Improving probabilistic ensemble forecasts of convection through the application of QPF-POP relationships

by

Christopher John Schaffer

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

Major: Meteorology

Program of Study Committee:
William A. Gallus Jr., Major Professor
Kristie Franz
William Gutowski

Iowa State University
Ames, Iowa
2010

TABLE OF CONTENTS

ABSTRACT

Quantitative precipitation forecasts provide an accumulated precipitation amount for a given time period, and accurate forecasts depend on the correct prediction of areal coverage, timing, and intensity of precipitation. These forecasts are important to a variety of people for many different purposes, so expressing a likelihood of precipitation is also useful. Most simply, probabilities of precipitation are determined by considering the percentage of ensemble members forecasting precipitation greater than a specified threshold amount. Probabilities of precipitation can also be formed from quantitative precipitation forecasts through statistical post-processing. Past research has shown that there are many ways to post-process precipitation data, such as by binning the precipitation amounts, applying statistical calibration, and/or considering the percentage of an area receiving precipitation.

The main goal of this study was to expand upon relationships between quantitative precipitation forecasts and probabilities of precipitation by developing new approaches that yield more accurate probabilities of precipitation than methods that are currently more commonly used. Ensemble forecasts from the 2007 and 2008 NOAA Hazardous Weather Testbed Spring Experiments were used to provide quantitative precipitation forecasts for various days. In the study, four main approaches were developed and tested extensively using Brier scores and other statistics. Brier scores for different approaches were compared to traditional methods of calculating probabilities of precipitation. It was shown at both 20 km and 4 km grid spacings that new approaches were able to produce statistically significantly better forecasts than a traditional method that relies upon calibration of POP forecasts derived using equal-weighting of ensemble members. A deterministic approach developed during the study was also able to produce forecasts comparable to those of the calibrated traditional method.

CHAPTER 1. GENERAL INTRODUCTION

Introduction

Quantitative precipitation forecasts (QPFs) provide an accumulated precipitation amount for a given location and time period. QPFs are among the most important and challenging forecasts a meteorologist can prepare, and QPFs strongly influence decisions by the US government and industry (Fritsch et al. 1998). Forecasting the areal coverage, intensity, and timing of precipitation events is a difficult task, and the consequences of an erroneous forecast can be anywhere from inconvenient to devastating. The aviation industry needs to know where and when areas will be experiencing rain, because storms with heavy rain (which are also often associated strong wind shear and low visibilities) can cause dangerous flight and landing conditions (Luers and Haines, 1983). In the Chicago Metropolitan area, traffic accidents were twice as likely to occur on rainy days, and 57% of the 30-minute flight delays at Chicago's O'Hare Airport occurred on rainy days (Changnon 1996). Large areas of hypoxic water exist in the Gulf of Mexico primarily due to nitrogen leached from land upstream, and the timing and amount of precipitation determines the amount of nitrogen leached from land and the amount of leached nitrogen that makes it downstream (Donner and Scavia, 2007). Accurate QPFs could help farmers minimize their fertilizer losses and aid natural resource managers in determining changes in hypoxia. Accurate QPFs are also needed to forecast flash floods which, in an average year, cause property damage that exceeds that for all other weather-related natural phenomena (Fritsch et al. 1998). Due to the seemingly unlimited impacts of precipitation, there is a need for increasingly accurate precipitation forecasts.

Forecasting an exact amount of precipitation is not always practical, so forecasts are sometimes made for a certain range of amounts, e.g. between 0.01 inch and 0.10 inch (Du et al. 1997). Forecasting for a range better expresses the inherent uncertainty of the forecast compared to forecasting a specific amount, e.g. 0.07 inch. Another way of expressing forecast uncertainty is through the use of probabilities of precipitation (POPs). POPs can be used to express the probability of exceeding a precipitation threshold, e.g. 70% chance of greater than 0.10 inch of precipitation. POPs are generated from ensemble forecasts, which are forecasts of an event with variations to initial states and/or model configurations, such as the forecasts' physical parameterizations. POPs are prepared by NOAA through the National Weather Service and the Hydrological Prediction Center as part of NOAA's mission to protect life and property. The HPC provides POP forecasts for thresholds of precipitation through their Excessive Rainfall and Winter Weather forecasts. The HPC provides deterministic QPFs, as well, though these forecasts do not provide a measure of uncertainty (Im et al. 2006).

The rate of improvement of QPF skill has been slow despite improvements in observations and numerical models in recent years (Fritsch et al. 1998). In particular, Olson et al. (1995) noted how QPFs were noticeably, consistently worse in the warm season compared to the cool season. Im et al. (2006) found that the skill of HPC deterministic QPFs deteriorated as the amount of precipitation increased. The use of ensemble forecasts, however, has advantages over the use of deterministic QPF forecasts. For example, using the mean ensemble QPF can reduce QPF errors (Ebert 2001). Stensrud and Yussouf (2007) stated that precipitation forecasts after a few hours should be viewed only from a probabilistic perspective, because there is so much uncertainty in forecasts after a few hours. Because probabilistic forecasts provide a measure of uncertainty, probabilistic forecasts are more useful than deterministic forecasts (Fritsch et al. 1998).

When numerical weather prediction models, such as the Weather Research and Forecasting Model (WRF) produce QPFs, these QPFs can then be interpreted statistically to form POPs through QPF-POP relationships in a procedure known as post-processing. Finding new QPF-POP relationships through post-processing techniques can help us produce more accurate POP forecasts.

Research Questions

The primary goal of this study is to build upon the QPF-POP relationship investigated by Gallus and Segal (2004) and Gallus et al. (2007). These studies used a post-processing technique on model QPFs which involved separating QPFs into precipitation “bins.” A bin is a range of precipitation amounts, such as 0.01 inch to 0.05 inch, and 0.05 inch to 0.10 inch. Using these bins, the two studies showed that at model grid points where the “binned” quantity of forecasted precipitation was larger, the probability that those grid points would receive at least a small amount of precipitation was greater than where the forecasted precipitation amount was smaller. In other words, areas that were forecasted to receive much rainfall were more likely to receive at least some rainfall compared to areas were little rainfall was forecasted. This relationship, called the Gallus-Segal approach, was applied to single deterministic forecasts, as opposed to ensemble forecasts. The goal of this work is to adapt the QPF-POP relationship just described to ensemble forecasts, thereby creating new relationships in this ensemble setting. These relationships are shown in the study through unique ensemble forecasting approaches.

In order to develop approaches in the ensemble setting, another parameter besides the binning parameter can be used. This second parameter considers the number of ensemble members with precipitation greater than a threshold. This “agreement” parameter, along with the binning parameter, will be used to create 2D tables of POPs. The tabular POP forecasts from the new post-processing approaches for an ensemble environment should have improved skill compared to the deterministic Gallus-Segal approach, and will be compared to more traditional ensemble POP forecasts. In order to be of value to forecasters, these new approaches will need to create more accurate forecasts than the approaches traditionally used. Traditional forecasts provided a reference by which to measure the success of the new approaches, and a motivation for testing more elaborate forecasting approaches. The study will apply statistical tests to the results to determine if the improvements are statistically significantly different compared to the more traditional ensemble methods.

The NOAA Hazardous Weather Testbed Spring Experiments from April to early June 2007 and 2008 provided the ensemble data used in the study. The Center for Analysis and

Prediction of Storms (CAPS) developed a 4 km grid spacing WRF-ARW system for the experiments (Kong et al. 2007), which used a 10-member ensemble. This study will use the 29 cases from 2008 to form the POP tables, and test the POPs against the 20 cases from 2007. The primary reason for training over the 2008 data was because it was the larger data set.

Recent studies have shown the benefits of using "neighborhood" approaches, which consider an area of grid points instead of a single point. This type of approach had not been tested in the context of the QPF-POP relationship used by Gallus and Segal (2004) and Gallus et al. (2007), so it remained to be seen how a neighborhood approach to forecasting would impact the QPF-POP relationship. It was also unclear how the grid spacing of the forecasts could affect the QPF-POP relationship. Mass et al. (2002) and Gallus (2002) had shown that traditional methods of determining QPF skill may not be as appropriate for fine-scale grid spacings compared the coarse grid spacings, so testing the POPs at different grid spacings may provide further insight into these grid-scale consequences. The Spring Experiment data had 4 km grid spacing, and the data was coarsened to a 20 km grid spacing in order to investigate the differences between these two spacings.

Thesis Organization

This thesis follows the journal paper format. Chapter 1 contains the general introduction to the thesis, and Chapter 2 is a brief literature review of post-processing techniques commonly used in recent years to create POPs. Chapter 3 is a paper which will be submitted to *Weather and Forecasting.* Chapter 4 contains material which was not included in the paper from Chapter 3, but can provide additional insight into the topics considered within Chapter 3. Chapter 5 is the general conclusion which reviews the major findings of the paper in Chapter 3 along with the additional information from Chapter 4. Related topics for future research are also recommended. The final parts of the thesis are the acknowledgements and references.

CHAPTER 2. LITERATURE REVIEW

Forecasts should include probabilistic information for the good of the public (Murphy and Winkler 1979), since probabilistic guidance can aid in risk guidance (Fritsch and Carbone 2004). For this reason, probabilistic forecasts (POP forecasts, specifically) have been provided to the public by the National Weather Service since 1965 (Murphy and Winkler 1979).

Hamill and Colucci (1997) and Hamill and Whitaker (2006) demonstrated how POP forecasts can be calibrated using probability distributions. Hamill and Colucci (1997) determined a cumulative distribution function for a Gumbel distribution fit to POP forecasts from a 15-member ensemble to obtain calibrated POPs. Calibration occurred after the uncalibrated POPs were found to have nonuniform rank distributions for the 15 cases considered. The calibrated forecasts showed better reliability and Brier scores than the uncalibrated forecasts. Eckel and Walters (1998) also had success in producing more accurate POPs with this calibration technique. Hamill and Whitaker (2006) used analog methods to calibrate POP forecasts, which involved using reforecasts (or hindcasts) for dates in the past with similar atmospheric conditions. Using a 25-year collection of reforecasts, the analog approaches were determined to be more skillful and required less computational-cost compared a traditional ensemble method which determines POPs by considering the percentage of ensemble members forecasting precipitation greater than a specified threshold amount.

Stensrud and Yussouf (2007) and Yussouf and Stensrud (2008) both show the benefits of using post-processing to make skillful POPs. In Stensrud and Yussouf (2007), a post-processing technique was developed to produce reliable POPs. The data from 1 June to 15 September 2004 contained 107 forecast days (each consisting of 48 hours) and included 16 ensemble members. Stage-II data was used for observations, and a binning procedure was used to process the forecasts. The forecasts were compared to the observations, and a given day's forecast was adjusted based on the previous 12 days of observed 3-h accumulated precipitation values. The method also involved using an ensemble mean forecast.

This method tended to lower the POPs at the lower threshold values. This lowering of the POPs increased the accuracy of the method, since the raw ensemble forecasts typically over-predicted the POP of all thresholds. Reliability diagrams showed that the adjusted POPs tended to be more accurate than the raw POPs at many forecast times and thresholds. At higher precipitation amounts especially, the adjusted POPs are much more skillful than those for the raw ensemble. When Brier skill scores were computed, the adjusted ensemble was shown to be generally more skillful than the raw ensemble for accumulation periods of 24 hours or less. Still, the adjusted ensemble skill decreased with increasing forecast lead time and increasing precipitation amounts. Relative operating characteristic (ROC) curves also showed that ensemble forecast skill decreased as the precipitation amounts increased.

Yussouf and Stensrud (2008) used reliability diagrams, Brier skill scores, and ROC areas to examine a binning technique similar to that used in Stensrud and Yussouf (2007) for an ensemble system during the cool season. The technique increased the skill of the POPs compared to the raw ensemble results. When compared to Stensrud and Yussouf (2007), it was determined that the cool season's Brier skill scores and ROC areas were more favorable than those for the warm season, indicating that this technique worked better during the cool season.

Two studies closely related to the proposed research are Gallus and Segal (2004) and Gallus et al. (2007). Gallus and Segal (2004) simulated 20 warm-season convective events in the upper Midwest using 10-km versions of the Eta and WRF models in order to investigate if rainfall probability of occurrence is a function of forecast intensity. Specifically, they wanted to determine if heavy forecasted rainfall was better associated with observed precipitation than lighter forecasted rainfall. The simulations were run for 24 hours over a domain of about 1000 km by 1000 km. Rainfall forecast skill was assessed at 6-hour periods, and compared to NCEP Stage-IV observations. Overall, 51 6-hour cases were obtained from the 20 events.

Probability of precipitation was determined by calculating the hit rate, which is also known as the correct-alarm ratio. The forecasted precipitation was placed in QPF bins (based on operational verification), and observation thresholds were used to determine whether rain occurred at a given point on the domain. They found that hit rates (equivalent to POPs) increased as the quantity of forecasted precipitation (indicated by bins) increased. The POPs increased quickly at the lowest bins, and then increased more gradually at higher bins. This trend was present for all observed thresholds, though the POPs at the higher thresholds were not as high as POPs at the lower thresholds. They also observed that probabilities were larger still if two different model versions showed an intersection of grid points. Their findings indicated that more specific QPF-probability relationships could yield more detailed probabilities, if these relationships were applied to ensemble forecasts.

In order to test the reliability and skill of their results, they created reliability, ROC, and relative operating level (ROL) diagrams. The reliability diagrams used 41 training cases and 10 test cases, and showed points close to the reliability line. The ROC and ROL curves were both above the no-skill line, and the areas under the curves were close to 0.7, which indicates a useful forecast.

Gallus et al. (2007), which elaborates on the study by Gallus and Segal (2004), uses a 1-year period to establish the QPF/POP relationship using the Eta and AVN models. This relationship was then tested against another 1-yr period. This study used 3-hr time periods instead of 6-hr periods as in Gallus and Segal (2004), and considered Brier scores in addition to reliability and ROC diagrams. This study also uses a binning procedure, though with a slightly different partitioning of the bins than in Gallus and Segal (2004).

In both models used, the probabilities of precipitation increased with increasing forecasted rainfall accumulations, which is in agreement with what was seen in Gallus and Segal (2004).

Brier scores tended to be slightly smaller for the AVN compared to the Eta, but the difference lessened with increasing observed thresholds. The Brier score calculation was broken down into components to show uncertainty, reliability, and resolution, as described in Murphy (1973). The reliability in the study was nearly perfect, with the associated diagrams showing the reliability curve only deviating slightly from the reliability line. The resolution term became much worse as the observation threshold increased. Also, the Brier scores for the day 2 forecast period were worse than the scores for the day 1 forecasts, which showed that forecast skill decreased as forecast range increased. Finally, the ROC diagrams showed that the method was useful for both models out to the second day of data, with areas under the ROC curve at or above 0.70.

Numerous studies in recent years have introduced “neighborhood” approaches to forecasting POPs (Theis et al. 2005, Ebert 2009, Roberts and Lean 2008, Schwartz et al. 2009, among others). Theis et al. (2005) used a deterministic forecasting approach that considered a spatial area, or neighborhood, around a grid point. The grid points in the neighborhood with forecasted precipitation greater than a threshold were counted, and this number was divided by the total number of points within the neighborhood to produce a POP for the center grid point. The approach also used a temporal neighborhood of 3 hours. The purpose of the spatial and temporal neighborhood was to allow for small inconsistencies in the time and space of the forecast and gain information about the general likelihood of precipitation for a time and place. This approach was made to have low implementation and running costs, which is why it does not use ensemble data or calibration over observations. Schwartz et al. (2009) used a similar (but purely spatial) neighborhood approach on the 2007 Hazardous Weather Testbed Spring Experiment ensemble output. When ROC areas and fractions skill scores were used to evaluate the neighborhood approach’s forecasts, it was shown to have superior skill than a traditional ensemble forecast with equal weighting of members. These neighborhood approaches and others are summarized in Ebert (2009).

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

by

Christopher J. Schaffer[1], William A. Gallus[1], Jr., Moti Segal[2]

[1]Dept. of Geological and Atmospheric Sciences
[2]Dept. of Agronomy

Iowa State University

Ames, IA

A paper to be submitted to Weather and Forecasting

Corresponding Author: Christopher J. Schaffer, 3018 Agronomy, Iowa State University,

Ames, IA 50011, schaffec@iastate.edu

Abstract

Several approaches of post-processing quantitative precipitation forecasts (QPFs) from an ensemble were used to generate probability of precipitation (POP) tables in order to develop a forecasting method that could outperform a traditional method that relies upon calibration of POP forecasts derived using equal-weighting of ensemble members. Warm season 10-member ensemble output from the NOAA Hazardous Weather Testbed Spring Experiments was used, with 29 cases serving as a training set to create the POP tables and 20 cases used as a test set. The new approaches use QPF-POP relationships based on two properties termed precipitation amount and agreement. In the first approach, POPs were based on a binned precipitation amount and the number of ensemble members with 6-hour precipitation accumulations greater than given thresholds. In a second approach, a neighborhood method was used to find the number of points in an area with precipitation greater than a threshold, while also considering the binned amount representative of the neighborhood. This approach for a single ensemble member yielded forecasts as good as those obtained by using a traditional calibrated 10-member ensemble. A third approach synthesized the previous methods and led to an increase in skill relative to the individual methods. After application of a correction for forecast overestimation, a fourth approach using a combination of methods produced forecasts that were improved statistically significantly compared to the calibrated traditional method's forecasts, both at 20 km and 4 km grid spacing. The second approach on its own showed skill comparable to that obtained

by a traditional calibrated ensemble, so adopting this approach alone should save computer resources which could then be used for model refinements, at the expense of the increased skill from including the other approaches used in the fourth approach.

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). Also, ensemble mean forecasts tend to be more skillful than any single member forecast (Smith and Mullen 1993, Ebert 2001, Chakraborty and Krishnamurti 2006).

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%. In this study, this method will be referred to as the uncalibrated traditional method (Uncali_trad, hereafter), because it is the simplest approach to determining POPs (Hamill and Whitaker 2006). Hamill and Colucci (1997) showed how calibration over observed data can improve POPs created using a Gumbel distribution fit to ensemble data, while Hamill and Whitaker (2006) described a method to calibrate POPs using reforecasts. 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.

More complicated methods, with or without ensemble output, can be used to obtain probability forecasts that are potentially superior to those from Cali_trad and Uncali_trad. For instance, separating quantitative precipitation forecasts (QPF) 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 used a neighborhood approach, which considers an area surrounding a grid point in order to gain additional insight (Theis et al. 2005, Ebert 2009, Roberts and Lean 2008, Schwartz et al. 2009, among others). Operational centers have also begun using techniques like spatial density plots that incorporate neighborhood approaches (D. Novak, National Centers for Environmental Prediction, 2010, personal communication). The present study describes an exploratory attempt to use variants of such approaches to provide *grid point related POPs that outperform more traditional approaches.* Previous studies have not addressed this specific objective.

GS04 and GBE07 used a precipitation-binning technique in a deterministic forecast to show that, at grid points where the "binned" quantity of forecasted precipitation was larger, the probability that those grid points would receive at least a small amount of precipitation was greater than where the forecasted precipitation amount was smaller. They attributed this to the fact that when the models predicted larger amounts of precipitation the atmospheric state was such that precipitation was more likely to occur. In GS04, 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 relationship was applied to ensemble forecasts.

The specific goal of this study is to apply post-processing techniques similar to the GS04 technique to ensemble forecasts and to examine how the results compare to those from more traditional approaches. Section 2 describes the general methodology and data used. Section 3 discusses the results from different post-processing methods and provides Brier scores (BS) that will primarily be used when comparing methods to the more traditional methods. Discussion and conclusions follow in section 4.

2. Methodology and Data

The new methods of determining POPs typically involved the creation of 2D POP tables based on forecasted precipitation amount within a bin (as in GS04) and the number of ensemble members forecasting agreement on amount of precipitation above a threshold amount (as traditionally used for ensemble-based POP forecasts). In this paper, the term "ensemble" will not only refer to the traditional definition of sets of model variants as defined previously but will also include a number of related grid points within an area (i.e. neighborhood). The 2D POP tables represent joint probability distributions, as discussed in Wilks (2006) and illustrated later. Conceptually, POP tables can be of higher dimensions if additional variants of the properties are considered. In the present study the first of the above two properties is given either by taking the maximum forecasted amount from any ensemble member at that point, or by taking the ensemble average. Considering GS04 and the higher deterministic skill for the ensemble-averaged precipitation field compared to any member, these two characteristics are likely to support a POP table with improved forecasting skill. Using a characteristic precipitation amount 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. In the tables, the second property used to construct POP forecasts was the percentage of ensemble members forecasting precipitation amounts above specified thresholds (agreement).

Ensemble forecast output for the early warm season 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. The experiments differed some between the first and second years. In the 2007 experiment, five of the ten members (including the control member) used both perturbed initial conditions and mixed physical parameterizations, and the remaining five members used only mixed physical parameterizations. In the 2008 experiment, eight of the ten members used both perturbed initial conditions and mixed physical parameterizations. Descriptions of the initial conditions and lateral boundary conditions used can be found in Kong et al. (2007) and Xue et al. (2008). The 2007 experiment was initialized at 2100 UTC, while the 2008 experiment was initialized at 0000 UTC. Because of the differences in initialization time, the first 3 hours of the 2007 data were excluded for each day, and five 6 hour accumulated precipitation periods, 00-06, 06-12, 12-18, 18-00, and 00-06 UTC, were used to create the probability forecasts. The 2008 output was also on a larger grid than the 2007 output (3600 km x 2700 km versus 3000 km x 2500 km), but the present study uses the sub-domain (Fig. 3-1) used in Clark et al.

(2009) with the dimensions 1980 km x 1840 km. Two grid spacings were considered on this sub-domain: (i) a 20 km grid generated by mapping the 4 km output from both years to the new domain, and (ii) the original 4 km grid. The 20 km grid was emphasized in the present study because effectively the averaging removes noise in the precipitation fields associated with wave lengths less than or equal to 5Δx (e.g. Tustison 2001, Skamarock 2004) thus providing potentially more accurate precipitation fields. Gallus (2002), among others, showed that skill measures are generally better for coarser grid spacings than finer ones, and analyzing 4 km data requires greater computational and time resources. Although our analysis focuses on the 20 km output, sensitivity tests for the most promising methods were performed using the 4 km output. At each grid spacing, the 2D POP tables were created from the 29 2008 cases, and were tested against the 20 cases from 2007. Sensitivity tests of training over the 2007 cases and testing against the 2008 cases showed low sensitively, so the results were not shown.

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. The POPs in the tables were assigned by finding the hit rate (or correct-alarm ratio) for each case in the training dataset. The hit rate is defined as h/f, where f is the number of grid points with precipitation forecasted for a given bin/member scenario, and h is the number of “hits”, or points where the observed precipitation also exceeded the specified threshold. NCEP Stage IV precipitation observations (Baldwin and Mitchell 1997) were used to designate hits at a forecast point if the observed rainfall amount was greater than a threshold. Stage IV data can have a slight dry bias at thresholds less than 0.25 inch compared to gauge-only observations due to an overestimation of rainfall (Schwartz and Benjamin 2000).

For each method, the probability forecasts were verified using decomposed BSs, Brier skill scores (BSSs), bias calculations, and ROC areas. Reliability diagrams, ROC diagrams, BS scatterplots, and additional illustrations of skill were examined. Differences were tested for statistical significance at the 95% confidence level using the Student's t-test, unless stated otherwise. A summary of the methods tested and approaches introduced in this study can be found in Table 3-1.

3. Results

a) Two-parameter point forecast approach

1) POP TABLES

The first forecasting approach analyzed made use of two parameters, characteristic precipitation amount and count of ensemble member agreement, both determined at each grid point from the ensemble output. For the count of ensemble member agreement two different methods were used, resulting in a different 2D POP table for each. The first method counted the number of ensemble members with precipitation above a threshold (hereafter "thr"), and the second the number of ensemble members with precipitation in the same bin as the characteristic amount (hereafter "bin"). Considering both parameters used to define the characteristic amount, four POP tables for each threshold were created, denoted as Max_bin, Max_thr, Ave_bin, and Ave_thr.

An illustrative 2D POP table created for the 0.01 inch observed precipitation threshold using the Max_thr method is shown in Table 3-2. Due to space considerations, tables for the 0.10 inch and 0.25 inch thresholds, as well as tables for the other three methods are not shown, but can be found at http://www.meteor.iastate.edu/~schaffec/poptables.html. As the amount of simulated precipitation increased, the POP 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, which may have accounted for the unusual behavior.

As the percentage of ensemble members with precipitation amounts greater than the threshold (a traditional way of defining POPs from ensembles) increased, the POPs also generally increased. The increase of POPs associated with both increasing accumulated precipitation and ensemble member agreement percentage resulted in the highest POPs (lower-right corner of the table). Conversely, a combination of low precipitation amounts and low member agreement percentages yielded low POPs (upper-left side of the table). Points in the second column of Table 3-2 are restricted by definition of the method; if the maximum precipitation was less than 0.01 inch (essentially no precipitation), then all members had 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 ensemble members are forecasting measurable precipitation.

The right-most column of Table 3-2 is a summation over all bins for each member agreement percentage, indicating what the POP would be for each member percentage if binning of the precipitation amount was not considered. These POPs increase with increasing member agreement percentage and are the values used for Cali_trad. Cali_trad can also be thought of as a traditional method (defined as equally-weighted forecasts yielding POPs of 0%, 10%, 20%, etc.) that has been adjusted using observations, and as the marginal distribution of the joint probability distribution. Max_thr and Ave_thr by definition provide a refinement of Cali_trad.

The bottom row of Table 3-2 is a summation over all member percentages, similar to what would be determined from the GS04 approach. This row provides a single POP representative of each precipitation bin (i.e. reflecting the ensemble average POP). The POPs increase with increasing bin amounts. These POPs are used when making reliability and ROC diagrams, because they allow for bin-representative points on the diagrams, like the diagrams in GS04 and GBE07.

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

2) RELIABILITY DIAGRAMS

The reliability of Uncali_trad was clearly poorer at all three thresholds (Fig. 3-2) than that for Max_thr, Cali_trad, and a forecast applying the previous Gallus-Segal deterministic (abbreviated GSD) one-parameter method to one of the ten ensemble members for comparison purposes. Cali_trad indicates better reliability than that for all other methods.

The Max_thr method appears to have reliability comparable to that of both Cali_trad and GSD, which showed fairly good reliability as in GS04 and GBE07.

3) BRIER SCORES

BSs were examined to quantitatively compare reliability among the methods. The BS is defined as

$$BS = \frac{1}{n}\sum_{k=1}^{n} (p_k - o_k)^2 \tag{1}$$

where p_k is the forecast probability for forecast k of n total forecasts, and o_k is the observed probability (either 0% or 100%) corresponding to each forecast. Using the method described by Murphy (1973) and Wilks (2006), BSs can be decomposed into three components: reliability, resolution, and uncertainty. The decomposition is mathematically described as:

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \bar{o}_i)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2 + \bar{o}(1 - \bar{o}) \tag{2}$$

where

$$\bar{o}_i = \frac{1}{N_i}\sum_{k \in N_i} o_k \quad , \tag{3}$$

$$\bar{o} = \frac{1}{n}\sum_{i=1}^{I} N_i \, \bar{o}_i \quad , \tag{4}$$

N_i is the number of forecasts in the ith forecast category, and n is the total number of forecasts. The first, second, and third terms on the right side of Eq. 2 represent reliability, resolution, and uncertainty components of the Brier score, respectively. The reliability component, like in reliability diagrams, compares forecasts to observed frequencies, while the resolution component quantifies how well a method discerns different types of events. The uncertainty component is independent of the forecast approach used, because it only considers observations. BSs are essentially a measure of mean squared error so smaller scores (preferably close to 0) are ideal. Brier skill scores were also computed:

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = 1 - \frac{BS}{BS_{ref}} \tag{5}$$

where the reference BS (BS$_{ref}$) is the sample climatology. When calculating the sample climatology, o is used for p$_i$ in the decomposition equation (Eq. 2), so the reference BS is reduced to the uncertainty. Large BSSs indicate better skill compared to the sample climatology. Finally, a bias statistic was also calculated, using:

$$BIAS = \frac{\sum_{k=1}^{n} p_k}{\sum_{k=1}^{n} o_k} \tag{6}$$

Table 3-3 shows the overall decomposed BSs at each threshold for the new methods, GSD, Uncali_trad, and Cali_trad. Instead of showing each of the ten GSD forecasts, the results in the table are the averaged results for the ten (GSD_ave10) and best five (GSD_ave5) members. For all thresholds, the BSs for the new methods were always smaller (closer to zero) than the GSD and Uncali_trad BSs. As thresholds increase, however, the degree by which the scores differ becomes small. The new methods always had higher BSSs and lower bias scores than GSD and Uncali_trad. When compared to the Cali_trad BSs, Max_thr, Ave_thr, and Ave_bin still have more favorable scores. When the BSs of Max_thr and the three related methods were compared to the Cali_trad scores, however, the differences were not statistically significant.

The p-value from analysis of variance tests of the 100 BSs for each method (20 cases with 5 time periods each) and t-tests showed that the Max_thr results were statistically significantly different at the 95% confidence level for the 0.01 inch threshold when compared to the best results from GSD (member 10), and significant at the 99.9% confidence level when compared to the Uncali_trad results for that threshold (Table 3-4). For the 0.10 inch and 0.25 inch thresholds, the Max_thr results were statistically significantly different from the Uncali_trad results at the 99% and 95% confidence levels, respectively. The decrease in statistical significance with increased thresholds reflects how differences between BSs decreased with increasing thresholds. The differences between methods continued to decrease for thresholds greater than 0.25 inch, so these results were not shown.

The decomposed BS equation shows that in order to have a low BS, the reliability and uncertainty terms should be small and the resolution term should be 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. The reliability diagrams (Fig. 3-2) clearly showed that Uncali_trad had worse reliability than the other methods, which the reliability component of the Brier decomposition confirmed. The Max_thr and Ave_thr methods performed better than Max_bin and Ave_bin due to the resolution component. Max_bin and Ave_bin had slightly better reliability components, but worse resolution (especially Max_bin). Finally, the uncertainty term decreased with increasing thresholds, but it did not differ between methods because the uncertainty is only a function of the sample climatology and is thus independent of forecast method.

4) ROC DIAGRAMS

ROC diagrams illustrate the ability of a forecast method to discern events and non-events, while the areas under the curves quantify this discernment. Both the diagrams and areas relate the probability of detection (POD) to the probability of false detection (POFD). An ideal ROC area is 1, with a curve that goes from the lower left corner (where POD=0 and POFD=0) to the upper left corner (where POD=1 and POFD=0), and on to the upper right corner (where POD=1 and POFD=1). Figure 3-3 shows the ROC curves for Max_thr, GSD, Cali_trad, and Uncali_trad, while ROC areas for all methods are given in Table 3-3. Overall, the ROC areas 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 four new methods, however, were also greater than the GSD values. The Cali_trad and Uncali_trad ROC areas were higher than all other areas except Ave_thr at the 0.01 inch threshold, but the new methods had larger ROC areas than Cali_trad and Uncali_trad at the 0.25 inch threshold (three of the four were already larger than the Cali_trad and Uncali_trad ROC areas at the 0.10 inch threshold). The new methods yielded approximately the same values at each threshold, from around 0.85 at the 0.01 inch threshold to near 0.90 at the 0.25 inch threshold (with the exception of Ave_thr). The increase in ROC areas shows that resolution increased as the thresholds increased.

All of the new methods showed an increase in ROC area with increased thresholds. GS04 and GBE07 also noted this trend, which also occurred in the GSD method (Table 3-3). Cali_trad and Uncali_trad are the only methods that have a decrease in ROC area 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 more traditional approaches.

b) Two-parameter neighborhood approach

A second forecasting approach was developed using neighborhood methods (e.g., Ebert 2009, Gilleland et al. 2009, Schwartz et al. 2009, among others). Within a specified square area around a center point (a 3x3 point area, 5x5, 7x7, etc.), the maximum or average precipitation amount was determined and placed in a bin. It is worth noting that specifying a 1x1 point "area" reduces the approach to binning precipitation at a single point, i.e. to the GSD approach.

This approach not only uses the binned precipitation amount but also the number of points within the neighborhood that have forecast precipitation amounts greater than a threshold. Max_thr and similar methods considered forecasts from 10 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 "ensemble" (e.g. Theis et al. 2005) yielding more than just 10 members and thus the potential for better results. The size of the neighborhood/square determines the number of points considered in the method, so different tables were created based on neighborhood size for each of the ten members. This approach is different than the method tested in Schwartz et al. (2009) which issued POP forecasts for entire neighborhood areas, while in the present study the POP forecast is for individual points. If a neighborhood intersected the domain's boundary, the agreement parameter was calculated as shown in Appendix A.

The 1x1 BSs for Max_nbh and Ave_nbh matched those of the GSD method (shown in Table 3-3), because a 1x1 “neighborhood” is a single point. As neighborhood size increased, the reliability decreased, but resolution increased to a larger extent. The best BSs generally occurred for a 15x15 point neighborhood for Ave_nbh (Table 3-5), after which the loss of reliability began to outweigh improvements in resolution. For Max_nbh, the best BSs occurred for a 13x13 neighborhood. The best BSs for Ave_nbh, however, were lower (better) than the best scores for Max_nbh suggesting that averaging of nearby points provides a more skillful forecast than selecting the maximum precipitation within the neighborhood.

The 15x15 Ave_nbh results (Table 3-5) showed that some BSs were greater than the Max_thr scores, while others were less. The lowest scores for Ave_nbh were below 0.1000, which was more skillful than the Max_thr and also Cali_trad values. This result is surprising, because Max_thr considered all 10 ensemble members when creating POPs, but Ave_nbh considered only an individual member. 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.

ROC areas for Ave_nbh again increased with increasing thresholds (Fig. 3-4). 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. Improvement over Cali_trad can also be seen in scatterplots of Ave_nbh 15x15 and Cali_trad’s BSs (Fig. 3-5) where each point is a BS comparison of the methods for a case and time. The majority of points are above the diagonal, indicating that the Ave_nbh forecasts had lower BSs and thus higher skill than Cali_trad.

c) Three-parameter Approach

A third forecasting approach was examined that combined approaches (a) and (b) discussed above. This three parameter, or 3P, approach used precipitation binning (either the maximum or average precipitation amount) and the 10-member forecasts like Max_thr and Ave_thr, but also the neighborhood approach used in Max_nbh and Ave_nbh. With three parameters, POPs could be generated in a variety of ways. The points considered in this method existed within a volume (i.e. 3-D matrix) composed of the parameters considered in (a) and (b) combined. The agreement parameter used in Max_thr and related methods considered the number of ensemble members with precipitation amounts greater than a threshold, and the agreement parameter used in Max_nbh and Ave_nbh considered the number of points within a neighborhood with precipitation amounts larger than a threshold. An example of a set of ensemble members can be seen in a shaded column in Fig. 3-6, with rows representing members 1 and 10 labeled M1 and M10, respectively. An example of a 3x3 neighborhood is shaded at the top of Fig. 3-6. The two agreement parameters could be redefined as one parameter which considered the number of points within the 3-D matrix with precipitation amounts greater than a threshold. The loss of a parameter, however, led to worse results, so the two agreement parameters needed to be preserved.

By finding the point with maximum precipitation (an example being the darkly shaded point in Fig. 3-6), the two parameters could be investigated like previously by focusing on this point (areas of investigation are shaded grey in Fig. 3-6). However, this technique could not be used while finding the average precipitation amount, because a single point would not be specified within the matrix. Also, this technique would only be considering the information from the entire volume in the binning parameter, because the agreement parameters are only considering certain points (shaded grey in Fig. 3-6) within the matrix. The opportunity to use information within the matrix would be limited by this technique.

In order to remedy these three problems, a data mapping approach was used to place the information gathered from the volume onto the previously used 1-D vector (for Max_thr and related methods) and 2-D neighborhood (for Ave_nbh and Max_nbh). This was done using the following equations:

$$F_1 = \frac{F_0 * 10}{V} \tag{7}$$

$$F_2 = \frac{F_0 * A}{V} \tag{8}$$

where F_0 is the number of forecasts within the volume with precipitation greater than a threshold, V is the number of points within the volume, and A is the number of points within the square neighborhood. F_1 represents the number of points (with precipitation greater than the threshold) from the vector used in Max_thr and related methods, and F_2 represents the number of points from the neighborhood used in Ave_nbh and Max_nbh.

The 3P approach, like the two-parameter neighborhood approach from (b), showed better skill when the average precipitation amount was determined in the 1D vector and 2D neighborhood, rather than the maximum amount, so only the averaging version was used. The 3P method's BSs were best for an 11x11 point neighborhood (Table 3-6), and were much better than the BSs for the approaches in (a) and (b), but still were not statistically significantly different from Cali_trad's scores with a 0.01 inch threshold p-value of 0.1091. As with the Ave_thr and Ave_nbh methods, the areas under the ROC curve for each threshold were higher than for Cali_trad and increased with increasing thresholds. The 0.01 inch threshold value was close to 0.87, and the 0.25 inch value was over 0.90.

d) Combination of methods

A final forecasting approach was examined that combined several of the previous methods. 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, we believed that a combination of the forecasts might result in a forecast superior to the individual methods. By averaging the POPs for Ave_nbh, Max_thr, and Cali_trad, and increasing the Ave_nbh neighborhood from 3x3 to 15x15 (Table 3-7), the BS improved from 0.0995 to 0.0959 for the 0.01 inch threshold. The 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 90% confidence level with a p-value of 0.07884. In addition to the improvement in BSs, the bias values also improved for each threshold.

When the neighborhood was increased from 3x3 to 15x15 grid points, the reliability worsened, but the resolution improved to a greater extent. This behavior was observed in Ave_nbh, as well. The reliability was lower (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 for larger neighborhoods. Thus, the combination of methods had low reliability (comparable to Max_thr), and high resolution (like Ave_nbh).

By including Ave_thr in this combination method, the skill increased marginally. When Max_bin and Ave_bin were added in, skill did not improve, likely because these methods had lower skill than Max_thr and Ave_thr.

By including the 3P approach as an additional member in the combination approach, the skill increased slightly. However, as noted previously, the skill of the 3P approach was very good, and sensitivity tests showed that enhancing its weight in the combination increased the skill further. By including versions of Ave_nbh, Cali_trad, and the 3P method where their precipitation fields were multiplied by a "reduction factor" in the combination method, skill was increased further. By multiplying the forecast precipitation amounts by scalars which were determined (through sensitivity tests) to improve BSs at the 0.01 inch threshold specifically, the reduction factor helped correct forecast overestimation. Effectively, the reduction is a readjustment of the original selection. The sensitivity tests involved first multiplying the characteristic precipitation amount used for the binning and agreement parameters by reduction factors such as 0.25, 0.50, and 0.75, and determining which factor yielded improved BSs. The factors were then tested in 0.05 increments, until an "ideal" factor was found. These ideal factors ranged from 0.15 to 0.30, depending on the method, and were meant only to improve the 0.01 inch threshold results. In most cases, the other thresholds experienced an increase in BSs as a result of the reduction factor.

A different version of the combination approach used the POPs from Max_thr, Ave_thr, the factorized Cali_trad, the factorized Ave_nbh (which consisted of ten forecasts), and the factorized 3P approach (given eight times as much weighting) using the equation:

$$Combination = Max\_thr + Ave\_thr + Cali\_trad + \sum_{i=1}^{10} Ave\_nbh_i + 8 * 3P \quad (9)$$

A 0.01 inch threshold BS of 0.0949 was obtained (Table 3-7) which is statistically significantly better at the 95% confidence level (p-value = 0.04831) than the Cali_trad results. Figure 3-7 contains boxplots for this comparison, based on the 100 BSs from all cases and times. By combining Max_thr, Ave_thr, the factorized Ave_nbh (which consisted of ten forecasts), and the factorized 3P method (weighted eight times), the Cali_trad forecast was no longer needed within the combination method because its impact on the BS was minimal (though Cali_trad was still indirectly included in Max_thr and Ave_thr). With Cali_trad excluded, the combination approach consisted of an average of 20 POPs generated from 3 unique approaches.

Figure 3-8 compares BSs for the different methods, and shows that some techniques outperform Cali_trad. The GSD and Ave_nbh scores are average values for the 10 members. It is again worth noting how close the 3P approach's BS was to the combination approach's BS. The 3P approach's results at the 0.01 inch threshold were not statistically significantly different compared to Cali_trad at the 90% confidence level, but the combination approach's results were significantly different from Cali_trad at the 95% confidence level, according to analysis of variance tests and t-tests.

Figure 3-8 also shows a reference forecast based on a simplified version of the method presented in Theis et al. (2005). Theis et al. (2005) considered an uncalibrated neighborhood approach in which the number of points in the neighborhood with precipitation above a threshold is divided by the total number of points in the neighborhood. Theis et al. (2005) also used a temporal neighborhood approach with 3 hour time periods. In the present study 6 hour time periods were used, and because convective systems change substantially over 6 hour periods, it was felt that the temporal neighborhood approach could not be used for the output available here. The simplified Theis results provided maximum skill for a 21x21 grid point neighborhood, and yielded a member-averaged Brier score of 0.1156. If a BS was computed while using a binary approach (using POPs of either 100% or 0%), then the Brier score would be 0.1930, much higher than the other BSs computed. This binary score would be computed through the use of a 1x1 neighborhood in the simplified Theis method.

e) Sensitivity of results to grid spacing

In order to evaluate the sensitivity of the methods to the grid spacing of the data set, the most promising of the 20 km 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 BSs for the methods improved with finer grid spacing (Fig. 3-9 compared to Fig. 3-8), though the differences in skill between methods were similar to what was observed with the 20 km results (Fig. 3-10). When applying the methods that use neighborhood approaches, the neighborhoods were scaled to fit with the 4 km grid spacing (i.e. A 5x5 point area in the 20 km results was now a 25x25 point area in the 4 km study). For this reason, there was an increase in computer resources and time required to verify the 4 km forecasts. The neighborhoods with the best skill (lowest BSs) at 20 km also had the best skill at 4 km, and the reduction factors chosen at 20 km were also effective at 4 km.

Finding improved BSs at 4 km grid spacing compared to 20 km grid spacing was unexpected because past studies have shown that standard measures of skill usually show deteriorating skill at fine grid spacings. 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. However, these studies didn't consider BSs, so it is unclear whether this statistic should follow the trends that ETS did. The changes to bias with increasing thresholds in the current study tended to agree with the Gallus (2002) BMJ control run bias comparisons. Bias was worse at finer grid spacings [4 km here, 10 km in Gallus (2002)] than at coarser grid spacing [20 km here, 30 km in Gallus (2002)], but as the threshold increased the trend was reversed. The ROC areas for Max_thr and Ave_thr remained in the 0.85-0.90 range for the 4km results, however Ave_thr had a decrease in ROC area from the 0.10 inch threshold to the 0.25 inch threshold which did not exist in the 20 km results. Finally, a comparison of BSs for the combination method and Cali_trad still showed statistically significant differences between the two methods as in the 20 km results, though at the 90% confidence level instead of the 95% confidence level.

4. Discussion and Conclusions

The present study is an extension of the POP approach used in GS04 and GBE07 to a 10-member WRF ensemble, while providing a comparison to a calibrated traditionally-used equal weighting approach used to determine 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. The POP was evaluated based on the performance at each domain grid point. Quantification of the skill of the new approaches emphasized the use of BSs and ROC areas.

Because the approaches are based on post-processing of simulated precipitation fields, tests were performed using both 20 km and 4 km grid representations of the precipitation field. Hamill and Colucci (1997) showed that calibration over observations can improve forecasts using a statistical technique, and the present paper found other techniques that improve forecasts.

For all methods, the most pronounced improvements in POP skill occurred for the lowest threshold, with skill diminishing 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. Comparison of POP maps against those generated by Cali_trad and Uncali_trad may provide additional guidance into relevant aspects of the forecasted precipitation.

By examining binned precipitation amounts and the number of ensemble members with precipitation greater than a threshold (the two-parameter point forecast approach), tabular POP forecasts Max_thr, Ave_thr, Max_bin, and Ave_bin were created. While most of these methods had lower BSs than Cali_trad (e.g. 0.1013 for Max_thr compared to 0.1040 for Cali_trad), the differences between these methods and Cali_trad were not statistically significantly different.

The two-parameter neighborhood approach (which is conceptually similar to that of Theis et al. (2005) when the calibration is not applied) provided skillful results that exceeded expectations. As discussed in Theis et al. (2005), the approach (consisting of the methods Max_nbh and Ave_nbh) is effectively an ensemble that is generated based on the spatial distribution of points in a neighborhood. Ensembles generated in this manner produced POPs as skillful as those from the 10-member ensemble forecast Cali_trad. This suggests that the approach is very attractive operationally, and we are currently testing options to refine it in order to improve its performance. Because post-processing of a single deterministic simulation can provide skill comparable to that obtained by Cali_trad, computer resources used for the ensemble simulation might be better used for further refinement of the model grid spacing or for improved model physical formulation. The ensemble information can be obtained from POPs using post-processing to generate spatially based ensembles. Still, it is possible that when using an ensemble with more than 10 members (as was used in the present study) or when using an ensemble with different design characteristics, the POP of a single member may not yield forecasts as good as that from Cali_trad.

The neighborhood approach can be thought of in more than one way. For instance, within a neighborhood with squares of (NxN) points, a set of grid points with the same relative orientation to the grid points (I,J) can be considered an ensemble of (NxN) members. Alternatively, the neighborhood may be viewed as shifting the grid (NxN) times relative to the observed point (I,J) by one grid point. Hence, for an example using N=3, the ensemble effectively represents the original grid (no displacement) and 8 displacements of the simulated domain by one grid point northward, westward, eastward, southward, southeast, southwest, northeast and northwest.

A limited comparison of skill between 20 km and 4 km gridded precipitation forecasts indicated better POP skill for the 4 km setting. While this pattern needs to be confirmed in future studies, it may suggest an additional consideration in evaluating the merit of fine grid resolution single runs versus coarse grid ensembles. Questions remain about the best usage of computer resources for predicting convective QPF. For instance, is it better to run a single deterministic refined grid simulation or a coarser grid ensemble? Clark et al.

(2009) found that finer grid spacings tended to provide more accurate forecasts than forecasts on coarser grid spacings. Faster error growth at finer grid spacings led to more reliable forecasts. In the present study, the finer grid spacing forecasts had better Brier scores than those of the coarser grid spacing.

A three-parameter approach considered binned precipitation amounts and a representation of the number of ensemble members (in the 10-member ensemble and the neighborhood ensemble) with precipitation greater than a threshold in order to produce 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 and their appropriate reduction factors, the resulting combination approach produced forecasts that were statistically significantly better (p-value = 0.04831) compared to Cali_trad's forecasts at the 95% confidence level.

Overall, the approaches introduced in this study suggest that three important techniques can be used to create useful POP forecasts. Two of the techniques are represented by the two general parameters used within the approaches: binning a characteristic QPF amount and determining the member agreement. The QPF binning technique was used in all methods, as well as in GSD. In this study, the benefits of the QPF binning technique were especially apparent when considering the ROC areas (for the approaches introduced and GSD) because the areas increased further than the areas for Cali_trad and Uncali_trad. The member agreement technique is used in Cali_trad and Uncali_trad, in addition to the approaches introduced in this study, since it is well-established as an important POP-forecasting technique. The third technique which can create useful POP forecasts is the neighborhood technique, which we showed can be used to derive probabilistic information from deterministic forecasts in order to produce POP forecasts that can potentially rival calibrated ensemble POP forecasts.

5. Acknowledgements

This research was funded in part by National Science Foundation grants ATM-0537043 and ATM-0848200. The authors would like to thank Adam Clark for providing the 20 km output and for helpful discussions, Jon Hobbs for his advice regarding the statistical analysis, and Daryl Herzmann for his computational assistance.

APPENDIX A

Computation of POPs in neighborhoods truncated by domain boundaries

When a neighborhood extended outside of the domain used in this study, the agreement parameter(s) from the neighborhood was/were extrapolated using the general equation:

$$F_1 = \frac{F_0 * A_t}{A} \qquad \text{(A1)}$$

where F_0 is the number of forecasts within the neighborhood with precipitation greater than the threshold, A_t is the number of points that should exist in the neighborhood, and A is the number of points that actually exist within a given neighborhood. This equation is similar to Eq. 7 and Eq. 8; Eq. 7 and 8 apply data from a larger area onto a smaller area, while Eq. 9 applies data from a smaller area onto a larger area. If the neighborhood is entirely within the domain, then A_t=A, so F_0=F_1. If the neighborhood is partially outside of the domain, then A<At, and Eq. 9 will approximate what agreement the neighborhood would likely have had if

an entire neighborhood were considered. With the agreement parameter determined, the POP would then be calculated according to the approach's specifications.

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.

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 convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121-1140.

Ebert, E. E., 2001: Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

——, 2009: Neighborhood verification: a strategy for rewarding close forecasts. *Wea. Forecasting,* **24**, 1498–1510.

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.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. *Wea. Forecasting*, **24**, 1416–1430.

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.

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

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.

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

Schwartz, B. E., and S. G. Benjamin, 2000: Verification of RUC2 precipitation forecasts using the NCEP multisensory analysis. Preprints. *Fourth Symp. On Integrated Observing Systems*, Long Beach, CA. Amer. Meteor. Soc., 182-185.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, M. C. Coniglio, and M. S. Wandishin, 2009: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, (In Press)

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032.

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

Tustison B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.*, **106**, 11775–11784.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd edition. Academic Press, 627 pp.

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.

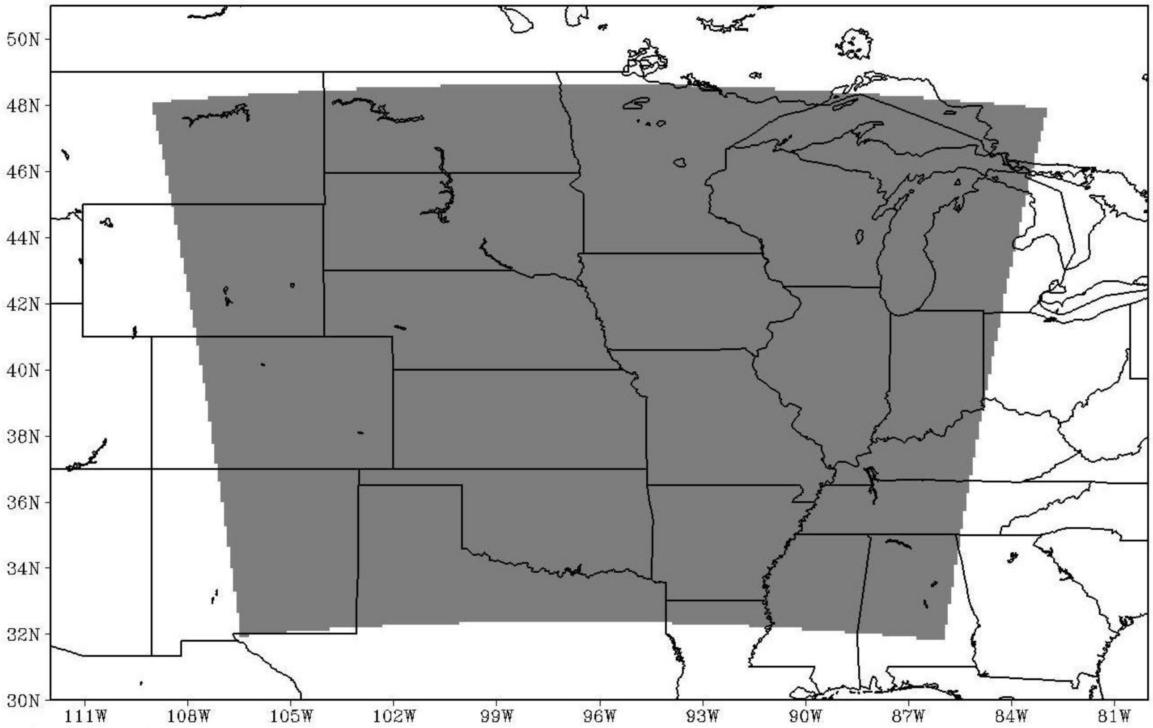

Figure 3-1. The sub-domain over which forecasts were tested.

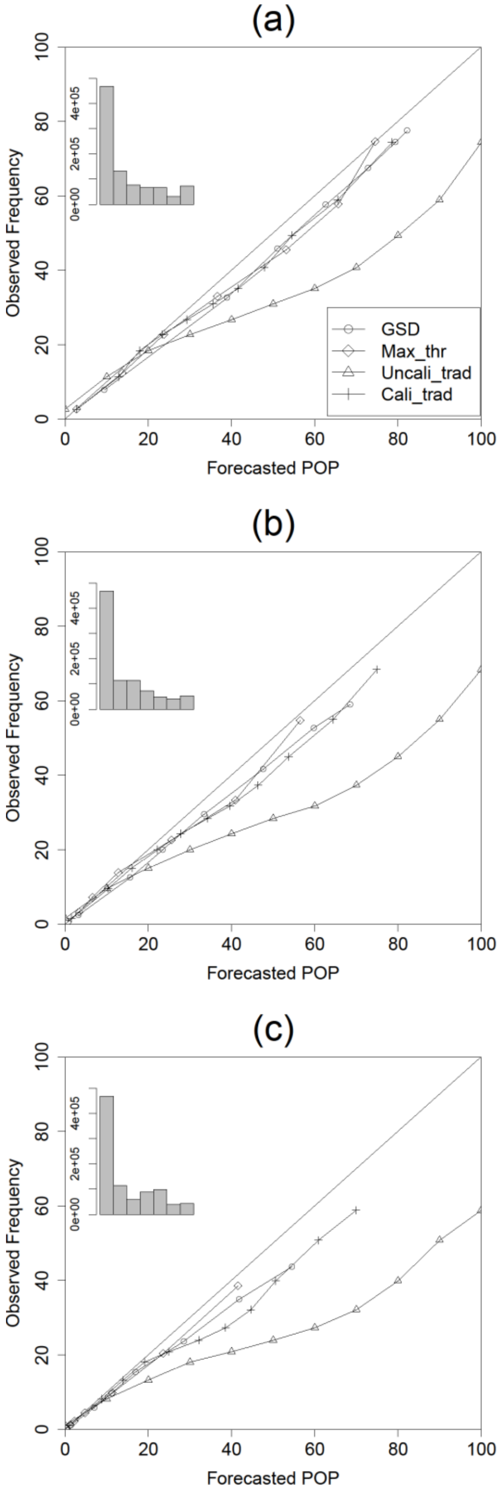

Figure 3-2. Reliability diagrams for GSD, Max_thr, Uncali_trad, and Cali_trad at thresholds a) 0.01 inch, b) 0.10 inch, and c) 0.25 inch. Histogram displays the distribution of Max_thr forecasts within the seven bins.

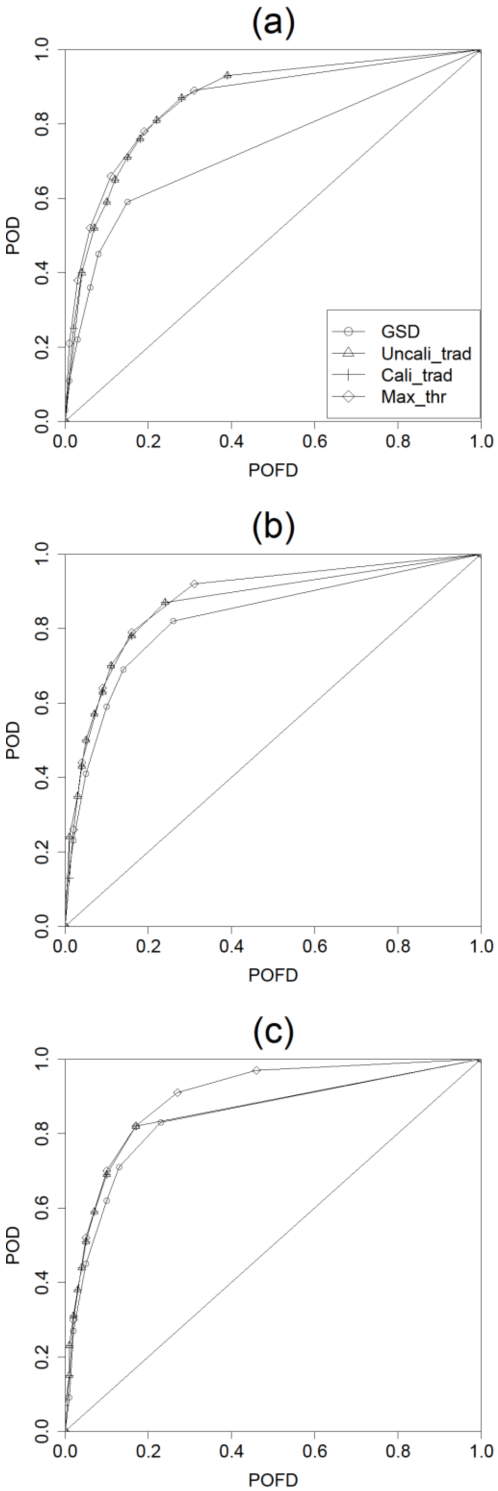

Figure 3-3. ROC diagrams for GSD, Max_thr, Uncali_trad, and Cali_trad at thresholds a) 0.01 inch, b) 0.10 inch, and c) 0.25 inch.

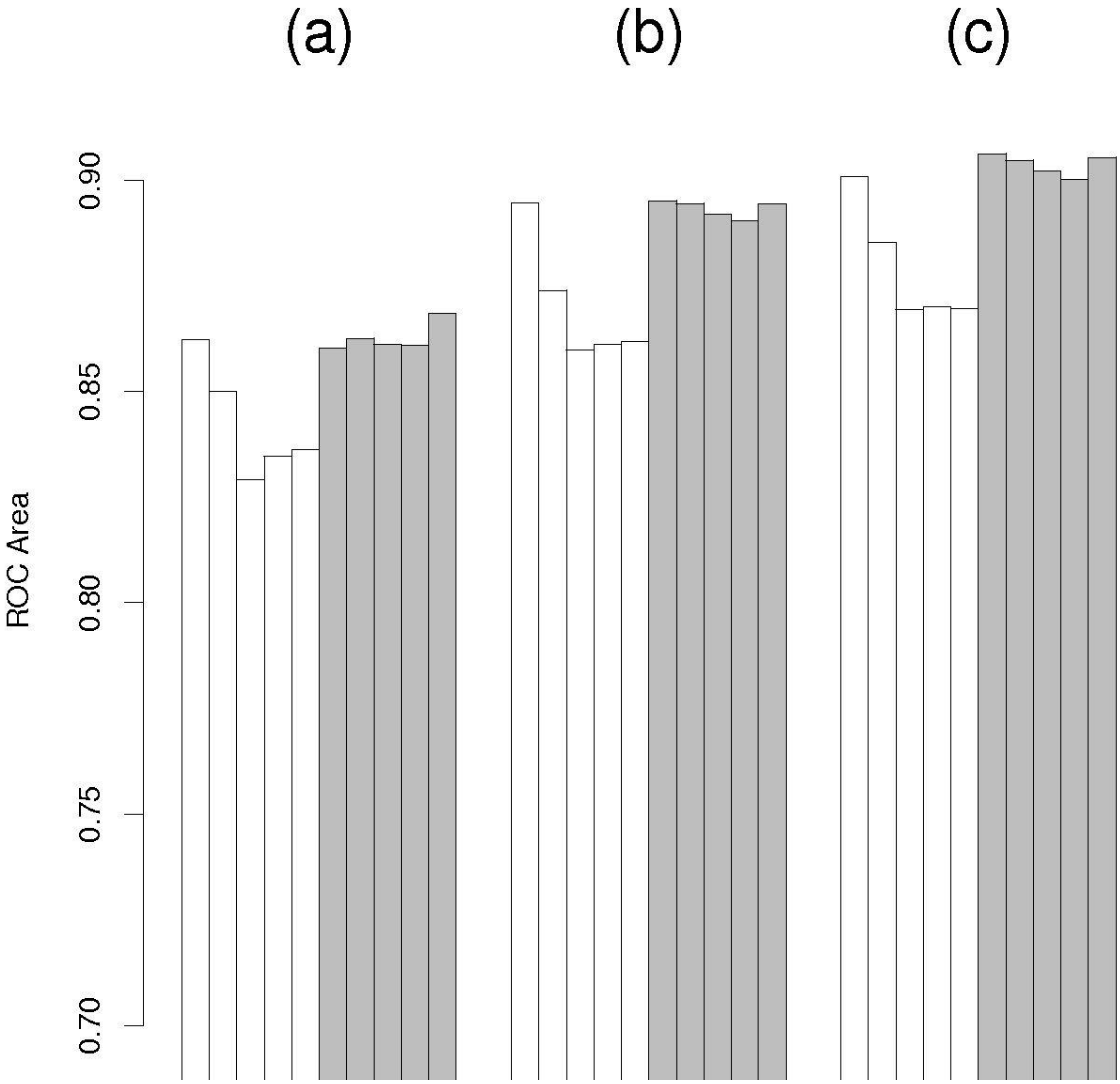

Figure 3-4. ROC areas for the 10 members of Ave_nbh (with a 15x15 neighborhood) at thresholds a) 0.01 inch, b) 0.10 inch, and c) 0.25 inch. The bar plots were truncated near 0.70 to emphasis differences. The first five members (including the control member) used mixed physics and perturbed initial conditions, and the last five members (the shaded bars) used only mixed physics.

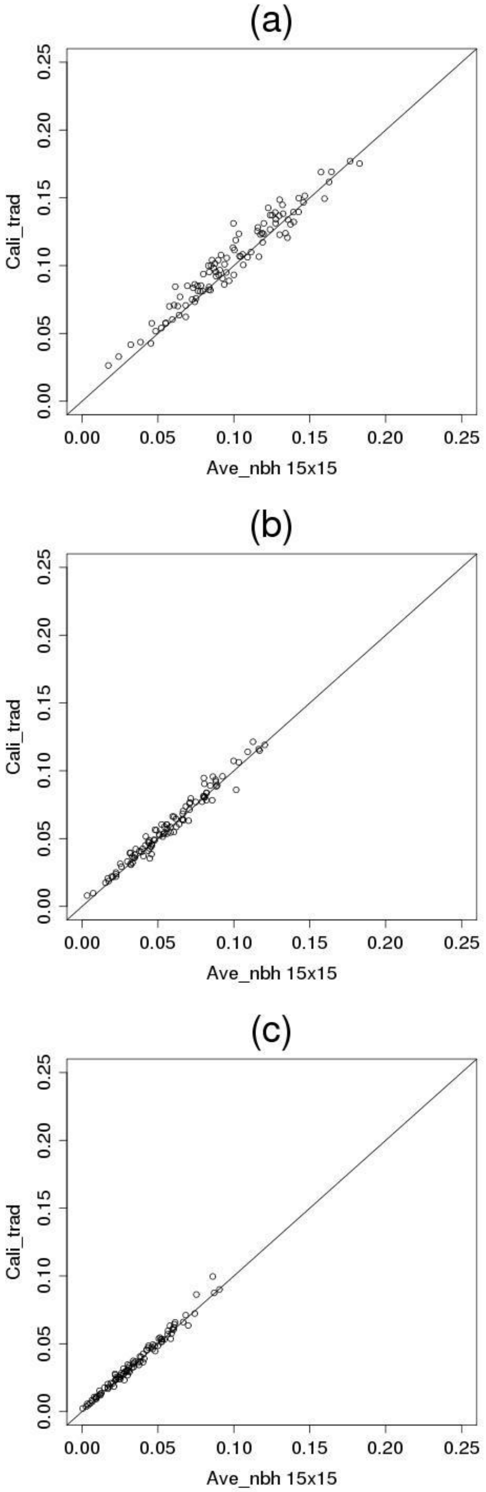

Figure 3-5. Scatterplots of BSs for Ave_nbh (with a 15x15 neighborhood) and Cali_trad at thresholds a) 0.01 inch, b) 0.10 inch, and c) 0.25 inch.

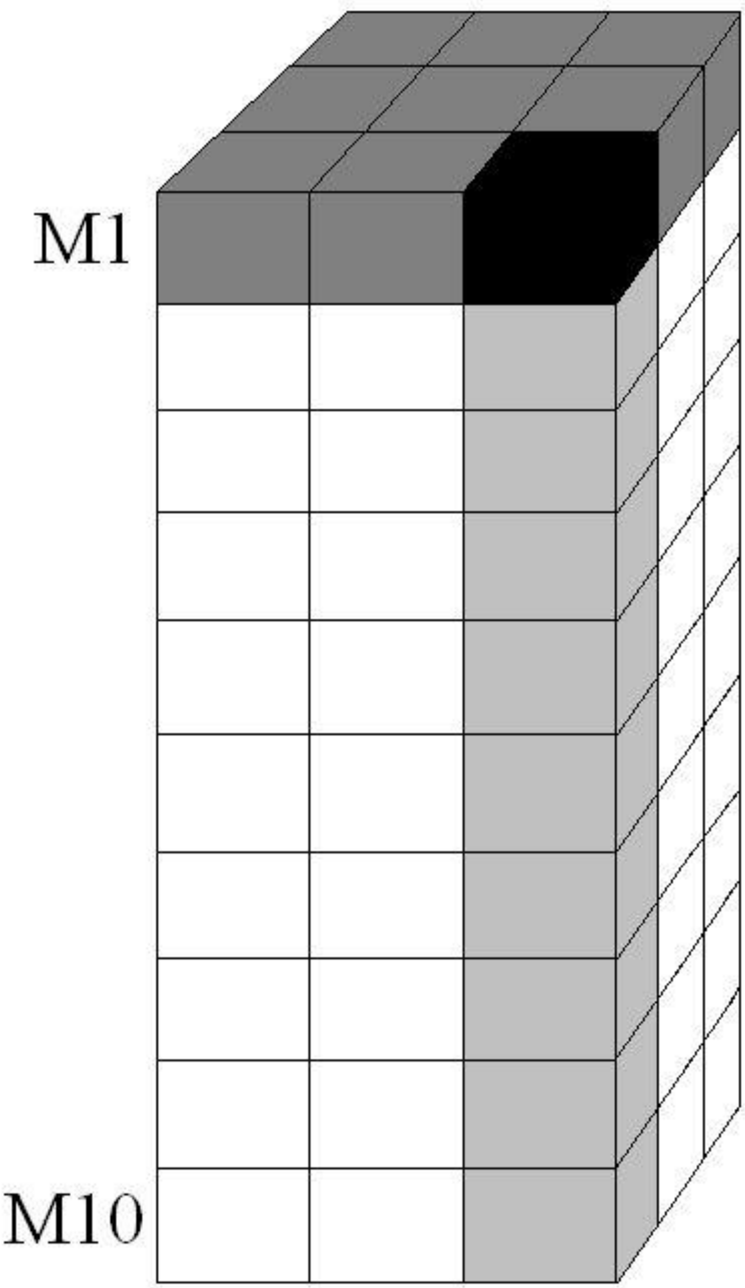

Figure 3-6. An example of a matrix of points that could be considered using a three-parameter approach to calculating POPs.

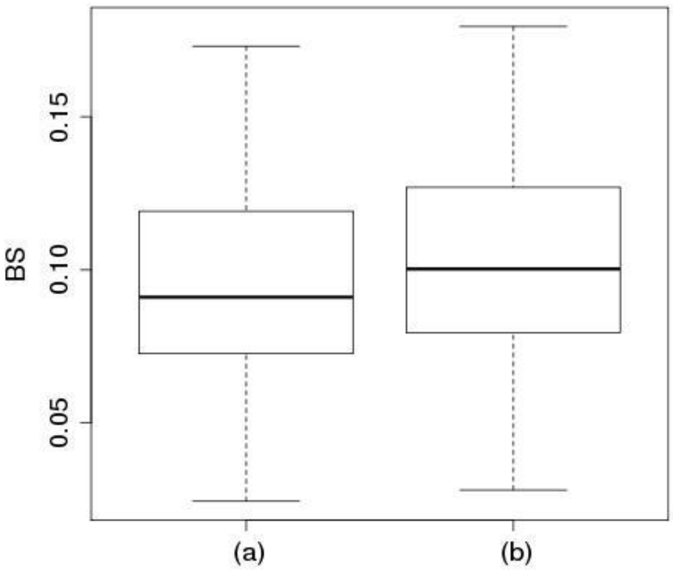

Figure 3-7. Box plots for the (a) best combination method compared to the box plot for (b) Cali_trad, created from the 100 BSs from all cases and times at the 0.01 inch threshold.

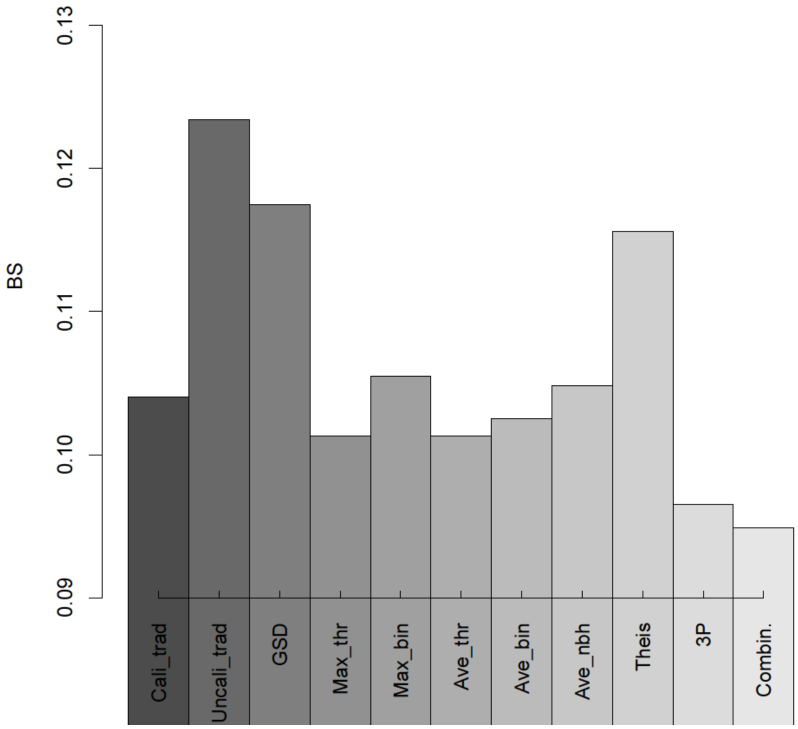

Figure 3-8. BSs for the 0.01 inch threshold for different methods at 20 km grid spacing. The BS of the combination approach shown here used Max_thr, Ave_thr, Cali_trad with a reduction factor of 0.30, 11x11 3P (included eight times) with a reduction factor of 0.20, and 15x15 Ave_nbh with reduction factors of 0.25 (on the binning parameter) and 0.15 (on the agreement parameter).

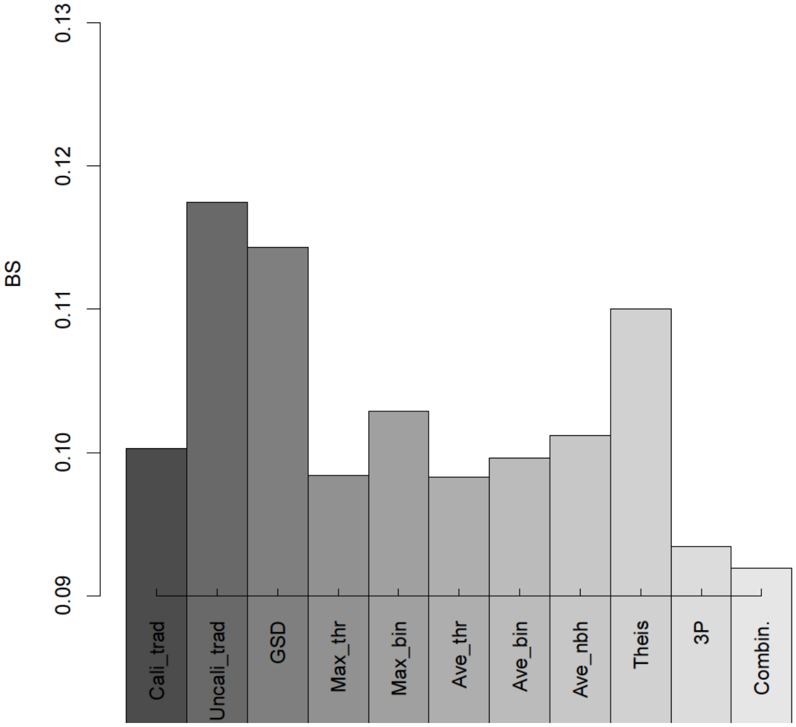

Figure 3-9. BS for the 0.01 inch threshold for different methods at 4 km grid spacing. The BS of the combination approach shown here used Max_thr, Ave_thr, Cali_trad with a reduction factor of 0.30, 55x55 3P (included eight times) with a reduction factor of 0.20, and 75x75 Ave_nbh with reduction factors of 0.25 (on the binning parameter) and 0.15 (on the agreement parameter).

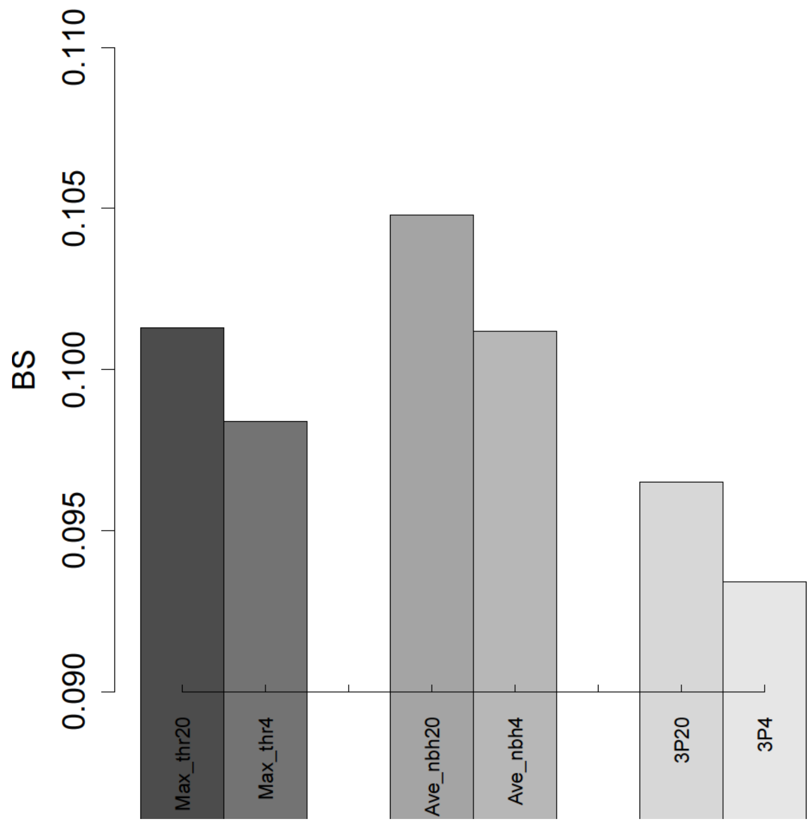

Figure 3-10. Comparison of BSs for selected methods at 20 km and 4 km grid spacing at the 0.01 inch threshold.

Table 3-1. Brief description of the various methods used in the study and their classification within the adopted approaches.

Approach/method	Description
a) Reference approaches	
	GSD - Single model run; one parameter point approach; as in GS04
	Uncali_trad - Ensemble; percentage of members forecasting precipitation ge. a specified threshold amount; uncalibrated
	Cali_trad - Ensemble; like Uncali_trad but calibrated using observed data
	Simplified Theis - Ensemble; considers number of neighborhood members forecasting precipitation amounts ge. a threshold, and divides this number by the total neighborhood members; uncalibrated
	Binary – Single model run; assigns a POP of 100% if the precipitation amount is ge. a threshold, otherwise the POP is 0%
b) Two parameter point forecast approach	
	Max_bin - Max of Ensemble; binned max # of ensemble members agreeing for various bins
	Max_thr - Like Max bin except using # of ensemble members forecasting precipitation amounts ge. a threshold
	Ave_bin - Ensemble; places the average of 10 members at a point into a bin, and considers number of members forecasting precipitation amounts in that same bin
	Ave_thr - Ensemble; like Ave_bin except considers number of members forecasting precipitation amounts ge. a threshold
c) Two-parameter neighborhood approach	
	Ave_nbh - Ensemble; places the average of a neighborhood into a bin, and considers number of members forecasting precipitation amounts ge. a threshold
	Max_nbh - Ensemble; like Ave_nbh, but finds the maximum precipitation amount instead of the average amount
d) Three dimensional approach	
	Ensemble; considers the three parameters used by Ave_thr and Ave_nbh within a volume
e) Combination approach	
	Ensemble of ensembles; considers the average of the resulting POPs from the other methods

Table 3-2. POP table (in %) for the 0.01 inch threshold in the Max_thr method with corresponding number of grid points in parentheses. Top row designates the accumulated precipitation bin, and the side column shows the percentage of ensemble members that forecasted precipitation greater than the 0.01 inch threshold.

Ensemble Agreement	**Bin ranges**							
%	**<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)

Table 3-3. Decomposed BSs, BSSs, bias scores, and ROC areas for the four new methods of the two-parameter point forecast approaches, the GSD ten-member and best five-member average, Uncali_trad, and Cali_trad. Statistically significantly different BSs for the new methods compared to the GSD ten-member method (Uncali_trad method) are shown in italics (bold).

		Score						
Method		BS	Reli	Resol	Uncert	BSS	Bias	ROC
GSD_ave10	0.01 inch	0.1175	0.0073	0.0354	0.1456	0.1932	1.3488	0.763
	0.10 inch	0.0653	0.0046	0.0161	0.0767	0.1489	1.6043	0.800
	0.25 inch	0.0386	0.0029	0.0072	0.0429	0.1006	1.9621	0.818
GSD_ave5	0.01 inch	0.1133	0.0063	0.0386	0.1456	0.2219	1.3234	0.777
	0.10 inch	0.0632	0.0041	0.0176	0.0767	0.1758	1.5696	0.816
	0.25 inch	0.0377	0.0026	0.0078	0.0429	0.1205	1.9259	0.834
Trad	0.01 inch	0.1234	0.0257	0.0480	0.1456	0.1530	1.4707	0.861
	0.10 inch	0.0705	0.0152	0.0214	0.0767	0.0810	1.6305	0.865
	0.25 inch	0.0440	0.0105	0.0095	0.0429	-0.0243	1.9159	0.854
Cali_trad	0.01 inch	0.1040	0.0064	0.0480	0.1456	0.2855	1.2609	0.862
	0.10 inch	0.0593	0.0040	0.0214	0.0767	0.2267	1.4582	0.866
	0.25 inch	0.0363	0.0028	0.0095	0.0429	0.1547	1.7572	0.854
Max_thr	0.01 inch	***0.1013***	0.0097	0.0540	0.1456	0.3041	1.2501	0.857
	0.10 inch	**0.0586**	0.0059	0.0240	0.0767	0.2357	1.4192	0.877
	0.25 inch	**0.0359**	0.0037	0.0108	0.0429	0.1633	1.6722	0.897
Ave_thr	0.01 inch	***0.1013***	0.0095	0.0538	0.1456	0.3041	1.2501	0.862
	0.10 inch	**0.0587**	0.0058	0.0238	0.0767	0.2345	1.4405	0.865
	0.25 inch	**0.0358**	0.0037	0.0108	0.0429	0.1655	1.6972	0.869
Max_bin	0.01 inch	**0.1055**	0.0087	0.0488	0.1456	0.2752	1.2622	0.851
	0.10 inch	**0.0607**	0.0058	0.0218	0.0767	0.2092	1.4312	0.880
	0.25 inch	**0.0364**	0.0036	0.0101	0.0429	0.1518	1.6718	0.897
Ave_bin	0.01 inch	**0.1025**	0.0088	0.0519	0.1456	0.2958	1.2572	0.861
	0.10 inch	**0.0592**	0.0059	0.0234	0.0767	0.2279	1.4284	0.884
	0.25 inch	**0.0359**	0.0039	0.0108	0.0429	0.1625	1.6795	0.896

Table 3-4. P-values for comparisons of the most skilled GSD member (#10) and Uncali_trad with Max_thr.

	0.01 inch	0.10 inch	0.25 inch
GSD member #10	0.02807	0.2573	0.5435
Uncali_trad	3.369e-05	0.003564	0.01258

Table 3-5. Decomposed BSs, BSSs, bias scores, and ROC areas for Ave_nbh 15x15 at thresholds 0.01 inch, 0.10 inch, and 0.25 inch.

Member	Score						
	BS	Reli	Resol	Uncert	BSS	Bias	ROC area
0.01 inch							
Mem1	0.1043	0.0279	0.0691	0.1456	0.2836	1.4890	0.862
Mem2	0.1113	0.0299	0.0642	0.1456	0.2354	1.4252	0.850
Mem3	0.1091	0.0261	0.0626	0.1456	0.2507	1.0603	0.829
Mem4	0.1102	0.0271	0.0625	0.1456	0.2430	1.1854	0.835
Mem5	0.1109	0.0280	0.0628	0.1456	0.2385	1.2827	0.836
Mem6	0.0990	0.0232	0.0699	0.1456	0.3203	1.1532	0.860
Mem7	0.1037	0.0270	0.0690	0.1456	0.2881	1.4137	0.863
Mem8	0.0988	0.0235	0.0703	0.1456	0.3218	1.1730	0.861
Mem9	0.0996	0.0239	0.0699	0.1456	0.3163	0.9587	0.861
Mem10	0.1007	0.0252	0.0701	0.1456	0.3085	1.3627	0.869
0.10 inch							
Mem1	0.0588	0.0140	0.0319	0.0767	0.2332	1.6641	0.895
Mem2	0.0644	0.0162	0.0286	0.0767	0.1611	1.5686	0.874
Mem3	0.0620	0.0136	0.0283	0.0767	0.1918	1.2104	0.860
Mem4	0.0629	0.0142	0.0279	0.0767	0.1798	1.3434	0.861
Mem5	0.0636	0.0149	0.0280	0.0767	0.1705	1.3901	0.862
Mem6	0.0572	0.0127	0.0322	0.0767	0.2543	1.3575	0.895
Mem7	0.0599	0.0149	0.0317	0.0767	0.2189	1.6494	0.895
Mem8	0.0576	0.0126	0.0318	0.0767	0.2497	1.3303	0.892
Mem9	0.0573	0.0125	0.0320	0.0767	0.2535	1.1045	0.890
Mem10	0.0577	0.0131	0.0321	0.0767	0.2482	1.4553	0.894
0.25 inch							
Mem1	0.0356	0.0081	0.0154	0.0429	0.1717	1.9218	0.901
Mem2	0.0386	0.0098	0.0141	0.0429	0.1003	1.7731	0.885
Mem3	0.0367	0.0076	0.0138	0.0429	0.1446	1.3544	0.869
Mem4	0.0378	0.0082	0.0133	0.0429	0.1183	1.5411	0.870
Mem5	0.0381	0.0083	0.0132	0.0429	0.1134	1.6317	0.870
Mem6	0.0351	0.0076	0.0154	0.0429	0.1822	1.5685	0.906
Mem7	0.0367	0.0091	0.0153	0.0429	0.1442	1.9184	0.905
Mem8	0.0351	0.0075	0.0153	0.0429	0.1814	1.6095	0.902
Mem9	0.0350	0.0075	0.0155	0.0429	0.1847	1.3440	0.900
Mem10	0.0355	0.0080	0.0155	0.0429	0.1740	1.7059	0.905

Table 3-6. Decomposed BSs, BSSs, bias scores, and ROC areas for the 3P method using 11x11 grid points.

	Score						
Threshold	BS	Reli	Resol	Uncert	BSS	Bias	ROC
0.01 inch	0.0965	0.0183	0.0674	0.1456	0.3371	1.2123	0.865
0.10 inch	0.0561	0.0099	0.0305	0.0767	0.2685	1.3399	0.887
0.25 inch	0.0346	0.0061	0.0144	0.0429	0.1935	1.5215	0.901

Table 3-7. Decomposed BSs, BSSs, bias scores, and ROC areas for the combination method using a) Max_thr, Cali_trad, and Ave_nbh 3x3, b) Max_thr, Cali_trad, and Ave_nbh 15x15, and c) Max_thr, Ave_thr, factorized Cali_trad, factorized Ave_nbh 15x15, and the factorized 3P method (included 8 times).

	Score						
Threshold	BS	Reli	Resol	Uncert	BSS	Bias	ROC area
a)							
0.01 inch	0.0995	0.0097	0.0559	0.1456	0.3170	1.3098	0.818
0.10 inch	0.0573	0.0061	0.0255	0.0767	0.2529	1.5435	0.872
0.25 inch	0.0349	0.0039	0.0119	0.0429	0.1870	1.8700	0.894
b)							
0.01 inch	0.0959	0.0104	0.0601	0.1456	0.3411	1.2512	0.875
0.10 inch	0.0556	0.0066	0.0278	0.0767	0.2759	1.4126	0.903
0.25 inch	0.0340	0.0042	0.0131	0.0429	0.2083	1.6498	0.916
c)							
0.01 inch	0.0949	0.0098	0.0606	0.1456	0.3485	1.2473	0.880
0.10 inch	0.0562	0.0070	0.0275	0.0767	0.2678	1.4285	0.879
0.25 inch	0.0341	0.0044	0.0133	0.0429	0.2062	1.6328	0.900

CHAPTER 4. ADDITIONAL RESULTS

POPs over the Domain

The POP fields from various methods can also be viewed over the domain, providing an indication of how the POPs are distributed in comparison to the areas that received precipitation. These images can also show how the ranges of POPs can differ between methods, which was not otherwise apparent without viewing the POP tables themselves. The POP fields are displayed for individual cases and times, so they are not meant to be a measure of any method's overall accuracy. Maps such as these provide supplementary insight into the POP analysis.

Figures 4-1, 4-2, and 4-3 show the domain plots for Max_thr, Cali_trad, and the difference in POPs between Max_thr and Cali_trad, respectively, on the 20 km grid. Comparisons of Figures 4-1 and 4-2 show that Cali_trad does not have any POPs greater than 80%, unlike Max_thr. Cali_trad's maximum POP (which is forecasted when all 10 members are greater than the 0.01 inch threshold) is less than 80%, while Max_thr has the option to choose higher POPs. Figure 4-3 shows that Cali_trad is doing poorly compared the Max_thr in the area around Nebraska, because it is forecasting a lower POP than Max_thr over a large area with precipitation (denoted by the contour). Cali_trad is forecasting higher POPs than Max_thr in areas in the north and south, however, even though those areas did not receive precipitation. Clearly, Cali_trad performed very poorly on this day. For most cases, it was difficult to visually identify which method was performing better, which is why we use statistics such as Brier scores to provide a quantitative measure of forecasts skill.

The combination method (Figure 4-4) required averaging the POP fields, so the POP fields were typically more broad and smooth compared to the Cali_trad POP fields. The smoothing allowed the combination method to predict precipitation around the edges of observed precipitation areas, though this also caused the non-zero values to sometimes spread beyond the observed areas. The averaging associated with the combination method tended to lower the method's POPs, so in areas where the ensemble members indicated precipitation was very likely, the POPs were usually higher for Cali_trad. In these areas, Cali_trad did better than the combination method when these higher POPs occurred within the observed areas, but Cali_trad did poorer than the combination method when the areas did not receive precipitation.

On the 4 km grid, Max_thr (Figure 4-5) and Cali_trad (Figure 4-6) appear slightly different than they did on the 20 km grid (Figure 4-1 and Figure 4-2, respectively) due to the finer grid spacing. The impact of this finer grid spacing was described quantitatively in Chapter 3. By averaging the precipitation data onto a 20 km grid, shortwaves due to wave lengths less than 5Δx were removed (Tustison 2001, Skamarock 2004). In the 4 km data, however, these shortwaves remain and cause noise in the precipitation fields, which causes noise in the POP fields as well. Considering the improved Brier scores for POPs at 4 km grid spacing compared to 20 km grid spacing, it is possible that problems due to noise were small compared to other benefits of finer grid spacings, such as an improved diurnal cycle of precipitation.

Temporal Variation of Brier Scores

The Brier scores discussed in Chapter 3 were time-averaged, though time-dependant scores were considered for selected methods. The change in Brier scores over the five time periods for a 5x5 grid point neighborhood for Max_nbh is shown in Figure 4-7. The Brier scores at all thresholds are lowest (best) from 06Z to 18Z, and worst from 18Z to 06Z (the final 12 hours), which suggests that there may be a diurnal effect on the scores. A diurnal oscillation in equitable threat scores was found by Clark et al. (2007), which used 5 km grid spacing WRF output on a similar domain from April through July of 2005.

The lowest Brier score for the 0.01 inch threshold occurs at the 06-12Z time period, and the 0.10 inch and 0.25 inch thresholds have their lowest Brier score during the 12Z-18Z time period. The differences in Brier scores between these two 6 hour time periods, however, are small compared to the differences between other time periods. This trend in Brier scores over time at each threshold was also observed for Max_thr. Clark et al. (2007) suggested that mesoscale convective systems (MCSs), which typically propagate through the central plains during the morning hours, are more predictable than convection later in the day and led to diurnal changes in ETSs. This diurnal predictability/unpredictability may also be the cause of the Brier score trend, with better skill associated with MCSs in the morning hours and worse skill associated with less-predictable convection in the afternoon and early evening hours.

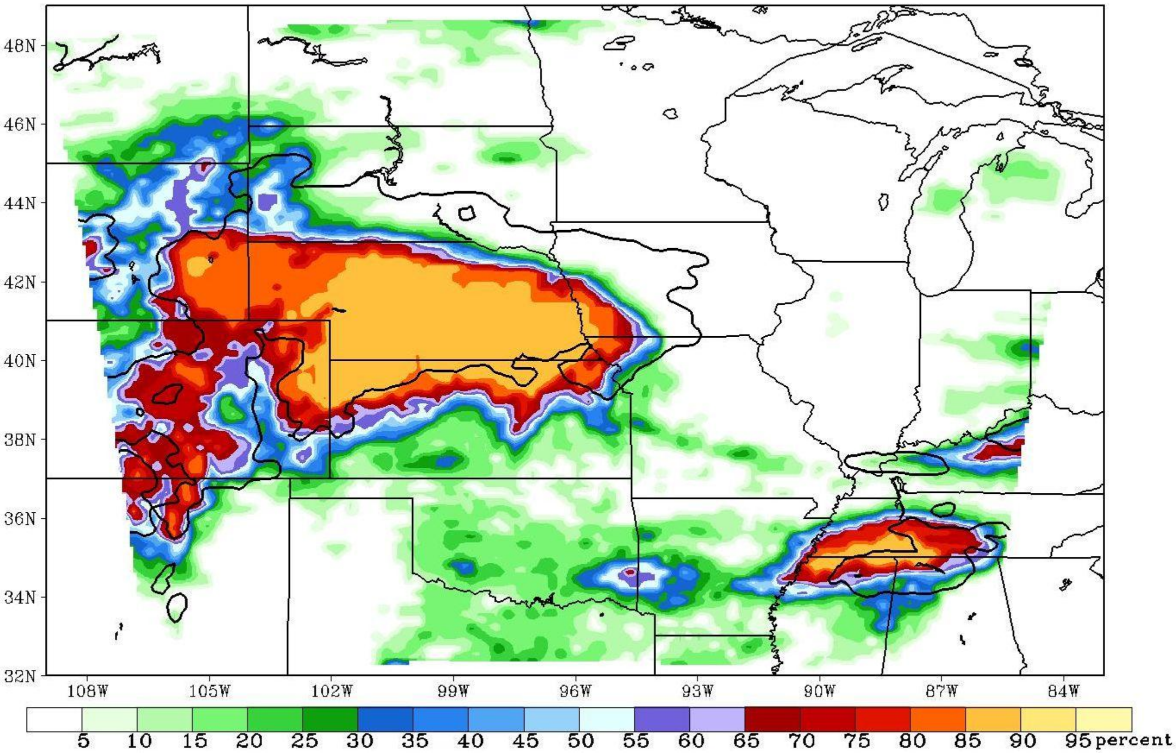

Figure 4-1. POPs (on the 20 km grid) for Cali_trad over the domain for 2007 April 23 for the period 06-12Z. The dark contour denotes observations at the 0.01 inch threshold.

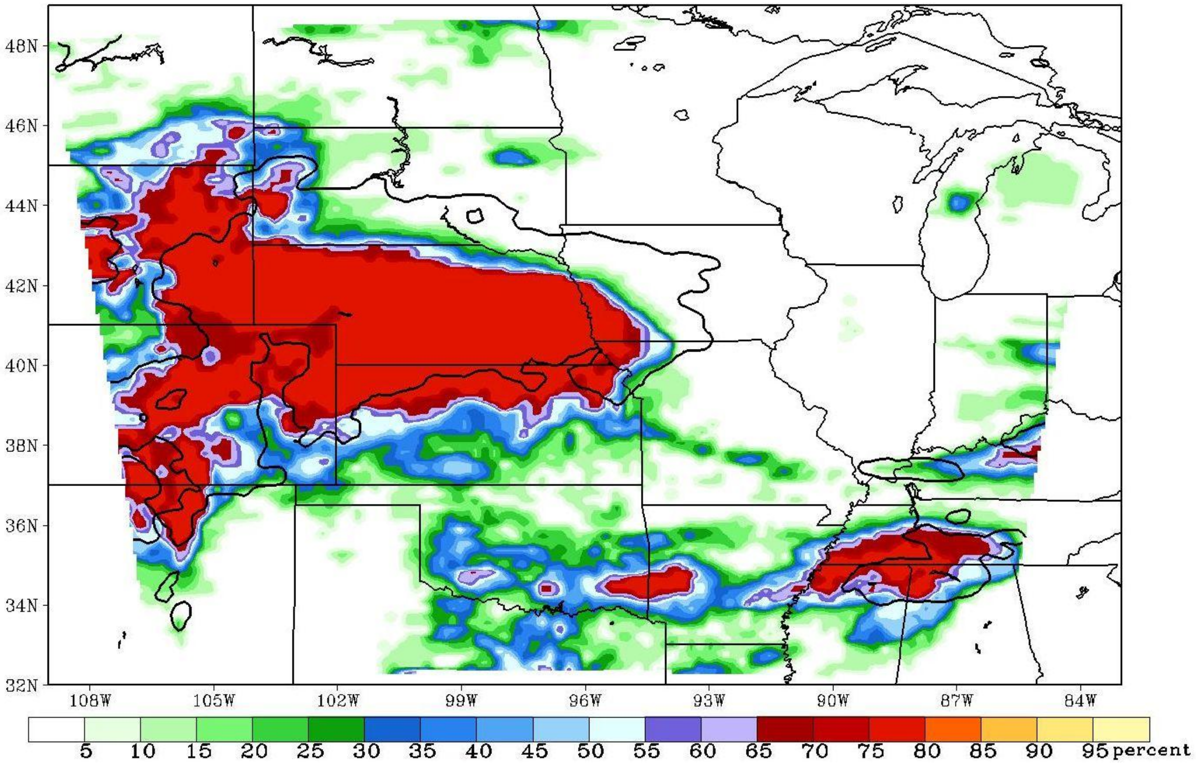

Figure 4-2. POPs (on the 20 km grid) for Cali_trad over the domain for 2007 April 23 for the period 06-12Z. The dark contour denotes observations at the 0.01 inch threshold.

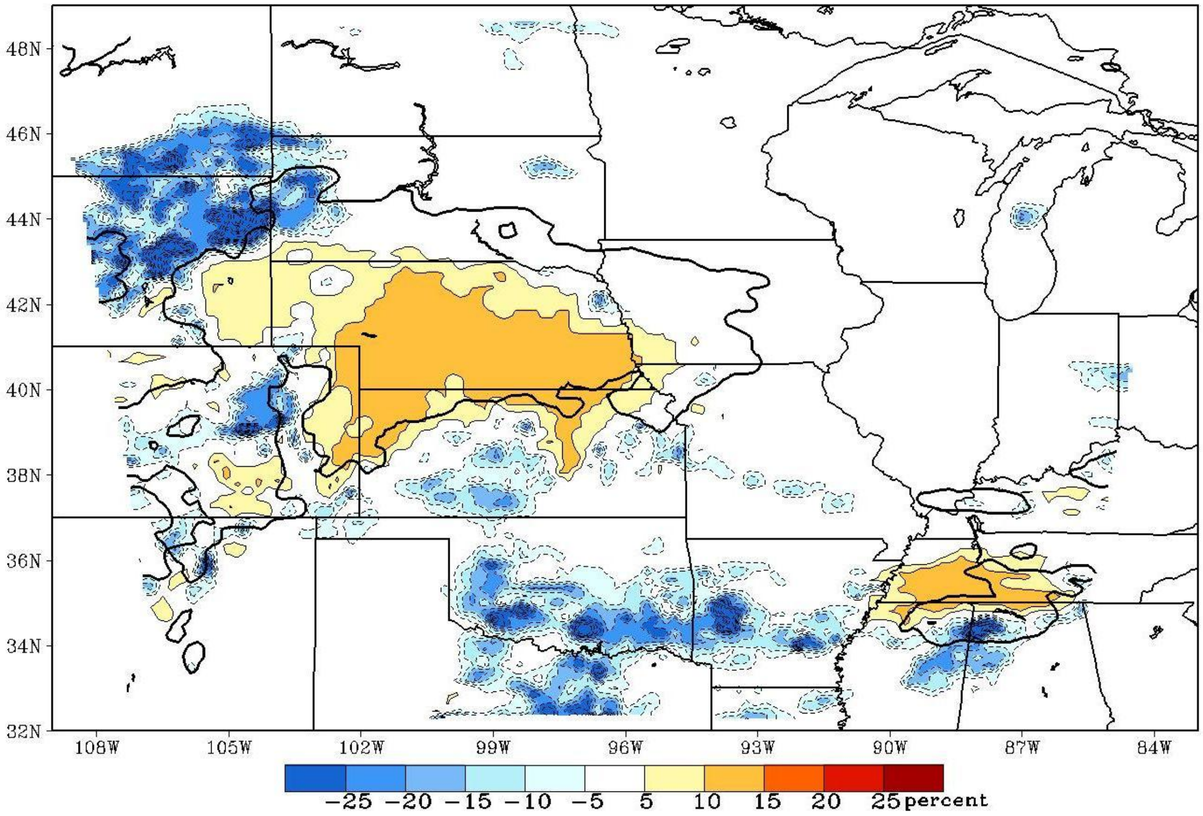

Figure 4-3. POP differences (Max_thr – Cali_trad) on the 20 km grid over the domain for 2007 April 23 for the period 06-12Z. The dark contour denotes observations at the 0.01 inch threshold, and the dotted contours denote where the POP differences were negative.

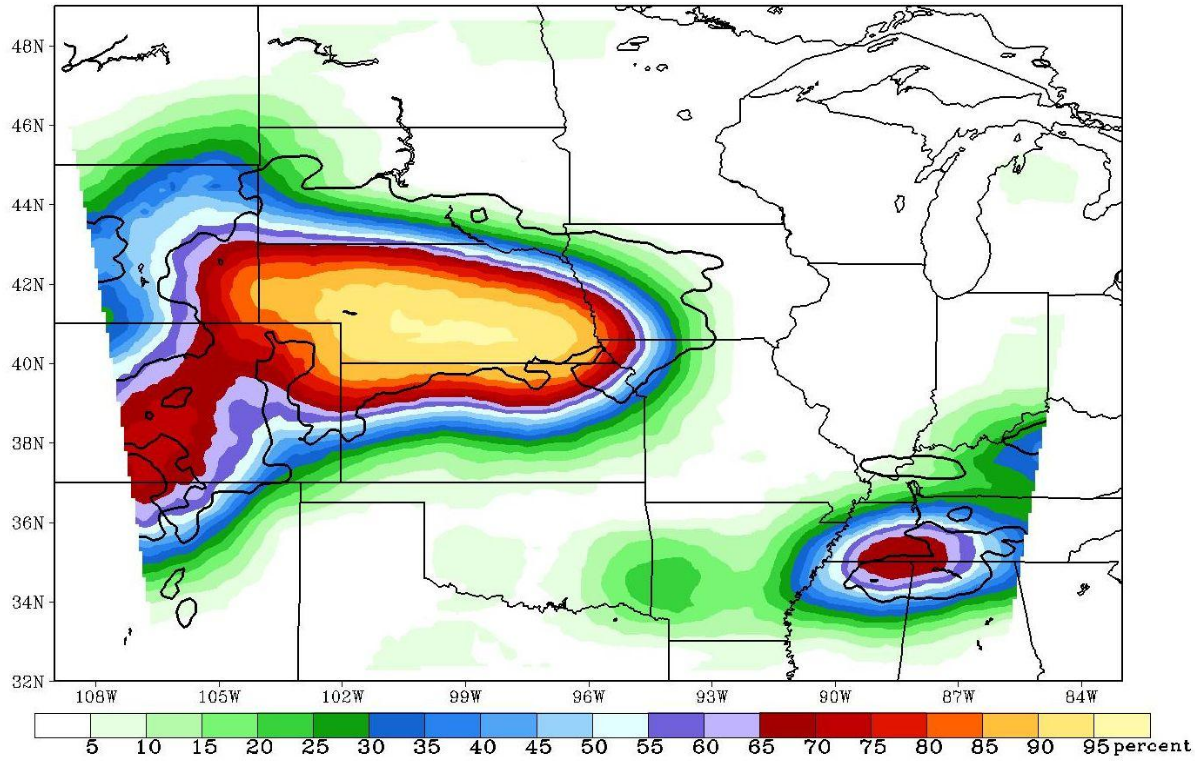

Figure 4-4. POPs (on the 20 km grid) for the combination method over the domain for 2007 April 23 for the period 06-12Z. The dark contour denotes observations at the 0.01 inch threshold.

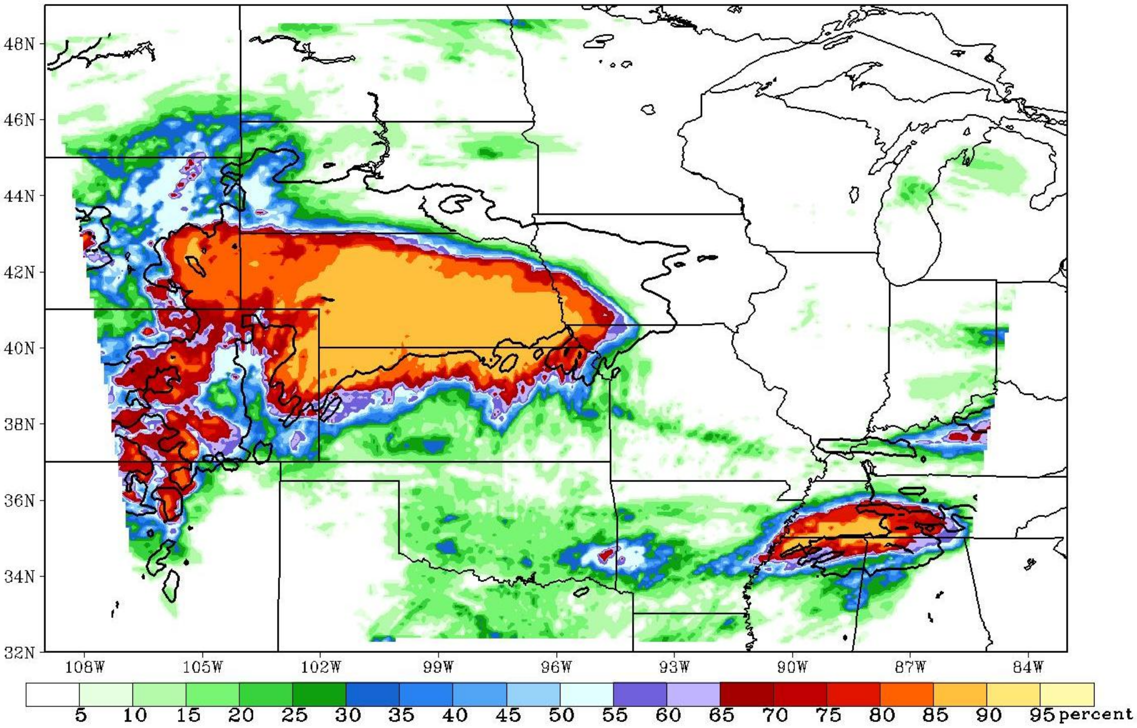

Figure 4-5. POPs (on the 4 km grid) for Max_thr over the domain for 2007 April 23 for the period 06-12Z. The dark contour denotes observations at the 0.01 inch threshold.

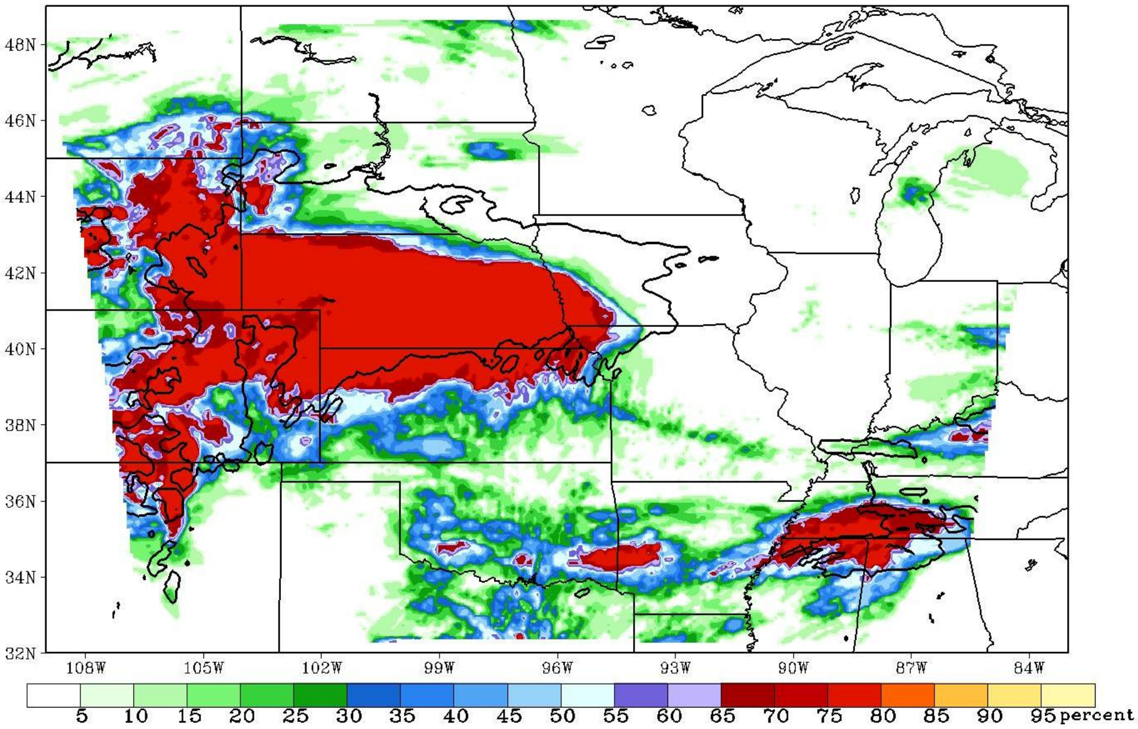

Figure 4-6. POPs (on the 4 km grid) for Cali_trad over the domain for 2007 April 23 for the period 06-12Z. The dark contour denotes observations at the 0.01 inch threshold.

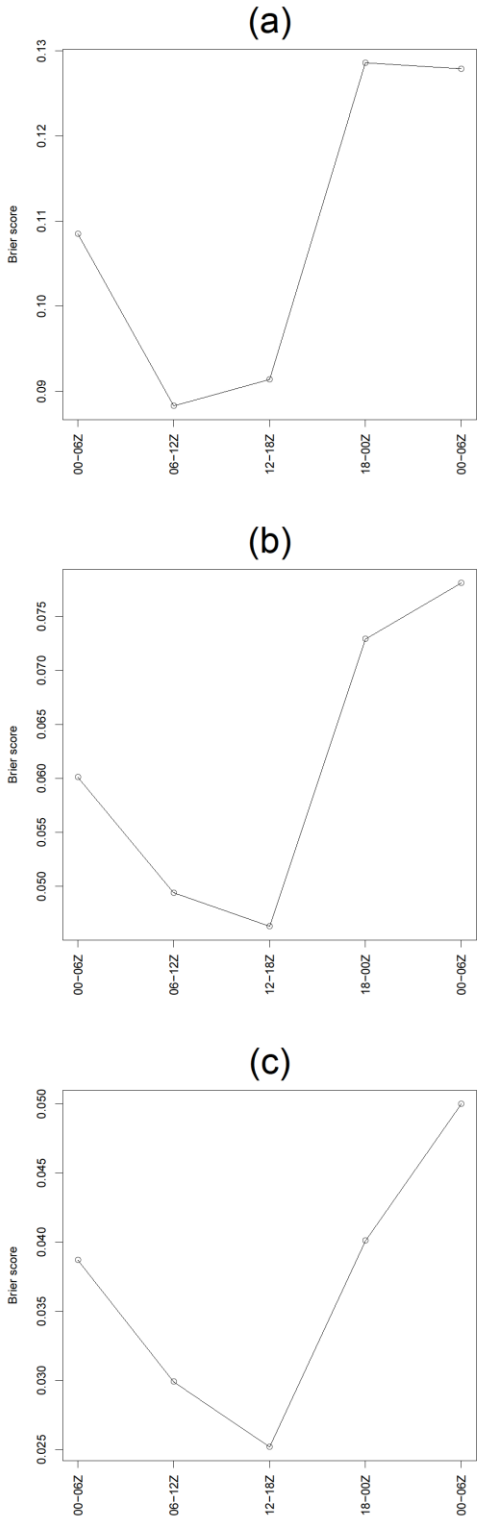

Figure 4-7. The average temporal variation of Brier scores for Max_nbh 5x5.

CHAPTER 5. CONCLUSIONS AND FUTURE WORK

Conclusions

The goal of this study was to find and test new QPF-POP relationships using ensemble forecasts, which led to the creation of four approaches. Each of these approaches has its own merits. The first of the four approaches presented in Chapter 3 can be thought of as an extension of the approach from Gallus and Segal (2004) into an ensemble environment. This first approach, which included the Max_thr and Ave_thr methods, was a stepping stone in creating the more complicated approaches, though it did not use a neighborhood approach. The forecasts of this approach showed improvements over Cali_trad, and encouraged the development of approaches using neighborhoods to obtain further improvements.

The second approach presented (consisting of the methods Max_nbh and Ave_nbh) was a two-parameter neighborhood approach. The improvements in BS from increasing the neighborhood size of this approach were surprising, because the methods within the approach did not use a traditional ensemble; the methods produced ten deterministic forecasts, which received data from each of their neighborhoods. This approach alone may be very useful in situations when traditional ensemble data is not available, because the deterministic forecasts were shown to sometimes outperform the ensemble forecast method Cali_trad and even the two-parameter point forecast methods (Max_thr and related methods). The Max_nbh and Ave_nbh methods within the approach sacrificed reliability in order to gain larger improvements to resolution.

The 3P approach acted as a synthesis of Ave_thr and Ave_nbh, and led to an increase in skill relative to the individual methods. The representations of the two agreement parameters from Ave_thr and Ave_nbh allowed for forecasts of much higher skill. This was probably because the representations allowed for much more data to be used in determining the value of each parameter, while the standard definition of each parameter was limited to either a ten grid point vector or a single neighborhood. Despite the improvements to the BSs from the 3P approach, a statistically significant difference compared to Cali_trad remained just out of reach at the 0.01 inch threshold.

The 3P approach continued the trend of improving BSs and moving further from the reference BS set by Cali_trad, primarily because the approach brought together the two previous approaches which we had proven could also produce more accurate POPs than Cali_trad. It is important to note that Cali_trad is, by definition, also contained in Max_thr and Ave_thr. While Cali_trad played an important role in motivating the research as a reference forecast, Cali_trad was never just a reference forecast; it was an important method in its own right in helping to achieve more accurate POPs. From the success of the 3P approach and considering the success of the approaches/methods before it, it was believed that combining the approaches in another manner would lead to further improvements. Plotting the POP fields over the domain, as was demonstrated in Chapter 4, also supported the idea that combining approaches could lead to improvements.

The combination approach, created by averaging the POP fields of Ave_nbh, Max_thr, and Cali_trad, showed that skill was indeed increased by bringing together methods in this manner. Extensive sensitivity tests showed the impact of including other methods, and that including Ave_thr and the 3P approach led to further improvements. In an effort to reduce QPF overestimation, a reduction factor was applied to the QPFs within the different approaches. Sensitivity tests were used to determine the reduction factors that were best for the 0.01 inch threshold of each method. The reduction factor was applied to QPFs before they were post-processed, so the factors can be considered an adjustment as part of the approaches. After the application of a reduction factor to the various methods to lessen forecast overestimation, some of the most skillful methods were used in the combination approach to produce forecasts that were statistically significantly better compared to Cali_trad's forecasts at the 95% confidence level.

The greatest improvements to the four approaches (compared the more traditional methods) occurred at the 0.01 inch threshold, which indicates that the four approaches were more likely than traditional methods to better delineate areas that received precipitation. Many of the problems caused by precipitation that were mentioned in Chapter 1 were the result of whether or not an area received rain, so forecasts from these new approaches may be better able to warn or protect against these problems. Though the greatest improvements were at the 0.01 inch threshold, the improvements made to the 0.10 inch and 0.25 inch thresholds are also of value. While not statistically significantly different from Cali_trad, these improvements could still be important. For example, while forecasting for an area that is close to reaching flood criteria, an additional 0.01 inch of precipitation may not be important, but an additional 0.25 inch could cause flooding. Even a slight improvement to a forecast could help protect life and property.

Future Work

The research for this study began at a time when only the 2007 Spring Experiment data was available, but the 2008 Spring Experiment data was included once it became available. Since this time, the 2009 Spring Experiment has taken place, and any future work related to these methods may want to incorporate this new data. There would be several advantages to doing so. There was evidence to suggest that some of the neighborhood approaches, which created a much larger number of POP forecasts, may have been limited by a lack of data. Adding the 2009 data could provide more data for the hit rate calculations, and lead to more accurate POPs. The addition of this data would also allow for more data to be used in POP tables for individual time periods and allow for an in-depth investigation of changes in skill with time.

The Theis et al. (2005) neighborhood approach used 3-hour time periods, which allowed for both spatial and temporal neighborhood approaches. If a future study used a 3-hour time period, instead of the 6-hour time period used in this study, the future study may be able to incorporate the temporal neighborhood approach and possibly improve forecasts further. This would be especially useful in the Ave_nbh method, which was already competitive with ensemble approaches. Using 3-hour time periods and the temporal neighborhood approach would also allow for direct comparisons to Theis et al. (2005).

Finally, finding a more sophisticated way to develop reduction factors for post-processing approaches would be beneficial. The reduction factors used in this study were constants determined through sensitivity tests, and were meant to improve only the 0.01 inch threshold forecasts. If a technique could be developed to apply reduction factors to individual grid points based on the QPF, this could lead to further improvements in the POP forecasts.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Gallus for the opportunity to work with him on his research, and for his guidance and expertise while developing and testing the forecasting approaches. I would also like to thank him for his time, revisions, and comments while editing the thesis.

I would like to thank Moti Segal for his guidance regarding the development of approaches and his advice about various ways to test the approaches. His insight into understanding the approaches was invaluable.

I thank Dr. Franz and Dr. Gutowski for their time, questions, and comments as part of my thesis committee. My thanks go to my thesis committee and the entire meteorology department for the education I have received at Iowa State University, without which my thesis would not have been possible. I also appreciate the assistance I have received from the graduate students of Iowa State University, especially Adam Clark, Eric Aligo, and Jon Hobbs.

Finally, I would like to thank my family for their support.

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.

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.

Changnon, S. A., 1996: Effects of summer precipitation on urban transportation. *Climatic Change*, **32**, 481-494.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121-1140.

Donner, S. D., and D. Scavia, 2007: How climate controls the flux of nitrogen by the Mississippi River and the development of hypoxia in the Gulf of Mexico. *Limnol. Oceanogr.*, **52**, 856-861.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459.

Ebert, E. E., 2001: Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

——, E. E., 2009: Neighborhood verification: a strategy for rewarding close forecasts. *Wea. Forecasting,* **24**, 1498–1510.

Fritsch, J. M., R. A. Houze, R. Adler, H. Bluestein, L. Bosart, J. Brown, F. Carr, C. Davis, R. H. Johnson, N. Junker, Y. H. Kuo, S. Rutledge, J. Smith, Z. Toth, J. W. Wilson, E. Zipser, and D. Zrnic, 1998: Quantitative precipitation forecasting: Report of the Eighth Prospectus Development Team, U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.*, **79**, 285-299.

——, and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: a USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85**, 955–965.

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.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. *Wea. Forecasting*, **24**, 1416–1430.

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.

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

Im, J. S., K. Brill, and E. Danaher, 2006: Confidence interval estimation for quantitative precipitation forecasts (QPF) using short-range ensemble forecasts (SREF). *Wea. Forecasting,* **21**, 24-41.

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.]

Luers, J., and P. Haines, 1983: Heavy rain influence on airplane accidents. *Journal of Aircraft.*, **20**, 187-191.

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.

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

——, and R. L. Winkler, 1979: Probabilistic temperature forecasts: the case for an operational program. *Bull. Amer. Meteor. Soc.*, **60**, 12–19.

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting*, **10**, 498-511.

Schwartz, B. E., and S. G. Benjamin, 2000: Verification of RUC2 precipitation forecasts using the NCEP multisensory analysis. Preprints. *Fourth Symp. On Integrated Observing Systems*, Long Beach, CA. Amer. Meteor. Soc., 182-185.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, M. C. Coniglio, and M. S. Wandishin, 2009: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, (In Press)

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032.

Stensrud, D. J., and N. Yussouf, 2007: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea. Forecasting*, **22**, 3–17.

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

Tustison B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.*, **106**, 11775–11784.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd edition. Academic Press, 627 pp.

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.