JP2.1 APPLICATION OF A QPF-POP RELATIONSHIP TO ENSEMBLES TO GENERATE PROBABILISTIC PRECIPITATION FORECASTS

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

1. INTRODUCTION

Ensemble forecasts have many advantages over deterministic forecasts. Ensemble forecasts facilitate probabilistic forecasts and provide a measure of uncertainty, unlike deterministic forecasts. Ensemble forecasts are more useful than single deterministic forecasts, because small errors in a single forecast's initial conditions will grow exponentially over time, making the forecast increasingly unreliable (Hamill and Colucci 1997) and the uncertainties provided by probabilistic forecasts increasingly meaningful. Numerous studies have shown that raw probabilistic methods can be improved (e.g. Yussouf and Stensrud 2008; Gallus et al. 2007). Post-processed probabilistic precipitation forecasts are more useful than raw ensemble probabilistic forecasts, because post-processing refines the results, improves reliability, and increases the quality of the forecasts.

Gallus and Segal (2004) and Gallus et al. (2007) used a precipitation-binning technique in a deterministic forecast to show that, at grid points where the "binned" quantity of precipitation was larger, the probability that those grid points would receive at least a small amount of precipitation was greater. It is believed that this is because when the models predicted larger amounts of precipitation the atmospheric state was such that precipitation was more likely to occur. In Gallus and Segal (2004), it was noted that POP values increased even further if two different models showed an intersection of grid points with rainfall in a specified bin. Their findings indicated that the QPF-POP relationship might yield an even better forecast if the relationships were applied to ensemble forecasts.

The goal of this study is to apply the Gallus-Segal method to ensemble forecasts and to show how this method compares to the traditional ensemble method, which only considers the percentage of ensemble members with accumulated precipitation above a threshold, and the Gallus-Segal (denoted GS) deterministic method, which generates POPs using precipitation bins alone.

* *Corresponding author address*: Christopher John Schaffer, Iowa State Univ., Dept. of Geol. and Atm. Sciences, Ames, IA; email: schaffec@iastate.edu

2. DATA AND METHODOLOGY

The new method of determining POPs involved the creation of 2D tables based on precipitation amount (like the GS method) and the number of ensemble members forecasting a certain amount of precipitation (like traditional ensemble precipitation forecasts).

Seven precipitation bins were used (with units in inches), 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. A characteristic amount of precipitation at each point was determined given the ensemble results, either by taking the maximum (Max) forecasted amount from any ensemble members at that point, or by taking the ensemble average (Ave). Finding a characteristic-precipitation amount was necessary because each of the ensemble members provides a precipitation amount, and a single representative precipitation amount was desired at each grid point in order to apply the binning-technique.

In the tables, the other parameter on which POP forecasts were based was the percent of members forecasting precipitation. Two types of member specification were used, resulting in a different 2D table for each. The first type of ensemble specification considered the number of ensemble members with precipitation above a threshold, and the second type considered the number of ensemble members with precipitation in the same bin as the characteristic bin amount. These two types of specifications will be referred to as "threshold" (abbreviated 'thr') or "bin." Considering both parameters, four types of POP tables were created and denoted as Max_bin, Max_thr, Ave_bin, and Ave_thr.

The POPs in the tables were assigned by finding the hit rate (or correct-alarm ratio) for each case in a training dataset. The hit rate is defined as h/f, where f is the number of grid points with precipitation forecasted within a given bin, and h is the number of "hits", or points where the observed precipitation exceeded a specified threshold. NCEP stage IV precipitation observations were used to designate hits at a forecasted point if the observed rainfall amount was greater than a

threshold (either 0.01in [T1], 0.10in [T2], or 0.25in [T3]).

Warm season ensemble data was generated by the 2007 and 2008 NOAA Hazardous Weather Testbed Spring Experiments, which took place from April-June of each year (Levit et al. 2008). The ensemble consisted of 10 WRF-ARW members with 4-km grid spacing run by the Center for Analysis and Predication of Storms (CAPS) located at the University of Oklahoma. There were a few differences between the two experiments, mainly that the 2007 experiment was initialized at 2100 UTC, while the 2008 experiment was initialized at 0000 UTC. For this reason, the first 3 hours of the 2007 data were excluded for each day. The 2008 data was also on a larger grid than the 2007 data (3600 x 2700 versus 3000 x 2500). The 4km data was mapped to a 20 km grid, as in Clark et al. (2009). The 2D tables were created from the 29 cases of 2008 data, and tested against the 20 cases from 2007.

3. RESULTS

3.1 2-D POP Table

Table 1 is the 2D POP table created for the 0.01in observed precipitation threshold using the Max_thr method. Due to space considerations, tables for the 0.10 and 0.25 thresholds, as well as tables for the other three methods, are not shown. As the amount of simulated precipitation increased in Table 1, the POP, defined by the hit rate, also tended to increase for each of the three thresholds. In the few instances where POPs decreased with increasing threshold, there were relatively few points associated with the percentage calculation,

Table 1. POP table for the 0.01 threshold of the Max_thr method. The upper part of each cell contains the POP, and the lower part contains the number of associated grid points.

	<0.01in	**0.01-0.05**	**0.05-0.10**	**0.10-0.25**	**0.25-0.50**	**0.50-1.0**	**>1.0in**	**Column Ave**
0%	2.8	-	-	-	-	-	-	2.8
	721837	0	0	0	0	0	0	721837
10%	-	11.4	15.4	18.1	18.6	19.7	28.7	12.8
	0	80102	13683	8873	2904	1091	369	107022
20%	-	14.3	19.2	22.3	23.8	26.7	31.3	18
	0	36475	15360	13010	4929	2192	803	72769
30%	-	16.1	23.5	26.2	30.5	30.6	39.1	23.4
	0	18532	13338	14282	6516	3383	1385	57436
40%	-	18.5	25	31.5	36.3	39.9	39.9	29.3
	0	10081	10863	14403	7969	4367	1872	49555
50%	-	19.4	27.6	36	42.5	45.8	46.8	35.5
	0	5282	8388	13593	8815	5411	2479	43968
60%	-	19.7	28.3	39.3	47.4	52.9	55.3	41.6
	0	2901	6379	12615	9379	6671	3504	41449
70%	-	23	31.4	42.5	53.1	57	61.7	47.9
	0	1821	4927	11463	10318	7998	4459	40986
80%	-	21.5	33	47.2	56.9	63.3	66.3	54.5
	0	922	3588	10510	11461	9931	5987	42399
90%	-	15.8	35.4	53.6	66.8	71.4	77.6	65.5
	0	438	2792	10506	14840	14738	10196	53510
100%	-	16.8	27.6	55.2	73.3	83.9	89.2	78.6
	0	167	1642	9954	21333	30985	25648	89729
Row Ave	2.8	13.7	23.7	36.6	53.1	65.6	74.5	19.4
	721837	156721	80960	119209	98464	86767	56702	1320660

which may have accounted for the discrepancy.

As the percentage of ensemble members increased, the POPs also generally increased. This was expected, because this is the basis for the traditional method of forecasting POPs. The combined increase in POPs with both increasing accumulated precipitation and member percentage resulted in the highest POPs. Similarly, low precipitation amounts and low member percentages yielded low POPs. Points in the second column of Table 1 were restricted by definition of the method; if the maximum precipitation is less than 0.01in (essentially no precipitation), then the rest of the members must also have accumulated precipitation less than the lowest threshold. This definition results in a very low POP value for this scenario, which is fitting, because we would expect a very low likelihood of precipitation when none of the members are anticipating measurable precipitation.

The right-most column of Table 1 is a summation over all bins for each member percentage, which indicates what the POP would be for each member percentage if precipitation amount was not considered. These POPs increase with increasing member percentage and are used in making forecasts for the calibrated traditional method.

The bottom row of Table 1 is a summation over all member percentages, similar to what would be determined from the GS approach. This row provides a single POP representative of each precipitation bin. The POPs increase with increasing bin amounts, and these POPs are useful when making reliability and ROC diagrams.

The lower-right value in Table 1 is the summation of the summation column/row, which represents a summation over the entire grid. Likewise, this probability of precipitation is representative of the entire grid, regardless of the forecast amount. This POP (known as the sample climatology from now on) tends to be greater than the POP associated with the no-precipitation bin, but less than the POP of the bin associated with 0.01 to 0.05 in precipitation (though not in Table 1). This result shows that rainfall was more likely to occur where the model predicts precipitation to occur on the domain, and less likely where the model doesn't predict precipitation, as observed in Gallus and Segal (2004).

A common trend in the tables was a decrease in the number of bin points with increasing precipitation amount and member. Few points had both high amounts and member percentages, though the POPs were generally very high (between 80% and 100%) for these points. The fact that the POPs were high indicates that precipitation was almost inevitable when most or all members forecasted heavy amounts.

3.2 Reliability Diagrams

Fig. 1 shows the reliability diagrams for thresholds 0.01in, 0.10in, and 0.25in. All of the methods were close to the perfect reliability line. The Max_bin method remained closest to the perfect reliability line for each threshold, though the other methods didn't deviate far. The Ave_bin and Ave_thr methods provided higher forecasted POPs than the Max_bin and Max_thr methods, and these points for the Average methods tended to be farther from perfect reliability than the related points for the Max methods.

Fig. 2 shows the same reliability diagrams as in Fig. 1, but with the addition of reliability curves for 10 deterministic forecasts formed by applying the previous GS method to each ensemble member. Reliability curves for the traditional and calibrated traditional methods were also added to Fig. 2. The traditional forecast clearly had poorer reliability than the new methods at each threshold, but the calibrated traditional curve showed better reliability. The GS deterministic (abbreviated GSD) method showed fairly good reliability as in Gallus and Segal (2004) and Gallus et al. (2007). In the next section, Brier scores will be examined for each method to quantify "better" reliability and gage which methods are more reliable, but Max_bin appears to be more reliable than the other methods. The other three new methods sometimes deviated far enough away from the perfect reliability line that it was difficult to say definitively that these forecasts were more reliable than some of the better GSD curves and the calibrated traditional method.

3.3 Brier Scores

Table 2 shows the overall decomposed Brier scores at each threshold for the new methods, deterministic GS method, and both the traditional and calibrated traditional methods. Instead of showing each of the 10 GSD forecasts, the results in the table are the averaged results. Using the method described by Murphy (1973), Brier scores were decomposed into three components: reliability, resolution, and uncertainty. Brier skill scores were also computed, using the sample climatology as a reference value. Brier scores are essentially a measure of mean squared error, so smaller scores (preferably close to 0) are ideal. On the other hand, large Brier skill scores

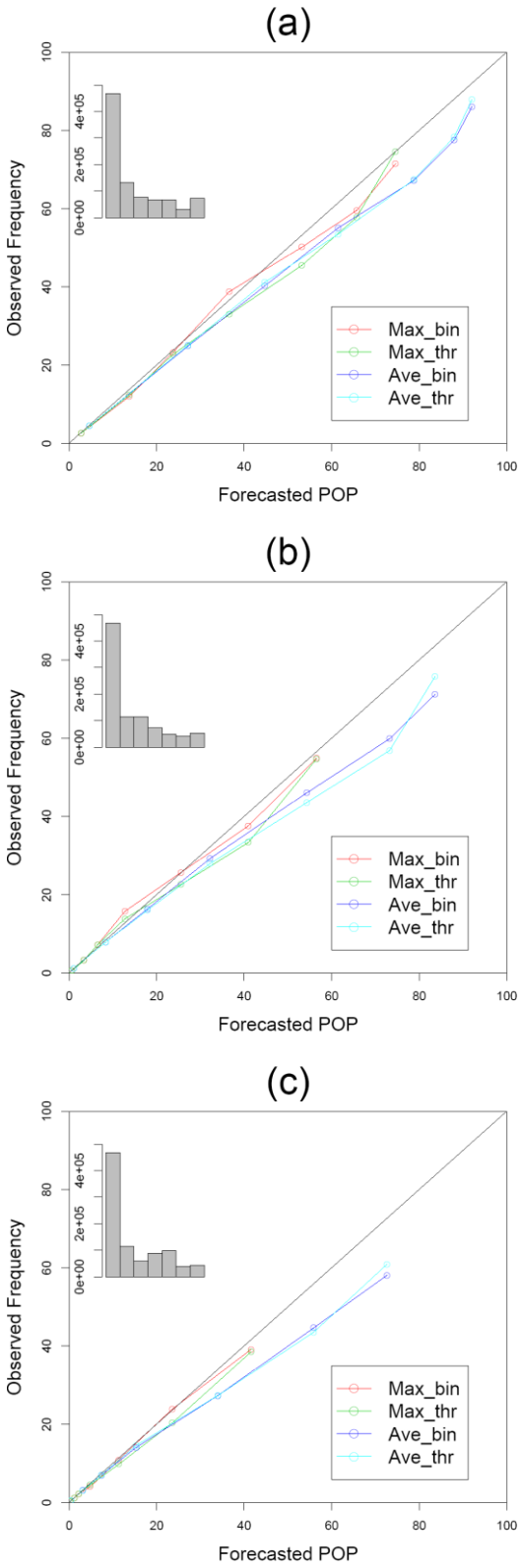


Figure 1. Reliability diagrams using the new methods for thresholds a) 0.01in, b) 0.10in, and c) 0.25in. Bar plot shows forecast point distribution in Max_thr.

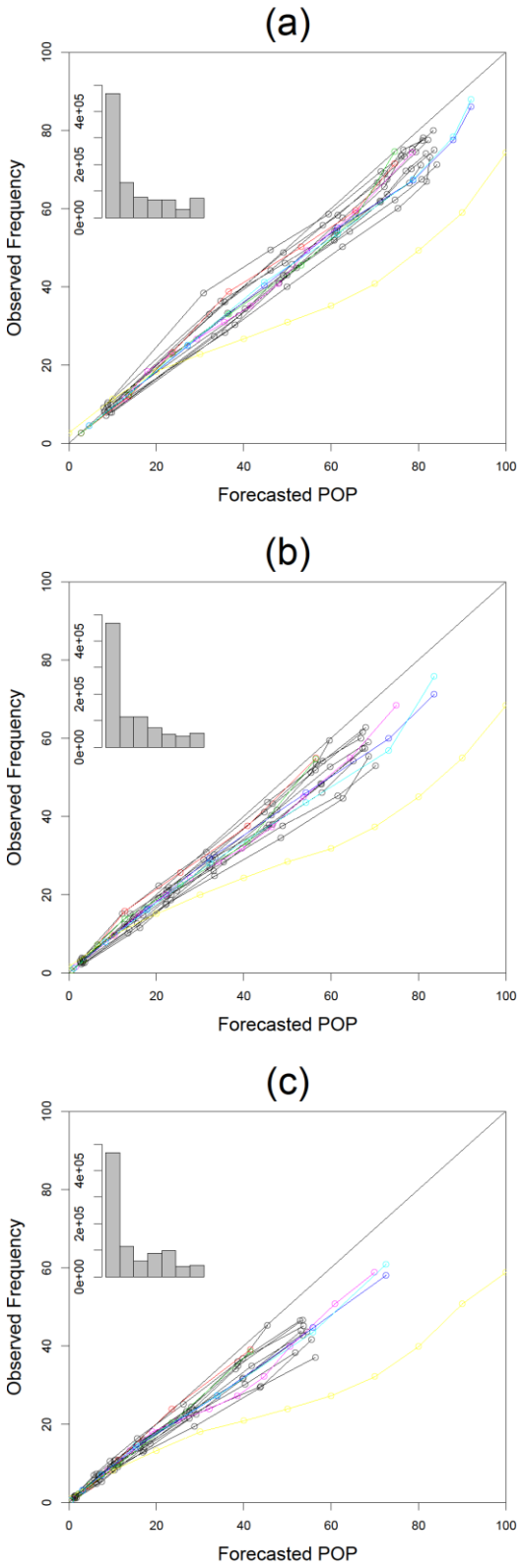


Figure 2. Reliability diagrams for thresholds a) 0.01in, b) 0.10in, and c) 0.25in as in Figure 1, except with 10 deterministic curves (black), the traditional curve (yellow), and calibrated traditional curve (pink).

Table 2. Decomposed Brier scores for the new methods, the GS-deterministic method (ten members averaged together), traditional method, and calibrated traditional method. (* One outlier removed)

Threshold and method	BS	Reli	Resol	Uncert	BSS
T1-Max_bin	0.105733	0.006038	0.045933	0.145628	0.232505
T1-Max_thr	0.102103	0.005919	0.049444	0.145628	0.251758
T1-Ave_bin	0.10486	0.007036	0.047805	0.145628	0.245133
T1-Ave_thr	0.104003	0.006963	0.048588	0.145628	0.249827
T1-GSD	0.117946	0.007693	0.035376	0.145628	0.178424
T1-Trad	0.122215	0.024612	0.048024	0.145628	0.068912
T1-Cali	0.104121	0.006517	0.048024	0.145628	0.231008
T2-Max_bin	0.060537	0.003594	0.019776	0.076718	0.148066
T2-Max_thr	0.05912	0.003582	0.02118	0.076718	0.153293
T2-Ave_bin	0.059489	0.004193	0.021422	0.076718	0.146729
T2-Ave_thr	0.059465	0.004438	0.021691	0.076718	0.142156
T2-GSD	0.065444	0.004786	0.016059	0.076718	0.085007
T2-Trad	0.070142	0.014857	0.021433	0.076718	-0.066058
T2-Cali	0.059472	0.004188	0.021433	0.076718	0.140518
T3-Max_bin	0.036499	0.002226	0.008646	0.042918	0.055115*
T3-Max_thr	0.036091	0.002234	0.009061	0.042918	0.047361*
T3-Ave_bin	0.036009	0.002776	0.009686	0.042918	0.032067*
T3-Ave_thr	0.035887	0.002679	0.009711	0.042918	0.035359*
T3-GSD	0.038632	0.002893	0.00718	0.042918	-0.058182
T3-Trad	0.043941	0.010504	0.009481	0.042918	-0.354376
T3-Cali	0.036438	0.003002	0.009481	0.042918	-0.015222

(close to 1) are desired.

For all thresholds, the Brier scores for the new methods were always smaller (closer to zero) than the GSD and traditional Brier scores. As thresholds increase, however, the degree by which the scores differ becomes small, so that the new methods' scores are smaller by only a very small margin by Threshold 3. The new methods' Brier skill scores are always higher than those of the GSD and traditional methods.

According to the decomposed Brier score equation, in order to have a low Brier score, the reliability and uncertainty terms should be small, and the resolution term should be large. For all methods and thresholds, the new methods' reliability terms were smaller than those of the GSD method and traditional method. This conclusion of better reliability agreed with the reliability diagrams (Fig. 1 and 2). The resolution components for the new methods were all greater than those of the GSD method, but were not always greater than the traditional method's resolution component. The differences were small and did not consistently favor one method or another, so it was difficult to judge resolution from the Brier score components alone. Resolution will be investigated further in the next section when ROC diagrams (which provide another means of measuring resolution) are examined. Finally, the uncertainty term decreases with increasing thresholds, but it does not differ between methods. This is because the uncertainty is only a function of the sample climatology, which is independent of forecast method.

When compared to the calibrated traditional Brier scores, Max_thr and Ave_thr still had more favorable scores, and Max_thr and Max_bin had the most favorable reliability components. The 0.25in (third) threshold's reliability results for Max_thr and Max_bin were significant at a 95% confidence interval, denoting better reliability for these new methods at this threshold.

3.4 ROC Diagrams

Table 3 shows ROC areas for all methods, while Fig. 3 and Fig. 4 show the ROC curves for each of the new methods and for all the methods at each threshold, respectively. Overall, the values were high; all values for all methods were greater than 0.70, which indicates a useful forecast (Buizza et al. 1999). All values for the new methods, however, were also greater than the GSD and traditional values. The calibrated traditional ROC area was higher than all but Ave_thr at threshold 1, but the new methods had the larger ROC areas at threshold 3 (three of the four were already larger by threshold 2). The new methods yielded approximately the same values for each threshold, from around 0.85 at threshold 1 to near 0.90 at threshold 3. The increase in ROC areas show that resolution increased as the thresholds increased. This trend was probably due to the decrease in forecasted areas, as forecasts of less than 0.10in and 0.25in would not be included at thresholds 2 and 3, respectively.

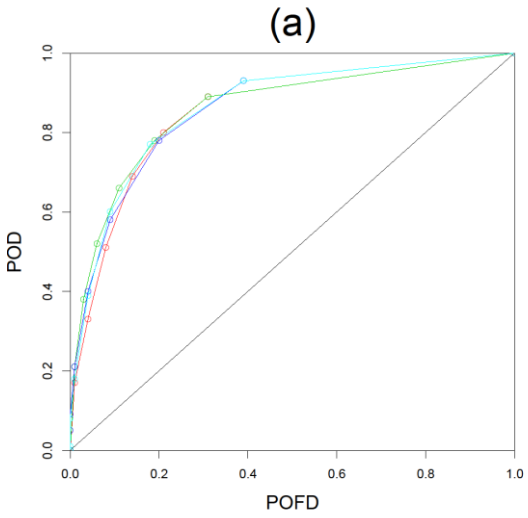
The traditional and calibrated traditional methods are the only methods that have a decrease in resolution as thresholds increased, so the increased resolution for forecasts of greater precipitation may be an added benefit of using the QPF-POP relationship compared to the traditional approach. Gallus and Segal (2004) and Gallus et al. (2007) also noted this trend.

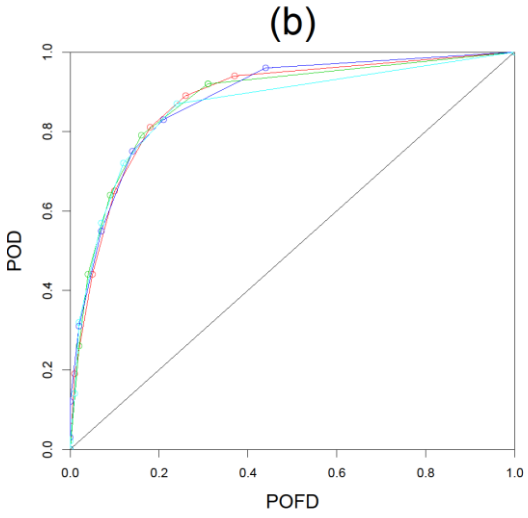
3.5 Forecast Illustration

Fig. 5a is a forecast (using the 0.01in observed threshold) for one of the 20 days for which the new methods (in this case, Max_thr) were tested. Fig. 5b is the same day and time period, but showing the calibrated traditional method instead of the new method. Fig. 5c shows the difference in POPs by subtracting the Max_thr POPs from the calibrated traditional POPs over the domain. Max_thr tends to forecast higher

Table 3. ROC areas for the new methods, the GS-deterministic method (ten members averaged together), traditional method, and the calibrated traditional method.

Threshold	Max_bin	Max_thr	Ave_bin	Ave_thr	GSD	Trad	Cali Trad
T1 (0.01)	**0.851**	**0.857**	**0.861**	**0.862**	**0.763**	**0.848**	**0.862**
T2 (0.10)	**0.880**	**0.877**	**0.884**	**0.865**	**0.800**	**0.835**	**0.866**
T3 (0.25)	**0.897**	**0.897**	**0.896**	**0.869**	**0.818**	**0.807**	**0.854**





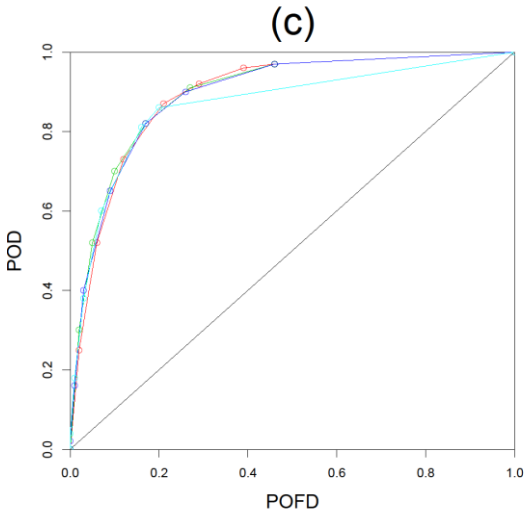
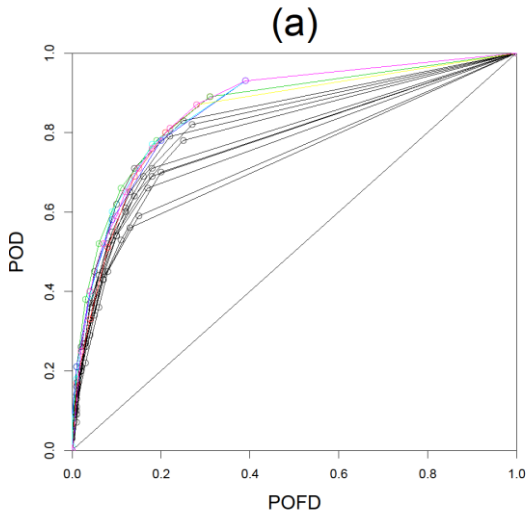
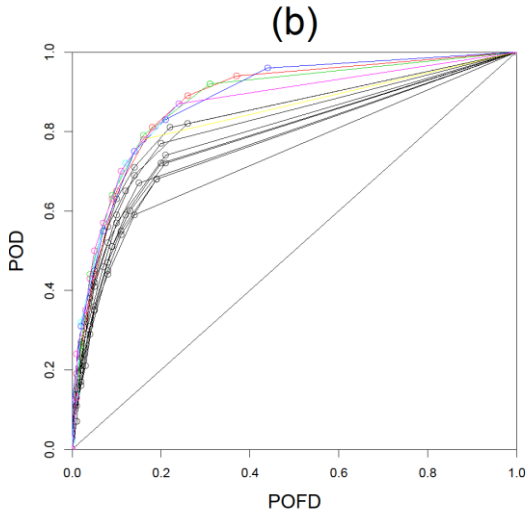


Figure 3. ROC diagrams for the new methods at thresholds a) 0.01, b) 0.10in, and c) 0.25in. Color scheme as in Fig. 1.





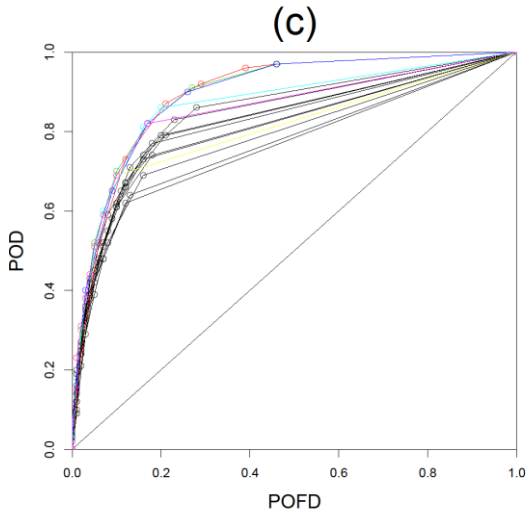


Figure 4. ROC diagrams for all methods at thresholds a) 0.01in, b) 0.10in, and c) 0.25in. Color scheme as in Fig. 2.

POPs in areas that receive precipitation (as denoted by the dark contours in each image), whereas the calibrated traditional method appears more likely to forecast higher POPs in areas that don't receive precipitation. Maps such as these provide supplementary insight into the QPF analysis.

4. CONCLUSION

By forming 2-D POP tables which consider the amount of accumulated precipitation and the percentage of ensemble members forecasting precipitation, new methods for forecasting POPs were created. Preliminary results from the new forecast methods, which apply the Gallus-Segal QPF-POP relationship, show that the new methods provide better reliability and resolution than the GS deterministic and traditional methods, according to the reliability diagrams, ROC diagrams, and Brier scores. The new methods' results are similar to the results of the calibrated traditional method, though the findings favor the new methods. The improvements of Max_thr and Max_bin for the 0.25 threshold's reliability component were statistically significant, and the methods' ROC areas continued to increase at this same threshold, so these new methods tend to do better in areas receiving large accumulated precipitation amounts. The ability to outperform the calibrated traditional method was visualized by plotting forecasted POPs and observations over a domain for each method. Future research will further investigate possible benefits of using the QPF-POP relationship and options to improve the performance of these new methods. For example, one of the new methods (such as Max_thr) may show further improvements over the calibrated traditional method when a neighborhood forecasting approach is used or when the number of ensemble members is reduced.

5. REFERENCES

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168–189.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small near-convection-permitting and large convection-parameterizing ensembles. *Wea. Forecasting* (In press).

Gallus, W. A., 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 S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

Levit, J.J., G.W. Carbin, D.R. Bright, J.S. Kain, S.J. Weiss, R.S. Schneider, M.C. Coniglio, M. Xue, K.W. Thomas, M.E. Pyle, and M.L. Weisman, 2008: The NOAA Hazardous Weather Testbed 2008 Spring Experiment: Technical and scientific challenges of creating a data visualization environment for storm-scale deterministic and ensemble forecasts. *Preprints*, 24th Conf. Severe Local Storms, Savannah GA.

Murphy, A.H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

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.

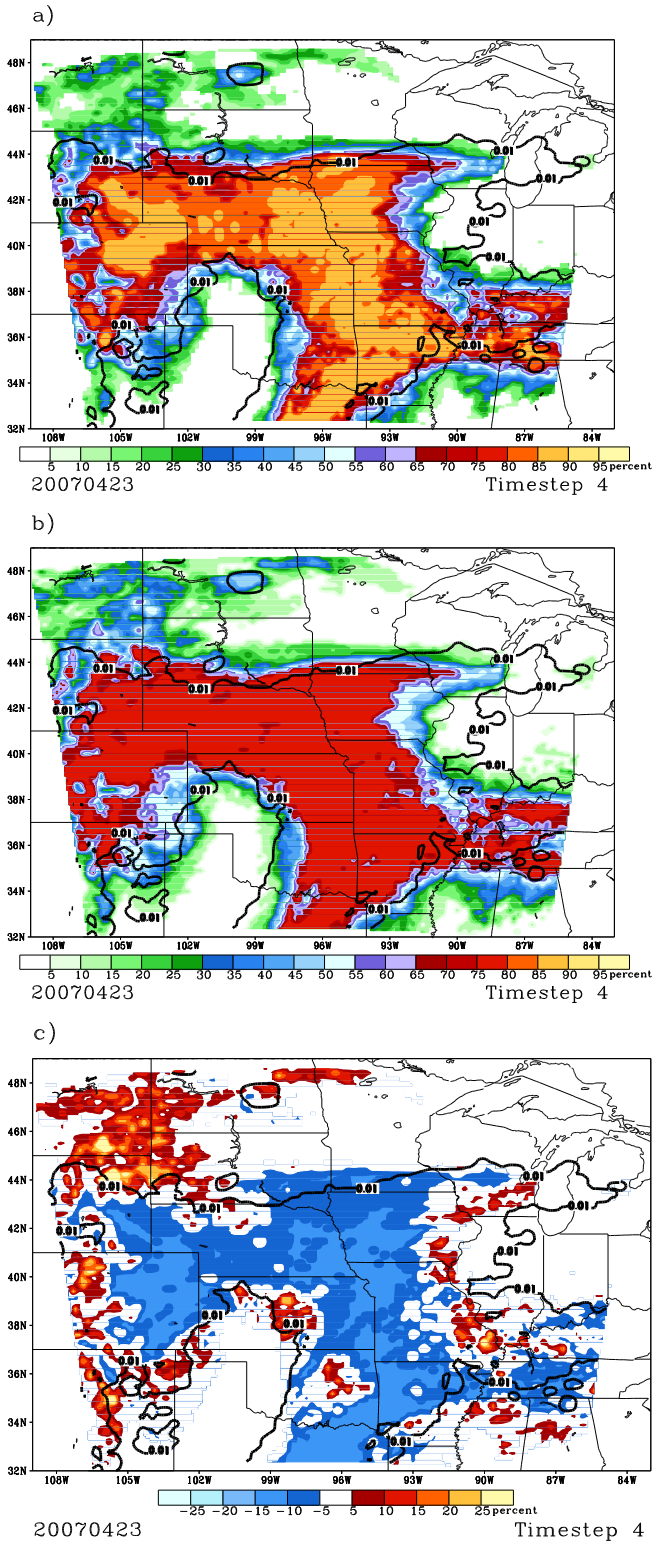


Figure 5. Forecasts for a) Max_thr, b) the traditional calibrated method, and c) the difference between the traditional calibrated and Max_thr. Observed rainfall is contoured and labeled.