123 Improving Probabilistic Ensemble Forecasts of Convection through the Application of QPF-POP Relationships

Christopher J. Schaffer*, William A. Gallus, Jr., and Moti Segal
Iowa State University, Ames, IA.

1. Introduction

Probabilities of precipitation (POPs) can be derived from ensemble forecasts in a variety of ways. Most simply, POPs are determined by considering the percentage of ensemble members forecasting precipitation greater than a specified threshold amount. For a ten member ensemble with equal weighting assigned to each member, the forecast probabilities of precipitation (POPs) would be 0%, 10%, 20%, up to 100%. This method can be thought of as an uncalibrated traditional method (Uncali_trad, hereafter), because it is the simplest approach to determining POPs (Hamill and Whitaker 2006). A calibrated version of the traditional method (Cali_trad, hereafter) formed by training over observed data can be used to provide improved forecasts, helping to correct for some biases.

It is hypothesized that more elaborate methods, using single or ensemble model output, can be used to obtain POP forecasts that are potentially superior to those from Cali_trad and Uncali_trad. For instance, separating quantitative precipitation forecasts (QPFs) into precipitation "bins" can provide new ways of obtaining useful probabilistic information (e.g., Gallus and Segal 2004, hereafter GS04; Gallus et al. 2007, hereafter GBE07; Yussouf and Stensrud 2008). Recently, various studies have used neighborhood approaches, which in their simplest form consider an area surrounding a grid point in order to gain POP improvements by accounting for spatial precipitation probability constraints (e.g. Theis et al. 2005, Ebert 2008).

The specific goal of this study is to identify and apply post-processing approaches similar to the GS04 and neighborhood approaches, as well as a hybrid of both, to traditional model ensemble forecasts, and to examine how the resulting POPs compare to those from traditional approaches.

2. Methodology and Data

The new approaches of determining POPs in the present study mostly involved the creation of 2D POP tables (termed "POP tables" hereafter) based on parameters related to the following two properties: (i) forecasted precipitation amount characteristic (PAC) within a bin (as in GS04) and (ii) the number of ensemble members forecasting agreement on occurrence of precipitation above a threshold amount (as traditionally used for ensemble-based POP forecasts). In this study, the term "ensemble" will not only refer to the traditional definition of sets of model forecasts as defined previously but will also be used for a number of related grid points within an area surrounding each

* *Corresponding author address*: Christopher John Schaffer, National Weather Service, WFO Goodland, Goodland, KS; email: christopher.schaffer@noaa.gov

grid point of the domain (earlier termed neighborhood). The term "method" will refer to a variant of an approach.

The POPs in the tables were assigned by finding the correct alarm ratio (referred to as the hit rate in GS04) for each case in the training dataset. The correct alarm ratio is defined as h/f, where f is the number of grid points with precipitation forecasted for a given combination of bin and member agreement (such a combination is termed a "scenario"), and h is the number of "hits", or points where the observed precipitation also exceeded the specified threshold.

In the present study, the first of the above two POP table properties (the PAC) is given at any grid point by either taking the maximum forecasted amount from any ensemble member, or by taking the ensemble average. Seven precipitation bins were used (with units in inches that are commonly used operationally; 1 inch = 25.4 mm), including <0.01, 0.01-0.05, 0.05-0.10, 0.10-0.25, 0.25-0.50, 0.50-1.00, and >1.0. Using a PAC was necessary because each of the ensemble members provides a precipitation amount, and a single representative precipitation amount was needed at each grid point to apply the binning approach as used by GS04.

Ensemble forecast output for the early warm season in the central U.S. was generated by the 2007 and 2008 NOAA Hazardous Weather Testbed Spring Experiments, which took place during April-June of both years (Kong et al. 2007 and Xue et al. 2008). The ensemble consisted of ten WRF-ARW members with 4-km grid spacing run by the Center for Analysis and Prediction of Storms (CAPS) located at the University of Oklahoma.

NCEP Stage IV precipitation analysis (Baldwin and Mitchell 1997) was used to designate hits at a forecast point if the observed rainfall amount was greater than a threshold. The Stage IV data, along with the forecasts, were interpolated onto 20 km grid spacing using NCEP procedures that conserve the total amount of liquid in the domain.

POP tables were created from the 29 cases from 2008, and were tested against the 20 cases from 2007. For each method, the probability forecasts were verified using decomposed Brier Scores (BSs), Brier skill scores (BSSs), bias calculations, and Relative Operating Characteristic (ROC) areas. Differences were tested for statistical significance using paired t-tests of BSs for all cases and time periods.

The first forecasting approach analyzed made use of two parameters: the 6-hour period PAC and the percentage/number of ensemble members forecasting precipitation above a given threshold, both determined at each grid point from the ten-member ensemble model output. Two definitions of the PAC were used when testing this approach, so two POP tables for each threshold were created, in methods denoted as Max_thr and Ave_thr.

A second forecasting approach was developed using neighborhood methods, as described briefly in the Introduction. Considering a square neighborhood with (NxN) grid points centered at domain grid point (I, J), then each of the (NxN) sets of grid points with the same relative orientation to the domain grid points (I, J) may be considered ensemble members. Within a specified square area around a center point representing a neighborhood, the maximum or average PAC was determined and placed in a bin.

This approach uses two parameters to generate POP tables: the binned PAC and the member agreement percentage, which provides the percentage of members/points within the neighborhood with forecast precipitation amounts greater than a given threshold. Max_thr and similar methods considered forecasts from ten ensemble members, but because this neighborhood approach (abbreviated as Max_nbh or Ave_nbh) uses each of the points within the neighborhood, all of these points can be thought of as a spatially generated pseudo-ensemble (e.g. Theis et al. 2005). This neighborhood approach was applied to deterministic QPFs because it uses a spatial ensemble instead of using the ten WRF-ARW members as an ensemble (a traditional ensemble). For this reason, the approach was used on each of the ten WRF-ARW members. Different POP tables were created by increasing the neighborhood size for each of the ten members until the optimal size was determined.

A third forecasting approach considered both definitions of ensemble members from the previous two approaches, the ten WRF-ARW members and the NxN spatial ensemble members, when determining the PAC and the member agreement parameter. Hence, a 3D neighborhood consisting of NxNxM members (termed hereafter "neighborhood-M"; M=10 in the present study) was formed. To establish a PAC, this approach averages the forecast precipitation amounts from all of the NxNxM members associated with each of the simulation domain grid points.

For large neighborhoods, the number of possible forecast scenarios would become very large, and this could have a negative impact on the efficiency of the approach. By introducing too many forecast scenarios, the grid points considered in the correct alarm ratio statistics could become overdispersed, which would degrade the approach's reliability. In order to decrease the number of forecast scenarios, members considered in the member agreement parameter were grouped such that 10 consecutive members were placed in each group. By grouping members in this way, NxN general member groups, each containing 10 specific members, could be considered, rather than considering NxNx10 specific members.

A final forecasting approach was examined that combined several of the previous methods by averaging their POPs. Considering each contributing method as an ensemble member that consists itself of ensemble members, this approach can be viewed as a "super-ensemble" generated by post-processing. Because POP fields over the domain for the different methods evidenced forecast spread, it was hypothesized that averaging the POPs of multiple methods might result in a forecast superior to the individual methods. This hypothesis was also supported by the common finding that an ensemble mean forecast tends to be more skillful than any single member forecast.

3. Results

3.1 Two-parameter point forecast approach

For all thresholds (0.01, 0.10, and 0.25 inch), the BSs for the new methods were always smaller (closer to zero) than the Uncali_trad BS and the BS from a forecast applying the previous GSD one-parameter precipitation-binning method (described in the introduction) to one of the ten ensemble members. As thresholds increase, however, the degree by which the scores differ tends to decrease. Max_thr and Ave_thr always had higher BSSs and lower bias scores than GSD and Uncali_trad. When compared to the Cali_trad BSs, Max_thr and Ave_thr still have more favorable scores (Fig. 1).

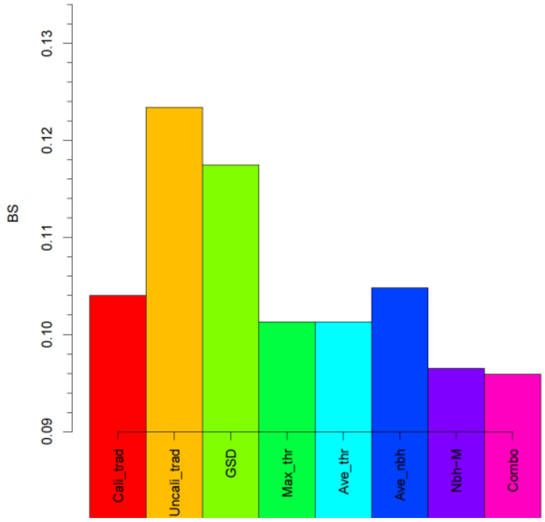

Figure 1: BSs for the 0.01 inch threshold for different methods at 20 km grid spacing.

By testing against the independent 2007 dataset, it was clear that the reliability of Uncali_trad was poorer at all three thresholds (Fig. 2) than that for Max_thr, Cali_trad, and GSD. Max_thr, Cali_trad, and GSD had similar reliability.

The p-values from the paired t-tests of the 100 BSs (20 cases with 5 time periods each) for each method showed that the Max_thr results were statistically significantly different at the 99.9% confidence level for all three thresholds when compared to the Uncali_trad results, the best results from GSD (member 10), and the Cali_trad results with p-values consistently < 0.001. The best BSs can be obtained if the reliability and uncertainty terms are both small and the resolution term large. All of the new presented methods using the two-parameter point forecast approach had larger resolution terms than GSD, Uncali_trad, and Cali_trad. Cali_trad had the smallest reliability term of all the methods.

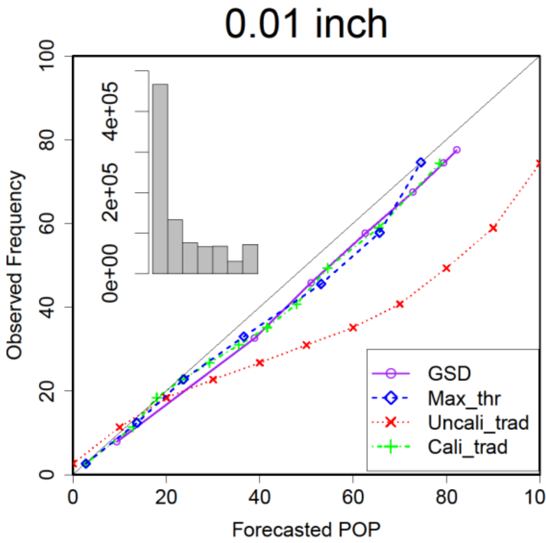

Figure 2: Reliability diagram for GSD, Max_thr, Uncali_trad, and Cali_trad at the 0.01 inch threshold.

The bias values for all methods (not shown) indicated an overestimation in the POP forecasts, though Max_thr and Ave_thr had values closest to 1, showing more favorable biases relative to the other methods. While Max_thr and Ave_thr had the same bias value at the 0.01 inch threshold, the Max_thr method had a slightly better bias value than Ave_thr at the 0.10 and 0.25 inch thresholds.

Both Max_thr and Ave_thr showed an increase in ROC area with increased thresholds (Fig. 3). GS04 and GBE07 also noted this trend, which also occurred in the GSD method. Cali_trad and Uncali_trad show a decrease in ROC area as thresholds increased, so the increased discrimination for forecasts of greater precipitation may be an added benefit of using the QPF-POP relationship compared to the more traditional approaches.

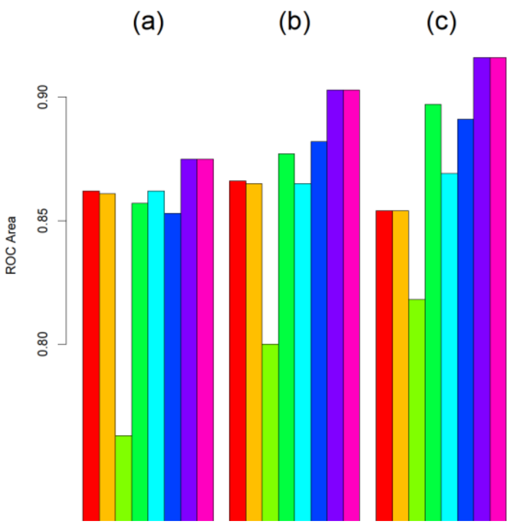

Figure 3: ROC areas for the methods from Fig. 2 (same color coding) at thresholds a) 0.01 inch, b) 0.10 inch, and c) 0.25 inch.

3.2 Two-parameter neighborhood approach

As neighborhood size increased, the reliability of Max_nbh and Ave_nbh deteriorated, but resolution improved to a larger extent. The best BSs generally occurred for a 15x15 point neighborhood for Ave_nbh, after which the loss of reliability began to outweigh improvements in resolution. For Max_nbh, the best BSs occurred for a 13x13 neighborhood (not shown). The best BSs for Ave_nbh, however, were lower (better) than the best scores for Max_nbh, suggesting that averaging the neighborhood points provides a more skillful forecast than selecting the maximum precipitation within the neighborhood.

The 15x15 Ave_nbh results showed that some BSs were greater than the Max_thr scores, while others were less, allowing for a member average BS (Fig. 1) near the value of Cali_trad. For the 0.01 inch threshold, the lowest scores for Ave_nbh were below 0.1000, which was more skillful than the Max_thr, Ave_thr, and also Cali_trad values. This result was surprising, because Max_thr, Ave_thr, and Cali_trad considered all ten ensemble members when creating POPs, but Ave_nbh *considered only an individual member*. However, the neighborhood approach provided additional information so that POP forecasts made from single deterministic forecasts were comparable (or sometimes superior) to POP forecasts made using Cali_trad. The BSs of Ave_nbh applied to WRF-ARW member 8 (which yielded the best average BS) were statistically significantly different from Cali_trad's scores at the 99.9% confidence level at all three thresholds, with p-values < 0.001. ROC areas for Ave_nbh (Fig. 3) again increased with increasing thresholds, and many of the members had ROC areas exceeding 0.90 at the 0.25 inch threshold, which was an improvement over the previous methods' ROC areas.

3.3 Two-parameter neighborhood-M approach

The two-parameter neighborhood-M approach, like the two-parameter neighborhood approach, showed better skill for the average PAC, rather than the maximum PAC, so only the averaging version is presented. The two-parameter neighborhood-M approach's BSs were best for an 11x11 point neighborhood (Fig. 1), were better than the BSs from the previous two sub-sections, and were statistically significantly different from Cali_trad's scores at the 99.9% confidence level at all three thresholds with all p-values < 0.001 (and nearly 0). The bias values were better than the Max_thr, Ave_thr, and the member-averaged Ave_nbh bias values. The areas under the ROC curve for each threshold were larger than for Cali_trad and increased with increasing thresholds (Fig. 3). The 0.01 inch threshold ROC area was 0.875, and the 0.25 inch area was 0.916, which was larger than any of the previous methods' areas. The large ROC areas indicate that this approach has better discrimination than the approaches presented earlier.

3.4 Combination approach

By averaging the POPs for Ave_nbh, Max_thr, and Cali_trad, and increasing the Ave_nbh neighborhood size from 3x3 to 15x15, the BS improved from 0.0995 to 0.0959 for the 0.01 inch threshold. This is a relatively large improvement over the BS of 0.1014 associated with the Max_thr method. The POP forecasts were superior to any of the forecasts from other methods examined thus far. When compared to Cali_trad, the results for this combination approach were statistically significantly improved at the 99.9% confidence level with p-values < 0.001 (and nearly 0) at all three thresholds.

When the neighborhood for Ave_nbh within the combination approach was increased from 3x3 to 15x15 grid points, the reliability worsened, but the resolution improved to a greater extent. This behavior was also observed for Ave_nbh alone. The reliability was better for the ensemble of methods compared to Ave_nbh, however, likely due to the contribution of Max_thr and Cali_trad, which had better reliability scores than Ave_nbh at larger neighborhoods. Thus, the combination of methods had reliability comparable to Max_thr and resolution similar to Ave_nbh. By including the ten forecasts from Ave_nbh, this combination approach used the ensemble average POP from the two-parameter neighborhood approach at each point, which alone yielded improved BSs compared to each of the individual Ave_nbh members. By including Ave_thr and the two-parameter neighborhood-M approach in this combination approach, the skill increased marginally.

3.5 Further comparisons of methods

Figure 4 contains POP domain plots for one 6 hour period from one case in the sample (the same case and time are used in both figures). The top panel shows the POP forecasts over the domain for Ave_nbh, and bottom panel shows the POP forecasts for the combination approach. The POP fields become smoother for the combination approach, and the BS improved as a result of the averaging (BSs indicated in the captions).

In order to evaluate the sensitivity of the approaches to the grid spacing of the data set, the methods were applied to an identical sub-domain, but using the original unsmoothed 4 km grid spacing instead of the smoothed 20 km spacing. The BS differences at 20 km were also present at 4 km grid spacing, indicating that the approaches could outperform traditional methods at both grid spacings. The 4 km BSs were better than the related 20 km BSs, but the differences in skill between methods were similar at 4 km to what was indicated with the 20 km results (Fig. 5).

Finding improved BSs at 4 km grid spacing compared to 20 km grid spacing for the two-parameter point approach was unexpected because past deterministic studies have found that standard measures of skill usually show deteriorating skill as grid spacings are refined. Mass et al. (2002) and Gallus (2002) show that the equitable threat score (ETS) was higher when evaluating QPF on coarser grid spacings compared to finer ones. It is possible that probability

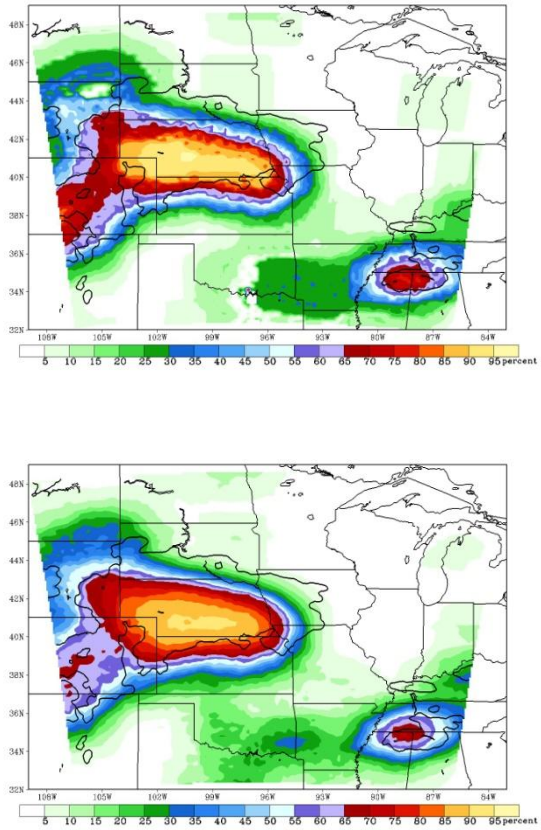

Figure 4: POPs (on the 20 km grid) for Ave_nbh 15x15 using member 8 (top panel) and the combination approach (bottom panel) over the domain for 2007 April 23 for the period 06-12Z. The dark contour denotes observations at the 0.01 inch threshold. The BS for this day, time, and threshold for Ave_nbh (Combination) was 0.0856 (0.0762).

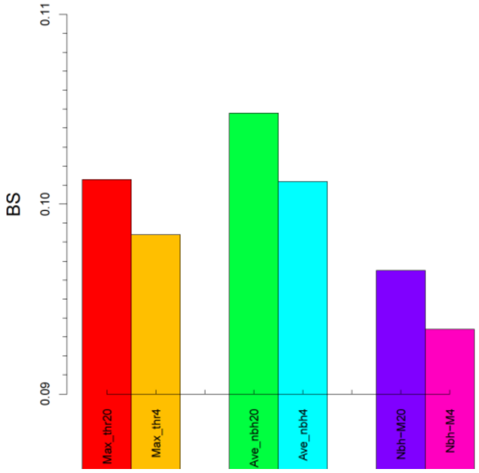

Figure 5: Comparison of Brier scores for selected methods at 20 km and 4 km grid spacing at the 0.01 inch threshold

trends sometimes behave differently than QPF trends when going to finer grid spacings, based on the metric being used. In order to test whether this was the case in our study, ETSs were computed at both 20 km and 4 km. The ETSs were better for the 20 km results compared to the 4 km results, which is what we would expect from past literature. Scatterplots of BS (Fig. 6) and ETS (Fig. 7) show the individual cases in the 20 km and 4 km datasets, where the BSs are for Ave_nbh using ensemble member 8, and the ETSs are evaluated using the ensemble member 8 QPFs. Values of ETS and BS are larger for the 20 km output but only differ slightly.

4. Discussion and Conclusions

The present study is an extension of the single forecast-based POP approach used in GS04 and GBE07 to a ten-member WRF-ARW ensemble, while providing a comparison to a calibrated traditionally-used equal weighting approach for determining POPs from ensembles. Exploratory tests were performed using a range of approaches, and some related variant methods were considered using data from early in the convection season in the central U.S. The POPs were evaluated based on performance at each domain grid point. Quantification of the skill of the new approaches emphasized the use of BSs and ROC areas.

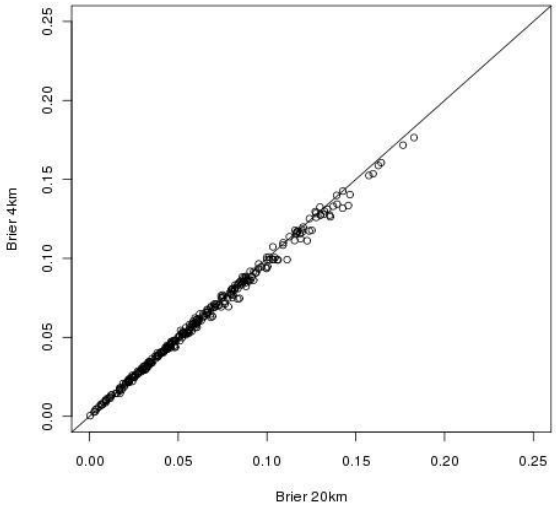

Figure 6: Scatterplot of Brier scores for Ave_nbh 15x15 (member 8) at the 20 km and 4 km grid spacings.

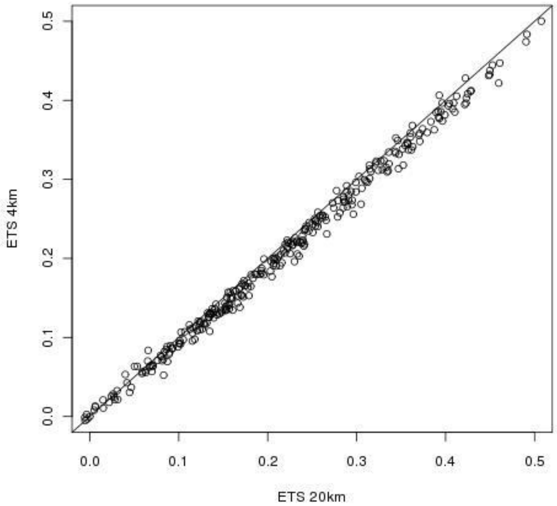

Figure 7: Scatterplot of ETSs for Ave_nbh 15x15 (member 8) at the 20 km and 4 km grid spacings.

For all methods, the most pronounced improvements in POP skill occurred for the lowest threshold, with diminishing improvements above a threshold of 0.25 inch. Hence, the methods may be better at delineating areas experiencing precipitation and determining the location and timing of convective initiation compared to Cali_trad and Uncali_trad, more so than generating better forecasts of excessive rainfall.

Ensembles generated using the neighborhood approach produced POPs as skillful as those from the ten-member ensemble forecast Cali_trad. This suggests that the approach is very attractive operationally. Because post-processing of a single deterministic simulation can provide skill comparable to that obtained by Cali_trad, computer resources used for the traditional ensemble simulation might be better used for further refinement of the model grid spacing or for improved model physical formulation.

A two-parameter neighborhood-M approach considered binned PAC and ensemble member agreements (in the ten-member model ensemble and the neighborhood ensemble) with precipitation greater than a threshold, and produced POP forecasts of even higher skill than the two-parameter point forecast approach and the two-parameter neighborhood approach. When all three approaches were considered together with Cali_trad, the resulting combination approach produced forecasts that were statistically significantly better compared to Cali_trad's forecasts at the 99.9% confidence level with p-values that were nearly 0 at all thresholds.

Overall, this study suggests that three of its evaluation techniques potentially can be used to provide useful POP forecasts.

Two of the evaluation techniques are represented by the two general parameters used within the approaches: binning a PAC and determining the member agreement percentage. The PAC-binning parameter was used in all of the new approaches, as well as in GSD. In this study, the benefits of the PAC-binning parameter were especially apparent when considering the ROC areas for the new approaches introduced and GSD, because these ROC areas increased further than the ROC areas for Cali_trad and Uncali_trad. The member agreement percentage parameter is used in Cali_trad and Uncali_trad in addition to the approaches introduced in this study because it is well-established as an important POP-forecasting technique. The third evaluation technique which can provide useful POP forecasts is the neighborhood approach. We showed that using a neighborhood of grid points can yield probabilistic information from deterministic forecasts, produce POP forecasts that may rival traditional calibrated ensemble POP forecasts, and also improve traditional ensemble forecasts.

5. Acknowledgements

This research was funded in part by National Science Foundation grants ATM-0537043 and ATM-0848200, with funds from the American Recovery and Reinvestment Act of 2009. The authors would like to thank Adam Clark for providing the model precipitation output and for helpful discussions, Jon Hobbs for his advice regarding the statistical analysis, and Daryl Herzmann for his computational assistance. We appreciate the constructive comments of Elizabeth Ebert and two anonymous reviewers that improved the related, submitted manuscript.

6. References

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

Ebert, E. E., 2008: Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, **15**, 51-64.

Gallus, W. A., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*, **17**, 1296-1302.

____, and M. Segal, 2004: Does increased predicted warm-season rainfall indicate enhanced likelihood of rain occurrence? *Wea. Forecasting,* **19**, 1127–1135.

——, M. E. Baldwin, and K. L. Elmore, 2007: Evaluation of probabilistic precipitation forecasts determined from Eta and AVN forecasted amounts. *Wea. Forecasting*, **22**, 207–215.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Mon. Wea. Rev.*, **134**, 3209–3229.

Kong, F., and Coauthors, 2007: Preliminary analysis on the real-time storm-scale ensemble forecasts produced as a part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. Preprints, 22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction, Park City, UT, Amer. Meteor. Soc., 3B.2.

[Available online at http://ams.confex.com/ams/pdfpapers/124667.pdf.]

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407-430.

Theis, S. E., A. Hense and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications*, **12**, 257-268.

Wilson, L. J., and M. Vallée, 2002: The Canadian Updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206–222.

Xue, M., and Coauthors, 2008: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2008 Spring Experiment. Preprints, 24th Conf. on Severe Local Storms, Savannah, GA, Amer. Meteor. Soc., 12.2. [Available online at http://ams.confex.com/ams/pdfpapers/142036.pdf.]

Yussouf, N., and D. J. Stensrud, 2008: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system during the 2005/06 cool season. *Mon. Wea. Rev.*, **136**, 2157–2172.