information that is helpful to the mean ensemble forecasts.

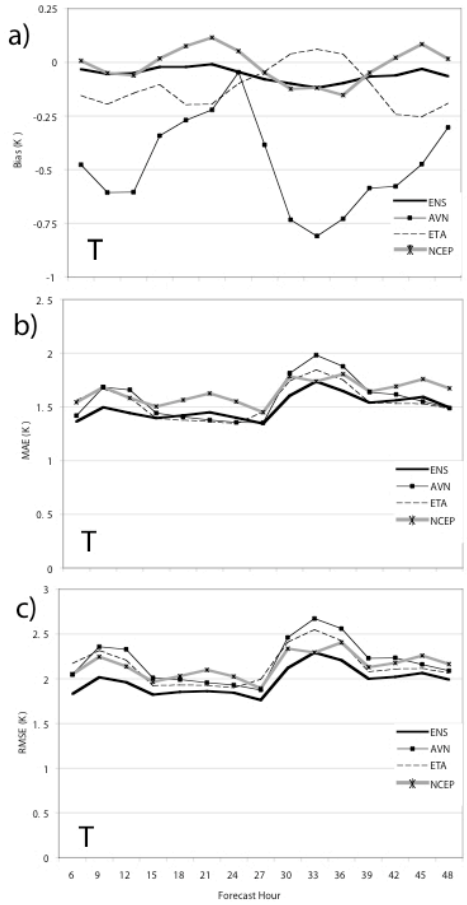


*Fig. 3. Values of (a) mean bias (K), (b) mean absolute error (K), and (c) root-mean-square error (K) plotted as a function of forecast hour for 2-m temperature from the full 31 member BCE (ENS), the NCEP-only BCE (NCEP), and the AVN and ETA MOS. Results are calculated at 1258 station locations. Further details are found in the legend.*

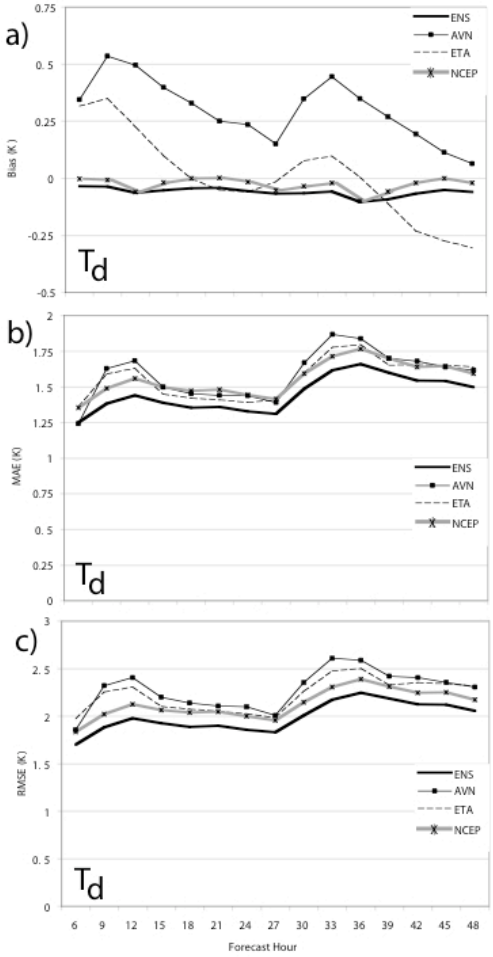


*Fig. 4. As in Fig. 3, but for 2-m dewpoint temperature forecasts.*

One of the main reasons to use an ensemble approach to forecasting the weather is to provide explicit guidance on the probabilities of various weather events. Murphy and Winkler (1979) suggest that forecasts need to be expressed in a probabilistic format in order to be used to their best advantage, while Richardson (2000) illustrates the potential economic value of even imperfect probabilistic forecasts. The value of raw forecast probabilities for 2-m temperatures and dewpoint temperatures are examined using the reliability, or conditional bias, of the temperatures exceeding selected threshold values. The probability is determined simply by calculating the number of forecast members that exceed (or are less than) the selected threshold, dividing this number by the total number of forecasts in the ensemble, and multiplying by 100. Results indicate that the BCE underestimates the frequency of occurrence of 2-m temperatures greater than or equal to 303 K

for probabilities less than 60% (Fig. 5). This underestimation is seen for both cooler and warmer threshold temperatures as well (not shown). While the raw forecast probabilities for 2-m dewpoint temperature greater than or equal to 285 K also indicate an underestimation of the frequency of occurrence for probabilities less than 50%, this underestimation is smaller than that for temperature (see Stensrud and Yussouf 2005). However, as the 2-m dewpoint temperature threshold value is decreased, the underestimation of the frequency of occurrence for the lower probabilities increases (not shown). Similar results are seen in the probabilities of 10-m wind speed equal to or exceeding 6 m $s^{-1}$ (not shown). The BCE underpredicts probabilities less than 25%, but then overpredicts probabilities greater than 25%. Thus, some calibration of these probability forecasts is needed.

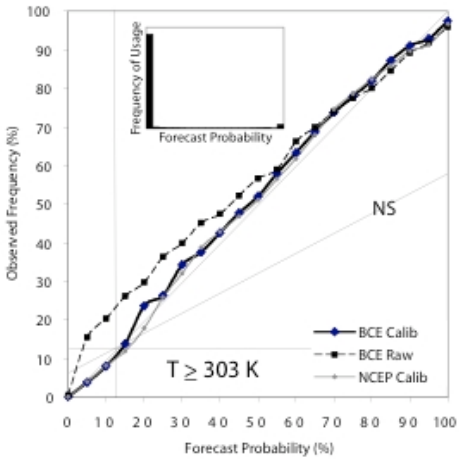


*Fig. 5. Attribute diagram for the BCE forecasts of 2-m temperature equal to or exceeding 303 K. The inset legend defines the various lines, while the inset histogram indicates the frequency of usage of each 5% interval forecast probability category for the uncalibrated (raw) ensemble. Horizontal line indicates the frequency of the event in the observed dataset, and the diagonal line is the no skill (NS) line. Lines above the diagonal indicate that the ensemble is underpredicting the probabilities, while lines below the diagonal indicate that the ensemble is overpredicting the probabilities.*

Hamill and Colucci (1998) suggest that the information from a rank histogram can be used to calibrate ensemble probability forecasts. Unlike the raw probability forecasts, in which each forecast member is assumed to have an equal probability of occurrence, we use the verification rank histogram calculated from all stations at each forecast time over the past 12 complete forecast days to calibrate the ensemble probabilities for each station location at each forecast time. This approach is described in Stensrud and Yussouf (2005). It uses the verification rank histogram to provide the past probabilities for each rank and then assumes that these probabilities are linearly distributed between the ensemble member forecasts. Maximum and minimum Gumbel distributions are used to calculate the probabilities of events that are either above the last rank or below the first rank of the BCE forecasts. For example, with a typical "U" shaped rank histogram, the first and last rank are more likely to be observed and this information is used in the calculation of the probabilities.

Owing to various computer problems, the number of ensemble members available on each day is not constant, which leads to rank histograms of different sizes during the 12 day calculation window. Having different sizes of the rank histograms makes using these data to adjust the probabilities very difficult. Thus, to overcome this problem, we assume a constant 20 ranks regardless of the number of ensemble members available. The location of the observation value within the joint, ranked distribution of model forecast values plus the observation value is determined and then scaled to a rank from 1 to 20 [see Stensrud and Yussouf (2005) for details]. This is done separately for each forecast variable examined.

Results from the calibrated probability forecasts show that the ensemble results are quite reliable, leading to consistently smaller Brier scores - a mean-square-error of the probability forecasts (Wilks 1995) – than the raw probability forecasts (Fig. 5). The Brier score for the BCE calibrated 2-m temperature forecasts, using a threshold temperature of 303 K, is 0.026, increasing very slightly to 0.027 for the raw BCE and to 0.028 for the NCEP bias-corrected ensemble. These Brier scores are slightly lower than those found from the 2002 pilot program (Stensrud and Yussouf 2003). As discussed in Stensrud and Yussouf (2005), the calibrated BCE for 2-m dewpoint temperature also shows that the calibration successfully improves the probability forecasts.

## 5. RESULTS FOR RAINFALL

The mean error (bias), mean absolute error (MAE), and the root-mean-square error (rmse)

(Stanski et al. 1989; Wilks 1995) at each forecast hour are calculated for both the adjusted and raw (original) ensemble mean QPFs. Statistical significance of the error values is determined using the bootstrap technique (Mullen and Buizza 2001), where resamples are randomly selected from the pool of 95 days for each forecast hour and error statistics for each of those resamples are generated. This resampling procedure is repeated 10,000 times to estimate the 90% confidence bounds of the error 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. 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 (not shown). This reflects higher error variances in the adjusted mean QPFs. Results also indicate that the differences in bias and MAE are often significant at the 95% level while the differences in rmse are significant at this level only for several of the forecast times. These results suggest that our approach is not producing significantly larger errors in the ensemble mean precipitation forecasts, even though we are randomly selecting observed precipitation amounts from the past 12 days and associated with grid points across the model domain and representing vastly different conditions.

To investigate the behavior of the ensemble system, attribute diagrams (Stanski et al. 1989; Wilks 1995) for the 3-, 6-, 12-, and 24-h precipitation totals with thresholds varying from 0.0125 to 5.0 cm (0.005 to 2.0 in) are generated along with the estimation of 90% confidence bounds. These diagrams (Fig. 6) show that the PQPFs from the adjusted ensemble system consistently outperform the PQPFs of the raw ensemble system, and these differences 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 for 3- and 6-h accumulation periods, while the adjusted ensemble shows good skill and generally very reliable PQPFs for lower forecast probability values. For longer accumulation periods, the raw ensemble typically is skillful for the smaller precipitation thresholds, but continues to have little skill for the higher amounts. In contrast, the adjusted ensemble is skillful even for the

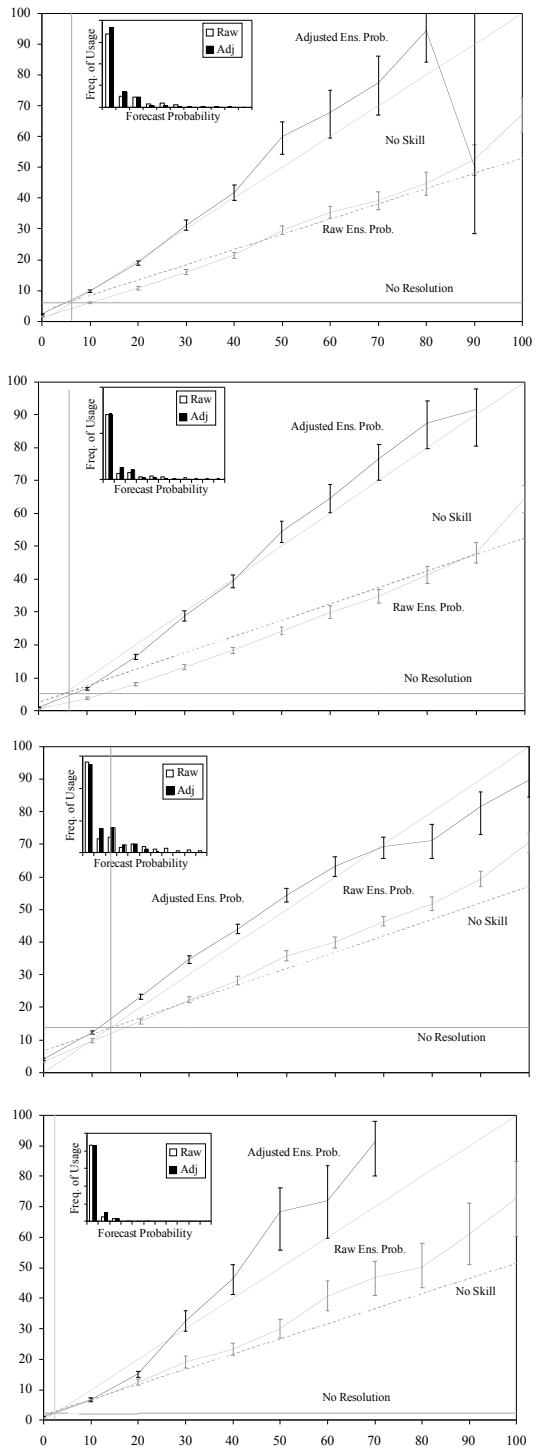


*Fig. 6. Attribute diagrams for the adjusted (solid black line) and raw (solid gray line) ensemble probabilities for precipitation equal to or exceeding 3-h accumulations of 0.04 inches valid at 18 h (top), 6-h accumulations of 0.10 in valid at 24 h (second from top), 12-h accumulations of 0.10 in at 36 h (third from top), and 24-h accumulations of 0.5 in at 24 h (bottom). Error*

*bars indicate the 90% confidence intervals for the adjusted (black) and raw (gray) ensembles.*

higher precipitation threshold values for lower forecast probability values. These results highlight the benefits gained through the simple post-processing technique. Note that the several large confidence intervals seen 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 data set.

## 6. DISCUSSION

A multimodel, multiphysics ensemble system consisting of up to 4 different models, with variations of physical parameterizations also specified within each model, and with a variety of different initial and boundary conditions, is used to assess the potential for a short-range ensemble forecasting system to provide improved near surface predictions when compared against the statistical postprocessing routinely available from present operational models and to provide reliable PQPFs. The ensemble forecasts are evaluated using routine NWS surface observations from over 1200 stations in the United States and NCEP Stage II analyses. Results from the short-range ensemble systems during the summers of 2003 and 2004 indicate that using the past complete 12 days of forecasts to adjust today's forecast yields ensemble mean forecasts that are better than the ETA, and AVN MOS during most of the forecast times for 2-m temperature, are better than the MOS forecasts at all forecast times for 2-m dewpoint temperature, and are comparable to the MOS forecasts for 10-m wind speed. This 12 day adjustment period also is sufficient to provide reliable PQPFs for forecasts from 3 h to 48 h in duration. Although results of this bias-correction approach during the cooler seasons have not been examined, although Woodcock and Engel (2005) show very good results for near surface variables using a similar technique over a 6-month period from summer into winter.

The probabilities produced by the adjusted ensembles are skillful and reliable, and previously have been found to be valuable when evaluated in a cost-loss model (Stensrud and Yussouf 2003). The ensembles further appear to provide better guidance for more unlikely events, such as very warm temperatures (see Stensrud and Yussouf 2005), that likely have the greatest economic significance. Thus, industries that are sensitive to the weather, such as power companies, transportation, and agriculture, may benefit from the probability information provided.

The results presented here indicate that it is possible to develop a robust post-processing system for new models, when used in conjunction with a reasonable short-range ensemble forecasting system, that is competitive with or better than traditional post-processing techniques, such as MOS, that take lengthy data archives to develop. This approach allows for the rapid production of useful and accurate guidance forecasts of many near surface variables and rainfall accumulations once an ensemble system is started operationally. And these ensemble-based approaches can be merged with the MOS forecasts to incorporate the strengths of each approach (Woodcock and Engel 2005).

In an era when model changes and updates are frequent, the use of relatively simple and computationally rapid postprocessing techniques with ensemble forecast model output makes good sense and needs to be pursued vigorously by operational agencies. These approaches make the best use of the available model forecast data and maximize the benefits of model forecasts to the public.

*Acknowledgments.* The authors are thankful to Jun Du, Jeff McQueen, Michael Baldwin, Jack Kain, Stan Benjamin, and Tracy Lorraine Smith for providing us with the output from the forecast models used in this experiment. We also are thankful to Paul Dallavalle for providing software to interpolate model forecast data to station locations, and to the reviewers for providing us with very helpful and constructive comments that led to improvements in this manuscript. We further appreciate the local computer support provided by Doug Kennedy, Steven Fletcher, and Brett Morrow. Discussions with Harold Brooks were very helpful. Partial funding for this research was provided under NOAA-OU Cooperative Agreement #NA17RJ1227.

## REFERENCES

Baldwin, M. E., and K.E. Mitchell, 1997: The NCEP hourly multi-sensor U.S. precipitation analysis for operations and GCIP research. *Preprints*, 13th Conf. on Hydrol., Long Beach, CA, Amer. Meteor. Soc., 54-55.

Benjamin, S. G., 1989: An isentropic meso-$\alpha$ scale analysis system and its sensitivity to

aircraft and surface observations. *Mon. Wea. Rev.,* **117**, 1586-1605.

_____, K. J. Brundage, P. A. Miller, T. L. Smith, G. A. Grell, D. Kim,J. M Brown, and T. W. Schlatter, 1994: The Rapid Update Cycle at NMC. Preprints, *10th Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 566–568.

_____, G.A. Grell, S.S. Weygandt, T.L. Smith, T.G. Smirnova, B.E. Schwartz, D. Kim, D. Devenyi, K.J. Brundage, J.M. Brown, and G.S. Manikin, 2001: The 20-km version of the RUC. Preprints, *14th Conf. on Numerical Weather Prediction*, Fort Lauderdale, FL, Amer. Meteor. Soc., J75-J79.

Black, T. L., 1994: The new NMC mesoscale eta model: description and forecast examples. *Wea. Forecasting*, 9, 265-278.

Chen, F., and Coauthors, 1996: Modeling of land-surface evaporation by four schemes and comparison with FIFE observations. *J. Geophys. Res.*, 101, 7251-7268.

Du, J., J. McQueen, G. DiMego, T. Black, H. Juang, E. Rogers, B. Ferrier, B. Zhou, Z. Toth, and S. Tracton, 2004: The NOAA/NWS/NCEP short range ensemble forecast (SREF) system: Evaluation of an initial condition vs multiple model physics ensemble approach. *Preprints,* 16th Conf. on Num. Wea. Prediction, Seattle, WA, Amer. Meteor. Soc., Paper 21.3, 10 pp.

Ferrier, B.S, 2004: Modifications of two convective schemes used in the NCEP Eta Model. *Preprints,* 16th Conf. on Num. Wea. Prediction, Seattle, WA, Amer. Meteor. Soc., Paper J4.2, 9 pp.

Fritsch, J. M., J. Hilliker, J. Ross, and R. L. Vislocky, 2000: Model consensus. *Wea. Forecasting,* **15,** 571-582.

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

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17**, 192-205.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327.

_____,and _____, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

Hart, K. A., W. J. Steenburgh, D. J. Onton, and A. J. Siffert, 2004: An evaluation of mesoscale-model-based model output statistics (MOS) during the 2002 Olympic and Paralympic winter games. *Wea. Forecasting,* **19,** 200-218.

Homleid, M., 1995: Diurnal corrections of short-term temperature forecasts using the Kalman filter. *Wea. Forecasting,* **10,** 689-707.

Jacks, E., J. B. Bower, V. J. dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting,* **5,** 128-138.

Juang, H.-M. H., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.,* **122,** 3-26.

Kain, J. S., M.E. Baldwin, P. Janish, and S.J. Weiss, 2001: Utilizing the Eta Model with two different convective parameterizations to predict convective initiation and evolution at the SPC. Preprints, *9th Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 91-95.

Klemp, J., 2004: Next-generation mesoscale modeling: A technical overview of WRF. *Preprints*, 20th Conf. Wea. Analysis and Forecasting, Seattle, Amer. Meteor. Soc., Paper 11.2.

Mao, Q., R. T. McNider, S. F. Mueller, H.-M. H. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea. Forecasting,* **14,** 190-202.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291-303.

McQueen J., J. Du, B. Zhou, G. Manikin, B. Ferrier, H.-Y. Chuang, G. DiMego, and Z, Toth, 2005: Recent Upgrades to the NCEP Short Range Ensemble Forecasting System (SREF) and Future Plans. Preprints, *17th Conf. on Num. Wea. Prediction,* Amer. Meteor. Soc., Paper 11.2, 7 pp.

Mullen, S. L., and R. Buizza, 2001: Quantitative Precipitation Forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.,* **129,** 638-663.

Murphy, A. H., R. L. Winkler, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bull. Amer. Meteor. Soc.,* **60,** 12-19.

Pan, H.-L., and L. Mahrt, 1987: Interaction between soil hydrology and boundary-layer development. *Bound. Layer Meteor.,* **38,** 185-202.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649-667.

Ross, G. H., 1989: Model output statistics – An updateable scheme. Preprints, *11th Conf. on Probability and Statistics in Atmospheric Sciences,* Monterey, CA, Amer. Meteor. Soc., 93-97.

Schaake, J., 1989: Importance of the HRAP grid for operational hydrology. *Preprints,* U.S/People's Republic of China Flood Forecasting Symp., Portland, OR,NOAA/NWS, 331-355.

Seo, D.J., 1998: Real-time estimation of rainfall fields using radar rainfall and rain gauge data. *J. of Hydrol.,* **208,** 37-52.

Smirnova, T. G., J. M. Brown, and S. G. Benjamin, 1997: Performance of different soil model configurations in simulating ground surface temperature and surface fluxes. *Mon. Wea. Rev.,* **125,** 1870-1884.

_____ , _____ , _____ , and D. Kim, 2000: Parameterization of cold season processes in the MAPS land-surface scheme. *J. Geophys. Res.,* **105** (D3), 4077-2086.

Smith, J.A., and W.F. Krajewski, 1991: Estimation of the Mean Field Bias of Radar Rainfall Estimates. *J. Appl. Meteor.,* **30,** 397-412.

Stanski, H., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO World Weather Watch Tech. Rep. 8, 114 pp.

Stensrud, D.J., and J.A. Skindlov, 1996: Grid point predictions of high temperature from a mesoscale model. *Wea. Forecasting*, **11**, 103-110.

_____, and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.,* **131,** 2510-2524.

_____, and _____, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.*, **12,** 217-230.

_____, J.-W., Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensembles of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077-2107.

Stull, R. L., 1988: *An Introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers, Boston, MA, 666 pp.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182**, 990-1000.

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.

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

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

Wilson, L. J., and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Wea. Forecasting,* **17,** 206-222.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting,* **20,** 101-111.