UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MACHINE LEARNING ENHANCEMENT OF STORM SCALE ENSEMBLE PRECIPITATION FORECASTS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

DAVID JOHN GAGNE II
Norman, Oklahoma
2012

MACHINE LEARNING ENHANCEMENT OF STORM SCALE ENSEMBLE PRECIPITATION FORECASTS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

Dr. Amy McGovern (Co-Chair)

Dr. Ming Xue (Co-Chair)

Dr. Fanyou Kong

Dr. Michael Richman

Acknowledgments

The help of many fellow colleagues and friends went into making this thesis a reality. First, I would like to thank my co-advisors Amy McGovern and Ming Xue. Amy has advised my research since my freshman year and guided me through the world of machine learning and data mining while providing many opportunities for collaboration with many facets of the meteorological community. Ming has provided valuable guidance from his extensive background in numerical weather prediction and helped secure the data and resources to make the project possible. Special thanks also go to committee members Michael Richman and Fanyou Kong for their guidance on this project and feedback on the thesis. Fanyou graciously provided the ensemble and verification data at the center of this project.

The people of the Center for the Analysis and Prediction of Storms were very helpful in providing resources for my project as well as providing feedback through personal interactions and the CAPS Science Meetings. In particular, I would like to thank Scott Hill for setting up and maintaining the hardware and software for this project, Youngsun Jung for help with the CAPS Science Meetings, and Nate Snook for his advice, stories, and ARPS-related humor.

My fellow current and former student colleagues in the IDEA lab were essential for feedback on the project as well as moral support and many good laughs. Scott Hellman developed a new algorithm to apply to my data source and always found new forms of entertainment for the lab. Matthew Collier brought his multidisciplinary perspective and treasure trove of puns. Ross Kimes provided inside information on all matters Apple related. Rachel Shadoan and Zach Tidwell made for fun travel companions to CIDU 2010. Derek Rosendahl has provided me with advice for navigating the graduate school process throughout the years. My mentees Tim Sliwinski, Jon Trueblood, and Braden Katona have

given me plenty of experience in teaching and guidance while impressing me with their own achievements.

My fellow meteorology graduate students have made many direct and indirect contributions to my thesis research and general livelihood. Zac Flamig provided RQI data and assistance with the NMQ. Tim Supinie, Andrew MacKenzie, and Greg Blumberg provided valuable friendship and moral support while keeping me mostly sane through this process.

I must also give generous thanks to my loved ones. My fiancée Cathy Bolene has been a wonderful source of support and encouragement. My parents David and Diana deserve special thanks for raising me and providing me with the fortitude to succeed.

Special recognition must also be given to the unsung heroes of this thesis, the computers. Stratus and its many processors and piles of RAM made running the extensive experiments repeatedly a mostly painless experience. Inspiration and synergy provided the space for all of my analysis and writing. Rollcloud allowed me to take my work anywhere. The computer science SVN server ensured that my files were correctly updated and distributed across multiple machines to lessen the chance of a catastrophic failure.

This study was funded by the NSF Graduate Research Fellowship under Grant 2011099434.

Table of Contents

List Of Tables

List Of Figures

Abstract

Precipitation forecasts provide both a crucial service for the general populace and a challenging forecasting problem due to the complex, multi-scale interactions required for precipitation formation. The Center for the Analysis and Prediction of Storms (CAPS) Storm Scale Ensemble Forecast (SSEF) system is a promising method of providing high resolution forecasts of the intensity and uncertainty in precipitation forecasts. The SSEF incorporates multiple models with multiple parameterization scheme combinations and produces forecasts every 4 km over the continental US. The SSEF precipitation forecasts exhibit significant negative biases and placement errors. In order to correct these issues, multiple machine learning algorithms have been applied to the SSEF precipitation forecasts to correct the forecasts using the NSSL National Mosaic and Multisensor QPE (NMQ) grid as verification. The 2010 runs of the SSEF were used for training and verification. Two levels of post-processing are performed. In the first, probabilities of any precipitation are determined and used to find optimal thresholds for the precipitation areas. Then, three types of forecasts are produced in those areas. First, the probability of the 1-hour accumulated precipitation exceeding a threshold is predicted with random forests, logistic regression, and multivariate adaptive regression splines (MARS). Second, deterministic forecasts based on a correction from the ensemble mean are made with linear regression, random forests, and MARS. Third, fixed probability interval forecasts are made with quantile regressions and quantile regression forests. Models are generated from points sampled from the western, central, and eastern sections of the domain. Verification statistics and case study results show improvements in the reliability and skill of the forecasts compared to the original ensemble while controlling for the over-prediction of the precipitation areas and without sacrificing smaller scale details from the model runs.

Chapter 1

Introduction

Precipitation forecasts are among the most challenging in meteorology due to the wide variability of precipitation over small areas, the dependence of precipitation amounts on factors at a wide range of scales, and the mixed discrete-continuous probability distribution of precipitation (Bremnes, 2004; Ebert, 2001; Doswell et al., 1996). The precipitation forecasting problem can be divided into three primary questions: where is it going to rain, when is it going to rain, and how much rain will occur? The answers to all three of those questions depend on the availability of precipitation ingredients and the placement and timing of the storms that can take advantage of those ingredients. Although light and moderate rain is often viewed as a mere annoyance for many people, heavy rains in short time periods can lead to flash floods that present risks to lives and property (Doswell et al., 1996). Improving the prediction of heavy precipitation events in particular is crucial for anticipating those events.

Numerical Weather Prediction (NWP) models are now being run experimentally and regularly at 1 to 4 km horizontal grid spacing, or storm scale, allowing for the formation of individual convective cells without the need of a separate scheme to determine if conditions are favorable for convection. This feature allows for a better representation of storm processes, but it adds additional uncertainty in placement and timing of precipitation compared to models with larger grid spacing. An ensemble of storm scale NWP models can provide

estimates of that uncertainty, but statistical post processing is needed to account for model biases and to increase reliability. The quality and design of the post-processing is constrained by many factors, including sample size, composition, variables used, number of points requiring a forecast, and predicted variable format.

This thesis evaluates multiple approaches to post-processing storm-scale ensemble precipitation forecasts with machine learning and statistical algorithms. The algorithms are designed to produce the following products: probabilities of exceeding a given threshold, deterministic precipitation forecasts, and a intervals of precipitation amounts give a fixed probability range. They are trained using aggregations of the raw ensemble precipitation predictions as well as other relevant variables. Our approach has improved precipitation quantity forecasts, provided better estimates of the uncertainty, and better defined the area covered by the precipitation forecasts.

Chapter 2

Related Work

2.1 Model Output Statistics

Since the early days of operational numerical weather prediction, statistical post-processing of model precipitation forecasts has been recognized as a necessity for producing accurate and useful predictions. The original method and basis for subsequent post-processing methods is Model Output Statistics (MOS; (Glahn and Lowry, 1972; Klein and Glahn, 1974)). The MOS approach fits a multivariate linear regression model between an observed quantity and a set of model variables selected using the screening process (Klein et al., 1959), which iteratively selects the combination of variables that is most correlated with the predicted quantity. For probability of precipitation MOS forecasts, one regression equation was applied to similar regions of observations in order to capture enough variability over the training time period (Klein and Glahn, 1974). Logistic regression, also known as the logit model, was used operationally for conditional probability of frozen precipitation forecasts (Klein and Glahn, 1974). Bermowitz (1975) applied MOS to predict the probability of precipitation amounts within 5 different categories. Most of the variables chosen by the screening process were either related to model precipitation or precipitable water. Incorporating zero precipitation events into the training data did not

seem to have a large impact on the performance. The model did have trouble predicting heavy precipitation due to the rareness of those events, although transforming the predictions to minimize the bias did help with this problem. Some approaches to MOS also included analyses of observed variables in addition to model output, which produced slight improvements in skill (Vislocky and Young, 1989).

2.2 Ensemble Postprocessing

With the advent of operational ensemble weather prediction (Toth and Kalnay, 1993; Tracton and Kalnay, 1993; Molteni et al., 1996), it soon became apparent that statistical post-processing was needed to produce accurate precipitation forecasts and uncertainty estimates from the ensemble forecasts. Verification of the operational ensemble forecasts used the rank histogram method (Hamill and Colucci, 1997), which is a histogram showing the distribution of the ranks of observations relative to ensemble member forecasts. It showed that ensemble precipitation forecasts tended to be underdispersive, so the observed precipitation amount did not fall within the range of ensemble forecasts. Initial calibration methods with linear regressions fit to a Gamma distribution (Hamill and Colucci, 1998) and calibration to the rank histogram itself (Hamill and Colucci, 1997; Eckel and Walters, 1998) did show improvements in skill compared to the uncorrected ensemble and traditional MOS.

As operational ensemble prediction systems matured, researchers investigated different ways of both combining individual ensemble members together as well as combining different ensemble prediction systems. Hamill and Colucci

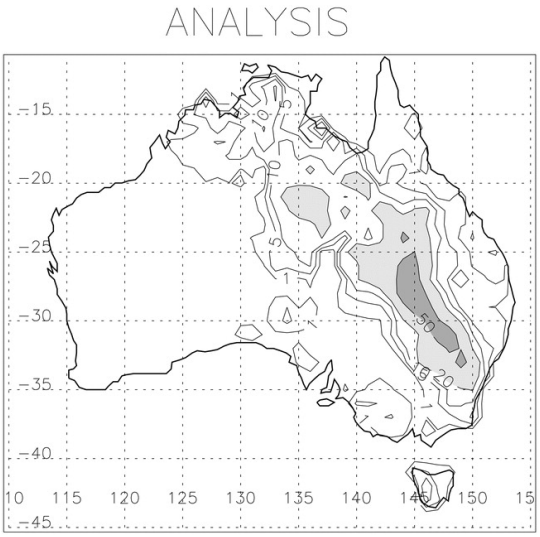

Figure 2.1: Analysis of 24-h accumulated precipitation on 21 April 1998 (Ebert, 2001).

(1997, 1998) used an ensemble consisting of ETA and RSM members. Krishnamurti et al. (1999) used a multiple linear regression technique at each grid point to combine the output from multiple operational models into a "super ensemble" and achieved significant improvements in skill, but the technique was very sensitive to changes in the model configurations. Ebert (2001) tested a wide range of approaches to combining multiple ensemble prediction systems made of different models to create a "poor man's ensemble" by comparing the spatial coverage of their precipitation forecasts and the range of predicted intensities (Fig. 2.1 and Fig. 2.2). Weighted means of the individual members based on past performance does not tend to improve precipitation forecasts because individual members are not consistently better than others from day to day. Using the ensemble median precipitation resulted in a smaller, more accurate rain area but intensities were underestimated. Bias correction estimated the rain area better but overestimated the rainfall amounts. Probability matching performed the best of the different methods tried. It uses the ensemble mean rain

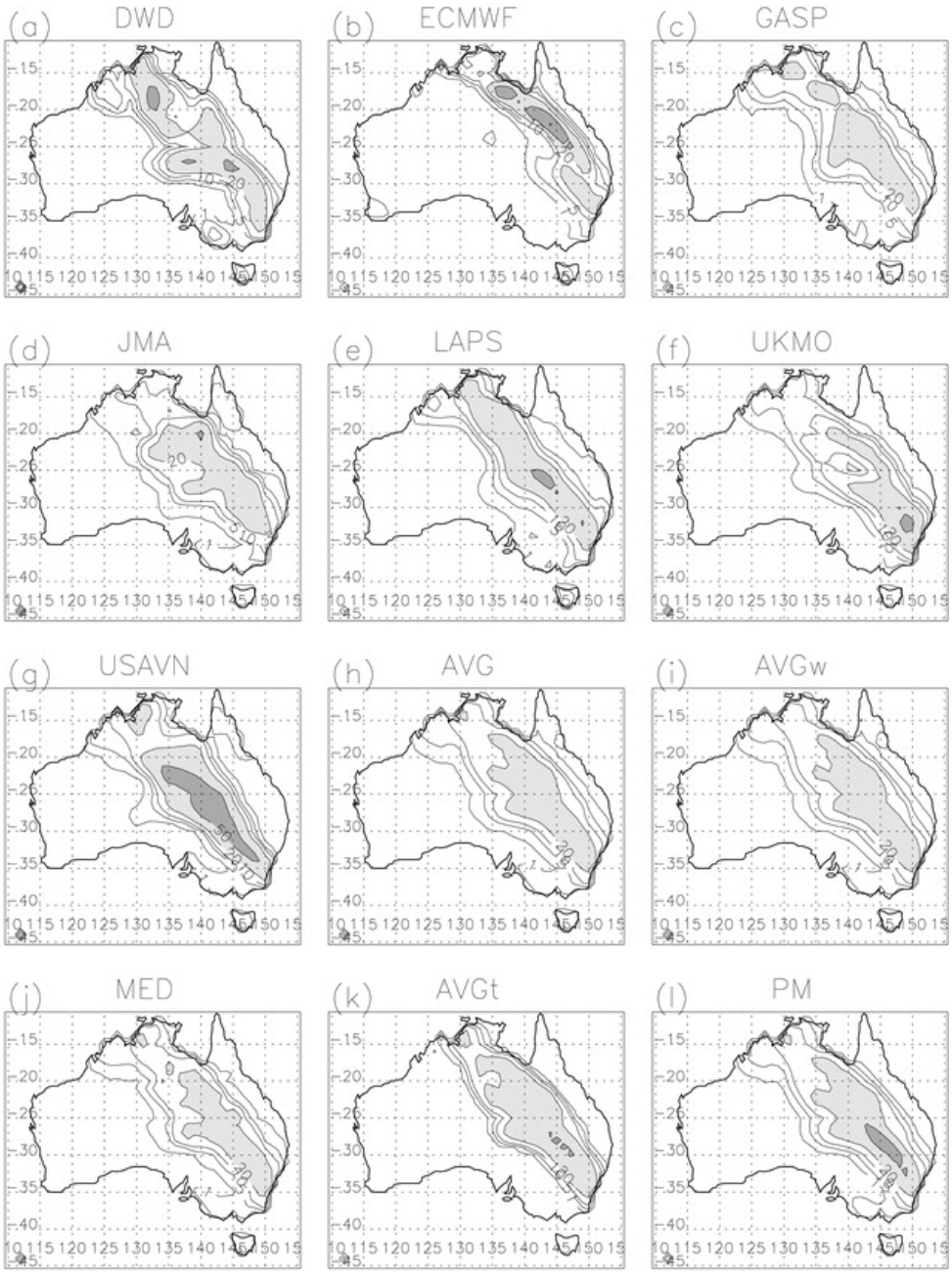

Figure 2.2: Comparison of 24-h quantitative precipitation forecasts for 21 April 1998 for seven models and different methods of generating an ensemble averaged forecast (Ebert, 2001).

area and then matched the ranked ensemble mean rain points with the ranked rainfall values from all of the ensemble members. The wider distribution of rain points produced predicted rain rates closer to those observed.

Many variations of models and datasets were used for precipitation post-processing in the last decade. Hall et al. (1999) and Koizumi (1999) applied the neural network technique to precipitation forecasts and found increases in performance over linear regression and persistence. Neural networks are machine learning models that consist of input, output, and hidden layers containing interconnected nodes. The weights at all the nodes are iteratively determined in order to fit the model as close to the training data as possible. Neural networks were applied to ensemble probabilistic QPF in Yuan et al. (2007) where they did improve calibration but produced more forecasts with low nonzero probabilities that moved the forecast distribution closer to climatology. Logistic regression, a transform of a linear regression to fit an S-shaped curve ranging from 0 to 1, has also been used extensively in multiple post-processing studies for deterministic and ensemble forecasts. Applequist et al. (2002) tested linear regression, logistic regression, neural networks, and genetic algorithms on the precipitation prediction problem and found that logistic regression outperformed the other methods at multiple thresholds for 24 hour probabilistic QPF. Since choice of model is not the only factor that affects the performance of statistical forecasts, methods to improve the input dataset were also explored. Hamill et al. (2004, 2008) introduced a procedure that used an older version of an operational global model to produce retrospective forecasts (reforcasts) over a multidecadal period in order to capture the model climatology of rainfall predictions that sample the full range of variation in rainfall amounts at every grid point. With the very large resulting training set, a simple model such as a logistic regression

can perform very well on probabilistic QPF. Both studies found that 2-week forecasts using the reforecast database outperformed 6-10 day forecasts from the operational ensembles.

More recent work explored different frameworks for producing models and expressing uncertainty from ensemble forecasts. Bremnes (2004) produced fixed probability interval forecasts with a quantile regression and found that training the post-processing algorithms on statistics about the ensemble forecast performed better than training on all ensemble members. This finding supported the development of the aggregation methods used in the pre-processing steps of this project. Bayesian Model Averaging (BMA) (Raftery et al., 2005), a weighted average of the probability distribution functions for bias-corrected ensemble member forecasts, has also been applied to precipitation forecasts (Soughter et al., 2007) with a Gamma distribution in place of the Gaussian distribution. Because it produces a model of the entire forecast PDF instead of just a single estimate, BMA can produce probabilistic, deterministic, and interval forecasts without the need for additional models. Both Bremnes (2004) and Soughter et al. (2007) use an additional algorithm to determine probability of precipitation before determining the conditional probability of precipitation exceeding a threshold or occurring within a range. Wilks (2009) attempted to address this deficiency in logistic regression approaches by including an additional term that specified the predicted threshold.

Chapter 3

Storm Scale Ensemble Forecast Data

3.1 Ensemble Specification

The post-processing algorithms for this project are being applied to the Center for the Analysis and Prediction of Storms (CAPS) 2010 Storm Scale Ensemble Forecast (SSEF) system (Kong et al., 2011; Xue et al., 2011). The 2010 SSEF is composed of 26 separate model runs from the Weather Research and Forecasting (WRF) Advanced Research WRF and Nonhydrostatic Mesoscale Model and the Advanced Regional Prediction System. Each model is run with different combinations of microphysics schemes, land surface models, and planetary boundary layer schemes. The SSEF ran every weekday at 0000 UTC in conjunction with the 2010 National Oceanic and Atmospheric Administration/Hazardous Weather Testbed Spring Experiment (Clark et al., 2012), which ran from 3 May to 18 June. The SSEF provides hourly model output over the continental United States at 4 km horizontal grid spacing out to 30 hours. Of the 26 models in the 2010 SSEF, the 14 models initialized from the Short Range Ensemble Forecast system are included in the post-processing procedure because they are the only models that combine perturbed initial and boundary conditions with perturbed physics. The 12 models not included use the same initial conditions from a control run, so they do not contribute additional information about the spread and could bias the mean toward the control run forecast. This dataset

is also being used in Hellman (2012) where the Ensembled Dynamic Bayesian Networks (EDBN) algorithm is introduced. They compare performance to the random forests presented here.

3.2 Verification Data

A storm scale verification dataset was paired with the storm scale ensemble forecasts. The National Mosaic Multi-Sensor QPE (NMQ; Vasiloff et al., 2007) derives precipitation estimates from a 3-dimensional mosaic of the NEXRAD radar network. The estimates are overlaid on a grid over the CONUS with 1 km horizontal spacing. The original grid has been interpolated to the same grid as the SSEF.

3.3 Data Selection and Aggregation

The relative performance of any machine learning algorithm is conditioned on the distribution of its training data. The sampling scheme for the SSEF is conditioned on the constraints of relatively few ensemble runs over a short, homogenous time period with nearly 1 million grid points from each time step. The short training period and large number of grid points preclude training a single model at each grid point, so a regional approach was used.

The SSEF domain was split into thirds, and points were selected with a stratified random sample from each subdomain in areas with quality radar coverage. Grid 1 corresponds to the western third of the CONUS, Grid 2 corresponds to the central third, and Grid 3 corresponds to the eastern third. Fig. 3.1 shows that most of the continental US was sampled with the exception of areas of the

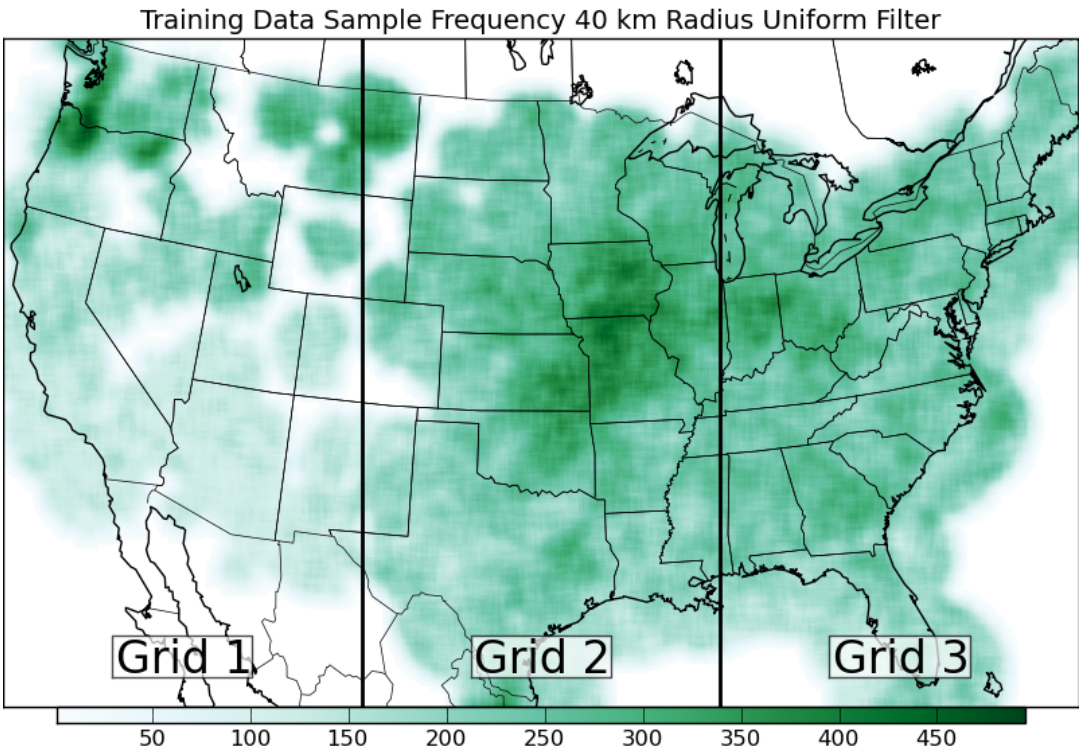

Figure 3.1: Map of spatial histogram showing how frequently the area within a 40 km radius of a particular point was sampled. The domain sub grids are also shown and labeled.

Rocky Mountains with poor radar coverage. The sampling of areas within radar coverage is not uniform and does have a slight bias toward areas that received large amounts of precipitation during the SSEF time period.

A comparison of the sampled rainfall distributions and the full rainfall distributions for each subgrid are shown in Fig. 3.2. Undersampling of the 0 and light precipitation points was necessary because of the large number of no-precipitation events, which overwhelmed the signal from the actual precipitation events. The upper bins were slightly oversampled since the heavy precipitation events were very rare and would not be handled well by the machine learning algorithms otherwise. The random sampling of grid points helps reduce the chance of sampling multiple grid points from the same storm without explicitly filtering parts of the domain as in Hamill et al. (2008).

Relevant model variables (Table 3.1) at each sampled grid point were also extracted from each ensemble member. These variables were selected to capture additional information about the mesoscale and synoptic conditions in each model. An additional variable called the Radar Quality Index (RQI) was sampled at each point to determine the trustworthiness of the verification data at that point. Only points with RQI greater than 0 were included in the training data.

Multiple ways of aggregating the ensemble variables were tested. For the probabilistic forecasts, ensemble variables were grouped into mean and standard deviation. The mean and standard deviation capture information about the predicted changes in a variable and the spread in those predictions. For the regression and quantile forecasts, the 5th, 50th and 95th percentiles of each

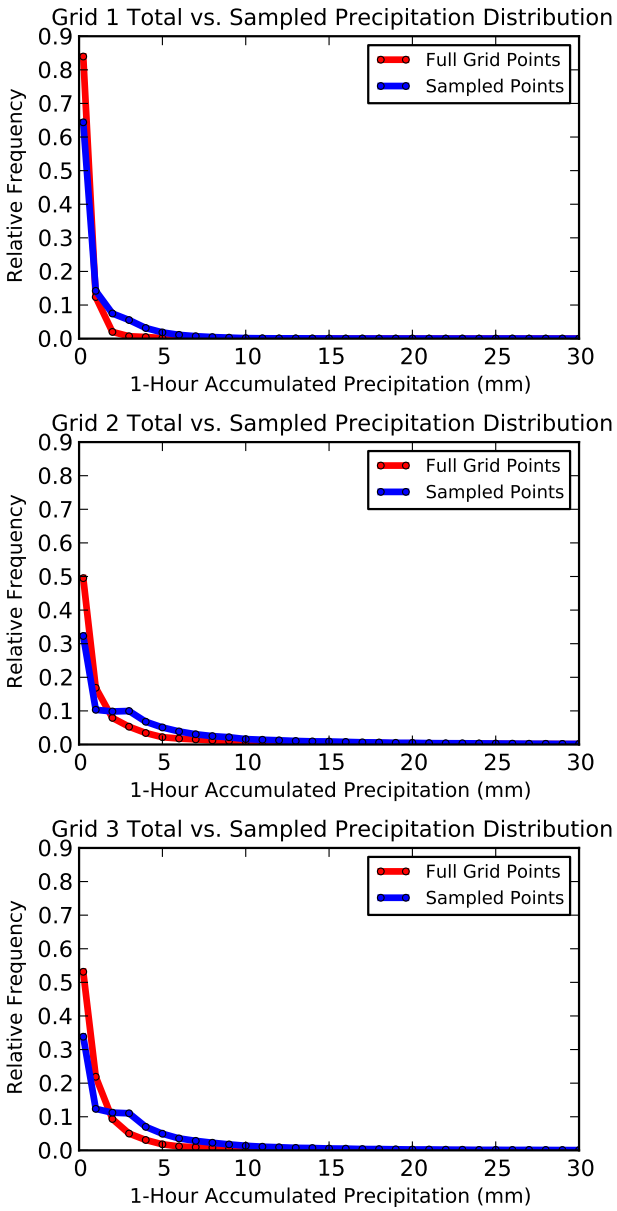

Figure 3.2: Histogram comparing the relative frequencies of the full precipitation distribution for the each subgrid to the sampled rainfall distributions.

variable were extracted. This format also provides information about the median forecast and spread while also sampling the range of forecast values and weakening the influence of outliers.

Table 3.1: The names and descriptions of model variables sampled from the SSEF runs. CAPE is Convective Available Potential Energy, and CIN is Convective Inhibition.

Variable	Description
accppt	1-hour accumulated precipitation
cmpref	Composite reflectivity
dewp2m	2 m dew point temperature
refmax	1-hour maximum reflectivity
mspres	Mean sea level pressure
sbcape	Surface-based CAPE
sbcins	Surface-based CIN
pwat	Precipitable water
temp2m	2 m air temperature
tmp700	700 mb temperature
u700	700 mb east-west wind
v700	700 mb north-south wind
hgt700	700 mb height
v500	500 mb north-south wind
wupmax	1-hour max upward vertical velocity
wdnmax	1-hour max downward vertical velocity

Chapter 4

Machine Learning Methods and Procedures

4.1 Machine Learning Algorithms

4.1.1 Logistic and Linear Regression

A mix of more traditional and complex machine learning algorithms were trained on the aggregated ensemble data. For the probabilistic domain, logistic regressions were used as the baseline algorithm. Logistic regressions are linear regression models fitted to a logit curve that ranges from 0 to 1. Logistic regressions were trained using just the raw ensemble probability and the mean and standard deviation of the ensemble accumulated precipitation forecasts. Two formulations of logistic regression are being used. The first formulation uses the ensemble mean precipitation forecast as expressed in Eqn. 4.1:

$$p(R \geq t|x_1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}} \tag{4.1}$$

where x_1 is the ensemble mean precipitation forecast and β_0 is the intercept term. This formulation will adjust the probability of the ensemble forecast based on systematic biases in the mean but will not change the areal coverage of the precipitation forecasts. The second formulation incorporates a multivariate

approach using variable selection similar to standard MOS approaches (Eqn. 4.2):

$$p(R \geq t|x_1, x_2, ...x_n) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)}} \tag{4.2}$$

The terms are chosen through an iterative process that finds the set of terms that minimize the Akaike Information Criterion (Akaike, 1974), which rewards goodness of fit but penalizes for large numbers of terms. This approach does produce the best fit regression model given the available parameters, but the searching process can take extensive time given the large number of terms and number of samples in the training set.

For the deterministic predictions, linear regressions were used as the baseline model. One was trained with the ensemble median precipitation forecast and the other was trained with the 5th, 50th, and 95th percentiles of the ensemble precipitation predictions.

4.1.2 Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS; (Friedman, 1991)) were also applied to the problem. MARS have a form similar to multiple linear regression but with an added layer of complexity. Each term consists of a hinge function, which consists of two linear functions that intersect at an intercept point. Linear combinations of these hinge functions produce a piecewise linear regression that can approximate complex functions with a relatively compact model. MARS performs variable selection using forward selection and backward elimination of variables. For computational timing purpose, the maximum number of variables

was limited to 15 and only linear relationships were considered. This project uses the open-source *earth*[1] library as its MARS implementation.

4.1.3 Random Forests

Random forests (Breiman, 2001) consist of an ensemble of classification and regression trees (Breiman, 1984) with two key modifications. First, the training data are bootstrap resampled with replacement for each tree in the ensemble. Then only a small random subset of the total number of variables are evaluated for splitting at each node in each tree. The final prediction from the forest is the mean of the predicted values from all the trees. Random forests can produce both probabilistic and regression predictions through this method. The random forest method contains a few advantages that often lead to performance increases over traditional regression methods. The averaging of the results from multiple trees produces a smoother range of values than individual decision trees while also reducing the variance of the predictions (Strobl et al., 2008). The random selection of variables within the tree-building process allows for less optimal variables to be included in the model and increases the likelihood of the discovery of interaction effects among variables that otherwise would be missed by a regression method that selects variables based on a global optimization, as is done with the logistic regression (Strobl et al., 2008).

Since the tree-building process selects a subset of variables for each tree, a technique called variable importance (Breiman, 2001) can be used to rank the relevance of the variables in the dataset. Variable importance is computed by first calculating the accuracy of each tree in the forest on classifying the cases that were not selected for training, known as the out-of-bag cases. Within the

[1] http://www.milbo.users.sonic.net/earth/

out-of-bag cases, the values of each variable are randomly rearranged, or permuted, and those cases are then re-evaluated by each tree. The mean variable importance score is then the difference in prediction accuracy on the out-of-bag cases averaged over all trees. The mean variable importance is then divided by the standard error of the variable importances, which is the standard deviation divided by the square root of the number of trees. Variable importance scores can vary randomly among forests trained on the same dataset, so the variable importance scores from each forest were averaged together for a more definite ranking. The statistical significance of the importance score can be assessed through a z-test comparing the distribution of the scores with 0 to determine if the difference is significant. The z-score approach calculated for a single random forest depends more on the number of trees than the sample size of the data (Strobl et al., 2008). By performing a z-test over the distribution of mean scores from multiple random forests of the same size, the variability in the significance depends on the number of forests tested instead of the number of trees in each forest. Variable importance scores can also be influenced by the correlations among variables in the dataset, so that if two highly correlated variables are included in the dataset, both may have high importance scores even if only one of the two is the true cause. Controlling for the causal variable would then cause the resulting importance of the second variable to decrease substantially (Strobl et al., 2008). A modified algorithm called conditional variable importance was developed to address this issue but was not implemented in the R random forest library used by the project (*randomForest*). While variable importance is effective at determining which variables the random forest model finds most relevant, it does not reveal what aspects of those variables make them relevant. There is no information provided about what ranges of values have the greatest

affect on performance. The importance of the interactions of multiple variables can be estimated by permuting multiple variables at once and comparing that importance score with the importance scores of the individual variables (Pappenberger et al., 2006), but this method would be computationally infeasible for datasets with large numbers of variables. Visualizing the distribution of predicted values versus the distribution of variable values could also provide additional information (Gagne II et al., 2012).

4.1.4 Quantile Regressions

In addition to producing a single probability or quantity, two methods can produce fixed probability intervals. Quantile regressions (Koenker and Hallock, 2001) estimate the conditional median and other quantiles by treating the procedure as an optimization problem. The conditional median is found by minimizing the sum of the absolute residuals while the other quantiles are found by minimizing the asymmetrically weighted sum of the absolute residuals. The residual in this instance is the difference between the observation and a linear function of predictor variables and can be solved by linear programming methods. Because the median and quantiles are being predicted, the model is less sensitive to outliers in the distribution and does not require the assumption of a particular empirical distribution.

4.1.5 Quantile Regression Forests

Quantile regression forests (Meinshausen, 2006) are a variation of random forests that use the distribution of values at the selected leaf node of each tree in the forest to estimate specified quantiles. Because of the focus on the median and

quantiles instead of the mean, quantile regression forests are less sensitive to outlier cases in the training data, which is beneficial on problems where that could lead the predictions awry.

4.2 Model Training and Evaluation

The training and evaluation procedure for the post processing algorithms used an approach that maximized the available training and testing data for the algorithms without compromising the independence of either set. Each machine learning algorithm was trained and evaluated using leave-one-model-run-out cross validation. The post-processing occurred in a two-step process. First, each probabilistic algorithm is trained to predict the probability of 1-hour precipitation exceeding 0.25 mm (0.01 in). This prediction is used to estimate the area where any precipitation will occur. This forecast can be treated as a probability of precipitation forecast or it can be thresholded to provide rain and no-rain areas. Second, another set of algorithms is trained on the conditional probability of 1-hour precipitation exceeding 6.54 mm (0.25 in) given that precipitation is occurred at that point. This precipitation threshold was chosen because it is a moderate amount of rain for the period and matches current SSEF forecast products, allowing for easier comparisons. Third, another set of algorithms is trained to predict a correction factor to the ensemble mean precipitation forecast. Although the algorithms could be used to predict the amount of precipitation directly, the distribution of corrections is closer to Gaussian whereas the distribution of precipitation amounts is heavily skewed to lighter amounts. The only downside of the correction approach is that it does allow

for negative precipitation predictions to occur, so any corrections that result in a forecast below 0 are corrected to 0.

Evaluation techniques depend on the type of forecast. The Brier Skill Score (BSS) (BSS; Brier, 1950) is one method used to evaluate probabilistic forecasts. The Brier Skill Score can be decomposed into three terms (Murphy, 1973), as shown in Eq. 4.3:

$$BSS = \frac{\frac{1}{N}\sum_{k=1}^{K} n_k(\overline{o}_k - \overline{o})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \overline{o}_k)^2}{\overline{o}(1-\overline{o})} \tag{4.3}$$

N is the number of forecasts, K is the number of probability bins, n_k is the number of forecasts in each probability bin, $\overline{o}_k$ is the observed relative frequency for each bin, $\overline{o}$ is the climatological frequency, and p_k is the forecast probability for a particular bin k. The first term describes the resolution of the forecast probability, which should be maximized and increases as the observed relative frequency differs more from climatology. The second term describes the reliability of the forecast probability, which should be minimized and decreases with smaller differences between the forecast probability and observed relative frequency. The third term is the uncertainty, which is fixed for a given dataset. Positive BSS indicates positive skill and vice versa. The components of the BSS can be displayed graphically with an attributes diagram (Wilks, 2011), in which the observed relative frequency of binned probability forecasts are plotted against lines showing perfect reliability, no skill where the reliability and resolution are equal, and no resolution where the observed relative frequency and probability equal climatology.

The Area Under the Relative Operating Characteristic (ROC) curve, or AUC (Mason, 1982) is another method used to evaluate probabilistic forecasts. To calculate AUC, first, the decision probability threshold is varied from 0 to 1 at regular intervals. At each interval, a contingency table is constructed by splitting the probabilities into two categories at the decision threshold. From the contingency table, the probability of detection (POD) and probability of false detection (POFD) are calculated and plotted against each other. This plot becomes the ROC curve. The AUC is the area between the lower right side of the curve and the curve itself. AUC above 0.5 has positive skill. AUC only determines how well the forecast discriminates between two categories, so it does not take the reliability of the forecast into account.

The ROC curve also serves as a way to set an optimal decision threshold. At low thresholds, the number of true positives and false positives is maximized and both decrease with increasing threshold value but at different rates. True and false negatives are minimized at low thresholds and both increase with increasing threshold value. The optimum threshold, assuming equal penalties for false positives and false negatives, would be where the two are most similar. This occurs where the ROC curve is farthest from the positive diagonal. The horizontal distance between the ROC curve and the positive diagonal (Fig. 4.2) is the Peirce Skill Score (PSS; Peirce, 1884; Hansen and Kuipers, 1965), and thus the PSS is maximized at the optimum decision threshold (Manzato, 2007). Finding the maximum PSS then is a matter of calculating it at each point along the ROC curve when POD and POFD are also calculated and finding the position of the maximum value. This same procedure could also be used to optimize to other 2x2 contingency table skill scores, including Heidke Skill Score, Threat

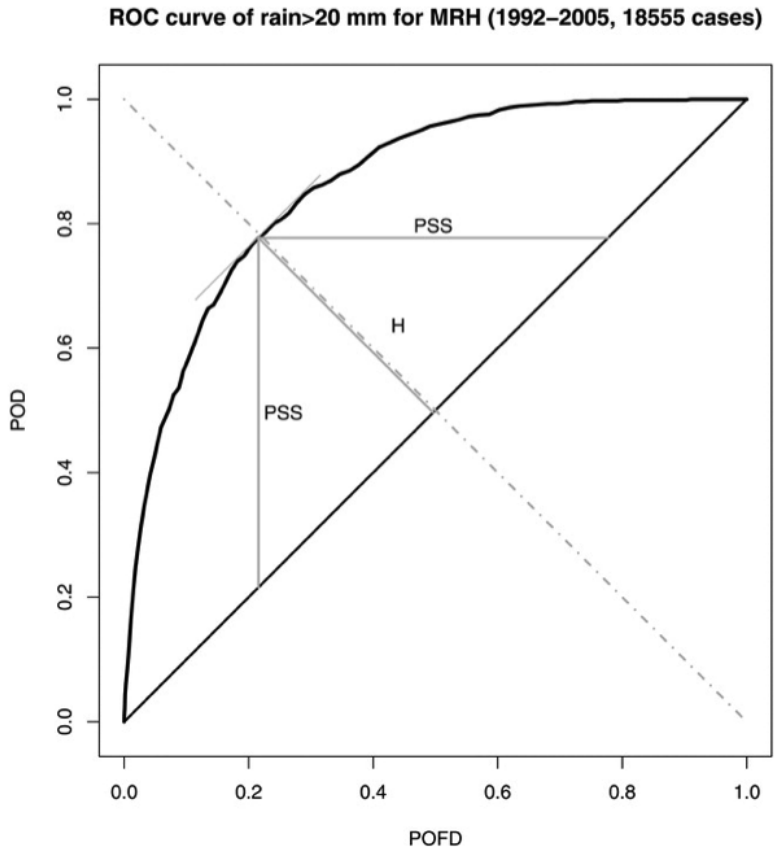

Figure 4.1: An example ROC curve showcasing the relationship between the PSS and the ROC curve from Manzato (2007).

Score, and Equitable Threat Score, but those scores do not necessarily maximize at the same point on the ROC curve. If the decision process is concerned with maximizing another score, then using another score would be more effective at the risk of hedging the forecasts. The optimal threshold is calculated separately for each sub grid and hour to account for the differences in the regional probability distributions and the increases in uncertainty with forecast hour. The performance of the threshold choice can be validated by treating the predictions as a binary classification problem. The contingency table is shown in Table 4.1. Using that as a reference, the following scores can be computed, as shown in Table 4.2. Probability of detection (POD) is the ratio of hits to the total number of observed events. The False Alarm Ratio (FAR) accounts for the number of false alarms compared to the total number of yes forecasts. The Equitable Threat Score (ETS) compares the ratio of hits to hits, misses, and false alarms with the hits of a random classifier. The Peirce Skill Score (PSS) is the difference between the POD and the POFD. The Bias compares the ratio of the sum of hits and false alarms to the sum of hits and misses.

Deterministic precipitation forecasts are verified with the mean error (ME) and the root mean squared error (RMSE). The mean error, shown in Eq. 4.4, is

Table 4.1: An example binary contingency table for whether or not rain is forecast.

		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

Table 4.2: The scores calculated from the binary contingency table for use with the optimal threshold predictions.

Score	Formula
POD	$\frac{a}{a+c}$
FAR	$\frac{b}{a+b}$
POFD	$\frac{b}{b+d}$
ETS	$\frac{a-a_{random}}{a+b+c-a_{random}}$
a_{random}	$\frac{(a+c)(a+b)}{a+b+c+d}$
PSS	$\frac{a}{a+c}-\frac{b}{b+d}$
Bias	$\frac{a+b}{a+c}$

indicative of whether the forecasts have a particular bias. A ME of 0 indicates that the error is evenly spread on both sides of the observed values.

$$ME = \frac{1}{P}\sum_{p=1}^{P} f_p - o_p \tag{4.4}$$

The RMSE is shown in Eq. 4.5.

$$RMSE = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(f_p - o_p)^2} \tag{4.5}$$

In both equations, P is the number of precipitation forecasts, f_p is a single precipitation forecast, and o_p is the matching observed precipitation. RMSE places additional penalties on very large errors and is more sensitive to outliers.

Forecasts of fixed probability intervals are evaluated based on the proportion of samples that fall within the quantile intervals as well as the width of the quantiles. The relative frequency within each quantile should match the percentile range. In addition, the widths of the quantiles should also appropriately match the uncertainty of the predictions. If the forecasted quantiles cover a large range of the possible distribution of a particular forecasted quantity, then the forecast is not likely to be particularly useful.

Chapter 5

Model Verification

5.1 Probabilistic Forecasts

5.1.1 Attributes Diagrams

The attributes diagrams for each subdomain show how the machine learning algorithms have all greatly improved the raw ensemble forecast. Grid 1 (Fig. 5.1) has the lowest climatological probability of three and had the fewest high precipitation events (Fig. 3.2). Because of that, there was a relatively large "skill area" on the attributes diagram, so even the raw ensemble forecast was skilled although it still had a large under-forecasting tendency. The other two grids had slightly higher climatological probabilities (Fig. 5.2 and Fig. 5.3). The calibration logistic regression did decrease the magnitude of the under-forecasting but did not fully correct it. The other four machine learning models did produce nearly perfect reliability with their probabilistic forecasts. Their Brier Skill Scores were nearly identical within each sub-grid. The highest BSSs for the machine learning models and the lowest BSS for the raw ensemble were in grid 2 (Fig. 5.2). This large improvement likely occurred due to the wide range of precipitation events contained within the training data for this area. Using 50 versus 100 trees did not cause any significant change in the scores of

the random forest models, which is consistent with previous work with random forests.

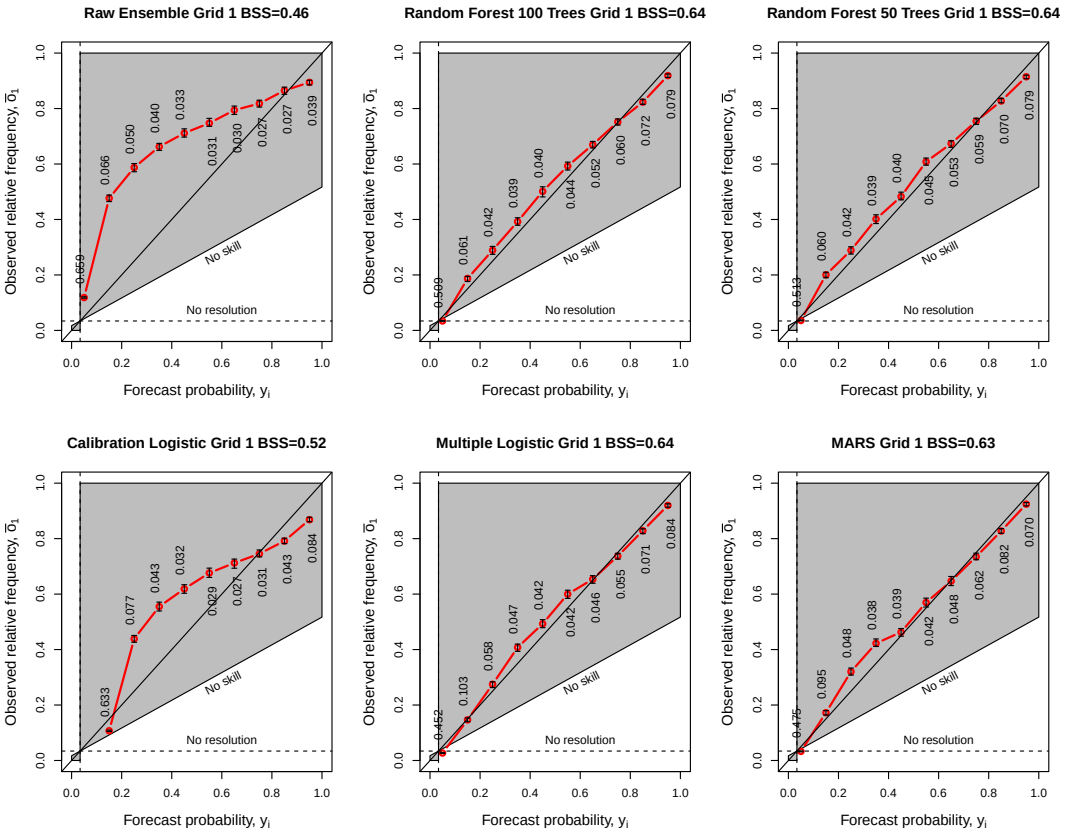

Figure 5.1: Attributes diagrams for the probability of 1-hour precipitation exceeding 0.25 mm from the raw ensemble and each machine learning model applied to grid 1.

Examination of the attributes diagrams for conditional probability of precipitation threshold exceedance shows smaller degrees of improvement than observed with probability of precipitation. In grid 1 (Fig. 5.4), the raw ensemble generally over predicts the probabilities, resulting in a large negative BSS. The machine learning models do correct for this, resulting in BSSs slightly above climatology. Random forests are best able to provide perfect reliability for the low

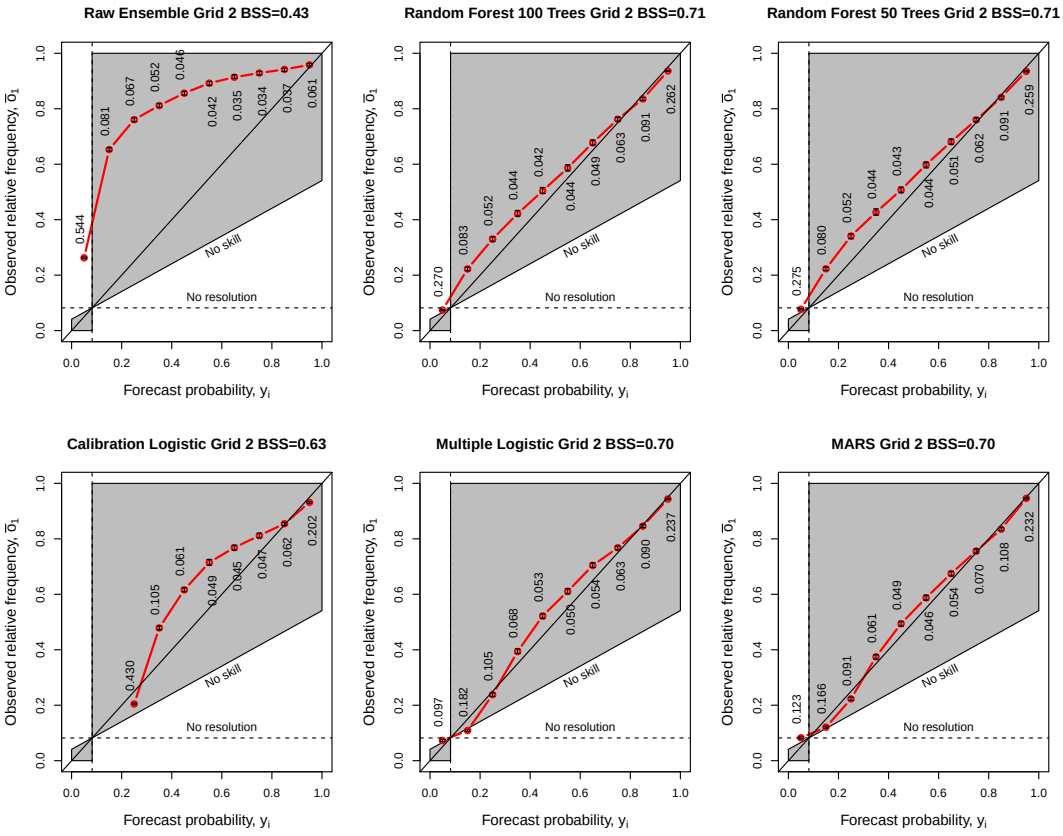

Figure 5.2: Attributes diagrams for the raw ensemble probabilities and each machine learning model applied to grid 2.

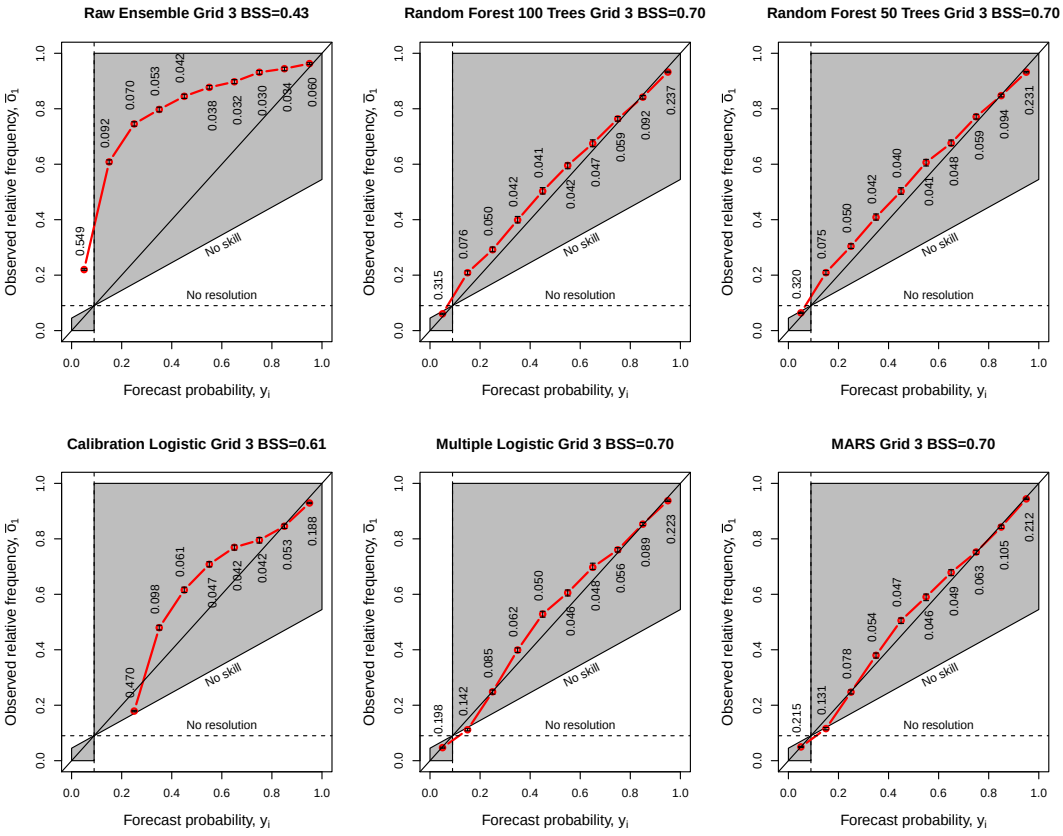

Figure 5.3: Attributes diagrams for the raw ensemble probabilities and each machine learning model applied to grid 3.

range of probabilities. In grid 2, the ensemble under-forecasts the lower probabilities and over forecasts the higher probabilities. Multiple logistic regression and MARS both perform better than random forest because their treatment of the lowest probabilities was more reliable. Random forest under-predicted them but did to a lesser extent than the raw ensemble.

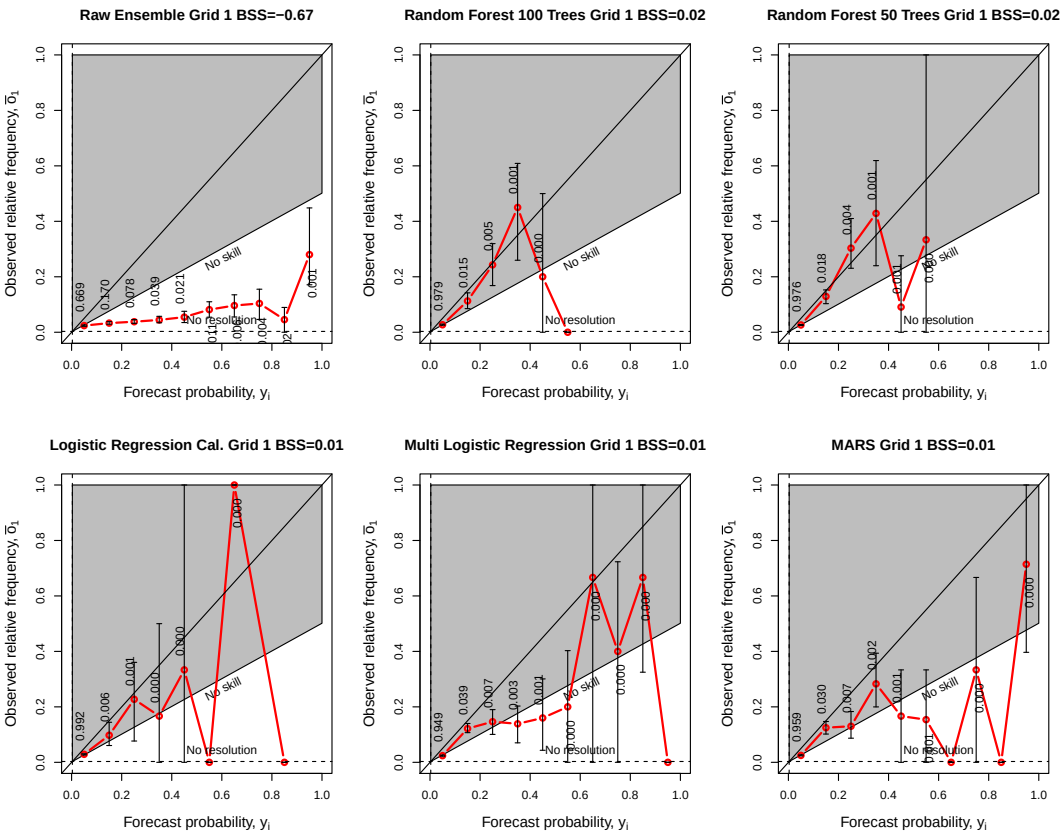

Figure 5.4: Attributes diagrams for the probability of 1-hour precipitation exceeding 6.35 mm from the raw ensemble and each machine learning model applied to grid 1.

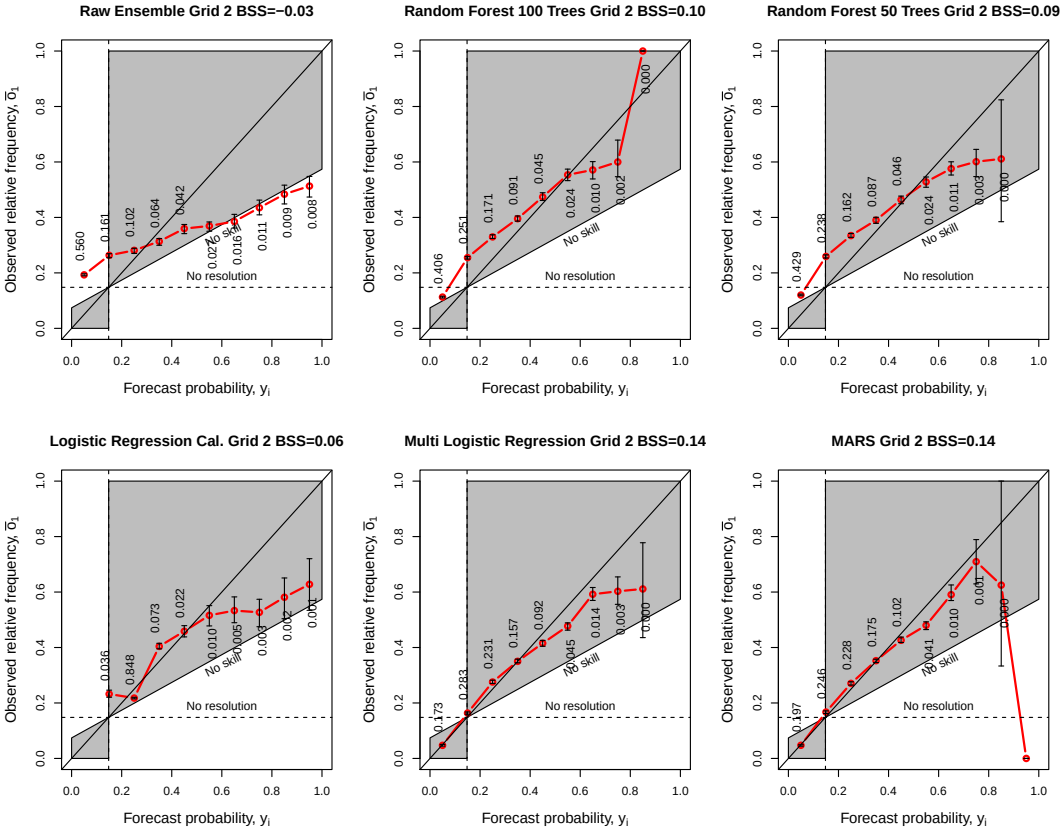

Figure 5.5: Attributes diagrams for the conditional probability of 1-hour precipitation exceeding 6.35 mm from the raw ensemble and each machine learning model applied to grid 2.

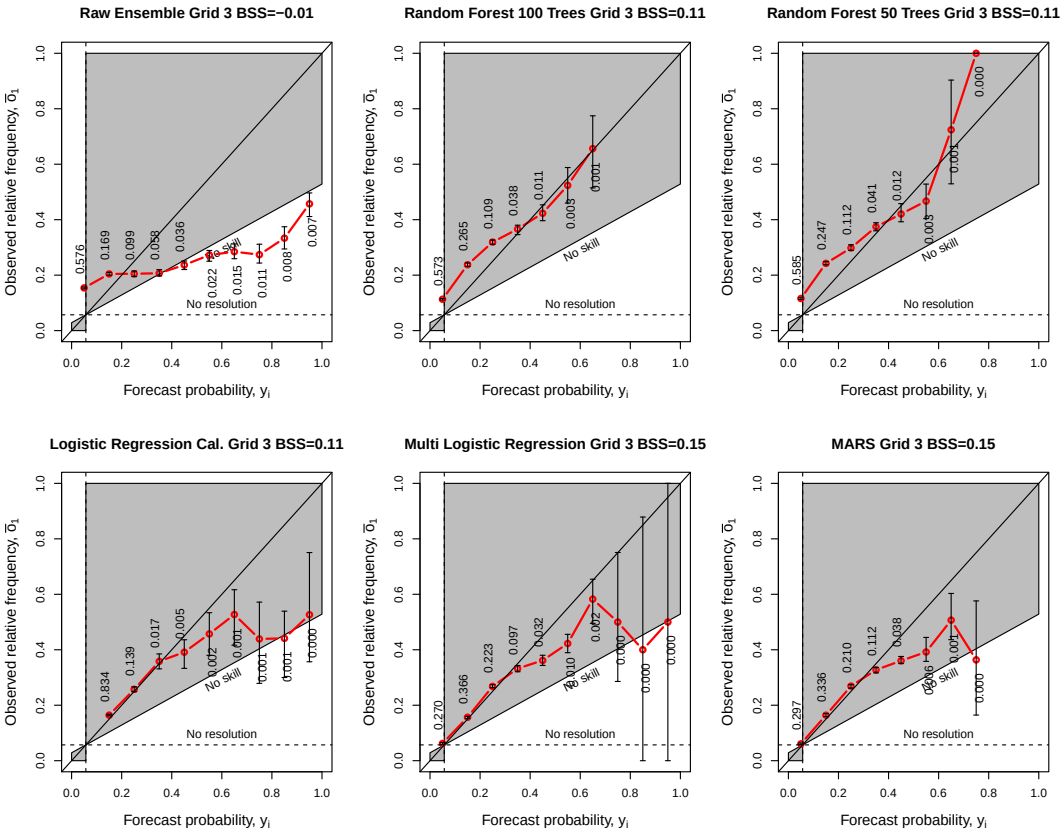

Figure 5.6: Attributes diagrams for the raw ensemble and each machine learning model applied to grid 3.

5.1.2 Forecast Hour Comparisons

Comparison of BSS by hour and by sub-grid show additional trends in the probability of precipitation forecasts. Within all three sub-grids, the raw ensemble has the worst BSS (Fig. 5.7). For all models, the best performance occurs at forecast hour 1 then decreases sharply until forecast hour 4 before stabilizing. This initial decrease is likely due to radar data assimilation, but that assimilation advantage appears to be quickly lost as the models spin up. There is a slight increase in performance between hours 6 and 12 for grid 2. This increase may be due to convection typically becoming more linear or dying during the early morning hours. There is another major decrease in performance in grid 2 between hours 18 and 24. This time period is when the greatest uncertainty exists due to convective initiation and the tendency for initial convection to be isolated. Both 100 and 50 tree random forests have slightly higher BSS than multiple logistic regression and MARS. The calibration logistic regression did still improve on the ensemble forecast but not to the same extent as the models that selected from the full set of model variables. The BSS hourly trends for the probability of exceeding 6.35 mm of precipitation forecasts follow similar patterns (Fig. 5.8), but the divide between multiple logistic regression and MARS versus random forests is consistent throughout all forecast hours. The ensemble forecast has negative BSS for most hours.

Examining AUC reveals how well the different algorithms distinguish precipitation and non-precipitation events over a range of probability thresholds. Fig. 5.9 shows how AUC varies by hour for each model and each sub-grid for the probability of precipitation forecasts. Random forests perform the best at most hours with multiple logistic regression and MARS providing comparable

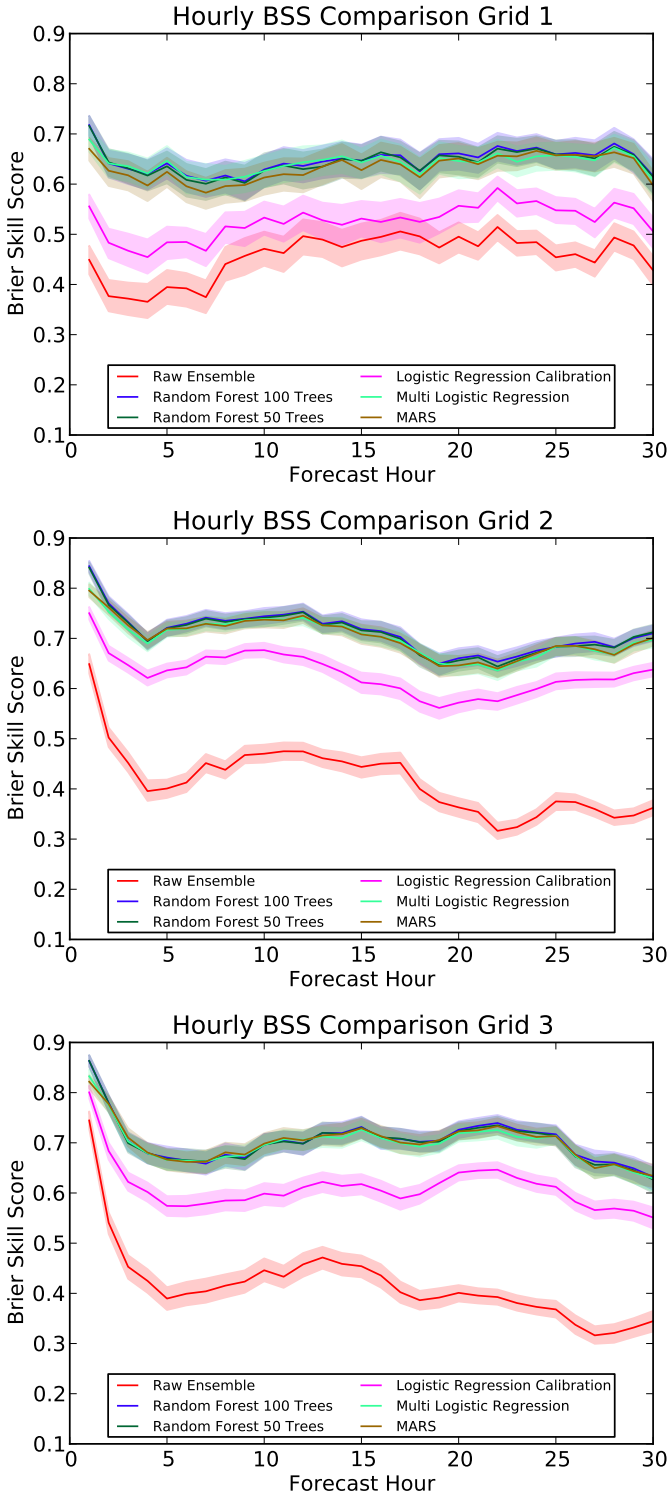

Figure 5.7: Brier Skill Score comparisons by hour for each model and each subgrid for probability of precipitation forecasts. The shaded area indicates the 95% bootstrap confidence interval around each value.

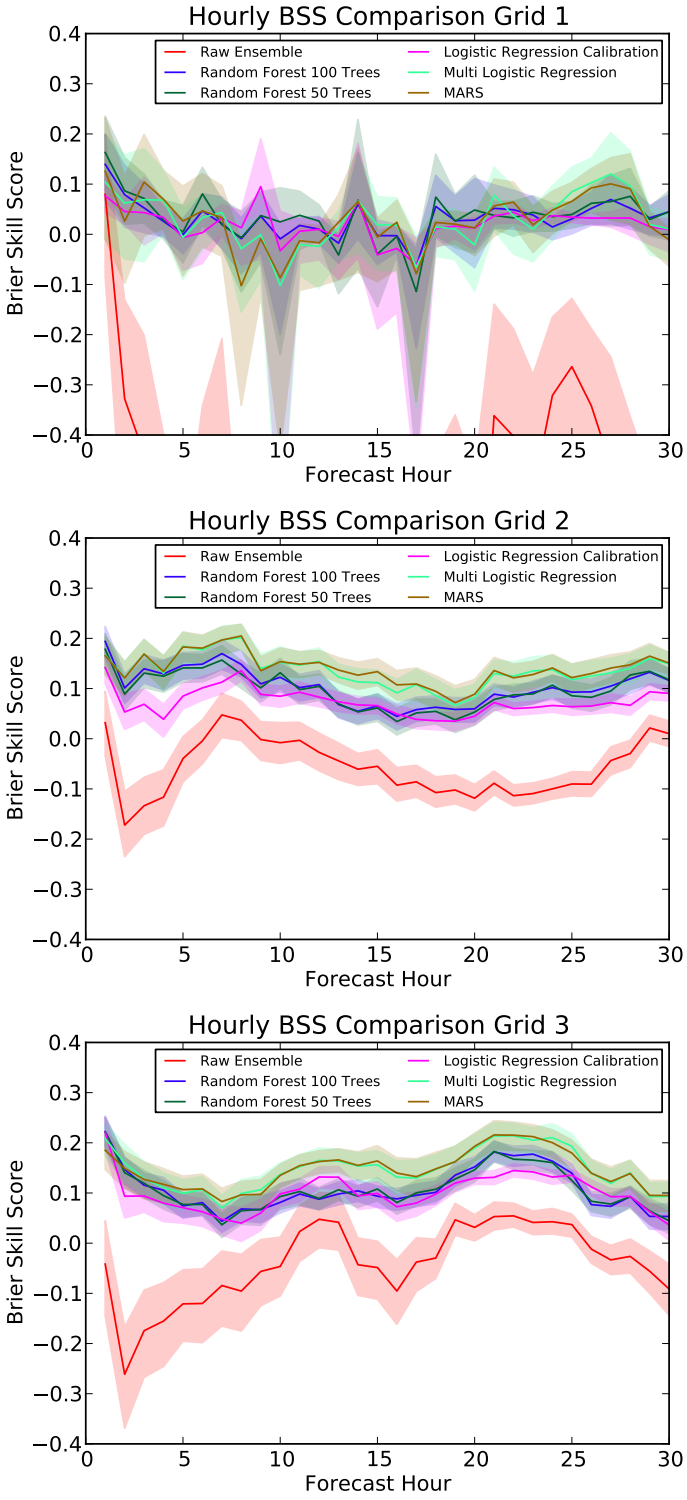

Figure 5.8: Brier Skill Score comparisons by hour for each model and each sub-grid for probability of precipitation exceeding 6.35 mm. The shaded area indicates the 95% bootstrap confidence interval around each value.

but slightly lower AUCs. The calibration logistic regression consistently has lower AUC than even the raw ensemble. The likely reason for the difference can be seen in the attributes diagrams (Fig. 5.4, 5.5, and 5.6). The probabilistic forecasts in the lowest bin for each sub-grid are consistently over-forecasted whereas with the raw ensemble they were under-forecasted. The other models provide nearly perfect reliability and a larger range of probabilities forecasted, so their higher AUCs are expected. Hourly trends in AUC are similar to those in BSS. There is a sharp decrease in AUC from hour 1 to 5 with a slight increase between 6 and 18 hours. There is another decrease between 18 and 24 hours but without the slight gain from 25 to 30 hours. In the 6.35 mm probability forecasts shown in Fig. 5.10, MARS and the multiple logistic regression generally have the highest AUC throughout all hours, although there is less of a discrepancy between them and random forest than in Fig. 5.8.

5.1.3 Optimal Threshold Verification

The choice of optimal threshold from the raw ensemble and each machine learning model is validated by evaluating the discrimination ability of the threshold with multiple binary verification scores. In Fig. 5.11, the optimal thresholds slightly increase with forecast hour. All of the machine learning thresholds fall within the 15 to 40 % range. The raw ensemble stays below 10 % for its threshold value, which is indicative of forecasting precipitation if at least one ensemble member is predicting it. PSS and ETS share similar hourly trends with AUC as the peak values are in the initial and 10 to 20 hour time frames. The multivariate machine learning models have improved performance compared to the raw

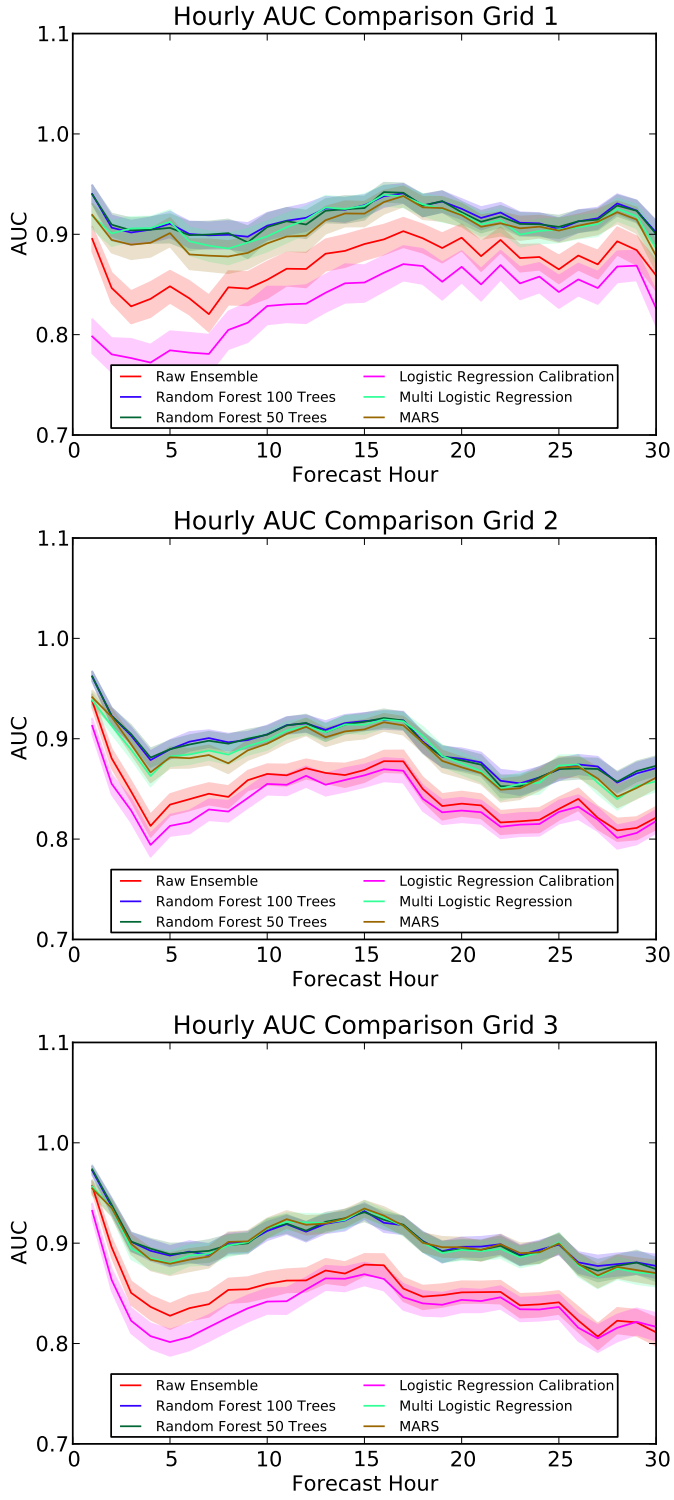

Figure 5.9: AUC comparisons by hour for each model and each sub-grid for probability of precipitation forecasts. The shaded area indicates the 95% bootstrap confidence interval around each value.

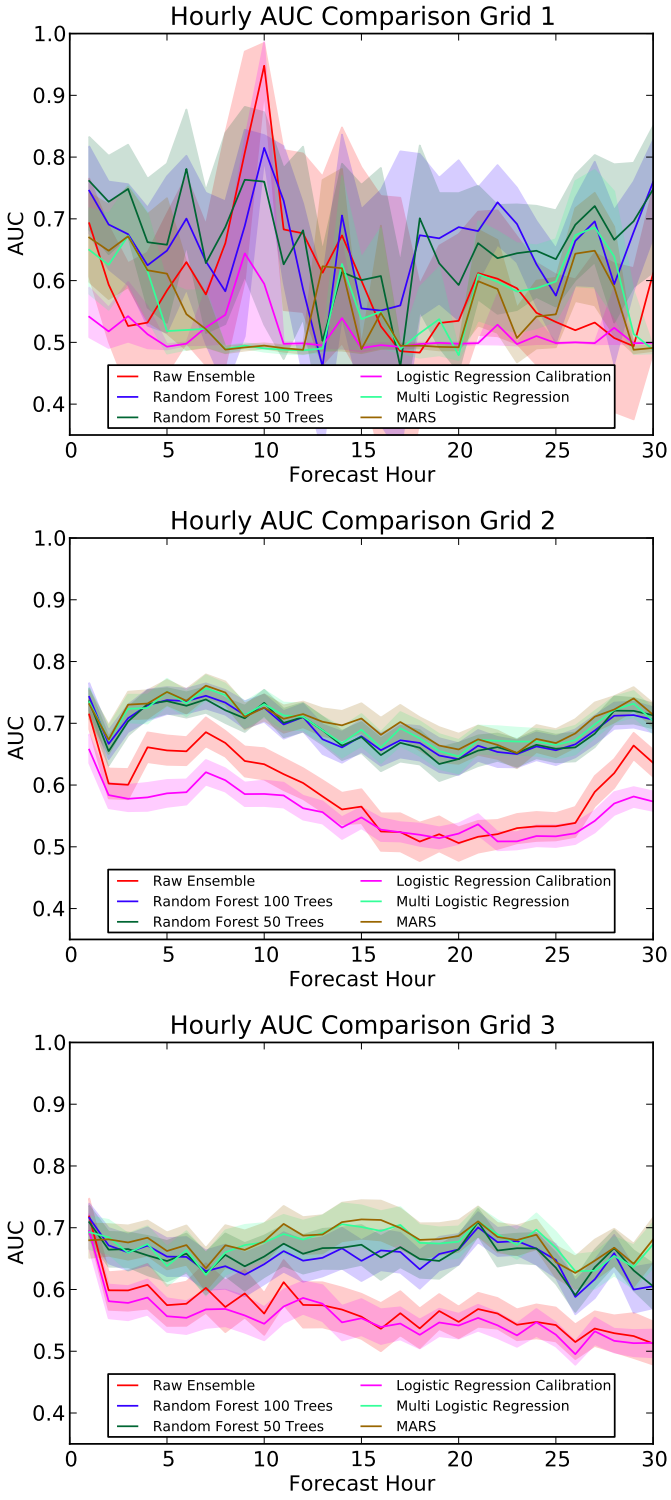

Figure 5.10: AUC comparisons by hour for each model and each sub-grid for probability of precipitation exceeding 6.35 mm forecasts. The shaded area indicates the 95% bootstrap confidence interval around each value.

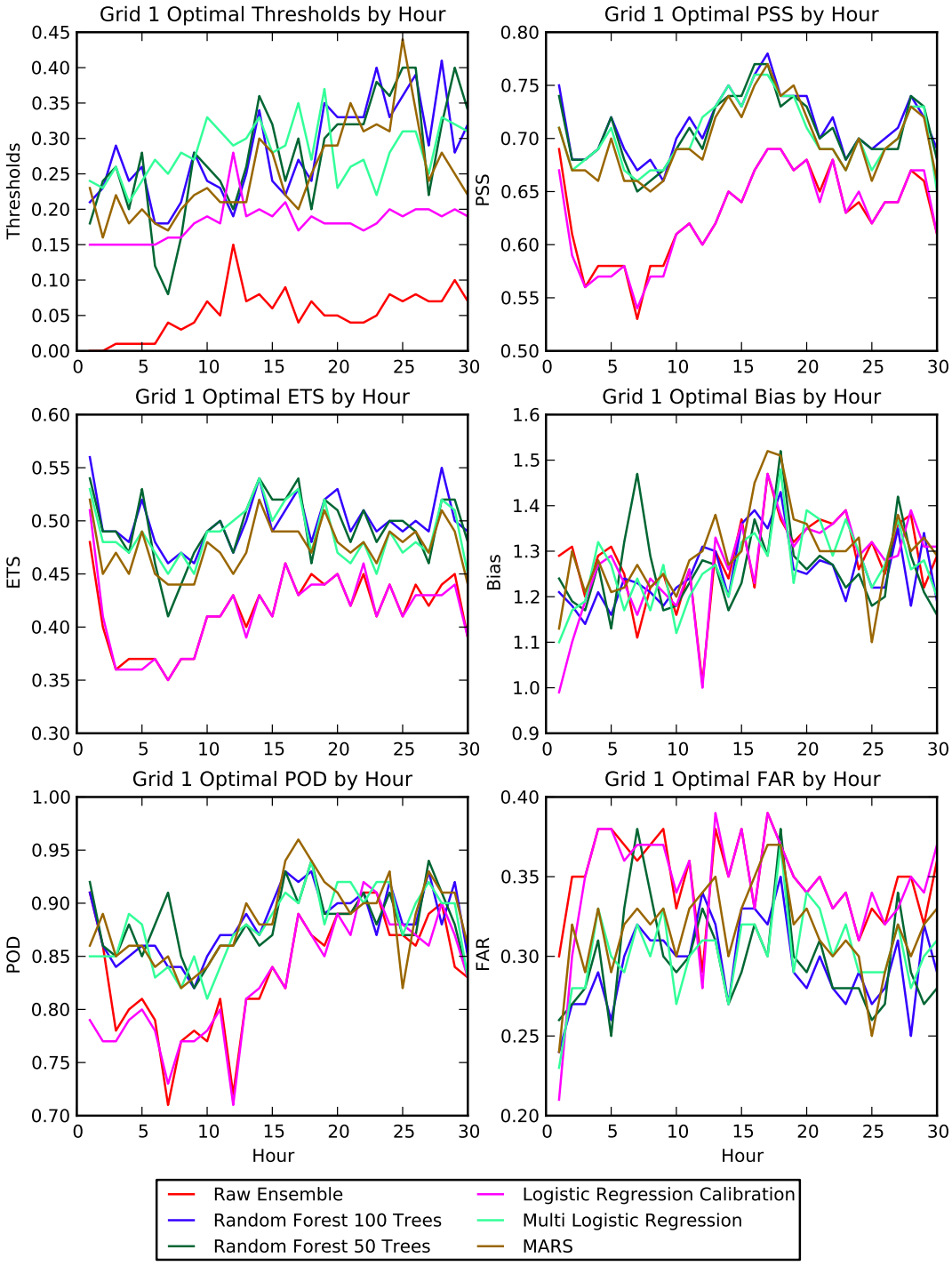

Figure 5.11: Optimal thresholds and verification statistics associated with that threshold for the raw ensemble and each machine learning model in grid 1.

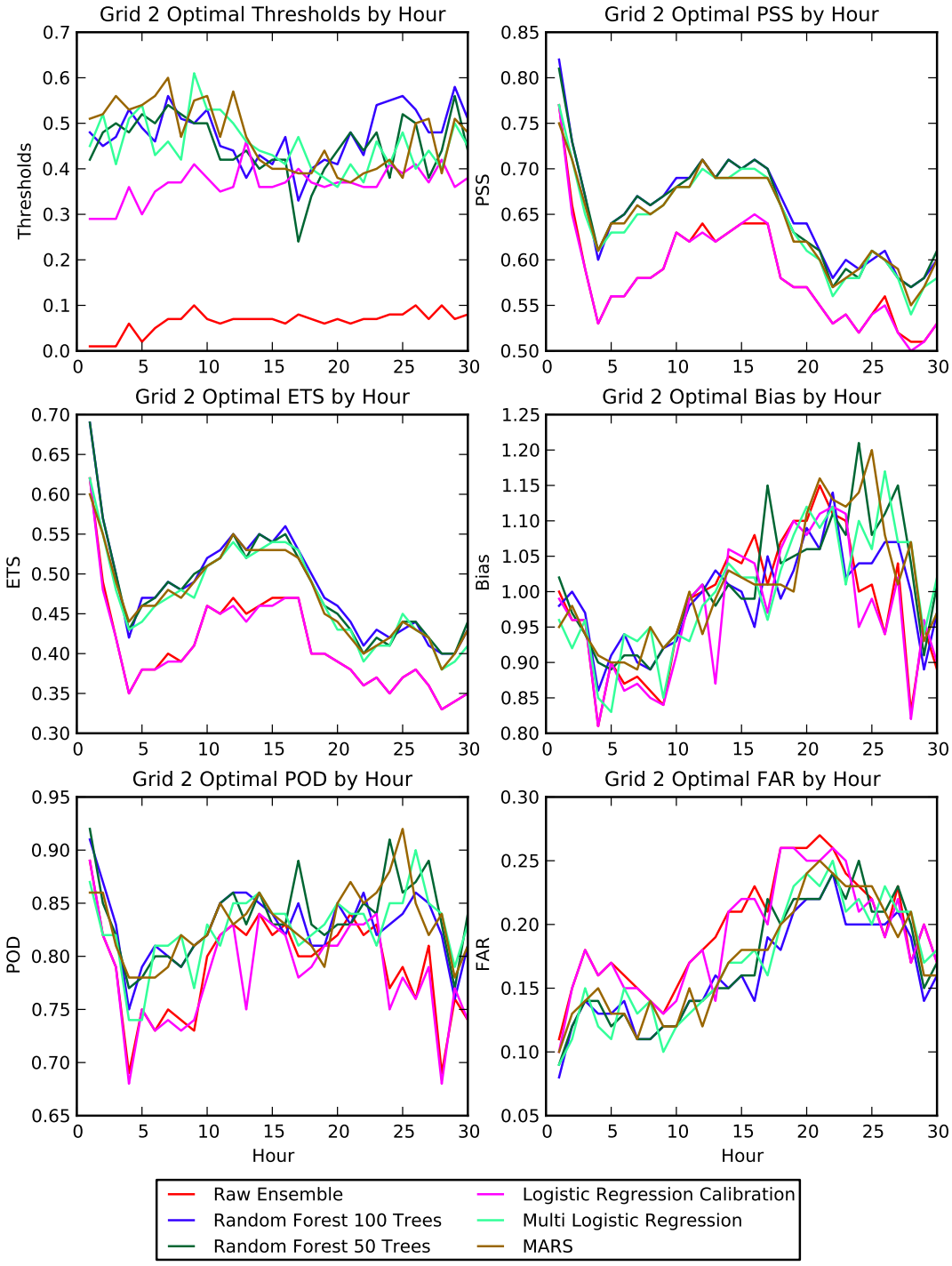

Figure 5.12: Optimal thresholds and verification statistics associated with that threshold for the raw ensemble and each machine learning model in grid 2.

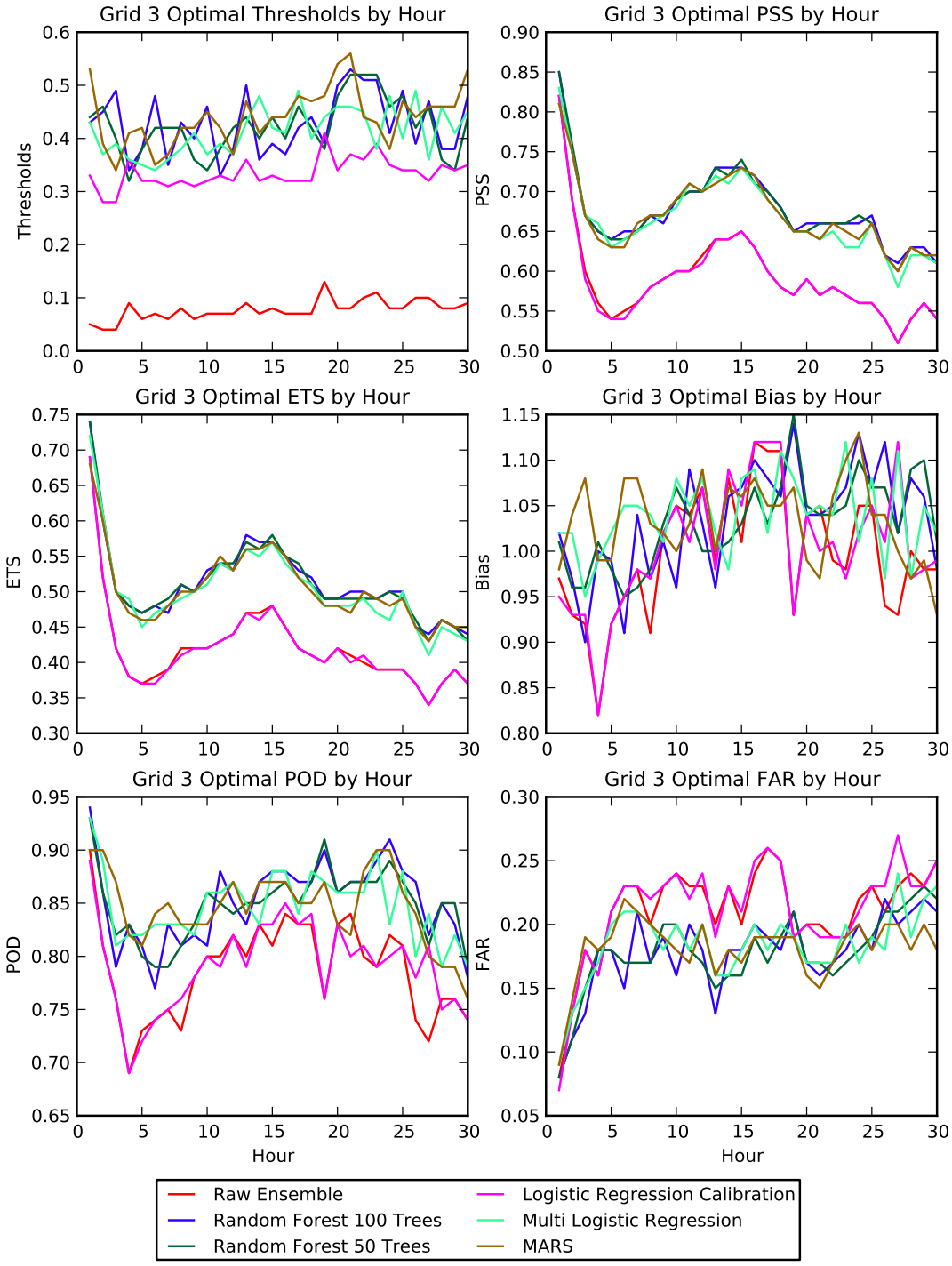

Figure 5.13: Optimal thresholds and verification statistics associated with that threshold for the raw ensemble and each machine learning model in grid 3.

ensemble and the calibration logistic regression. All of the models show a bias score greater than 1, which indicates that the models are all over forecasting the precipitation area. Because of this over forecasting bias, POD is very high, but FAR is also relatively larger than for the other sub grids.

The verification statistics for the grid 2 are shown in Fig. 5.12. The thresholds are similar for all of the multivariate machine learning models and generally range between 0.4 and 0.6. The raw ensemble again stays below 0.1. The PSS and ETS are both highest at hour 1 with sharp decreases until hour 4 before peaking locally near hour 18 and decreasing for the rest of the period. All of the models showed no bias initially then an under forecasting bias until hour 18, and an over forecasting bias after that. The over forecasting bias resulted in a significant jump in FAR although POD was not significantly affected. Grid 3 shows similar trends (Fig. 5.13) although the over forecasting bias is more consistent as well as the trends in POD and FAR. Similar results to these are found in the SSEF verification from Kong et al. (2011) although the ETS is lower in that paper due to that study verifying against all grid points and including points with no radar coverage.

5.1.4 Forecast Day Comparisons

Skill scores for each day varied greatly among the different models. Since the model skill varies in similar fashion from day to day, the most likely cause of variance in skill is the distribution of precipitation events. Fig. 5.14 displays the distribution of BSS vs. the mean and standard deviation of the observed precipitation over each region. In most cases, the BSS increases when the mean precipitation and the standard deviation of the precipitation increase. The

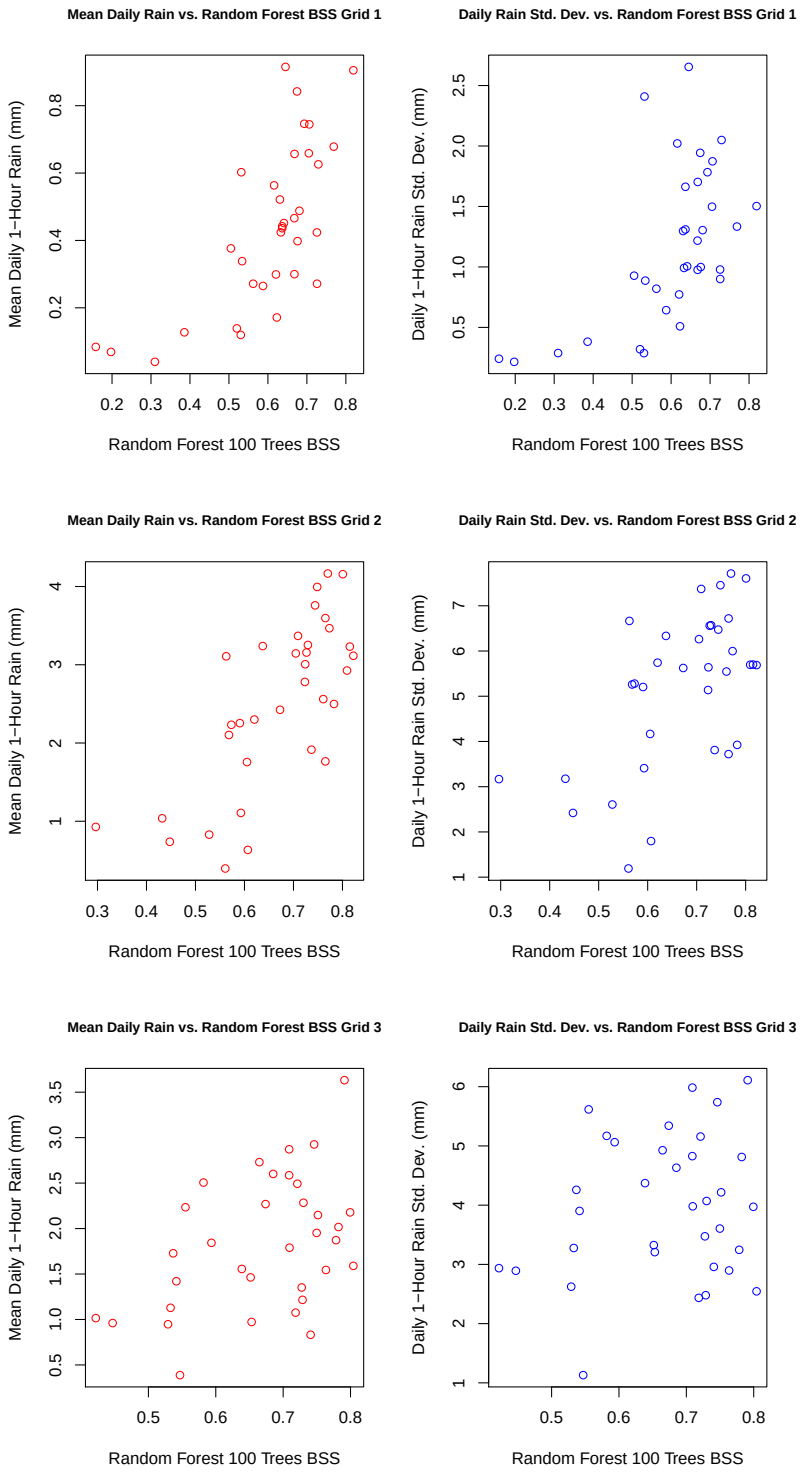

Figure 5.14: Plot of the mean and standard deviation of the 1-hour precipitation of the samples from each SSEF run vs. the BSS for each of those runs.

differences are most pronounced in grid 2 where all of the days with a mean 1-hour precipitation less than 1 mm also have a BSS less than 0.6. Grid 3 exhibits the weakest correlation, but the BSSs are all above 0.4. The connection between low rainfall and low BSS is probably due to the lack of higher probabilities being forecasted within that sample, which would decrease the resolution term (Eqn. 4.3).

5.1.5 Variable Importance

Variable importance z-scores were calculated and averaged over each random forest and sub-grid to determine if the random forests were choosing relevant variables and how the choice of variables was affected by region. Variable importance is only indicative of how the changing the ordering of the distribution of each variable affects the random forest performance. It accounts for how often a variable is used in the model, the depth of the variable in the tree, and the number of cases that transit through the branch containing that variable, but the importance score cannot be decomposed into those factors. The top 10 variable importance z-scores for the 100-tree random forests predicting the probability of precipitation are shown in Table 5.1. Not included among the top ten most important variables for any of the sub-grids are any of the variables describing the ensemble precipitation forecasts. More emphasis is placed on the 700 and 500 mb winds, heights, temperatures, and surface pressures than on the actual precipitation forecasts. The precipitable water, surface dewpoint, composite reflectivity, and SBCIN are also considered highly important in at least one of the sub-grids. The mean quantities were also generally more important than the standard deviations. Most of the variables selected highly are consistently

Table 5.1: Top ten variables ranked by the importance z-score along with mean and standard deviation of the decrease in accuracy from 100 tree random forests predicting probability of precipitation.

Variable	Z-score	Mean	Std. Error
Grid 1			
RQI	728.14	0.6021	0.0008
Max Reflectivity Mean	591.45	0.5541	0.0009
Precipitable Water Mean	571.45	0.5926	0.0010
MSLP Mean	491.79	0.5920	0.0012
Dewpoint 2 m Mean	381.57	0.5588	0.0015
V 500 mb Mean	377.65	0.5807	0.0015
Height 700 mb Mean	330.32	0.5739	0.0017
Temperature 2 m Std. Dev.	329.21	0.5526	0.0017
Precipitable Water Std. Dev.	321.71	0.5567	0.0017
V 700 mb Mean	315.57	0.5621	0.0018
Grid 2			
Temperature 700 mb Mean	1154.49	0.4424	0.0004
Height 700 mb Mean	1003.93	0.4343	0.0004
MSLP Mean	971.64	0.4354	0.0004
U 700 mb Mean	895.25	0.4343	0.0005
V 500 mb Mean	780.43	0.4324	0.0006
Precipitable Water Mean	736.19	0.4313	0.0006
SBCIN Mean	702.70	0.4300	0.0006
Temperature 2 m Std. Dev.	696.69	0.4258	0.0006
Dewpoint 2 m Std. Dev.	639.61	0.4227	0.0007
Precipitable Water Std. Dev.	618.40	0.4222	0.0007
Grid 3			
Precipitable Water Mean	951.44	0.4751	0.0005
V 700 mb Mean	910.62	0.4818	0.0005
V 500 mb Mean	825.60	0.4783	0.0006
U 700 mb Mean	697.78	0.4751	0.0007
Temperature 700 mb Mean	695.39	0.4763	0.0007
MSLP Mean	680.28	0.4751	0.0007
Composite Reflectivity Mean	671.66	0.4662	0.0007
Temperature 2 m Std. Dev.	468.08	0.4575	0.0010
Height 700 mb Std. Dev.	459.82	0.4604	0.0010
Height 700 mb Mean	457.27	0.4697	0.0010

Table 5.2: Top ten variables ranked by the importance z-score along with mean and standard deviation of the decrease in accuracy from 100 tree random forests predicting probability of precipitation exceeding 6.35 mm.

Variable	Z-score	Mean	Std. Error
Grid 1			
Temperature 2 m Std. Dev.	103.57	0.6336	0.0061
Max Downward Vertical Velocity Mean	95.78	0.6422	0.0067
V 700 mb Std. Dev.	95.56	0.6427	0.0067
V 500 mb Mean	91.09	0.6360	0.0070
U 700 mb Mean	88.00	0.6788	0.0077
Temperature 700 mb Std. Dev.	77.14	0.5975	0.0077
Precipitable Water Std. Dev.	76.96	0.6457	0.0084
Temperature 700 mb Mean	75.51	0.6193	0.0082
Dewpoint 2 m Std. Dev.	73.88	0.6302	0.0085
Max Upward Vertical Velocity Std. Dev.	71.35	0.6035	0.0085
Grid 2			
V 700 mb Mean	405.70	0.5655	0.0014
V 500 mb Mean	337.18	0.5523	0.0016
Temperature 700 mb Mean	318.23	0.5533	0.0017
Precipitable Water Mean	304.57	0.5531	0.0018
MSLP Mean	281.48	0.5492	0.0020
Max Upward Vertical Velocity Mean	270.98	0.5185	0.0019
Max Downward Vertical Velocity Mean	268.19	0.5241	0.0020
U 700 mb Mean	267.67	0.5450	0.0020
Height 700 mb Mean	266.23	0.5456	0.0020
Max Downward Vertical Velocity Std. Dev.	236.13	0.5189	0.0022
Grid 3			
U 700 mb Mean	315.94	0.6178	0.0020
Temperature 700 mb Mean	278.79	0.6146	0.0022
MSLP Mean	260.53	0.6046	0.0023
Composite Reflectivity Mean	257.19	0.6046	0.0024
Precipitable Water Mean	242.52	0.5900	0.0024
Max Downward Vertical Velocity Mean	240.27	0.6037	0.0025
V 500 mb Mean	234.92	0.5834	0.0025
MSLP Std. Dev.	234.05	0.6154	0.0026
Height 700 mb Std. Dev.	228.56	0.5864	0.0026
Max Upward Vertical Velocity Std. Dev.	217.54	0.5599	0.0026

higher near low pressure centers and in the warm sectors. The precipitation-related variables received a low z-score because while some of them may have had a high mean importance score, the standard error of their scores was significantly larger. In many cases the permutation of the precipitation values helped nearly as often as it hurt the predictions. This result is likely due to the spatial and temporal errors in the ensemble precipitation forecasts causing very large errors if the model did not put the precipitation in the right place. The most important variables all have smoother continuous fields that are not as subject to the nonlinear variations of precipitation. For convective precipitation CIN is a good indicator of where storms will not form. RQI was the most important variable in grid 1 most likely due to the variability of radar coverage in the mountain ranges of the western US. While the mean z-scores are fairly similar for the variables in each sub-grid, the standard errors increased with decreasing rank, leading to the wider range of overall z-scores across the 34 forests.

The variable importance scores for the 100-tree random forests predicting the conditional probability of 1-hour precipitation exceeding 6.35 mm are shown in Table 5.2. Many of the same variables have been selected for the top ten, but there are some notable contrasts between Tables 5.2 and 5.1. The are many more standard deviation variables being selected, which is likely due to the probability for larger precipitation being higher in areas with more uncertainty about their forecasts. Grid 2 included more vertical velocity variables, grid 1 had the largest increase in spread variables, and grid 3 includes composite reflectivity as well as vertical velocities. The vertical velocity variables are indicative of likely locations for thunderstorms and associated heavy precipitation. As compared across the variable importance z-scores from all 34 forests, all variables have statistically significant importance ($p < 0.01$).

5.2 Deterministic Forecasts

The anomaly prediction method for deterministic forecasts produced improved forecasts for all three sub-grids. The variation of the mean error by hour (Fig. 5.15) shows how successfully each model corrected the inherent bias of the ensemble mean. The random forest, MARS, and both linear regression models eliminated the bias from the raw ensemble mean, which had a consistent negative bias. The quantile regression and quantile regression forest have a smaller mean error than the raw ensemble mean, which is due to the fact that they are solving for the median of the distribution while the random forest, MARS, and linear regressions all converge toward the mean. Since the distribution is skewed by a few outlier heavy precipitation events, the median will generally be smaller than the mean.

The Root Mean Squared Errors (RMSE) for each model are shown in Fig. 5.16. Again, all of the models consistently improve on the raw ensemble. Random forests and MARS have the lowest RMSE consistently, but MARS has a greater tendency to predict extreme values, which results in the large jumps in RMSE at some hours. Random forests are a non-parametric model, and its predictions are not directly affected by the values of predictors, so it is less prone to extreme errors. Both linear regressions and the quantile regression forest had similar RMSE values while quantile regressions were slightly worse. The additional quantile information about the precipitation distribution did not appear to have any impact on the linear regression predictions since the calibration and 3 value linear regressions have the same ME and RMSE throughout all hours. The RMSEs were all lowest in grid 1 and highest in grid 2 due to the widest

range of precipitation values being in grid 2. Unlike the probabilistic forecasts, the hourly trends of the deterministic machine learning models closely matched the raw ensemble because the final predictions for the deterministic models are adjusted from the ensemble mean.

Tendencies in the machine learning models can be examined more closely by viewing the full distribution of the error. Due to the large number of sample points, a traditional scatterplot would not reveal a significant amount of information, so a 2-dimensional histogram with 1 mm bins was used instead.

5.3 Interval Forecasts

The quantile regression and quantile regression forecasts produced forecasts for a fixed probability width around the deterministic forecast. The 5th and 95th percentiles were used to produce a 90% confidence interval for the precipitation forecast. The relative frequency of the confidence intervals is shown in Table 5.3. In all three grids, the majority of the rain events were outside the ensemble confidence interval. Both Quantile Regressions and Quantile Regression Forests corrected the distribution so that the interval contained 90% of the cases. Quantile Regression in all three grids provided a more balanced interval than the Quantile Regression Forest. Quantile Regression Forest still had a slight low bias in the distribution.

The best fixed probability interval forecasts should minimize the widths of their uncertainty forecasts while still including the expected relative frequency associated with that probability range. The mean and standard deviation of the full 90% probability interval for quantile regressions and quantile regression forests is shown in Fig. 5.17. Quantile regression forests have higher

mean widths and smaller standard deviations than quantile regressions, but the standard deviation varies more with quantile regression forests. The quantile regression mean fluctuates more by hour and is highest at the beginning of the forecast and at 24 hours, which correspond to where the probabilistic forecasts had the lowest BSSs (Fig. 5.8). The largest intervals are in grid 2, which is expected due to the larger range of precipitation values. Fig. 5.18 and 5.19 show how the upper and lower halves of the full interval vary by hour. The upper half of the interval contains most of the range on average. The quantile regression forest has a slightly smaller lower interval for all hours and a larger upper interval for most hours. This unevenness correlates with the slight unbalance seen in the results from Table 5.3. In this instance, the quantile regression provides a more balanced uncertainty forecast that varies on average with the uncertainty of the ensemble.

Table 5.3: Percentage of samples that fall within each percentile range for the SSEF, Quantile Regression, and Quantile Regression Forest.

Grid	Model	<5	5-50	50-95	>95
1	Raw Ensemble	2.3	12.5	36.4	48.9
	Quantile Regression Forest	1.6	49.2	42.9	6.3
	Quantile Regression	5.0	44.9	44.9	5.2
2	Raw Ensemble	1.2	7.6	34.0	57.3
	Quantile Regression Forest	3.2	47.4	43.7	5.7
	Quantile Regression	5.2	44.7	45.0	5.1
3	Raw Ensemble	1.5	8.2	34.7	55.5
	Quantile Regression Forest	2.9	46.8	44.0	6.3
	Quantile Regression	5.2	44.8	44.9	5.1

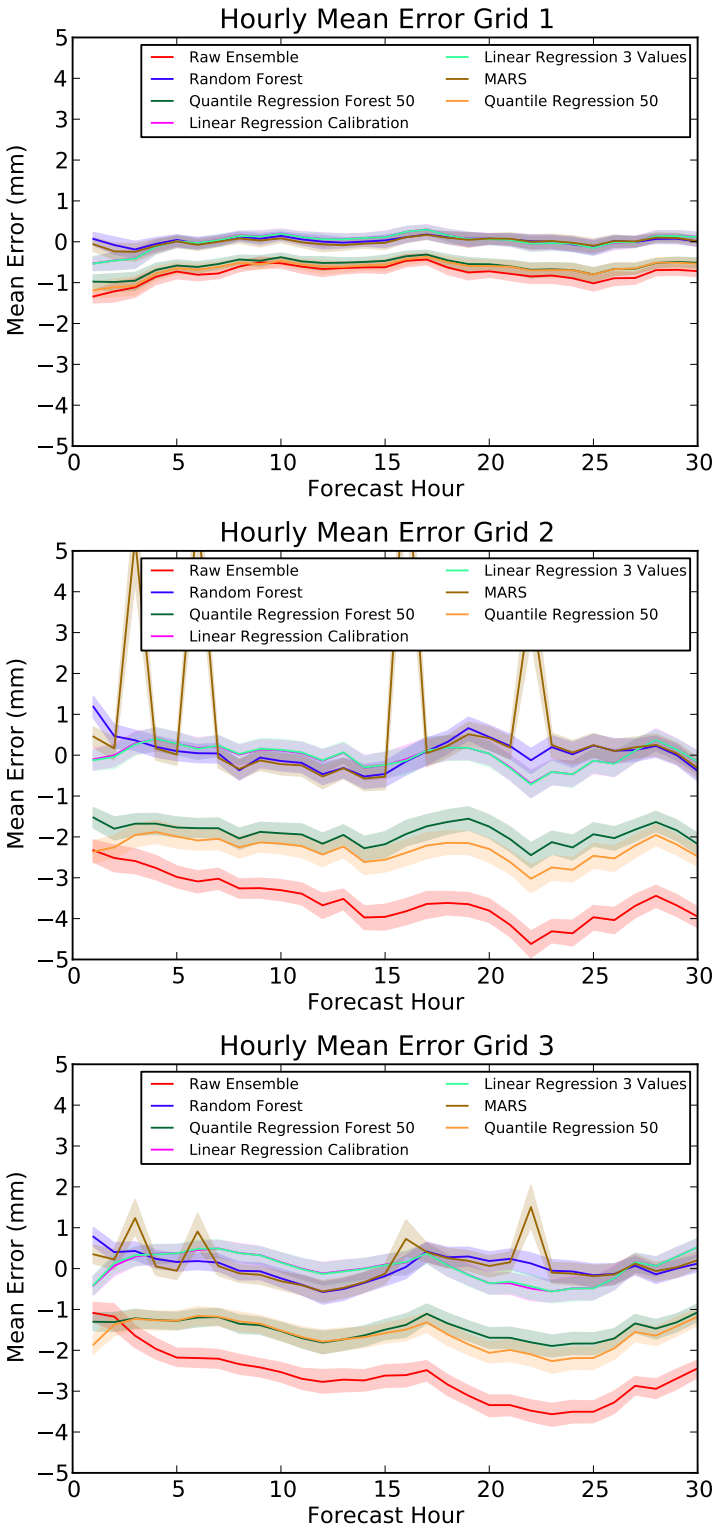

Figure 5.15: Comparison of Mean Error (ME) aggregated by forecast hour for each machine learning model being evaluated. The shaded area indicates the 95% bootstrap confidence interval around each value.

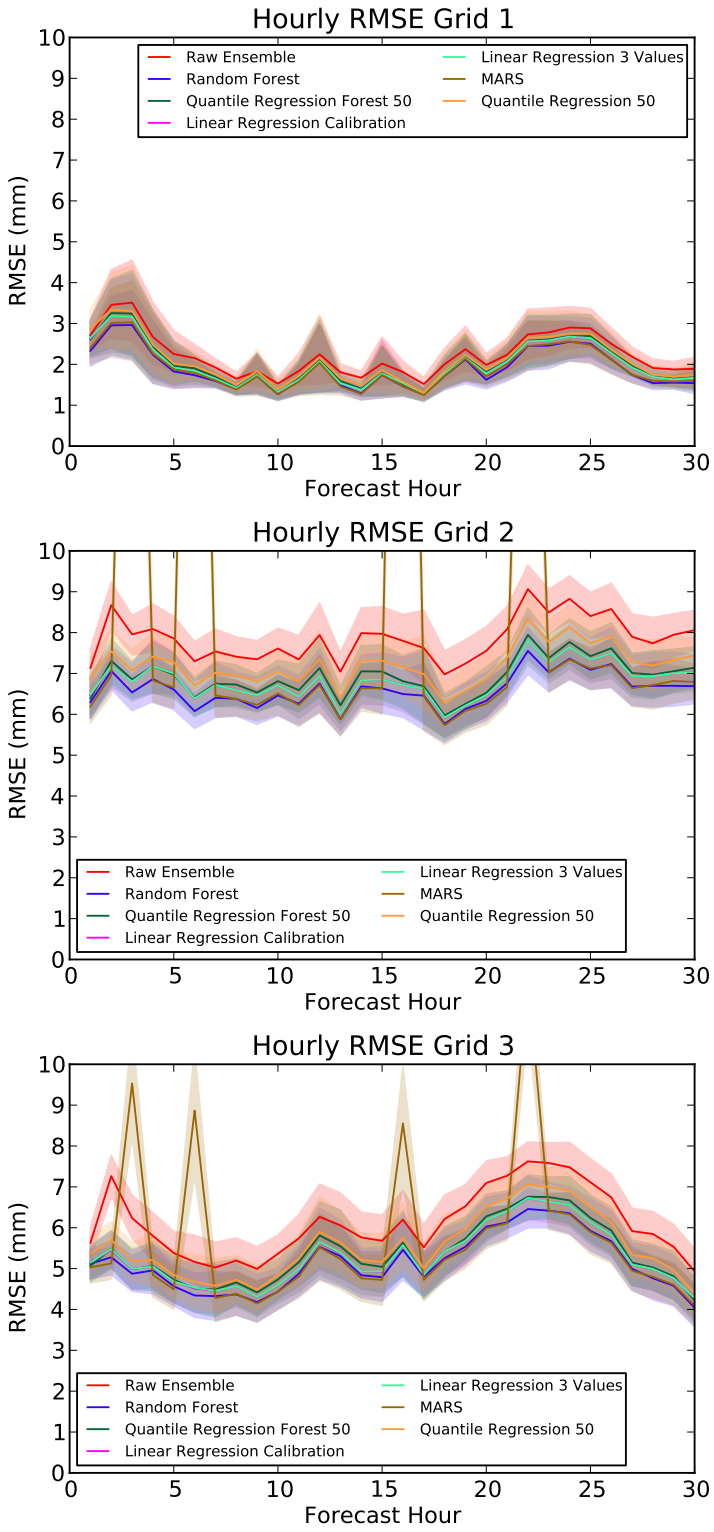

Figure 5.16: Comparison of Root Mean Squared Error (RMSE) aggregated by forecast hour for each machine learning model being evaluated. The shaded area indicates the 95% bootstrap confidence interval around each value.

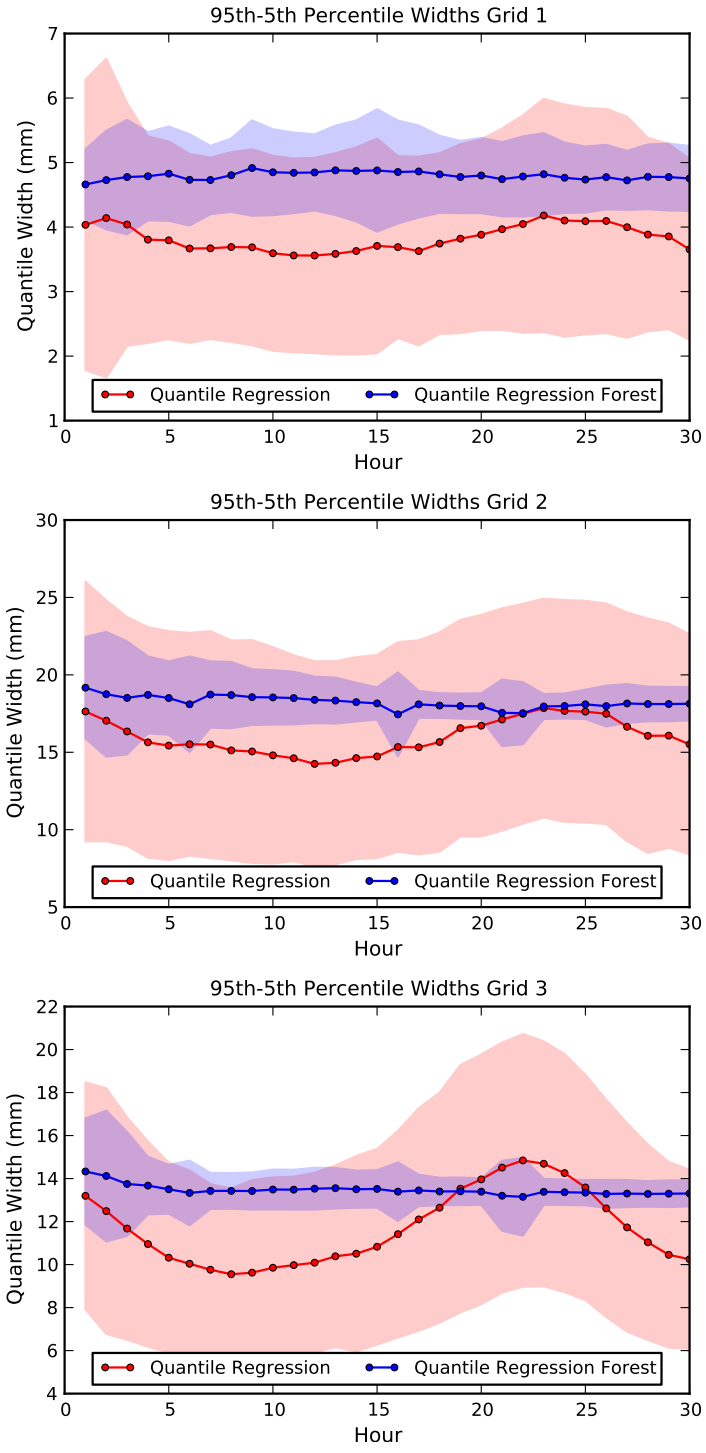

Figure 5.17: Variability of the mean widths of the 95th to 5th percentile fixed probability intervals by hour for the quantile regression and quantile regression forest. The shaded areas correspond to 1 standard deviation on each side of a point.

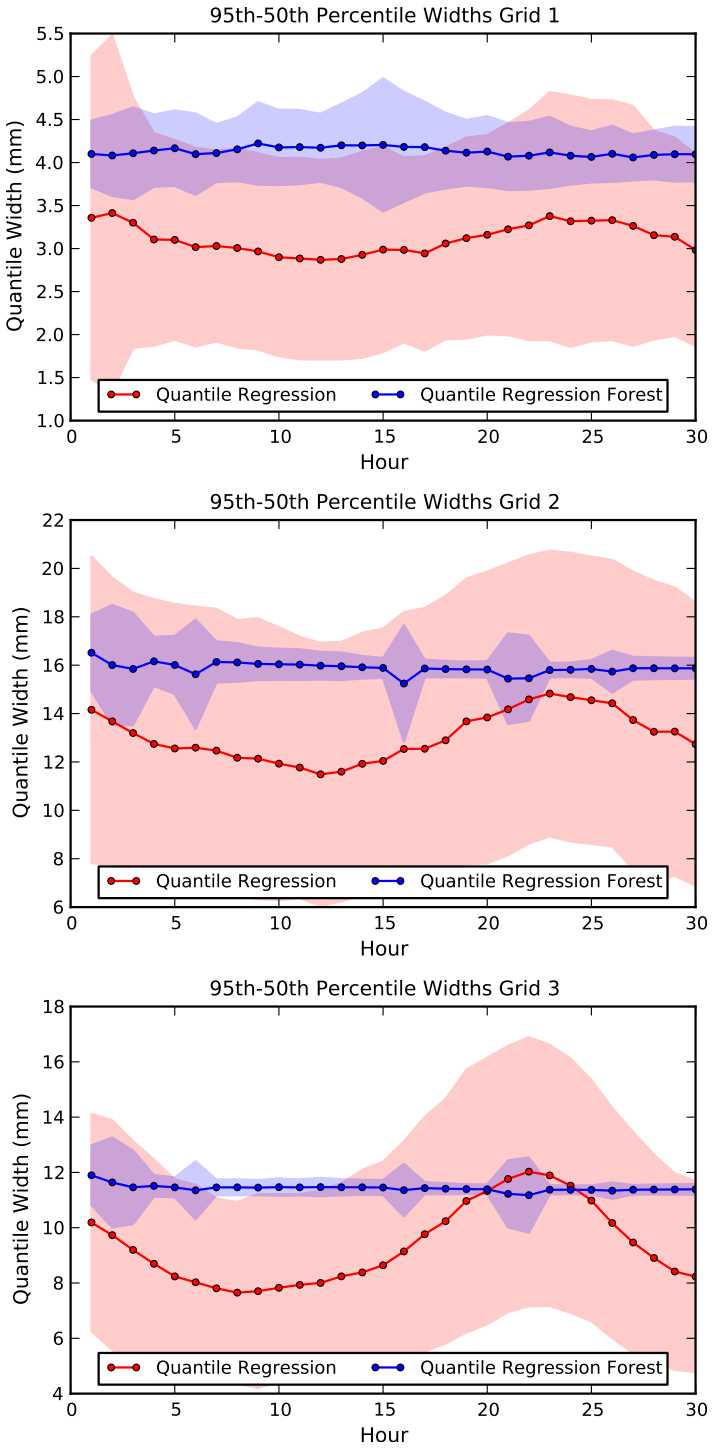

Figure 5.18: Variability of the mean widths of the 95th to 50th percentile fixed probability intervals by hour for the quantile regression and quantile regression forest. The shaded areas correspond to 1 standard deviation on each side of a point.

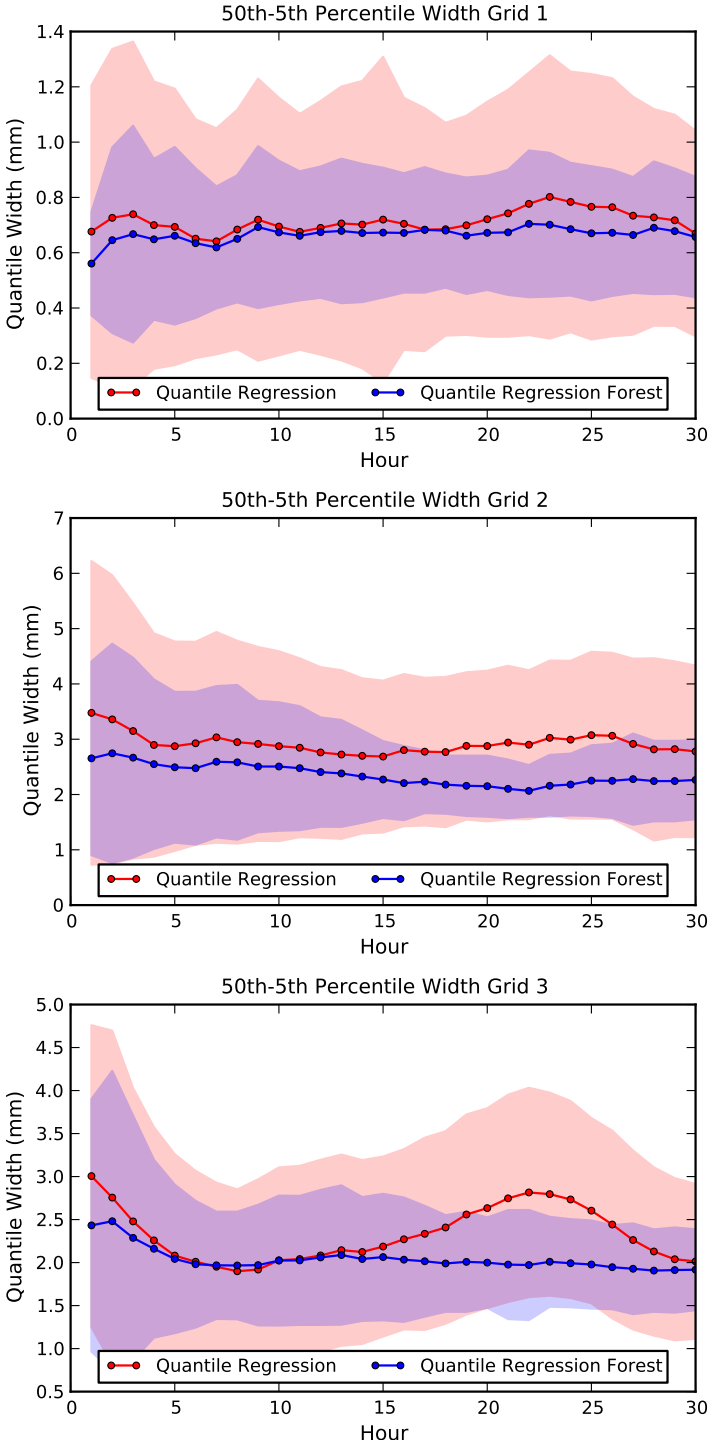

Figure 5.19: Variability of the mean widths of the 50th to 5th percentile fixed probability intervals by hour for the quantile regression and quantile regression forest. The shaded areas correspond to 1 standard deviation on each side of a point.

Chapter 6

Case Study

The objective verification analysis of the predictions from the machine learning models demonstrated that the algorithms all provided improvements to the ensemble precipitation forecasts on a general basis. That analysis gave a thorough look at many aspects of the forecasts as a whole, but it provided limited to no information about other crucial aspects of the forecast product. Since the forecasts and the verification focused on individual grid points, it did not take any spatial correlations of the error into account beyond differences among the three sub grids. It also could not show the smoothness of the forecasts or the spatial concentration of errors. This section will attempt to address these limitations by providing a subjective analysis of the forecasts from the machine learning algorithms for a case day from the study period.

6.1 19 May 2010

6.1.1 Synoptic Setup

19 May 2010 provides a test of the algorithm on multiple forms of precipitation across all three sub grids. The 500 mb maps for 1200 and 0000 UTC show three main troughs (Fig. 6.1). One is an intensifying trough coming ashore

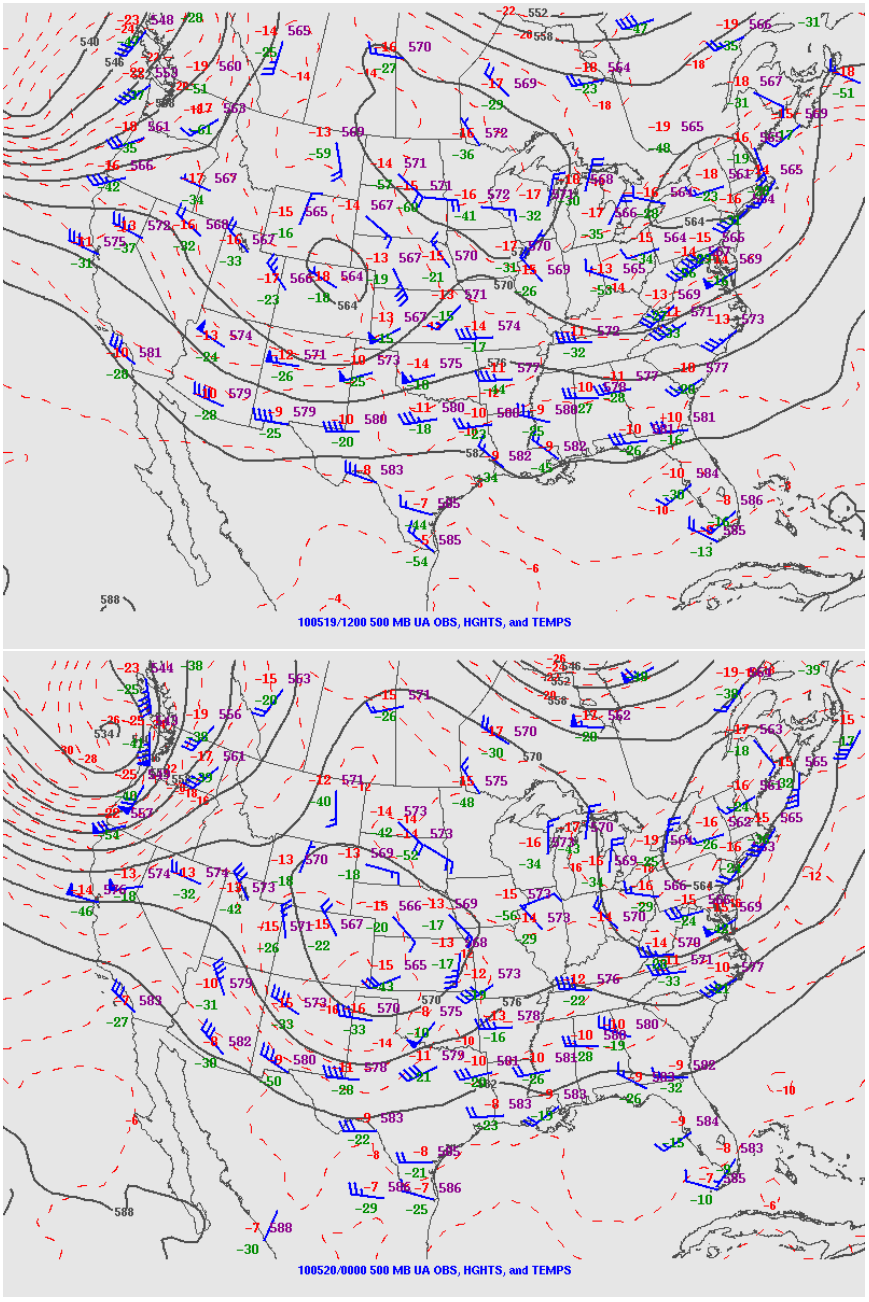

Figure 6.1: SPC 500 mb analysis for 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

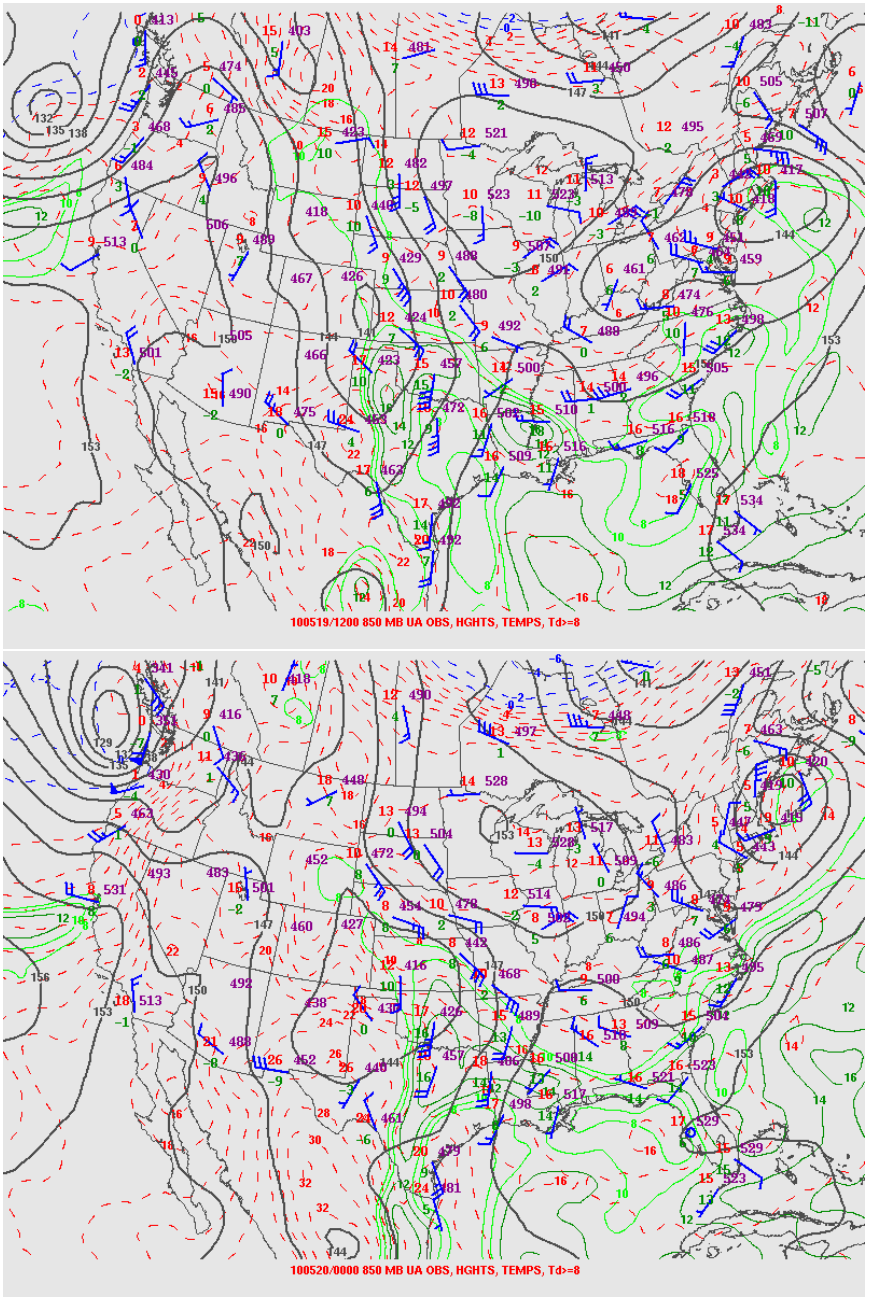

Figure 6.2: SPC 850 mb analysis for 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

in Washington. Another is a closed low centered over northeastern Colorado that would produce multiple severe weather events across the Southern Plains. Another closed low centered over upstate New York produced precipitation in Maine and off the coast of North Carolina. The 850 mb analysis (Fig. 6.2) shows a plume of moist air extending from the Gulf of Mexico into Texas and Oklahoma. Strong southerly winds were also present in the layer, which aided the advection of moisture into the region.

The observed precipitation from the NMQ (Fig. 6.3) showed a mesoscale convective system advancing through southern Kansas and northern Oklahoma with additional precipitation near the coasts of North Carolina and Maine at 1200 UTC. By 0000 UTC on 20 May, a line of discrete supercells had formed in central Oklahoma with a larger area of stratiform precipitation in Kansas and Missouri. Additional isolated storms were found in eastern Louisiana and off the coast of the North Carolina and Florida.

6.1.2 Probabilistic Predictions

The predictions from the 100 tree random forest for 19 May are shown in Fig. 6.4. At 1200 UTC, the random forest does capture the placement of the Oklahoma portion of the MCS very well and places the highest probabilities in that area. The additional storms further south are well covered and receive slightly lower probabilities. The probabilities for the storms in Kansas are underestimated given the observed rainfall totals, which is likely due to the precipitation in Kansas being influenced by a warm front, so measures of the ingredients for convection, such as CAPE, would be lower. Most of the northern section of the MCS is not contained in the precipitation area, and the heavy rain in western

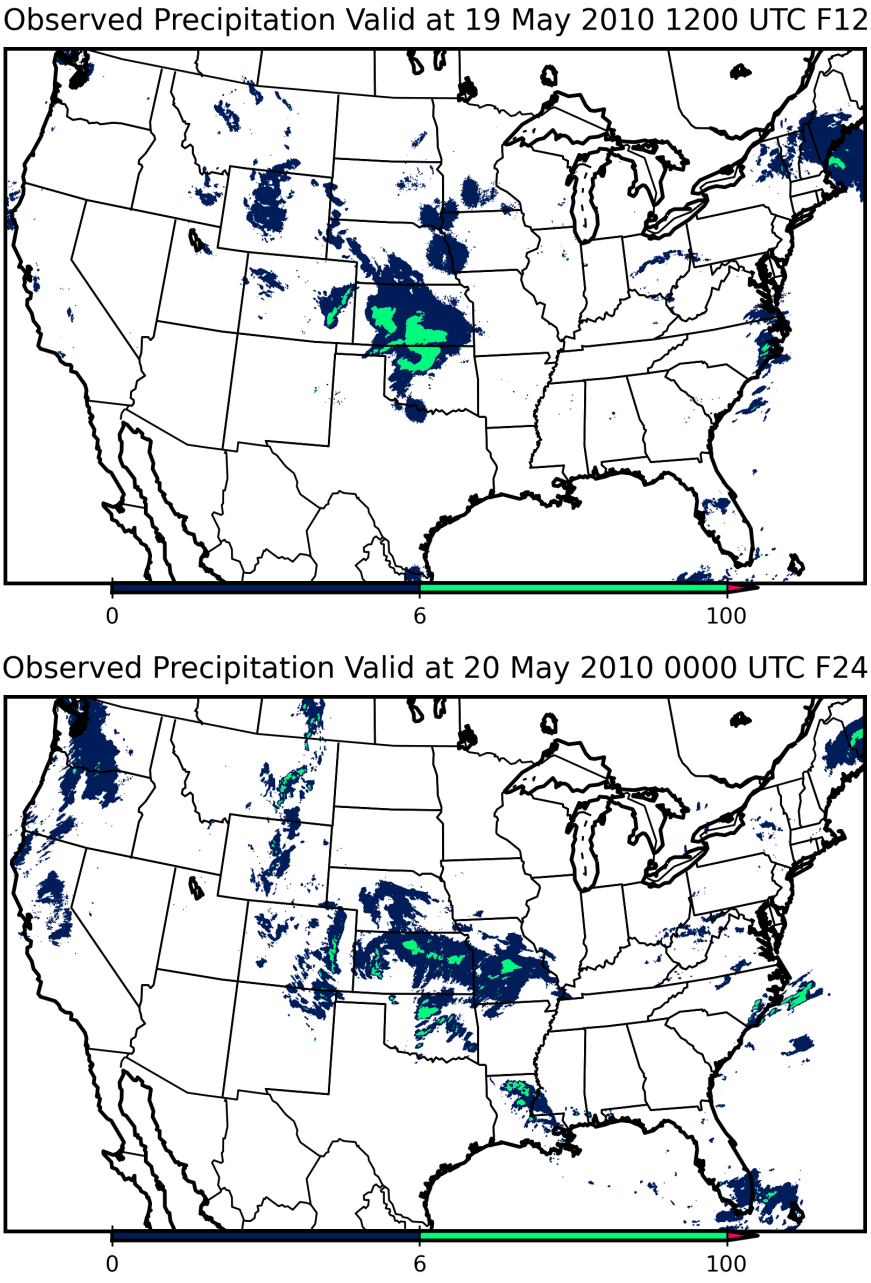

Figure 6.3: Observed precipitation on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

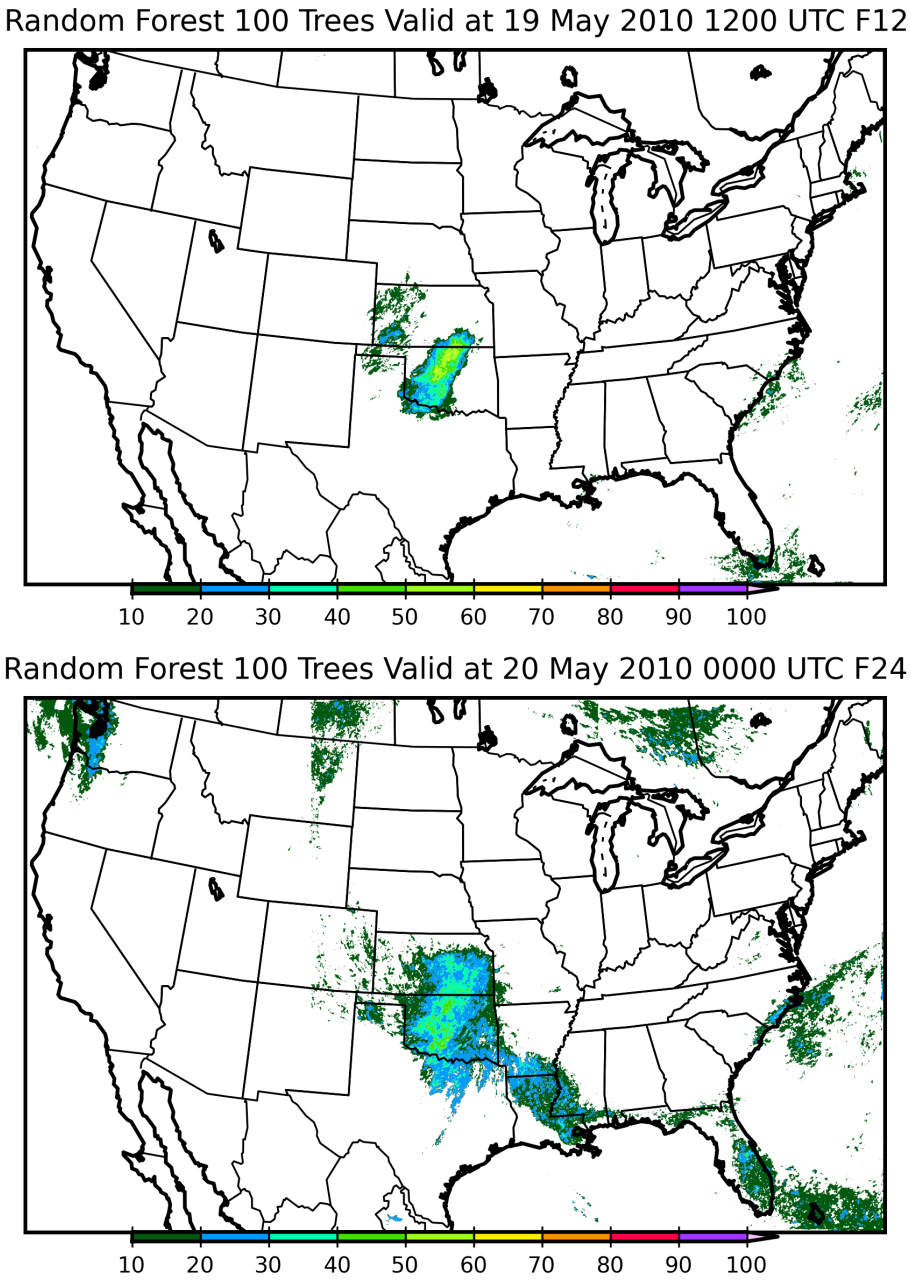

Figure 6.4: Predictions from the 100 tree random forest on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

Kansas, while given nonzero probabilities, is not given as high a probability as the observations would suggest. The random forest does produce small, spotty areas of nonzero probability near North Carolina, Florida, and Maine. The optimal thresholding, in this case, appears to remove too much of the precipitation probabilities. At 0000 UTC on 20 May, the random forest also does a good job capturing the precipitation areas and generally places the higher probabilities near the locations of observed storms. It highlights the heavy rain areas in Oklahoma, Florida, North Carolina, Montana, and Washington. The probability area does extend further into Texas than was observed, and the Washington and Montana areas are both smaller than observed.

The calibration logistic regression does show higher probabilities than the original ensemble, but there are clear differences in the probabilities between regions (Fig. 6.5). Much of the area in grid 2 contained the baseline probability for the calibration logistic regression,which was higher than grid 2 due to the larger amount of missed precipitation events in grid 2.

The multiple logistic regression provided many advantageous features from its forecast (Fig. 6.6). At 1200 UTC, the multiple logistic regression places the highest probabilities in Oklahoma and lower probabilities in Kansas. The distribution of the higher probabilities seem to indicate isolated convection as opposed to the MCS that was actually there at the time. At 0000 UTC, the logistic regression output places a series of high probability areas across central Oklahoma, indicating a line of isolated storms. The area forecast also covers the precipitation in Kansas unlike the other algorithms, but no high probabilities are provided in that location.

The MARS algorithm produces similar distributions of probabilities to the raw ensemble but with a larger range of values (Fig. 6.7). It has higher probabilities throughout central Oklahoma but does not give indication of the isolated nature of the convection.

6.1.3 Deterministic Predictions

The deterministic precipitation forecasts provided a good estimate of precipitation area and some guidance of where relatively larger amounts of precipitation would fall, but none of the algorithms predicted the extreme amounts seen in Fig. 6.8. The raw ensemble (Fig. 6.9), predicts precipitation over a wider area but smoothes out any intense precipitation values. The random forest (Fig. 6.10) rainfall totals have similar spatial patterns to the probability forecasts. Higher precipitation values are found farther south than the raw ensemble. The forecasts still have a smoothed quality to them and do not account for the sharp variations in precipitation amounts seen in the verification. This result is most likely due to the smoothing caused by the averaging of the predicted values from all of the trees. MARS (Fig. 6.11) produced higher precipitation amounts than the raw ensemble as well but matched the spatial distribution of the raw ensemble closely. Linear regression (Fig. 6.12) had a higher baseline for its precipitation areas but the maximum precipitation amounts were less than the raw ensemble.

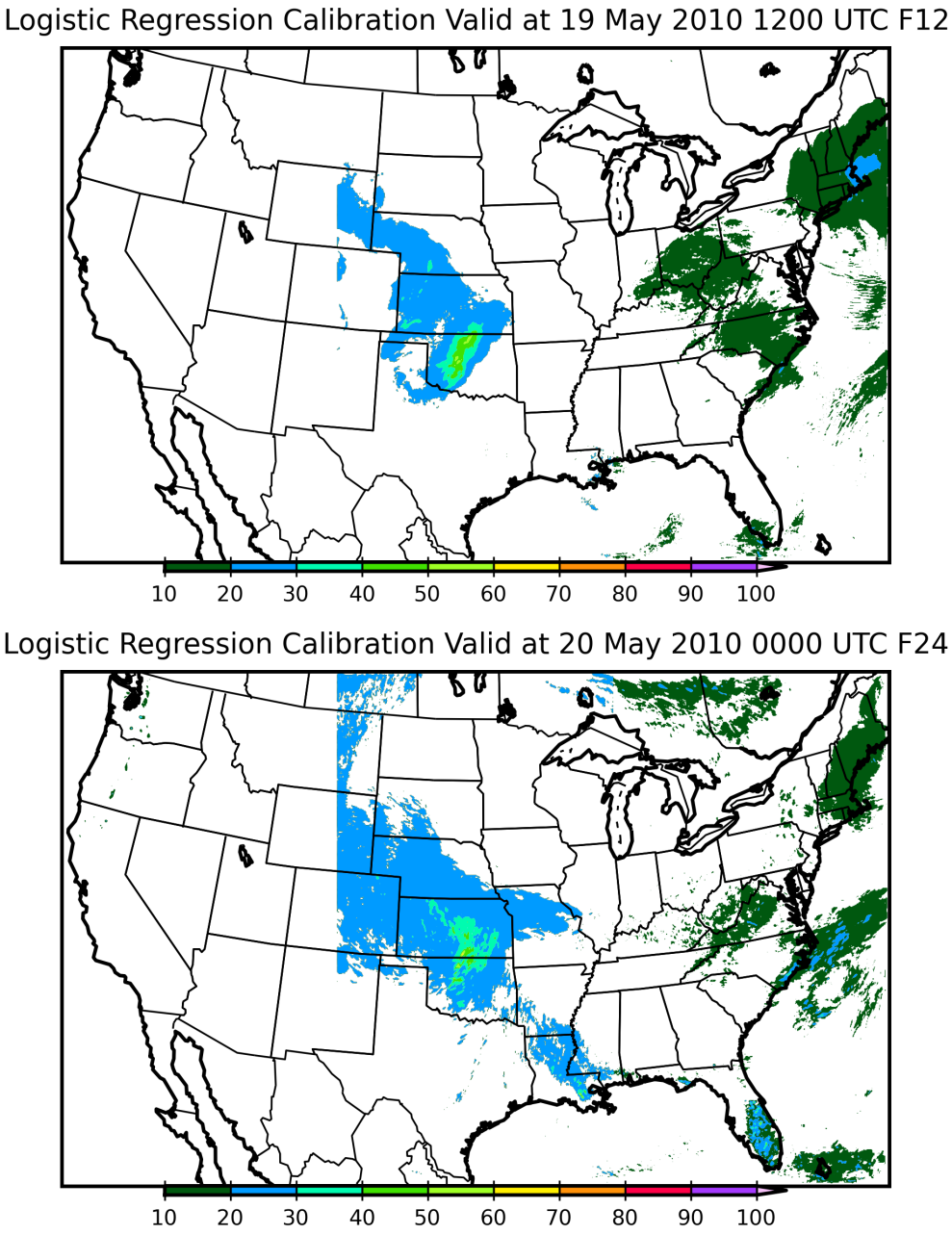

Figure 6.5: Predictions from the calibration logistic regression on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

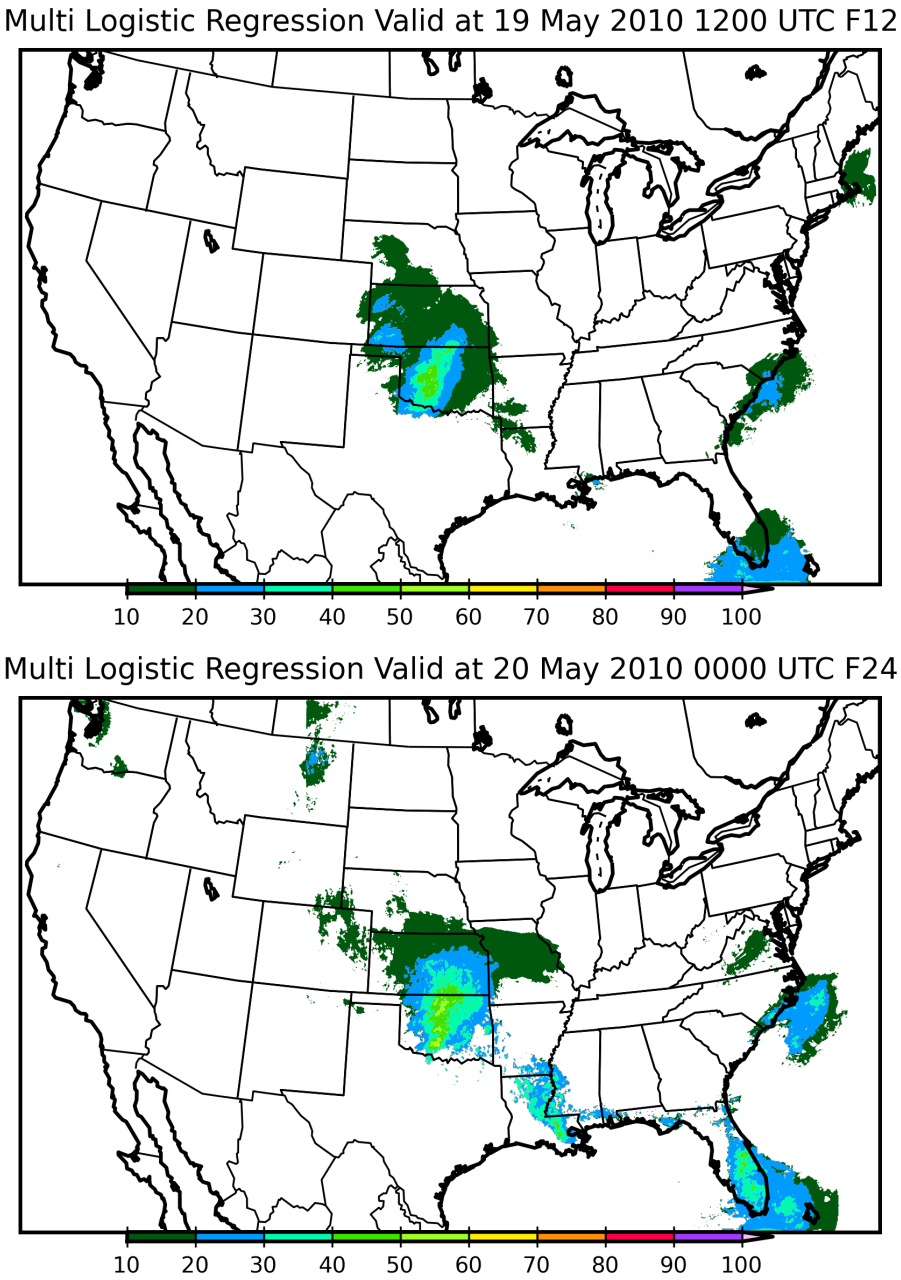

Figure 6.6: Predictions from the multiple logistic regression on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

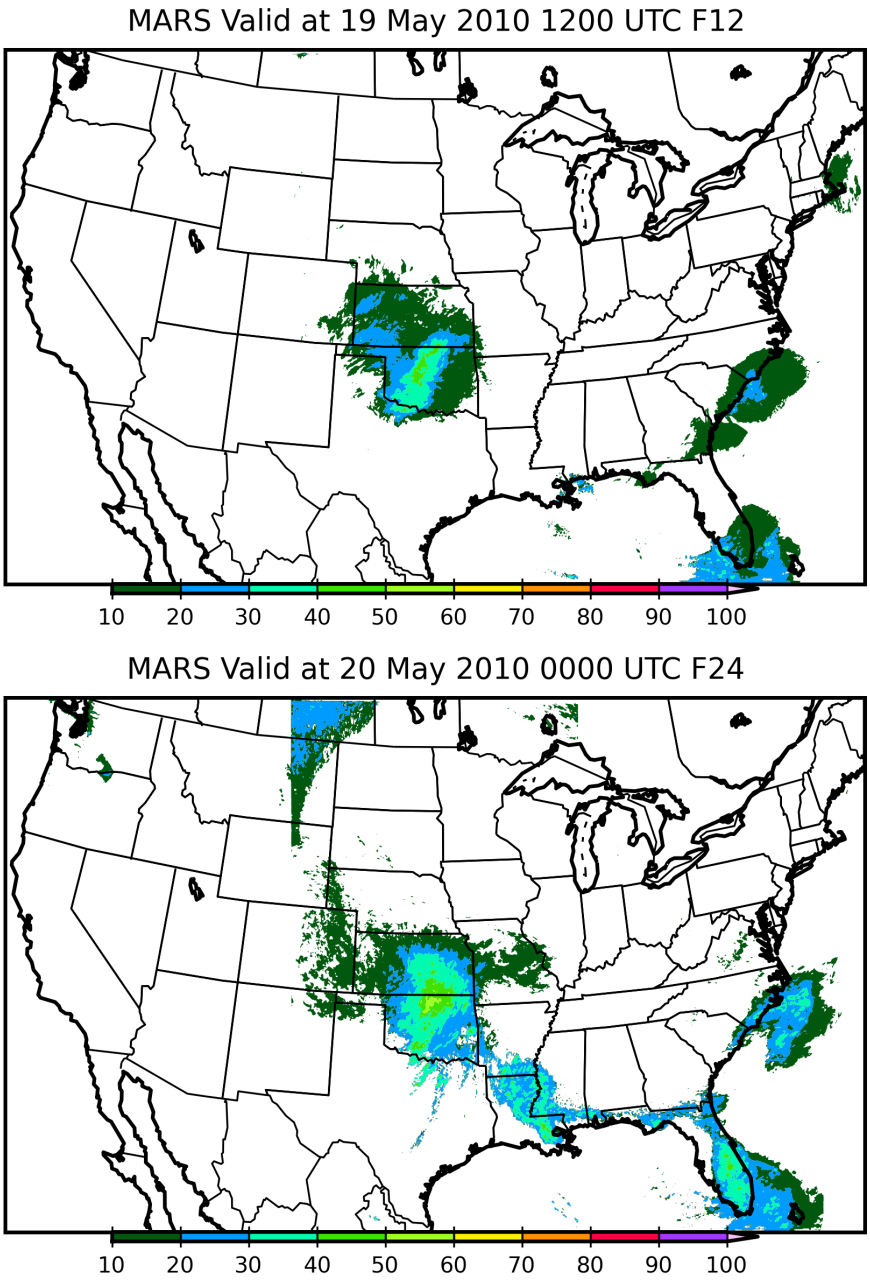

Figure 6.7: Predictions from the multiple logistic regression on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

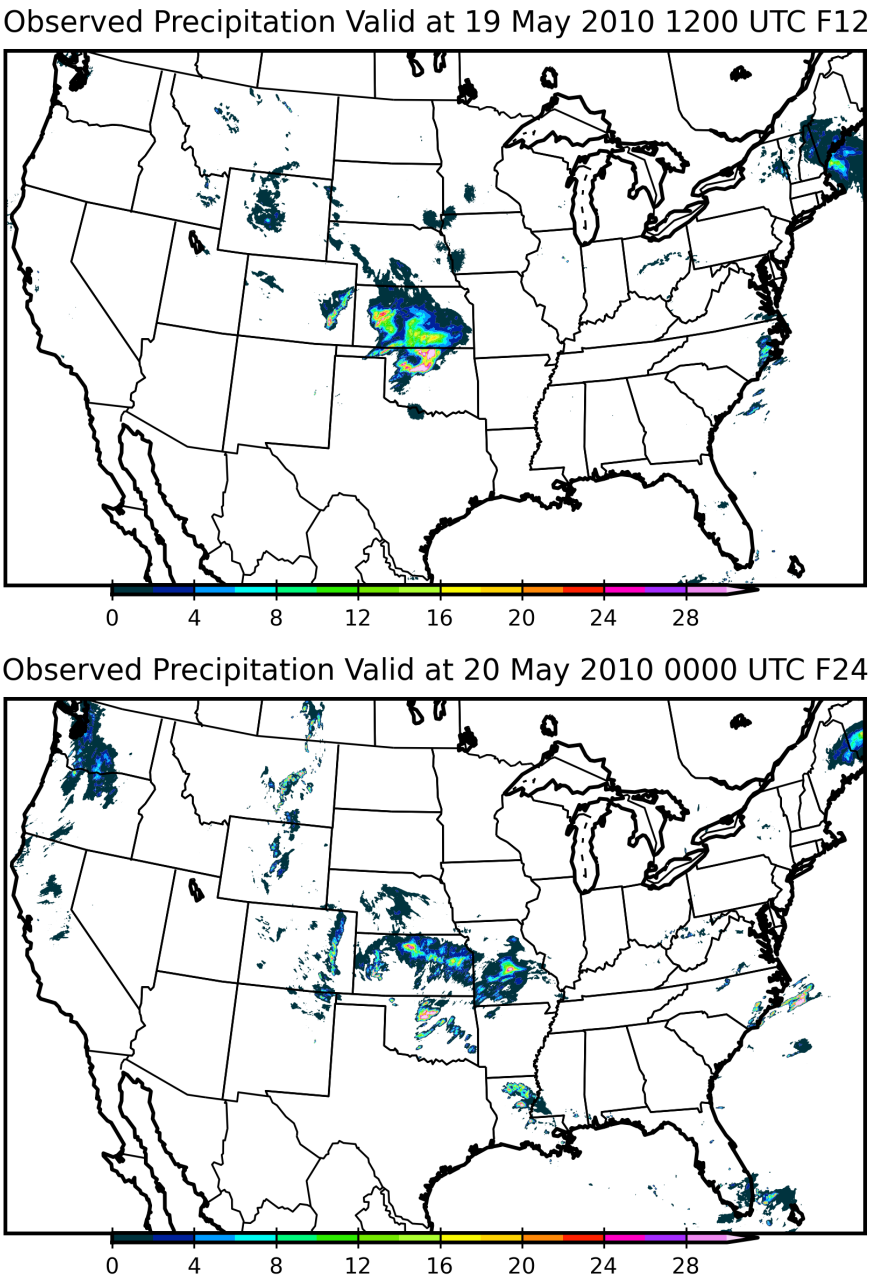

Figure 6.8: Observed precipitation on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

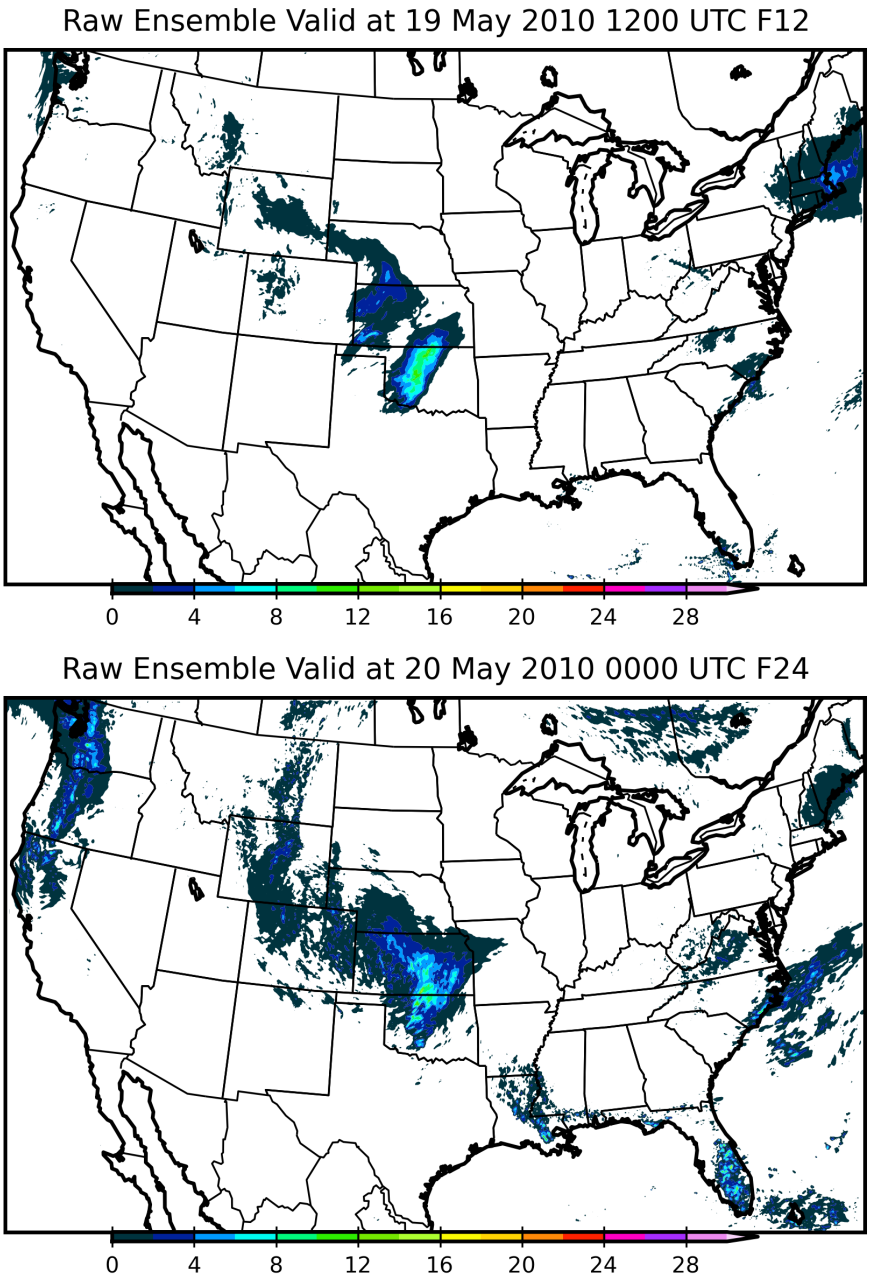

Figure 6.9: The ensemble mean 1-hour precipitation forecast on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

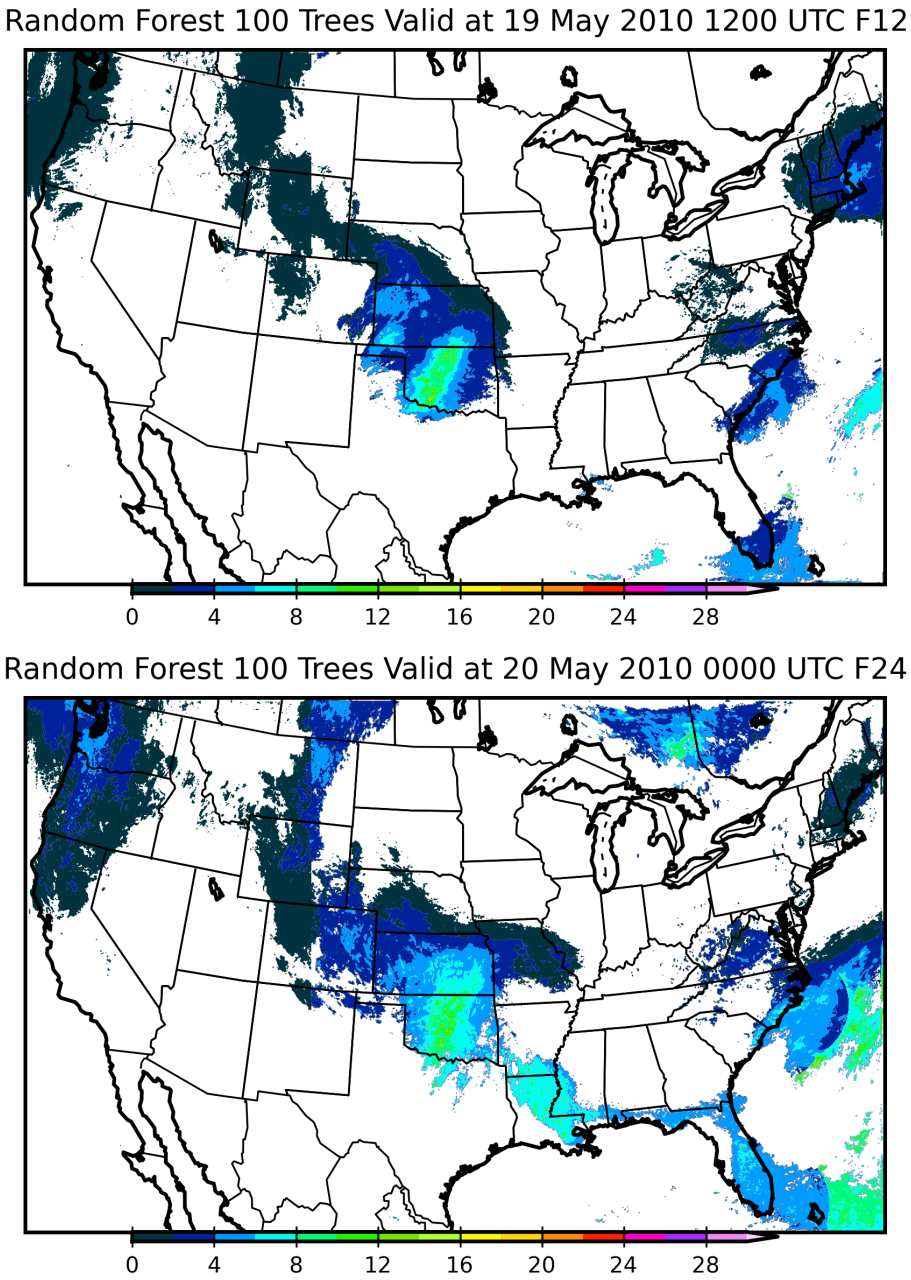

Figure 6.10: The 100 tree random forest 1-hour precipitation forecast on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

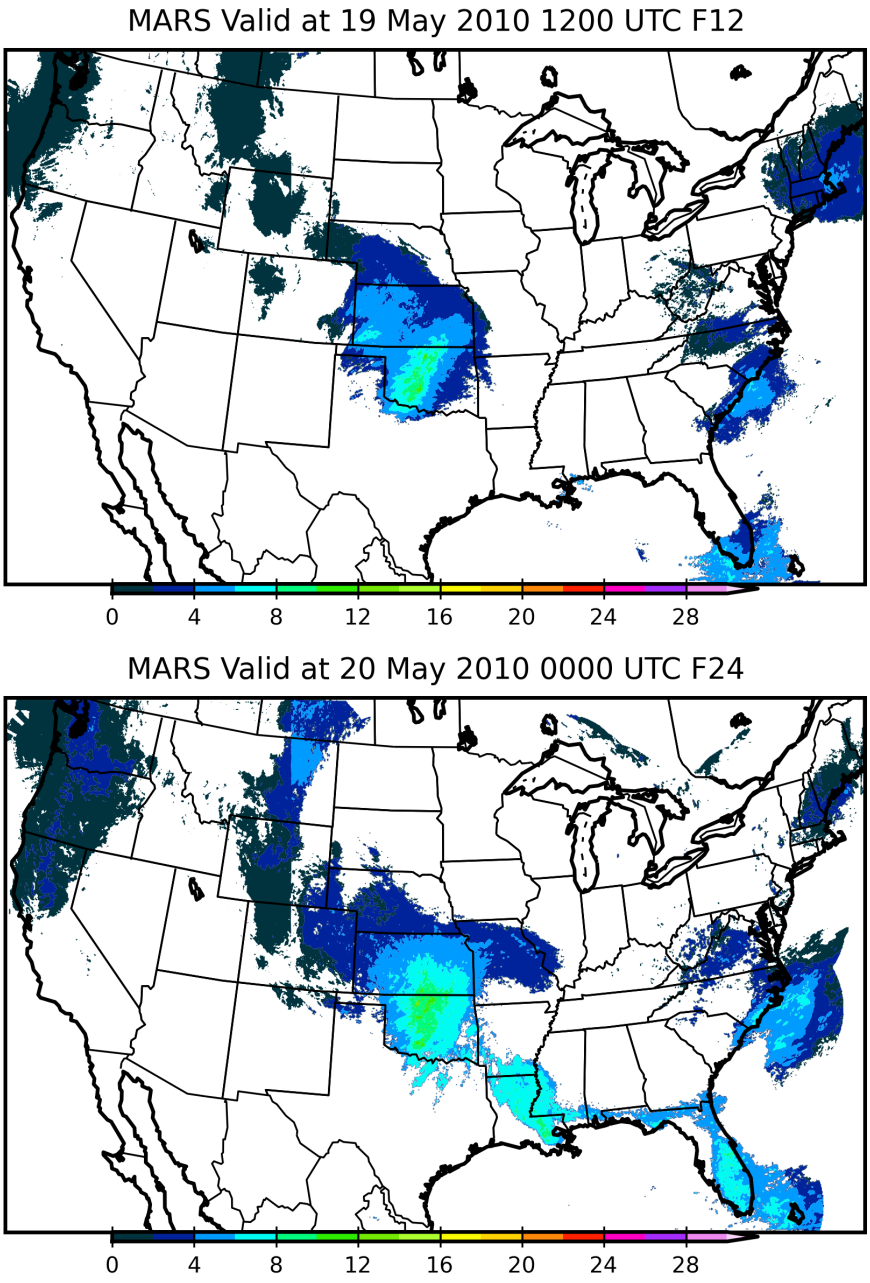

Figure 6.11: The MARS 1-hour precipitation forecast on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

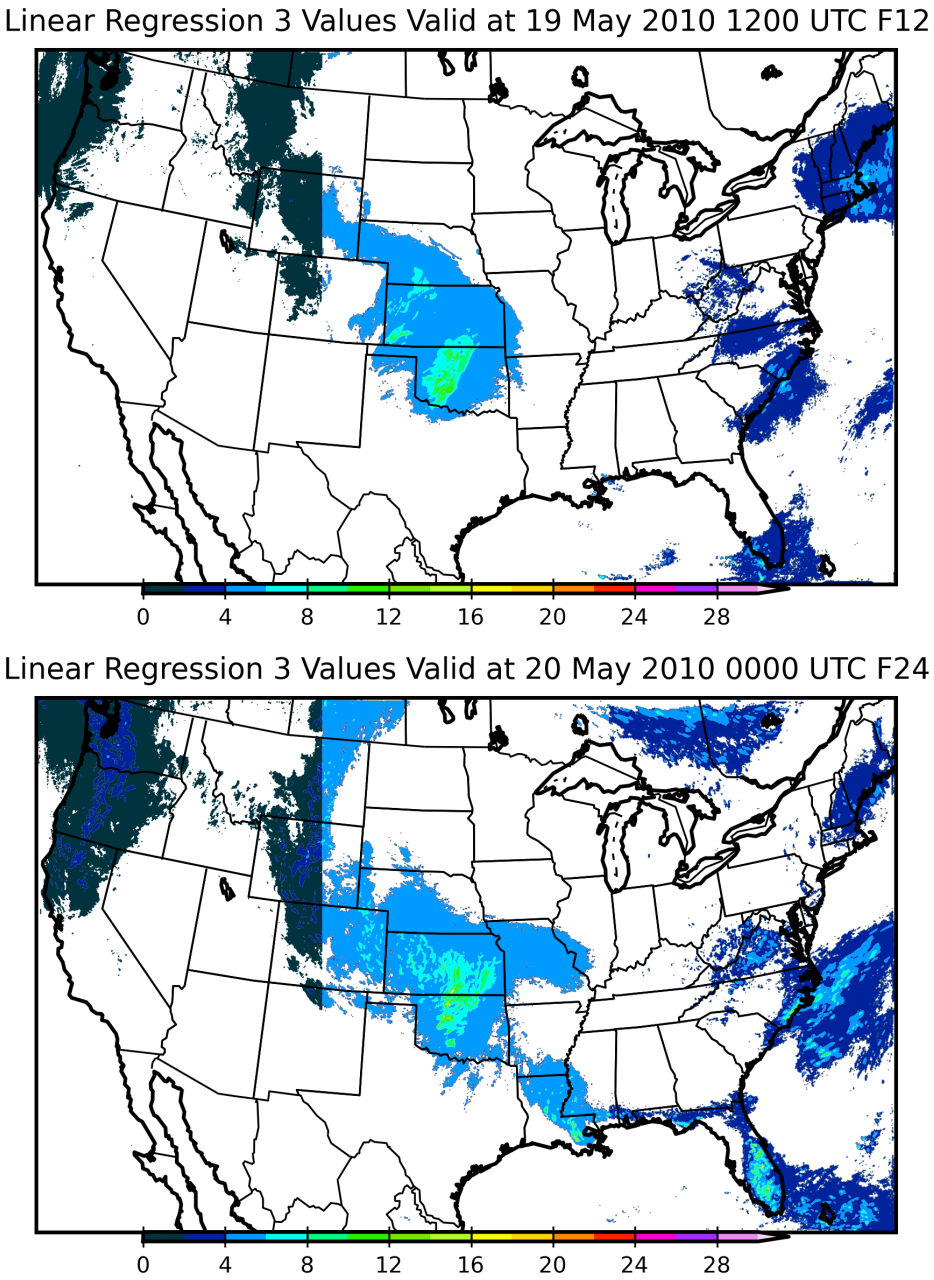

Figure 6.12: The linear regression 1-hour precipitation forecast on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

6.1.4 Interval Predictions

The quantile regression produced a range of precipitation values given the 5th, 50th, and 95th percentiles of the ensemble rainfall predictions. The 95th percentile for the transformed distribution, which is the 5th percentile in the forecast distribution, has nonzero probabilities in the areas where precipitation is most likely, and produces 0 probabilities otherwise (Fig. 6.13). The 50th percentile forecast (Fig. 6.14) has many similarities to the ensemble mean forecast. The 5th percentile forecast (Fig. 6.15) did cover the extreme precipitation forecasts and had a slightly higher baseline than the 50th percentile forecasts. The differences in the regions were clearest in the 5th percentile forecasts with the central US baseline being higher than the maxima from the east and west coasts. The forecasts did capture the range of precipitation well. The width of the interval was particularly wide in Oklahoma, but this was appropriate given the isolated nature of the convection.

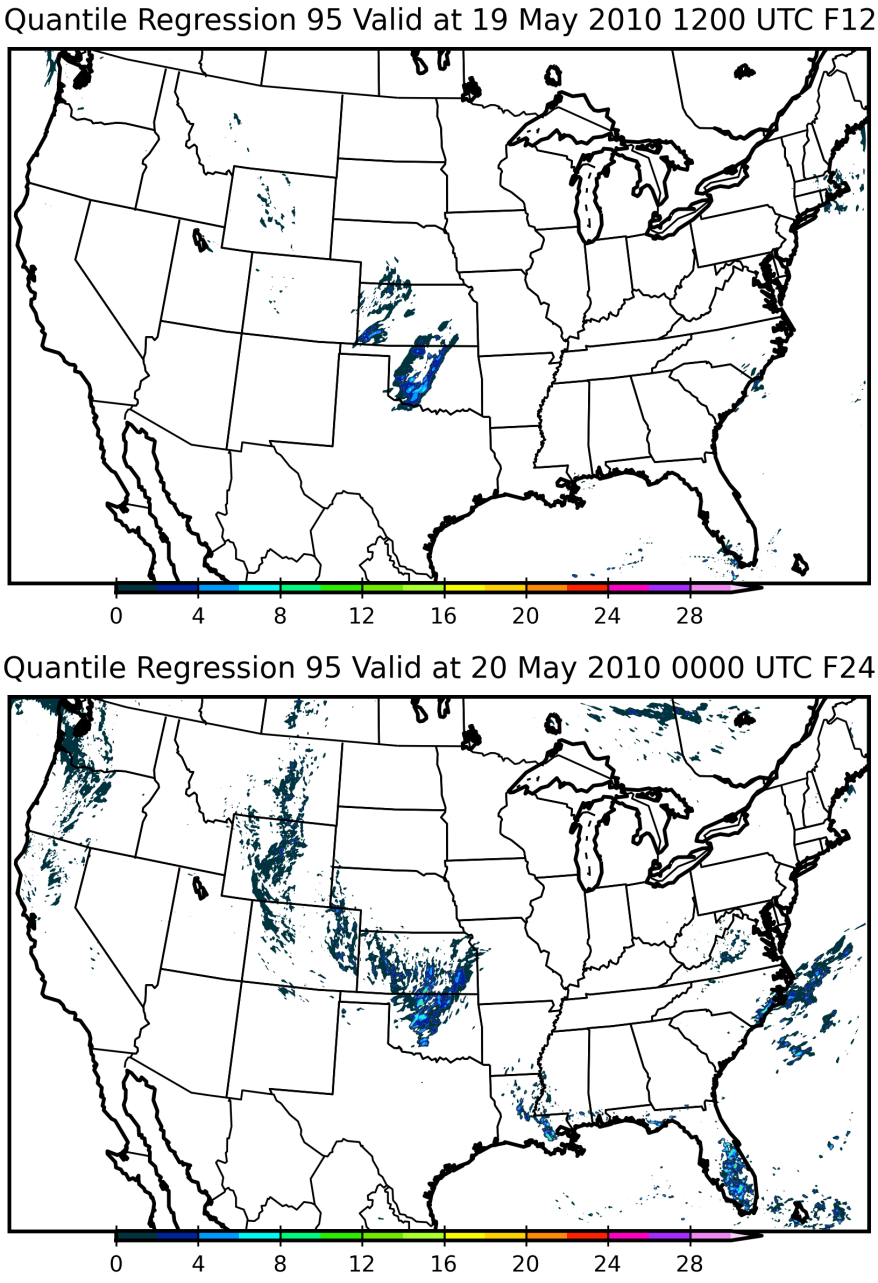

Figure 6.13: The quantile regression 95th percentile correction to the ensemble mean 1-hour precipitation forecast on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

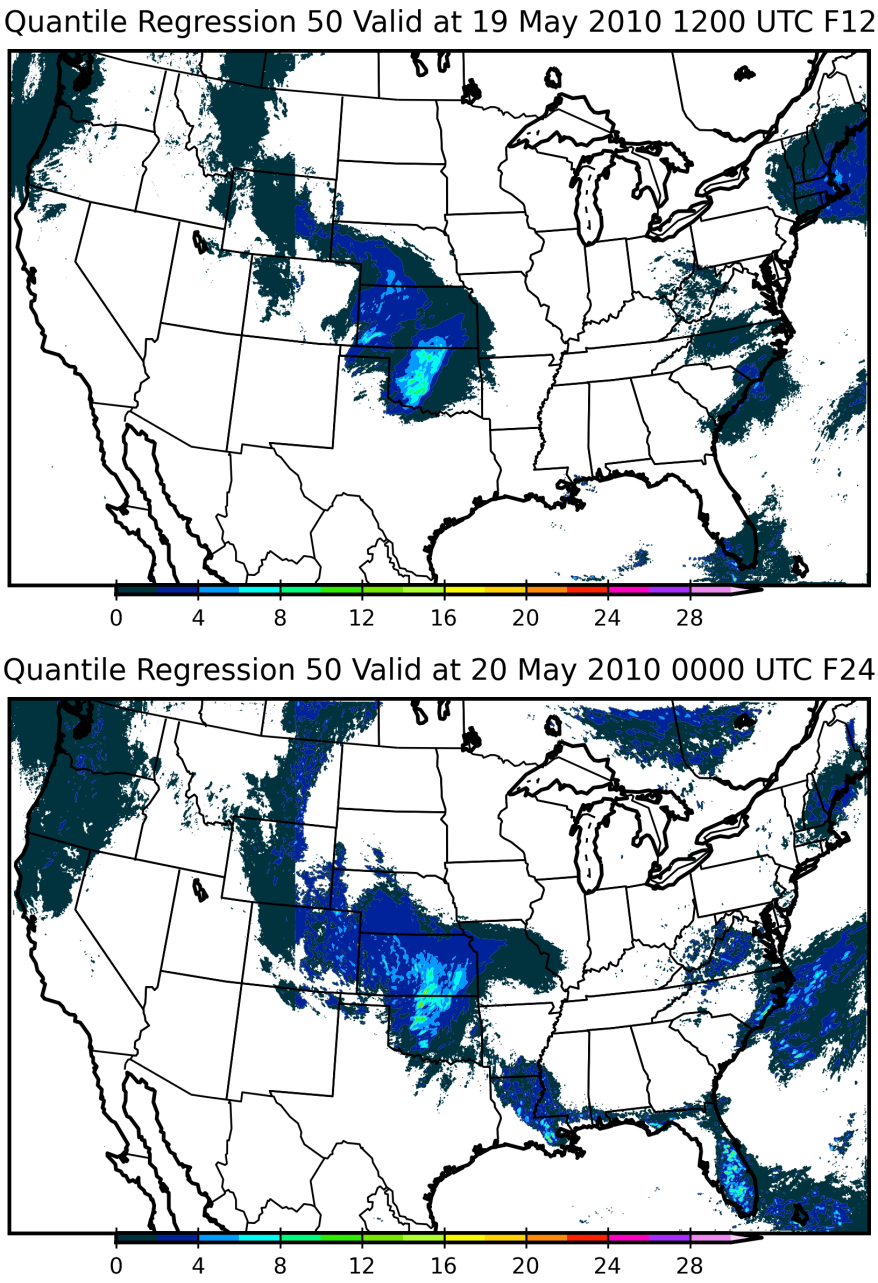

Figure 6.14: The quantile regression 50th percentile correction to the ensemble mean 1-hour precipitation forecast on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

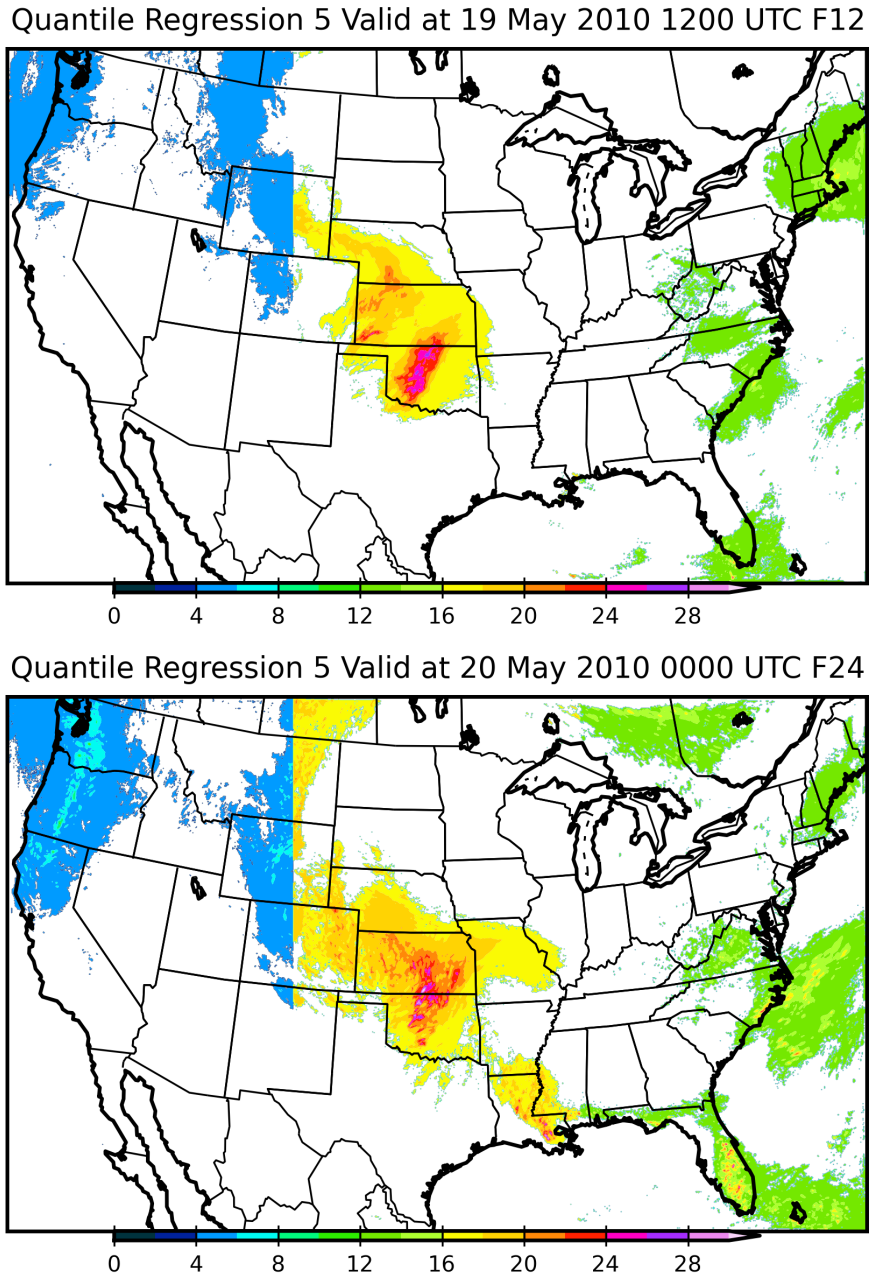

Figure 6.15: The quantile regression 5th percentile correction to the ensemble mean 1-hour precipitation forecast on 19 May 2010 at 1200 UTC and 20 May 2010 at 0000 UTC.

Chapter 7

Conclusions

This study demonstrated that a storm scale ensemble post-processing system based on ensemble machine learning algorithms, radar mosaic verification, and ensemble variable statistics can provide improved precipitation forecasts. Multiple machine learning models of varying complexity were applied to forecasts from the 2010 SSEF over the continental US for the period from 3 May to 18 June 2010 and verified against a radar-derived precipitation mosaic. Probabilistic, deterministic, and interval forecasts of 1-hour precipitation accumulation were created with the different models. Verification statistics showed that random forests, MARS, and multiple logistic regression provided significant improvements for probabilistic and continuous forecasts by improving the reliability, the skill, and the bias in the ensemble probabilistic and deterministic forecasts. Quantile regression forests produce more accurate median forecasts than quantile regressions, but quantile regressions are better at distributing their intervals to capture the correct probability densities. The models were applied to a case study to illustrate the geographic variability of the forecasts and tendencies in the predictions.

While the machine learning algorithms were able to demonstrate improvement in the aggregate over the original ensemble, there are still outstanding issues with the forecasts that the machine learning algorithms were not able to

address. The chief issue is the ability to handle both the high and low precipitation amounts. All of the approaches tested here tend to optimize their forecasts to either the mean or median of the distribution, so they tend to overestimate the low cases and underestimate the high cases. Including more extreme values in the training dataset will increase the extreme predictions, but they may also bias the probabilistic and deterministic forecasts away from the true distribution of the population. An additional calibration step could be performed using a simpler approach such as probability matching, but having three levels of post processing may not be advisable from a computational or meteorological perspective.

While the additional ensemble variables were able to provide additional information to supplement the precipitation forecasts, none of them showed a statically significant variable importance score. There were still many cases where the randomization of a variables values improved the accuracy. This finding necessitates the need for more variables to be included in future post-processing algorithms and for different transformations to be used. Focusing on a single grid point and a single time is a great way to save computational resources, but incorporating variables that take the spatial and temporal uncertainties of the forecast into account may provide further improvement.

Even with the limitations mentioned, the use of multivariate machine learning algorithms is worth the extra computational effort. The two-step approach has shown its worth in terms of highlighting the area of interest. The choice of which machine learning algorithm to use depends on what the user values the most for their particular system as they all greatly improved on the raw ensemble but none consistently stood out from the rest. Logistic regressions and MARS provided smoother forecasts spatially than random forest and were more

apt to predict extreme values given this dataset, but both are tied to empirical distributions, so datasets not fitting those distributions will not perform as well. Random forests are non-parametric and forecast only within the bounds of their training set, which allows them to handle noisy data better. Because of these qualities, the choice of algorithm should also depend on the decisions made with the data selection and pre-processing. The choice of output predictions should depend on the preferences of the user. Finally, the use of this type of ensemble post-processing is not limited to rain. The prediction of other high impact weather events could also be enhanced by machine learning approaches.

Reference List

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting*, **17**, 783–799.

Bermowitz, R. J., 1975: An application of model output statistics to forecasting quantitative precipitation. *Mon. Wea. Rev.*, **103**, 149–153.

Breiman, L., 1984: *Classification and regression trees*. Wadsworth International Group, 358 pp.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5 – 32.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Clark, A. J., S. J. Weiss, J. S. Kain, and Coauthors, 2012: An overview of the 2010 hazardous weather testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, **11**, 560–581.

Ebert, E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

Eckel, F. A. and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132–1147.

Friedman, J. H., 1991: Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–67.

Gagne II, D. J., A. McGovern, J. B. Basara, and R. A. Brown, 2012: Tornadic supercell environments analyzed using surface and reanalysis data: A spatiotemporal relational data mining approach. *J. Appl. Meteorol.*, In press.

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

Hall, T., H. E. Brooks, and C. A. Doswell, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, **14**, 338–345.

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

Hamill, T. M. and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447.

Hansen, A. W. and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verhand.*, **81**, 2–15.

Hellman, S., 2012: Learning ensembles of dynamic continuous bayesian networks. M.S. thesis, School of Computer Science, University of Oklahoma.

Klein, W. H. and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55 (10)**, 1217–1227.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16**, 672–682.

Koenker, R. and K. F. Hallock, 2001: Quantile regression. *Journal of Economic Perspectives*, **15**, 143–156.

Koizumi, K., 1999: An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network. *Wea. Forecasting*, **14**, 109–118.

Kong, F., et al., 2011: Evaluation of CAPS multi-model storm-scale ensemble forecast for the NOAA HWT 2010 spring experiment. *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.*, Seattle, WA, Amer. Meteor. Soc., Paper 457.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550.

Manzato, A., 2007: A note on the maximum Peirce skill score. *Wea. Forecasting*, **22**, 1148–1154.

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

Meinshausen, N., 2006: Quantile regression forests. *Journal of Machine Learning Research*, **7**, 983–999.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.

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

Pappenberger, F., I. Iorgulescu, and K. J. Beven, 2006: Sensitivity analysis based on regional splits and regression trees (SARS-RT). *Environmental Modeling and Software*, **21**, 976–990.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174.

Soughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probilistic quantitative precipitation forecasting using bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC Bioinformatics*, **9 (1)**, 307, doi:10.1186/1471-2105-9-307, URL `http://www.biomedcentral.com/1471-2105/9/307`.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Tracton, M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–398.

Vasiloff, S., et al., 2007: Improving QPE and very short term QPF: An initiative for a community-wide integrated approach. *Bull. Amer. Meteor. Soc.*, **88**, 1899–1911.

Vislocky, R. L. and G. S. Young, 1989: The use of perfect prog forecasts to improve model output statistics forecasts of precipitation probability. *Wea. Forecasting*, **4**, 202–209.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability distribution MOS forecasts. *Meteorol. Appl.*, **16**, 361–368.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3d ed., Academic Press, 676 pp.

Xue, M., et al., 2011: CAPS realtime storm scale ensemble and high resolution forecasts for the NOAA hazardous weather testbed 2010 spring experiment. *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.*, Seattle, WA, Amer. Meteor. Soc., PAPER 9A.2.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network. *Wea. Forecasting*, **22**, 1287–1303.