# Machine Learning Enhancement of Storm Scale Ensemble Probabilistic Quantitative Precipitation Forecasts

David John Gagne II*

*School of Meteorology, University of Oklahoma, Norman, Oklahoma*

Amy McGovern

*School of Computer Science, University of Oklahoma, Norman, Oklahoma*

Ming Xue

*Center for the Analysis and Prediction of Storms/School of Meteorology*

*University of Oklahoma, Norman, Oklahoma*

* *Corresponding author address:* David John Gagne II, 120 David L. Boren Blvd., Suite 5900, Norman, OK 73072. Email: djgagne@ou.edu

# ABSTRACT

Probabilistic quantitative precipitation forecasts challenge meteorologists due to the wide variability of precipitation amounts over small areas and their dependence on conditions at multiple spatial and temporal scales. Ensembles of convection-allowing numerical weather prediction models offer a way to produce improved precipitation forecasts and estimates of the forecast uncertainty. These models allow for the prediction of individual convective storms on the model grid, but they often displace the storms in space, time, and intensity, which results in added uncertainty. Machine learning methods can produce calibrated probabilistic forecasts from the raw ensemble data that correct for systemic biases in the ensemble precipitation forecast and incorporate additional uncertainty information from aggregations of the ensemble members and additional model variables. This study utilizes the 2010 Center for the Analysis and Prediction of Storms Storm Scale Ensemble Forecast system and the National Severe Storms Laboratory National Mosaic and Multisensor Quantitative Precipitation Estimate as input data for training logistic regressions and random forests to produce a calibrated probabilistic quantitative precipitation forecast. The reliability and discrimination of the forecasts are compared through verification statistics and a case study.

# 1. Introduction

Most flooding fatalities occur due to flash floods, in which waters rise and fall rapidly due to concentrated rainfall over a small area (Ashley and Ashley 2008). The first step to anticipating flash floods is quantitative forecasting of the amount, location, and timing of precipitation. These quantitative precipitation forecasts are challenging due to the wide variability of precipitation amounts over small areas, the dependence of precipitation amounts on processes at a wide range of scales, and the dependence of extreme precipitation on any precipitation actually occurring (Bremnes 2004; Ebert 2001; Doswell et al. 1996). Recent advances in numerical modeling and machine learning are working to address these challenges.

Numerical Weather Prediction (NWP) models are now being run experimentally at 4 km horizontal grid spacing, or storm scale, allowing for the formation of individual convective cells without a convective parameterization scheme. These models better represent storm processes and output hourly predictions, but they have the challenge of correctly placing and timing the precipitation compared to models with coarser grid spacing and temporal resolution. An ensemble of storm scale NWP models can provide improved estimates of uncertainty compared to coarser ensembles due to better sampling of the spatiotemporal errors associated with individual storms (Clark et al. 2009). As each ensemble member produces predictions of precipitation and precipitation ingredients, the question then becomes how best to combine those predictions into the most accurate and useful consensus guidance.

The final product should highlight areas most likely to be impacted by heavy rain and also provide an uncertainty estimate for that impact. Probabilistic Quantitative Precipi-

tation Forecasts (PQPFs) incorporate both of these qualities. A good PQPF should have reliable probabilities, such that a 40% chance of rain verifies 40% of the time over a large sample (Murphy 1977). PQPFs should also discriminate between extreme and trace precipitation events consistently, so most extreme events occur with higher probabilities, and most trace precipitation events are associated with low probabilities. Since individual ensemble members may be biased in different situations, a simple count of the ensemble members that exceed a threshold will often result in unreliable forecasts that discriminate poorly. Incorporating trends from past forecasts and additional information from other model variables can offset these biases and produce an enhanced PQPF.

Ensemble weather prediction (Toth and Kalnay 1993; Tracton and Kalnay 1993; Molteni et al. 1996) has required various forms of statistical post-processing to produce accurate precipitation forecasts and uncertainty estimates from the ensembles, but most previous studies used coarser ensembles and longer forecast windows for calibration. The rank histogram method (Hamill and Colucci 1997) showed that ensemble precipitation forecasts tended to be underdispersive. Linear regression calibration methods have shown some skill improvements in Hamill and Colucci (1998), Eckel and Walters (1998), Krishnamurti et al. (1999), and Ebert (2001). Hall et al. (1999), Koizumi (1999), and Yuan et al. (2007) applied neural networks to precipitation forecasts and found increases in performance over linear regression. Logistic regression, a transform of a linear regression to fit an S-shaped curve ranging from 0 to 1, has shown more promise, as in Applequist et al. (2002), which tested linear regression, logistic regression, neural networks, and genetic algorithms on 24-hour PQPF and found that logistic regression consistently outperformed the other methods. Hamill et al. (2004, 2008) also utilized the logistic regression with an extended training period for added skill.

Storm scale ensemble precipitation forecasts have been post-processed with smoothing algorithms that produce reliable probabilities within longer forecast windows. Clark and Coauthors (2011) applied smoothing algorithms at different spatial scales to the SSEF precipitation forecasts and compared verification scores. Johnson and Wang (2012) compared the skill of multiple calibration methods on neighborhood and object-based probabilistic forecasts from the 2009 SSEF. Marsh et al. (2012) applied a Gaussian kernel density estimation function to the NSSL 4 km Weather Research and Forecasting (WRF) to derive probabilities from deterministic forecasts. These methods are helpful for predicting larger scale events but smooth out the threats from extreme precipitation in individual convective cells. Gagne II et al. (2012) took the first step in examining how multiple machine learning approaches performed in producing probabilistic, deterministic, and quantile precipitation forecasts over the central United States at individual grid points.

The purpose of this paper is to analyze PQPF predictions produced by multiple machine learning techniques incorporating data from the Center for the Analysis and Prediction of Storms (CAPS) 2010 Storm Scale Ensemble Forecast (SSEF) system (Xue et al. 2011; Kong et al. 2011). In addition to the choice of algorithm, some variations in the algorithm setups are also examined. The strengths and weaknesses of the machine learning algorithms are shown through the analysis of verification statistics, variables chosen by the machine learning models, and a case study. This paper expands on the work presented in Gagne II et al. (2012) by including statistics for the eastern US, a larger training set more representative of the precipitation probabilities, a different case study day with comparisons of multiple runs, multiple precipitation thresholds, and more physical justification for the performance of the machine learning algorithms.

# 2. Data

*a. Ensemble Data*

The Center for the Analysis and Prediction of Storms (CAPS) 2010 Storm Scale Ensemble Forecast (SSEF) system (Xue et al. 2011; Kong et al. 2011) provides the input data for the machine learning algorithms. The 2010 SSEF consists of 19 individual model members from the WRF Advanced Research WRF (ARW), 5 members from the WRF Nonhydrostatic Mesoscale Model (NMM), and 2 members from the CAPS Advanced Research Prediction System (ARPS; Xue et al. 2000, 2001, 2003). Each member has a varied combination of microphysics schemes, land surface models, and planetary boundary layer schemes. The SSEF ran every weekday at 0000 UTC in support of the 2010 National Oceanic and Atmospheric Administration/Hazardous Weather Testbed Spring Experiment (Clark et al. 2012), which ran from May 3 to June 18, for a total of 34 runs. The SSEF provides hourly model output over the contiguous United States at 4 km horizontal grid spacing out to 30 hours. Of the 26 members, the 14 members included initial condition perturbations derived from the National Centers for Environmental Prediction Short Range Ensemble Forecast (SREF; Du et al. 2006) members, and only they are included in our post-processing procedure. The 12 other members used the same control initial conditions although with different physics options and models; designed to examine forecast sensitivities to physics parameterizations, they do not contain the full set of initial conditions and model uncertainties and are therefore excluded from our study.

*b. Verification Data*

A radar-based verification dataset was used as the verifying observations for the SSEF. The National Mosaic and Multi-Sensor QPE (NMQ; Vasiloff et al. 2007) derives precipitation estimates from the reflectivity data of the NEXRAD radar network. The estimates are made on a grid with 1 km horizontal spacing over the continental US (CONUS). The original grid has been bi-linearly interpolated to the same grid as the SSEF.

*c. 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 34 ensemble runs over a short, homogenous time period with 840,849 grid points from each of the 30 time steps. 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 (280,283 points per time step), and points were selected with a uniform random sample from each subdomain in areas with quality radar coverage. The gridded Radar Quality Index (RQI; Zhang et al. 2011) was evaluated at each grid point to determine the trustworthiness of the verification data. Points with an RQI$> 0$ were located within the useful range of a NEXRAD radar and included in the sampling. For points with precipitation values less than 0.25 mm, 0.04% were sampled, and for points with more precipitation, 0.4% were sampled. Grid 0 corresponds to the western third of the CONUS, Grid 1 corresponds to the central third, and Grid 2 corresponds to the eastern third. Fig. 1 shows that the north central section of the United States and the states south

of the Great Lakes were most heavily sampled. Grid 0 was excluded from post-processing due to the low frequency of heavy precipitation for most of the region.

A comparison of the sampled rainfall distributions and the full rainfall distributions for each subgrid are shown in Fig. 2 and Table 1. Undersampling of the 0 and trace (below 0.25 mm) precipitation points was necessary because of the large number of no-precipitation events, which overwhelmed the signal from the actual precipitation events. The random sampling of grid points helps reduce the chance of sampling multiple grid points from the same storm without explicitly filtering subsets of the domain, which was performed by Hamill et al. (2008). Filtering grid points surrounding each sampled point is extremely computationally expensive because filtering requires multiple passes over the grid while random sampling only requires 1 grid reordering.

Relevant model output variables, called predictors (Table 2), were also extracted from each ensemble member at each sampled point. These predictors captured additional information about the mesoscale and synoptic conditions in each model. The predictors from each ensemble member were then aggregated into 4 descriptive statistics for each predictor. The mean provided the most likely forecast value, and the standard deviation estimated the spread, and the minimum and maximum showed the extent of the forecasted values.

# 3. Methods

*a. Machine Learning Methods*

1) Logistic Regression

One of the goals of this study is to compare the skill of more advanced machine learning methods with more traditional statistical methods. Logistic regression was used as the baseline statistical method. Logistic regressions are linear regression models in which a logit transformation is applied to the data so that the predicted values will range between 0 and 1. Two formulations of logistic regression were used: simple and multiple. The first formulation uses the ensemble mean precipitation forecast as expressed in Eqn. 1:

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

in which R is the precipitation amount, t is the threshold, $x_1$ is the ensemble mean precipitation forecast, $\beta_0$ is the intercept term, and $\beta_1$ is the weight given to the ensemble mean precipitation forecast. Observed precipitation with an amount greater than or equal to the chosen threshold is assigned a probability of 1, and precipitation amounts below the threshold are assigned a probability of 0. The $\beta$ coefficients are estimated by iteratively maximizing the log-likelihood of the transformed training data. 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 multiple variables with stepwise variable selection similar to standard Model Output

Statistics (Glahn and Lowry 1972) approaches (Eqn. 2):

$$p(R \geq t|x_1, x_2, ...x_n) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{n=1}^{N} \beta_n x_n\right)\right]} \quad (2)$$

in which $x_1$ through $x_n$ are the predictors chosen from Table 2 and $\beta_1$ through $\beta_n$ are the weights for those predictors. The terms are determined through a forward selection process that finds the set of up to 15 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 dimensionality of the training set. The *glm* (generalized linear model) and *step* (stepwise variable selection) functions in the *R* statistical package are used to generate the logistic regressions. A more detailed description of logistic regression and the fitting process can be found in James et al. (2013).

### 2) Random Forest

An non-parametric, nonlinear alternative to linear regression is the classification and regression decision tree (Breiman 1984). Decision trees recursively partition a multidimensional dataset into successively smaller subdomains by selecting variables and decision thresholds that maximize a dissimilarity metric. At each node in the tree, every predictor is evaluated with the dissimilarity metric, and the predictor and threshold with the highest metric value are selected as the splitting criteria for that node. After enough partitions, each subdomain is similar enough that the prediction can be approximated with a single value. The primary

advantages of decision trees are that they can be human-readable and perform variable selection as part of the model growing process. The disadvantages lie in the brittleness of the trees. Trees can undergo significant structural shifts due to small variations in the training data, which results in large error variance.

Random forests (Breiman 2001) consist of an ensemble of classification and regression trees (Breiman 1984) with two key modifications. First, the training data cases are bootstrap resampled with replacement for each tree in the ensemble. Second, a random subset of the predictors are selected for evaluation at each node. The final prediction from the forest is the mean of the predicted probabilities from each tree. 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 sensitivity of the model predictions to minor differences in the training set (Strobl et al. 2008). The random selection of predictors within the tree-building process allows for less optimal predictors to be included in the model and increases the likelihood of the discovery of interaction effects among predictors that would be missed by the stepwise selection method used in logistic regression (Strobl et al. 2008). Random forests have been shown to improve predictive performance on multiple problem domains in meteorology, including storm classification (Gagne II et al. 2009), aviation turbulence (Williams 2013), and wind energy forecasting (Kusiak and Verma 2011). For this project, we used the R *randomForest* library, which implements the original approach (Breiman 2001). For the parameter settings, we chose to use 100 trees, a minimum node size of 20, and the default values for all other parameters. A more detailed description

of random forest can be found in James et al. (2013).

In addition to gains in performance, the random forest can also be used to rank the importance of each input variable (Breiman 2001). 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 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. Variable importance scores can vary randomly among forests trained on the same dataset, so the variable importance scores from each of the 34 forests trained for cross-validation were averaged together for a more robust ranking.

*b. Evaluation Methods*

Two scores were used to assess the probabilistic forecasts. 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. 3:

$$BSS = \frac{\frac{1}{N}\sum_{k=1}^{K} n_k(\overline{o}_k - \overline{o})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \overline{o}_k)^2}{\overline{o}(1-\overline{o})} \tag{3}$$

$N$ 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 in the numerator 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 in the numerator 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 denominator term is the uncertainty, which is based on the climatological probability of precipitation and cannot be reduced through calibration. BSS increases with skill from 0. 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) evaluates how well a probabilistic forecast correctly identifies heavy precipitation events compared to identifying the events based on random chance. To calculate AUC, first, the decision probability threshold is varied from 0 to 1 in equal steps. At each step, a contingency table is constructed by splitting the probabilities into two categories with the decision threshold (Table 3). Using Table 3, the following scores can be computed (Table 4). 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 total yes forecasts. The Probability of False Detection (POFD) is the ratio of false alarms to total no forecasts. The Equitable Threat Score (ETS) is the skill score officially used to assess the skill of 24 hour precipitation forecasts, and is sensitive to the climatological base rate of the validation data. The Peirce Skill Score (PSS) is another skill score that is insensitive to the climatological base rate but is sensitive to hedging by over forecasting rare events (Doswell

et al. 1990). The bias ratio compares determines if false alarms or misses are more prevalent. From the contingency table, the POD and POFD are calculated and plotted against each other, forming a ROC curve. The AUC is the area between the right and bottom sides of the plot and the curve itself. AUC with values 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 can be used to determine an "optimal" decision threshold to convert a probabilistic forecast to a deterministic forecast. As the decision probability increases, the POD decreases from 1 to 0 while 1-POFD increases from 0 to 1. At some decision threshold, POD and 1-POFD should be equal. At this optimal threshold (OT), the user has an equal chance of correctly detecting rain events and no-rain events. This point occurs where the ROC curve is farthest from the positive diagonal "no-skill" line. In addition, the vertical distance between the ROC curve and the positive diagonal is the Peirce Skill Score (PSS; Peirce 1884; Hansen and Kuipers 1965) since PSS=POD−POFD (Manzato 2007). Since the PSS can be calculated from the contingency table at each decision threshold, finding the maximum PSS also finds the OT. The performance of the threshold choice can be validated by treating the predictions as a binary classification problem and using the binary verification statistics derived from the contingency table (Table 4).

### *c. Experimental Procedure*

Each model was trained and evaluated using a leave-one-day-out cross validation procedure. For the 34 daily model runs in the training set, each machine learning model was

trained on 33 days and tested on 1. Separate models were trained for each 6 hour forecast period and each subgrid to better capture the trends in the weather and ensemble dispersiveness. Models were primarily trained at two thresholds, 0.25 mm $h^{-1}$ and 6.35 mm $h^{-1}$. The 0.25 mm $h^{-1}$ threshold models were used to determine if rain was to occur at a point or not, and the 6.35 mm $h^{-1}$ models were trained and evaluated only on the conditions that either rain occurred or the ensemble mean precipitation forecast predicted rain. Deterministic rain forecasts were derived from the 0.25 mm $h^{-1}$ models with the ROC curve optimal threshold technique to evaluate any biases in the areal forecasts. Additional models were trained at the 2.54 and 12.70 mm $h^{-1}$ thresholds to evaluate the skill in predicting lighter and heavier precipitation, but only the 6.35 mm $h^{-1}$ models were physically evaluated with variable importance and the case study. The 0.25, 2.54, 6.35, and 12.70 mm $h^{-1}$ were chosen because they correspond to the 0.01, 0.1, 0.25, and 0.5 in $h^{-1}$ thresholds, respectively, that are used for determining trace and heavy precipitation amounts. The probabilities shown in the case study are the joint probabilities of precipitation greater than or equal to 6.35 and 0.25 mm $h^{-1}$. They were calculated by multiplying the conditional probability of precipitation greater than or equal to 6.35 mm $h^{-1}$ with the probability of precipitation greater than or equal to 0.25 mm $h^{-1}$.

# 4. Results

*a. Deterministic Rain Model Evaluation*

Evaluation of the deterministic forecasts of precipitation location shows that the multiple logistic regression and random forest do add skill compared to the uncalibrated ensemble probability and the simple logistic regression. Fig. 3 shows only slight variations in the optimal threshold with forecast hour and that the thresholds for each machine learning algorithm are similar. The low threshold for the raw ensemble indicates that the best precipitation coverage is found when any ensemble member forecasts rain. PSS and ETS show similar hourly trends, but PSS is higher than ETS since ETS tends to maximize at a higher probability threshold. The multi-predictor machine learning algorithms (random forest and multiple logistic regression) provide the most improvement for the first 15 hours of the forecast with only a slight improvement for the afternoon and evening. The improvement comes from an increased POD without also increasing the FAR. Similar results are seen in grid 2 in terms of thresholds and score values, and the multi-predictor algorithms are able to provide consistent improvement over all forecast hour (Fig. 4). 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, including points with no radar coverage.

*b. Probabilistic Threshold Exceedance Model Evaluation*

### 1) Evaluation over All Forecast Hours

The attributes diagrams for the raw ensemble and machine learning algorithms show how the forecasts were altered to improve their calibration and skill. Fig. 5 contains the attributes diagrams for the ensemble and algorithms applied to grids 1 and 2 at the 6.35 mm $h^{-1}$ threshold for all forecast hours. The raw ensemble in both domains tends to be overconfident with probabilistic forecasts above climatology and under confident with forecasts below climatology, resulting in a negative BSS. In spatial terms, the ensemble probabilities are too high for areas where it forecasts heavy rain, and it is missing some areas where rain actually occurred. The simple logistic regression improves the ensemble forecast by rescaling the probabilities based on the ensemble mean rain amount. In grid 1 and grid 2 this generally results in all of the probabilities being scaled closer to the climatological probability to address the general overconfidence of the ensemble. This scaling brings the observed relative frequencies above climatology into the positive skill zone, but they are still overconfident.

The multiple logistic regression and random forest incorporate additional ensemble variables to add more information to the probability estimates. This does result in a large improvement in BSS over the simple logistic regression. The multiple logistic regression performs the best in grid 1 because it produces nearly perfect reliability between 0 and 50% while maintaining positive skill from 60 to 90%. The random forest also places its forecasts close to the perfect reliability line but slightly further away than the stepwise logistic regression, and the forecast probabilities do not exceed 70%, so the random forest has a lower maximum resolution than the multiple logistic regression.

In grid 2, the BSS for all three models decreases. The observed relative frequency for the random forest lies just above the no skill line and then increases above 50%, which may be a due to chance from the low sample size in the higher percentage range. The multiple logistic regression performs better with the lower probability forecasts but has negative skill with the higher probabilities. The weaker performance may be due to having fewer heavy rain events in grid 2 during the study time period.

The ROC curves show that the multiple-predictor machine learning algorithms enhance the discrimination abilities of the ensemble. The raw ensemble in both subgrids has slightly positive skill in terms of AUC (Fig. 6). At the optimal PSS threshold, the raw ensemble detects only 67% of heavy rain events (POD) in grid 1, and 86% of its positive forecasts are false alarms (FAR). In grid 2, the detection ability is worse with a 66% POD and a 91% FAR. The Bias scores greater than 1 indicate a larger proportion of false alarms than misses. Since the simple logistic regression rescales the predictions over a smaller probability range, it has a slightly lower AUC than the raw ensemble. Its optimal threshold is higher than the raw ensemble, so it has a correspondingly lower POD, POFD, and FAR. It also has a smaller bias score at the optimal threshold. The multiple logistic regression and random forest have a much larger AUC and POD and similar FAR compared to the raw ensemble. The multiple logistic regression and random forest have very similar scores with only slight differences in the POFD, FAR, and bias. The same relationships hold in grid 2 except that there is a slightly larger difference in the scores of the multiple logistic regression and random forest. The overall scores are also slightly lower for grid 2, which is likely due to the lower frequency of convective precipitation events in the training data.

### 2) Evaluation by Forecast Hour

Comparisons of BSS and AUC by hour and by sub-grid show additional trends in the probability of precipitation forecasts. The raw ensemble consistently has the worst BSS (Fig. 7). For all models, the best performance occurs at forecast hour 1 then decreases sharply at hour 2 before stabilizing. This initial decrease is likely due to the radar data assimilation placing the storms in the same place initially and then having the individual storms diverge from the predicted storm motions. There is a slight increase in performance between hours 6 and 12 for grid 2. This increase may be due to the diurnal cycle of convection resulting in larger storm clusters during this time period, or it may be due to the model fully spinning up. There is another major decrease in performance in grid 1 between hours 12 and 24. This time period is when the greatest uncertainty exists due to convective initiation and the tendency for initial convection to be isolated. The multiple logistic regression and the random forest do not have any statistically significant ($\alpha < 0.05$) differences in their AUC and BSS. The biggest departure from the ensemble mean in terms of AUC occurred in the 18 to 24 hour range for both grids, which corresponds with peak convective activity. The simple logistic regression did still improve on the ensemble forecast in terms of reliability but not to the same extent as the multiple-predictor models, and it made the AUC worse. The temporal trends in the BSS match those found in the 2009 SSEF by Johnson and Wang (2012). The multiple-predictor methods appeared to produce similar increases in BSS to the single predictor neighborhood and object-based methods.

### c. *Evaluation of Multiple Precipitation Thresholds*

Machine learning models were trained at 3 precipitation thresholds (2.54, 6.35, and 12.70 mm $h^{-1}$) in order to evaluate how skill varies with precipitation intensity. Fig. 8 shows the random forest BSS by hour for the three precipitation thresholds. All three follow the same diurnal patterns, but the 2.54 mm $h^{-1}$ forecasts have consistently much higher skill than the other thresholds. While the 2.54 and 6.35 mm $h^{-1}$ show skill for all forecast hours, the 12.70 mm $h^{-1}$ forecasts do not show any skill from 14 to 26 hours. The decreasing skill with threshold size is likely due to the smaller heavy precipitation areas and spatial and timing errors in placement of convection in the ensembles. Similar results were found for the other machine learning methods (not shown).

### d. *Variable Importance*

Variable importance 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 indicative of how randomizing the values 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 8 variable importance scores for the random forests trained on each 6-hour period in Grid 1 are shown in Table 5. In the first 6 hours, all of the top 5 variables are aggregations of the hour precipitation or precipitable water with the rest being hour max upward wind. In this time period, the ensemble members

are very similar, so there is great overlap among the precipitation regions. By hours 7-12, the max upward wind becomes more dominant in the rankings, although the precipitation max and standard deviation are still found among the top 5 variables. Vertical velocities become the most common feature in hours 13 through 18 with only precipitable water and specific humidity contributing moisture information. For hours 19 through 30, the standard deviation of Surface-Based CAPE appears, which is likely associated with the presence of nearby boundaries.

The grid 2 variable importances highlight the greater importance of moisture and lower importance of vertical velocities in the Eastern United States. The predicted precipitation is again only important in hours 1 through 12, but the precipitable water and specific humidity show high importance through the entire forecast period. Upward and Downward winds are in the rankings for each time period, but they tend to be on toward the bottom of the top 8. Surface-Based CAPE is also important in the latter hours of the forecasts, but the mean and max are selected instead of the standard deviation.

### *e. Case Study: 13 May 2010*

The case of 13 May 2010 illustrates the spatial characteristics, strengths, and weaknesses of the precipitation forecasts from the SSEF and the machine learning methods. Since the SSEF is run through forecast hour 30, the last six forecast hours of one run overlap with the first six hours of the next run. This overlap allows for the comparison of two runs on the same observations and illustrates the effects of lead time on the forecast probabilities. Fig. 9 shows the distribution of the observed 1-hour precipitation at 02 UTC on May 13. The bulk of the

precipitation originates from a broken line of discrete supercell thunderstorms positioned along a stationary front stretching from northern Oklahoma into northern Missouri with additional storms in Iowa and ahead of the dry line in Oklahoma and Texas. Additional precipitation is falling in Virginia from a mesoscale convective system. A comparison of the 2-hour and 26-hour dew point and temperature SSEF forecasts (Fig. 10) shows major differences in the placement and strength of the fronts and dry line. The cold front is further north and more diffuse in the 26 hour forecast while the dry line is further east, moving from the central Texas panhandle to western Oklahoma. The differences in the placement of the surface boundaries also affect the placement of the precipitation areas in both forecasts.

Comparisons of the raw ensemble probabilities and the simple logistic regression (Fig. 11) show the effects of calibrating the probabilities only on precipitation. All of the precipitation probabilities shown are generated by multiplying the conditional PQPF of greater than 6.35 mm $h^{-1}$ with the probability of precipitation, which removed spurious low conditional probabilities of precipitation. In the 2 hour forecast, there is little dispersion among the ensemble members, so there is very high confidence in the raw ensemble. In this case though, the high confidence areas are displaced west from the actual heavy precipitation regions, resulting in some missed precipitation areas. The simple logistic regression corrects for this overconfidence by lowering the SSEF probabilities. The area with a greater than 10% probability is noticeably smaller and approximates the area occupied by the observed rain areas. Since it rescales the probabilities, it does not translate the predicted rain areas, resulting in more misses. At 26 hours, the raw ensemble probabilities are lower and spread over a wider area, indicating greater ensemble dispersion. The probabilities in Oklahoma and Kansas are shifted further east due to the positioning of the surface boundaries. In this

case, it results in the observed precipitation being captured better than they were in the 2 hour forecast with fewer misses but more false alarms. The simple logistic regression again reduces the probabilities and the area covered by them so that it does not have significantly more false alarms than at 2 hours, but it misses some precipitation that was captured by the raw ensemble. Of the machine learning methods, the simple logistic is best for accurately depicting areal coverage, but it overcorrected in terms of downscaling the raw probabilities.

The results from the multiple-predictor methods show the advantages of incorporating additional predictors. In Fig. 12, the multiple logistic regression is compared with the random forest at 2 and 26 hours. In the 2 hour forecasts, both the multiple logistic regression and random forest expand their probabilities over a wider area and shift them slightly east compared to the raw ensemble, enabling them to capture the heavy precipitation fully. The random forest also captures the precipitation areas north of the cold front better than the logistic regression, which may weight SBCAPE too highly and not handle cold sector precipitation as well. At 26 hours, the areas increase and the probabilities are generally lower. Both algorithms capture nearly all precipitation areas, but they also produce more false alarms. Precipitation did extend through central Texas into Mexico ahead of the dry line, but none of that precipitation was heavy. The multiple logistic regression highlighted that precipitation as well as ahead of the dry line. Unlike the 2 hour forecast, it did cover the precipitation north of the front in Iowa. The random forest covered all of the precipitation areas in the Plains and seemed to handle the east coast precipitation coverage better.

The extended probability areas of the random forest and multiple logistic regression can be explained by examining some of the variables deemed important by the random forest. In Fig. 13, the Hour Max Upward Wind, SBCAPE, Precipitable Water, and 850 mb Specific

Humidity are shown. The Hour Max Upward Wind was the most important variable for the 25-30 hour random forest (Table 5), and its spatial distribution closely matches that of the random forest, including the positive areas in southwest Texas. It also supports the precipitation north of the cold front in Iowa. The SBCAPE was highest in central Texas in the area where both machine learning models placed probabilities but no heavy rain occurred. The machine learning algorithms were able to account for situations in which high CAPE and no precipitation did not occur together by not assigning probabilities in the Gulf of Mexico. Precipitable water and 850 mb specific humidity were also high over a wide area and are generally correlated well with precipitation but occur over a much larger area than the actual precipitation. By weighting these variables together, the multi-predictor algorithms were able to grow the forecasted area to include points at which the mix of ingredients was favorable for precipitation.

# 5. Discussion

The results of the machine-learning post-processing of storm scale ensemble precipitation forecasts displayed not only improvements to the forecasts but also limitations of the ensemble, the algorithms, and grid-point-based framework. First, the post-processing model performance is constrained by the information available from the ensemble. If most of the ensemble members are predicting precipitation in the wrong place or not at all, and the environmental conditions are also displaced, then the machine learning algorithm will not be able to provide much additional skill. Second, the machine learning model will only make predictions based on the range of cases it has previously seen. For higher rain thresholds, the

algorithms will need more independent samples in order to make skilled predictions. Third, the grid-point framework does not fully account for the spatial information and error in the ensemble. The spatial error could be incorporated further by smoothing the verification grid with a neighborhood filter (Ebert 2009) or warping the ensemble forecast to more closely match the observed precipitation (Gilleland et al. 2010) and then training. Machine learning algorithms could also be used to improve the calibration from object-based frameworks (Johnson and Wang 2012) and could incorporate information from the additional predictors within the bounds of the objects.

# 6. Conclusions

Multiple machine learning algorithms were applied to the 2010 CAPS Storm Scale Ensemble Forecast (SSEF) system in order to improve the calibration and skill of its probabilistic heavy precipitation forecasts. Two types of machine learning methods were compared over a period from May 3 through June 18 with both verification statistics and a case study. Verification statistics showed that all of the machine learning methods improved the calibration of the SSEF precipitation forecasts but only the multiple-predictor methods were able to calibrate the models better and discriminate more skillfully between light and heavy precipitation cases. Hourly performance varied with diurnal storm cycles and the increasing dispersiveness of the ensemble members. Comparisons of the rankings of predictors indicated that the ensemble predicted precipitation was only important for the first 12 hours of the model runs. After that period, the upward wind and atmospheric moisture variables became better indicators of the placement of precipitation. The case study showed that the

multiple-predictor machine learning methods could shift the probability maxima to better match the actual precipitation areas, but they would also produce more false alarm areas in the process. For shorter-term forecasts, the false corrections were made without a significant increase in the false alarm area. Calibrating the probabilities with only ensemble rainfall predictions results in predicted areas that are too small and still displaced from the observed precipitation. The multiple-predictor machine learning algorithms did prove especially beneficial in that situation. Ultimately, machine learning techniques can provide an enhancement to precipitation forecasts by consistently maximizing the potential of the available information.

*Acknowledgments.*

Special thanks go to my Master's committee members Fanyou Kong and Michael Richman. Zac Flamig provided assistance with the NMQ data. CAPS SSEF forecasts were supported by a grant (NWSPO-2010-201696) from the NOAA Collaborative Science, Technology, and Applied Research (CSTAR) Program and the forecasts were produced at the National Institute for Computational Science (http://www.nics.tennessee.edu/). Scientists at CAPS, including Fanyou Kong, Kevin Thomas, Yunheng Wang, and Keith Brewster contributed to the design and production of the CAPS ensemble forecasts. This study was funded by the NSF Graduate Research Fellowship under Grant 2011099434 and by NSF grant AGS-0802888. The editor, Michael Baldwin, and the two anonymous reviewers provided very helpful feedback that strengthened the quality of the paper.

# REFERENCES


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.

Ashley, S. T. and W. S. Ashley, 2008: Flood fatalities in the united states. *J. Appl. Meteorol.*, **47**, 805–818.

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. and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418.

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

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

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

546 Doswell, C. A., R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in
547 rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585.

548 Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006:
549 New dimension of NCEP short-range ensemble forecasting (SREF) system: Inclusion
550 of wrf members. *Preprint, WMO Expert Team Meeting on Ensemble Prediction System*,
551 Exeter, United Kingdom, WMO, 5, URL `http://www.emc.ncep.noaa.gov/mmb/SREF/`
552 `reference.html`.

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

555 Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts.
556 *Weather and Forecasting*, **24 (6)**, 1498–1510.

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

559 Gagne II, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using
560 decision trees. *Journal of Atmospheric and Oceanic Technology*, **26 (7)**, 1341–1353.

561 Gagne II, D. J., A. McGovern, and M. Xue, 2012: Machine learning enhancement of storm
562 scale ensemble precipitation forecasts. *Proceedings of the Conference on Intelligent Data*
563 *Understanding (CIDU)*, Boulder, CO, IEEE-CIS, 39–46.

564 Gilleland, E., et al., 2010: Spatial forecast verification: image warping. Tech. Rep.
565 NCAR/TN-482+STR, NCAR.

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

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

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

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

574 Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration
575 using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*,
576 **136**, 2620–2632.

577 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.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013: *An Introduction to Statistical Learning with Applications in R*. Springer, 430 pp.

Johnson, A. and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077.

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.

Kusiak, A. and A. Verma, 2011: Prediction of status patterns of wind turbines: A data-mining approach. *J. Sol. Energy Eng.*, **133**, doi:10.1115/1.4003188.

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

Marsh, P. T., J. S. Kain, V. Lakshmanan, A. J. Clark, N. Hitchens, and J. Hardy, 2012: A method for calibrating deterministic forecasts of rare events. *Wea. Forecasting*, **27**, 531–538.

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

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.

Murphy, A. H., 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.

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

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.

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

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

621 Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3d ed., Academic Press,
622 676 pp.

623 Williams, J. K., 2013: Using random forests to diagnose aviation turbulence. *Mach. Learn.*,
624 doi:10.1007/s10 994–013–5346–7.

625 Xue, M., K. K. Droegemeier, and V. Wong, 2000: The Advanced Regional Prediction System
626 (ARPS) - A multiscale nonhydrostatic atmospheric simulation and prediction model. Part
627 I: Model dynamics and verification. *Meteor. Atmos. Phys.*, **75**, 161–193.

628 Xue, M., D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced
629 Regional Prediction System (ARPS), storm-scale numerical weather prediction and data
630 assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170.

631 Xue, M., et al., 2001: The Advanced Regional Prediction System (ARPS) - A multiscale
632 nonhydrostatic atmospheric simulation and prediction model. Part II: Model physics and
633 applications. *Meteor. Atmos. Phys.*, **76**, 134–165.

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

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

640 Zhang, J., Y. Qi, K. Howard, C. Langston, and B. Kaney, 2011: Radar quality index (RQI)—
641 a combined measure of beam blockage and vpr effects in a national network. *Proc. Eighth*
642 *Int. Symp. on Weather Radar and Hydrology*, Exeter, United Kingdom, Royal Meteoro-
643 logical Society.

# List of Tables

TABLE 1. Frequencies of total and sampled grid points for the different precipitation thresholds.

| Grid | 0-0.25 mm | 0.25-6 mm | 6 mm |
|---|---|---|---|
| Grid 1 Total | 218,052,271 | 13,062,623 | 1,884,725 |
| Grid 1 Sampled | 86,704 | 51,802 | 7,490 |
| Grid 2 Total | 194,343,572 | 10,923,189 | 1,137,694 |
| Grid 2 Sampled | 77,210 | 4,3218 | 4,497 |

TABLE 2. The names and descriptions of model predictors sampled from the SSEF runs. CAPE is Convective Available Potential Energy, and CIN is Convective Inhibition.

| Variable | Levels |
|---|---|
| Hour Precipitation | Surface |
| Surface-Based CAPE | Surface |
| Surface-Based CIN | Surface |
| Dewpoint | 2 m |
| Pressure | Mean Sea Level |
| Composite Radar Reflectivity | Column Maximum |
| Precipitable Water | Column Sum |
| Height | 700 mb |
| U-Wind | 700 mb |
| V-Wind | 700 mb, 500 mb |
| Specific Humidity | 850 mb, 700 mb, 500 mb |
| Temperature | 2 m, 850 mb, 700 mb, 500 mb |
| Hour Max Reflectivity | Column and Time Maximum |
| Hour Max Upward Wind | Column and Time Maximum |
| Hour Max Downward Wind | Column and Time Maximum |

TABLE 3. An example binary contingency table for whether or not rain is forecast.

| | | Observed | |
|---|---|---|---|
| | | Yes | No |
| Forecast | Yes | $a$ (hit) | $b$ (false alarm) |
| | No | $c$ (miss) | $d$ (true negative) |

TABLE 4. 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 | $\frac{a+b}{a+c}$ |

TABLE 5. Top 8 variable importance scores for the random forests trained over 6-hourly periods in Grid 1.

| Variable | Mean | Variable | Mean |
|---|---|---|---|
| Grid 1 Hours 1-6 | | Grid 1 Hours 7-12 | |
| Hour Precipitation SD | 0.0118 | Hour Max Upward Wind Mean | 0.0129 |
| Precipitable Water Mean | 0.0107 | Hour Max Upward Wind Max | 0.0126 |
| Hour Precipitation Mean | 0.0105 | Hour Max Upward Wind SD | 0.0113 |
| Precipitable Water Max | 0.0103 | Hour Precipitation SD | 0.0105 |
| Hour Precipitation Max | 0.0096 | Hour Precipitation Max | 0.0088 |
| Hour Max Upward Wind Mean | 0.0088 | Hour Max Downward Wind SD | 0.0087 |
| Hour Max Upward Wind Max | 0.0088 | Hour Max Downward Wind Min | 0.0084 |
| Hour Max Upward Wind SD | 0.0081 | Specific Humidity 700 mb Max | 0.0076 |
| Grid 1 Hours 13-18 | | Grid 1 Hours 19-24 | |
| Hour Max Upward Wind Mean | 0.0096 | Hour Max Downward Wind Min | 0.0114 |
| Hour Max Downward Wind SD | 0.0088 | Surface-Based CAPE SD | 0.0112 |
| Hour Max Downward Wind Mean | 0.0086 | Hour Max Downward Wind Mean | 0.0110 |
| Hour Max Downward Wind Min | 0.0084 | Hour Max Downward Wind SD | 0.0105 |
| Hour Max Upward Wind SD | 0.0071 | Specific Humidity 850 mb Mean | 0.0092 |
| Precipitable Water Max | 0.0070 | Temperature 700 mb Min | 0.0089 |
| Specific Humidity 700 mb Max | 0.0069 | Hour Max Upward Wind Mean | 0.0088 |
| Hour Max Upward Wind Max | 0.0062 | Hour Max Upward Wind Max | 0.0084 |
| Grid 1 Hours 25-30 | | | |
| Hour Max Upward Wind Mean | 0.0123 | | |
| Hour Max Downward Wind Mean | 0.0119 | | |
| Hour Max Upward Wind SD | 0.0112 | | |
| Hour Max Upward Wind Max | 0.0108 | | |
| Hour Max Downward Wind SD | 0.0105 | | |
| Hour Max Downward Wind Min | 0.0104 | | |
| Specific Humidity 850 mb Mean | 0.0078 | | |
| Surface-Based CAPE SD | 0.0077 | | |

TABLE 6. Top 8 variable importance scores for the random forests trained over 6-hourly periods in Grid 2.

| Variable | Mean | Variable | Mean |
|---|---|---|---|
| Grid 2 Hours 1-6 | | Grid 2 Hours 7-12 | |
| Hour Precipitation Mean | 0.0116 | Precipitable Water Mean | 0.0111 |
| Precipitable Water Mean | 0.0114 | Precipitable Water Min | 0.0085 |
| Precipitable Water Min | 0.0104 | Precipitable Water Max | 0.0079 |
| Precipitable Water Max | 0.0097 | Hour Max Upward Wind Mean | 0.0077 |
| Hour Max Upward Wind Mean | 0.0094 | Hour Max Upward Wind SD | 0.0066 |
| Hour Max Reflectivity Mean | 0.0091 | Hour Precipitation Mean | 0.0065 |
| Hour Max Upward Wind Max | 0.0082 | Hour Precipitation SD | 0.0062 |
| Hour Precipitation Max | 0.0080 | Hour Max Upward Wind Max | 0.0062 |
| Grid 2 Hours 13-18 | | Grid 2 Hours 19-24 | |
| Precipitable Water Mean | 0.0077 | Specific Humidity 850 mb Mean | 0.0162 |
| Hour Max Upward Wind Mean | 0.0075 | Specific Humidity 850 mb Max | 0.0128 |
| Specific Humidity 850 mb Mean | 0.0065 | Hour Max Upward Wind Mean | 0.0110 |
| Precipitable Water Min | 0.0065 | Surface-Based CAPE Max | 0.0106 |
| Precipitable Water Max | 0.0056 | Surface-Based CAPE Mean | 0.0103 |
| Temperature 700 mb Mean | 0.0056 | Hour Max Upward Wind SD | 0.0101 |
| Hour Max Upward Wind Max | 0.0055 | Hour Max Downward Wind Mean | 0.0098 |
| Temperature 500 mb Min | 0.0054 | Hour Max Upward Wind Max | 0.0090 |
| Grid 2 Hours 25-30 | | | |
| Specific Humidity 850 mb Mean | 0.0114 | | |
| Hour Max Upward Wind Mean | 0.0094 | | |
| Precipitable Water Mean | 0.0084 | | |
| Specific Humidity 850 mb Max | 0.0083 | | |
| Hour Precipitation SD | 0.0072 | | |
| Hour Max Upward Wind Max | 0.0071 | | |
| Hour Max Downward Wind Mean | 0.0070 | | |
| Surface-Based CAPE Max | 0.0069 | | |

# List of Figures

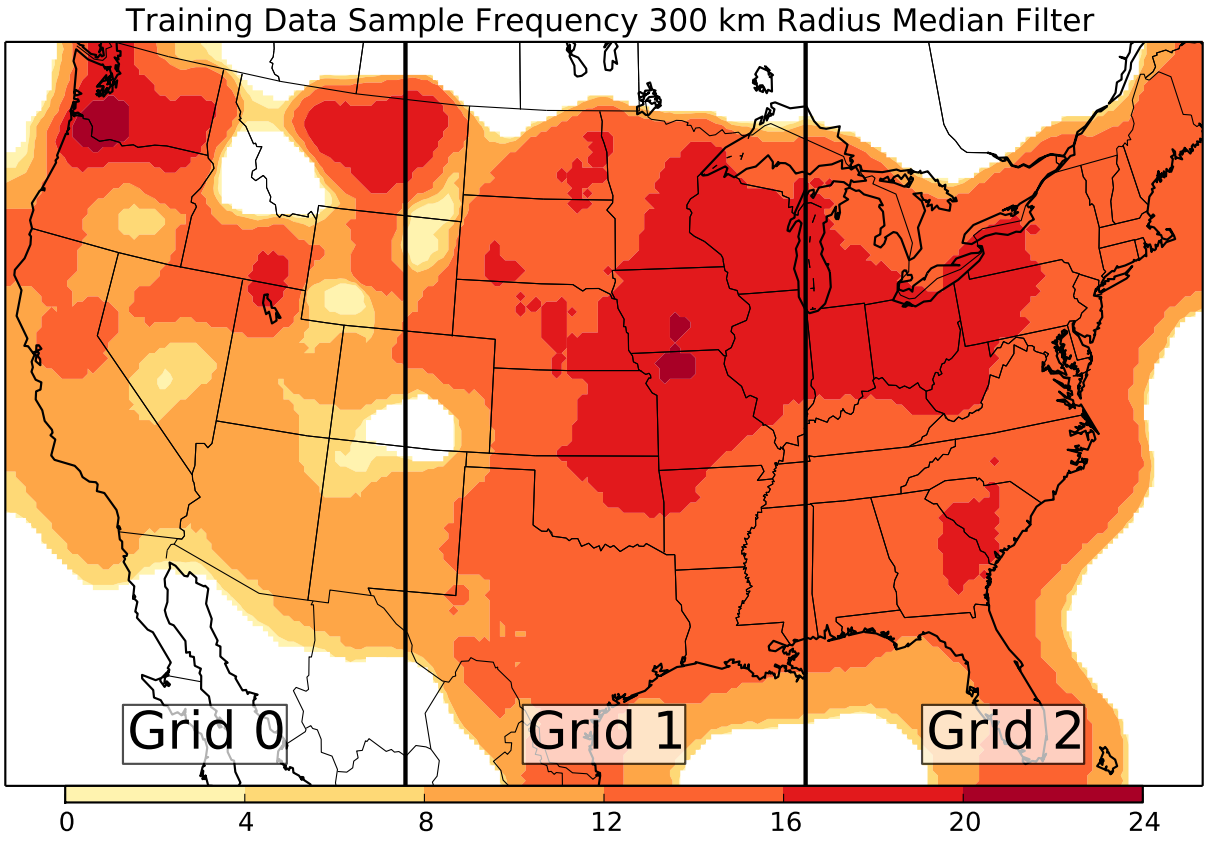


FIG. 1. Map of the number of grid points sampled within a 400 $km^2$ box spatially averaged with a 300 km radius median filter. The domain sub grids are also shown and labeled.

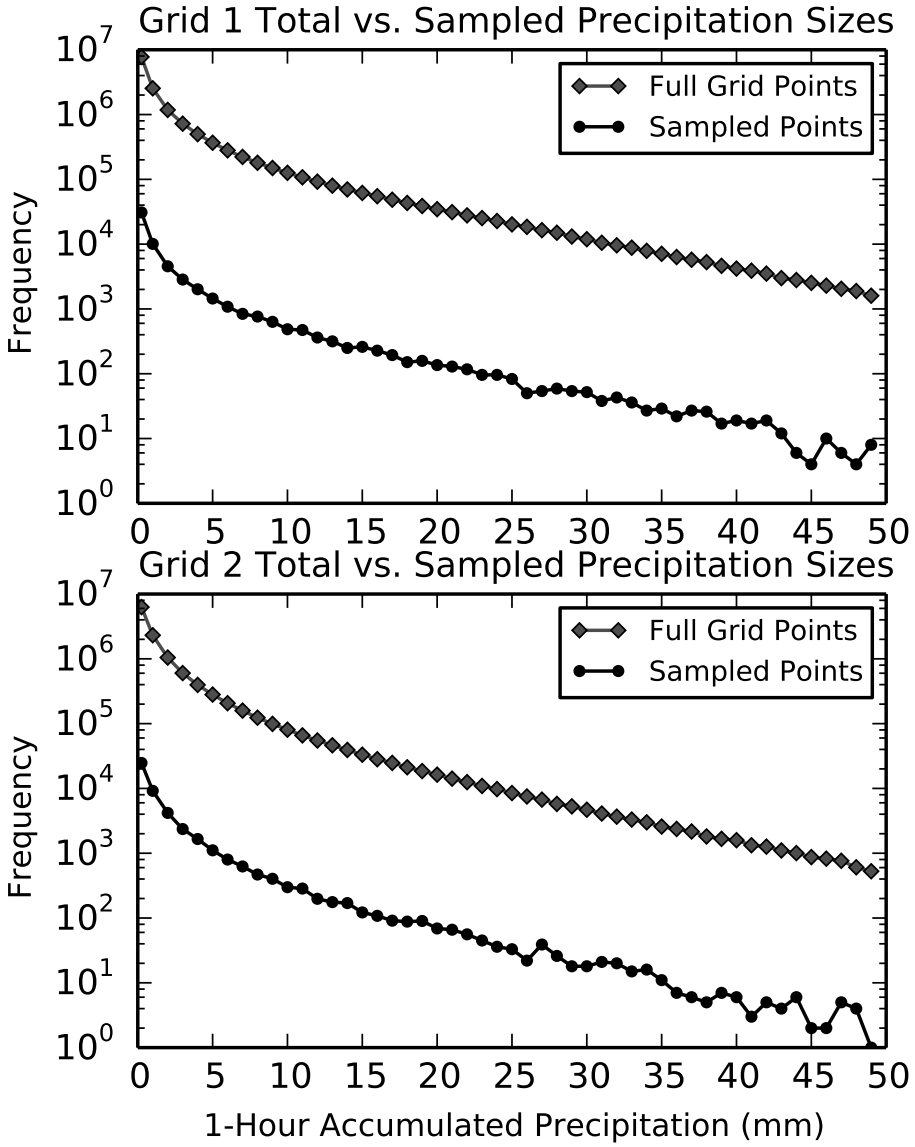


FIG. 2. Histogram comparing the relative frequencies of the full precipitation distribution for the each subgrid to the sampled rainfall distributions.

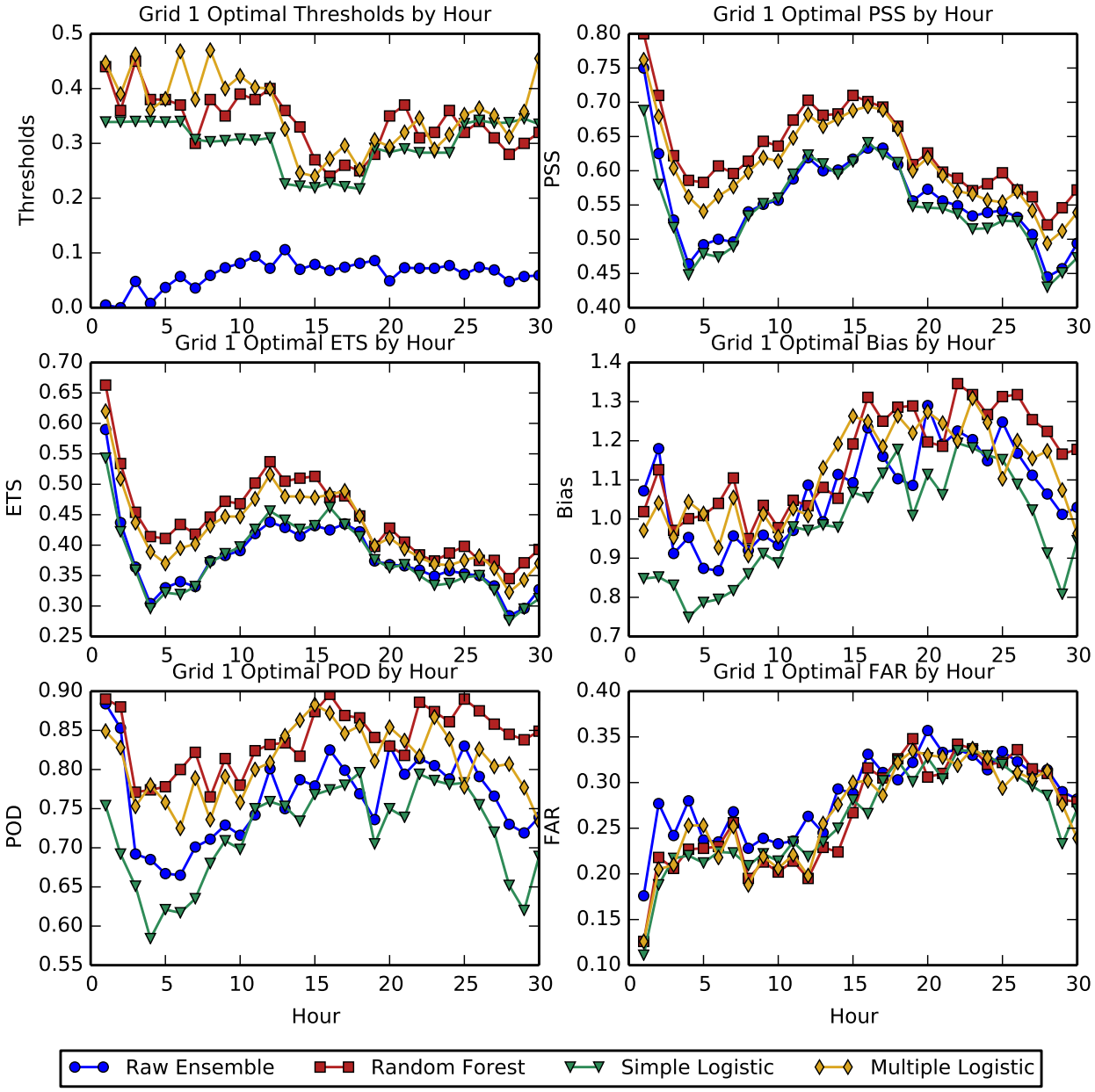


FIG. 3. For a precipitation threshold of 0.25 mm $h^{-1}$, the optimal probability thresholds (OT) and binary verification statistics calculated at the OT for the SSEF and each machine learning model in grid 1.

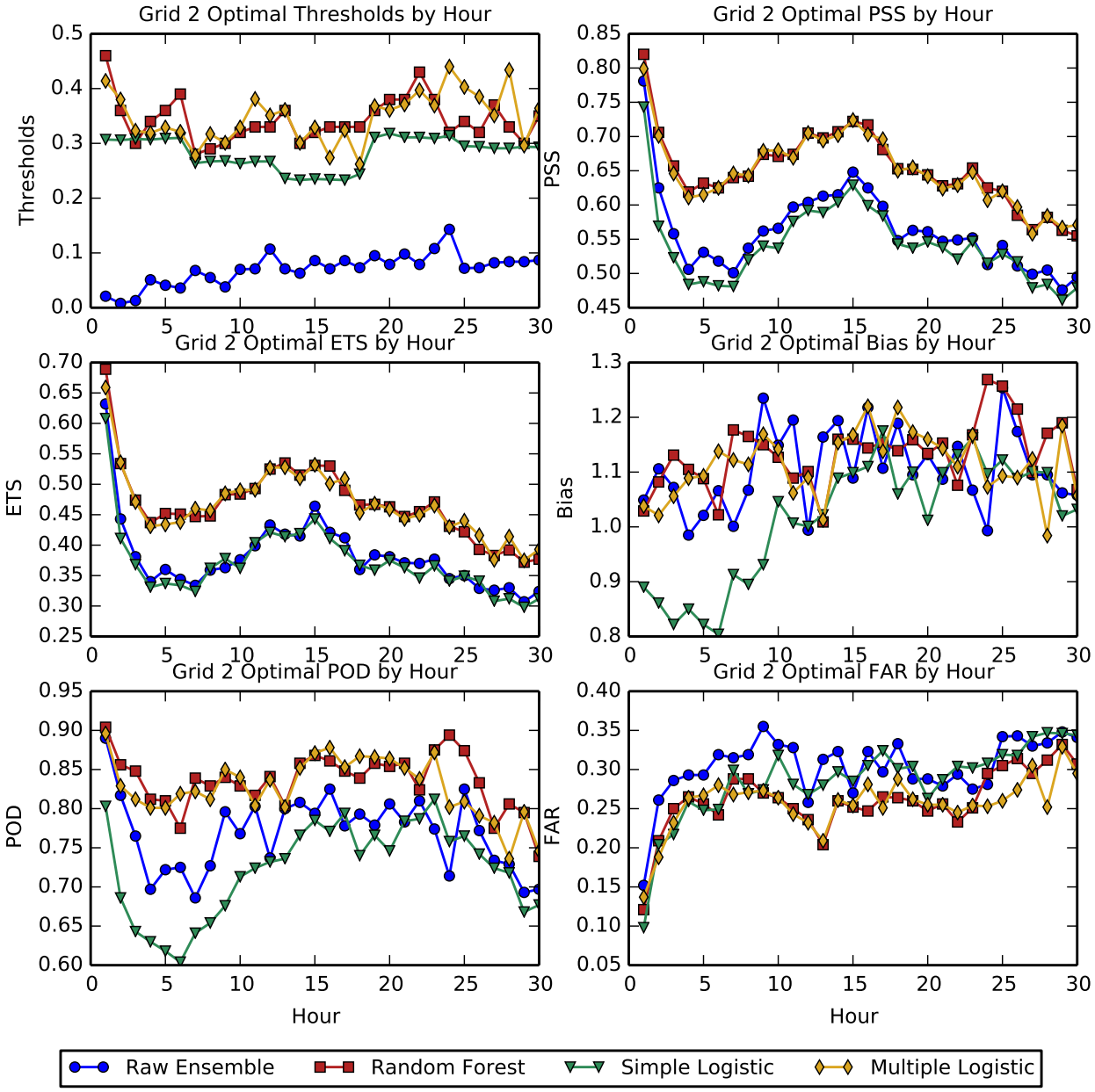


FIG. 4. For a precipitation threshold of 0.25 mm $h^{-1}$, the optimal probability thresholds (OT) and binary verification statistics calculated at the OT for the SSEF and each machine learning model in grid 2.

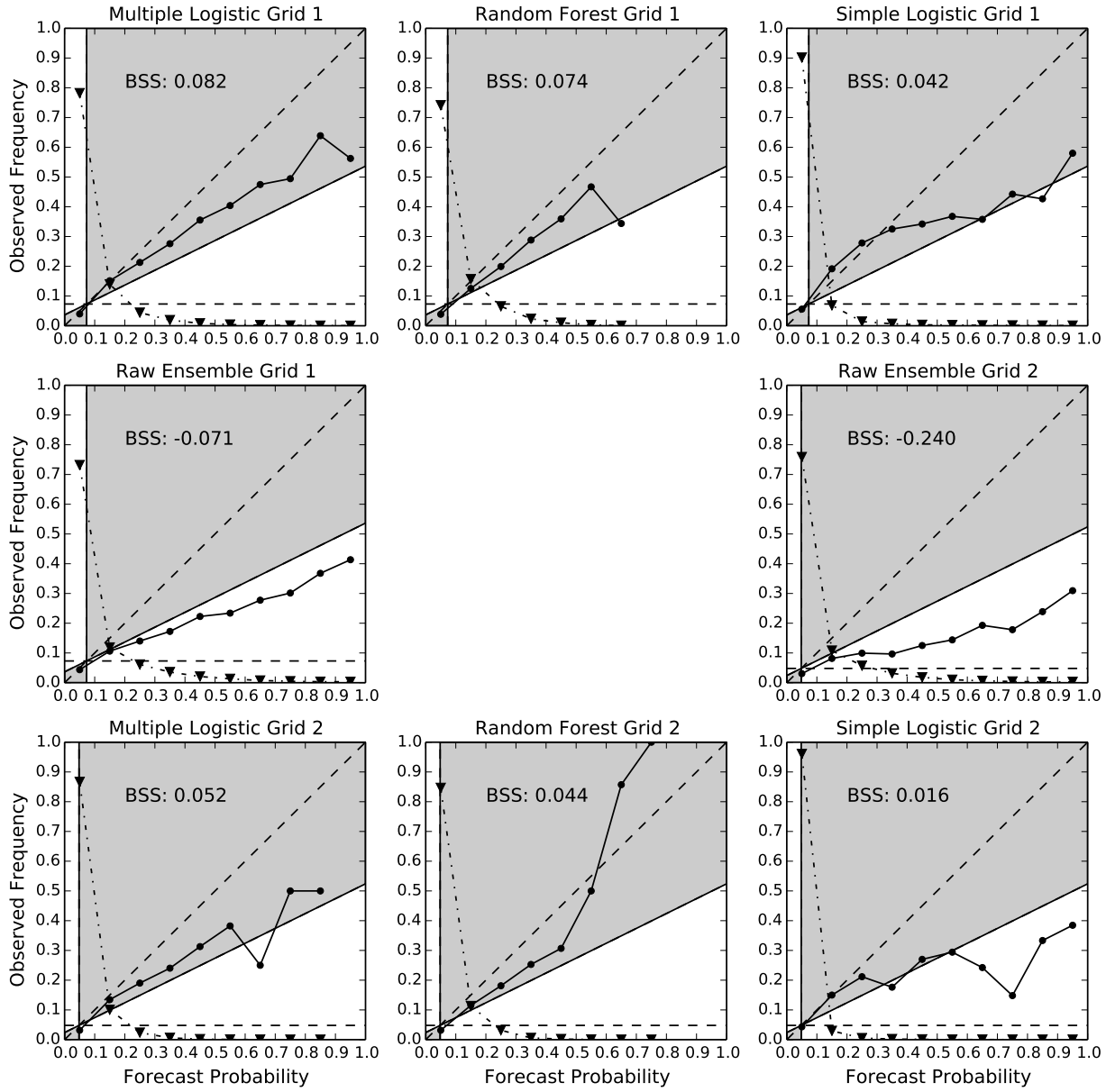


FIG. 5. Attributes diagrams for the raw ensemble and each machine learning model. The solid line with circles is the reliability curve, which indicates the observed relative frequency of rain events above the 6.35 mm threshold for each probability bin. The dot-dashed line with triangles shows the relative frequency of all forecasts that fall in each probability bin. If a point on the reliability curve falls in the gray area, it contributes positively to the BSS. The horizontal and vertical dashed lines are located at the climatological frequency for a particular sub grid. The diagonal dashed line indicates perfect reliability.

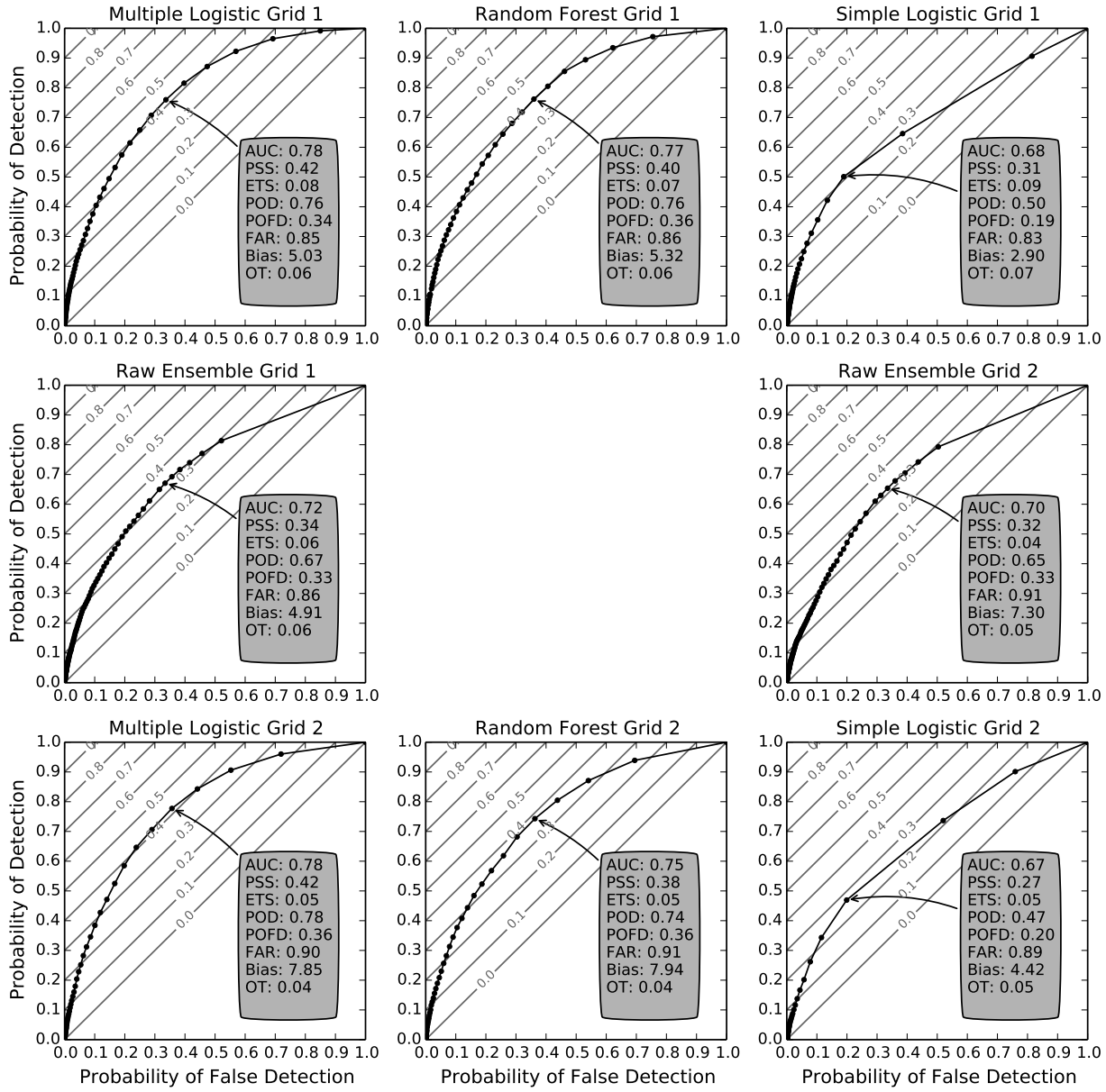


FIG. 6. Relative Operating Characteristic (ROC) curves for the raw ensemble and each machine learning model. The diagonal lines indicate PSS.

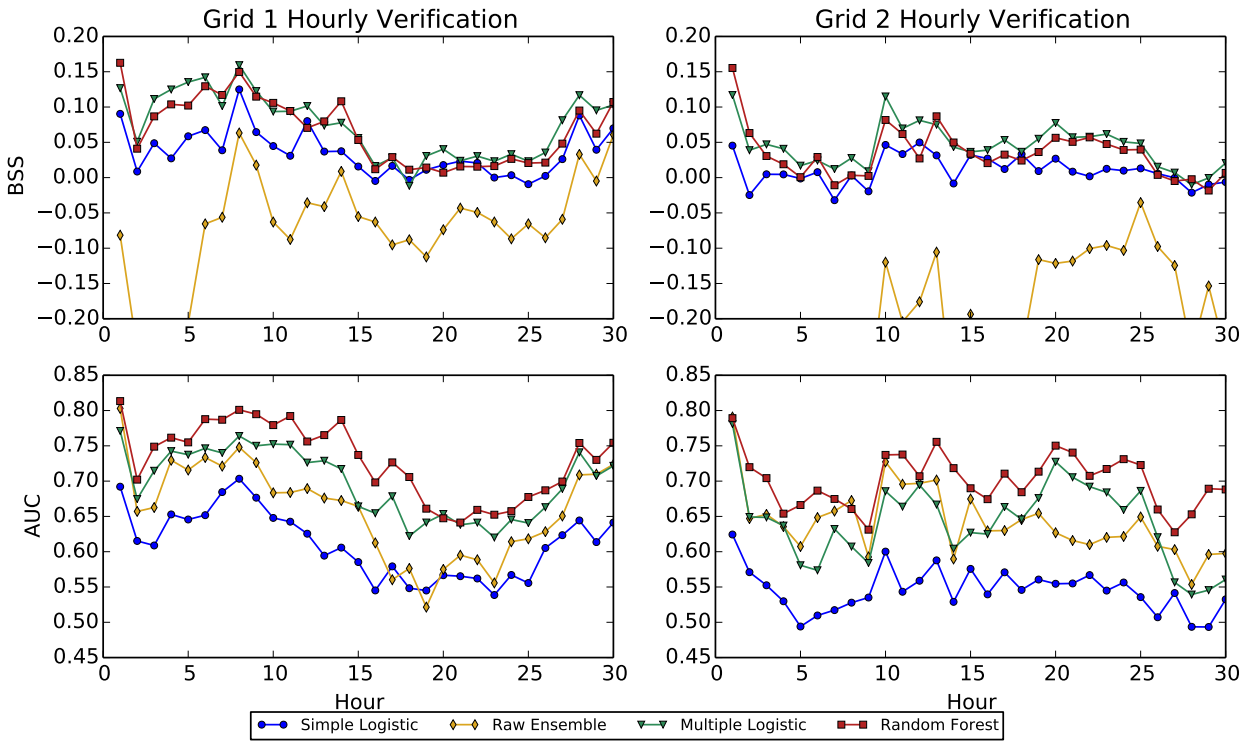


FIG. 7. BSS and AUC comparisons by hour.

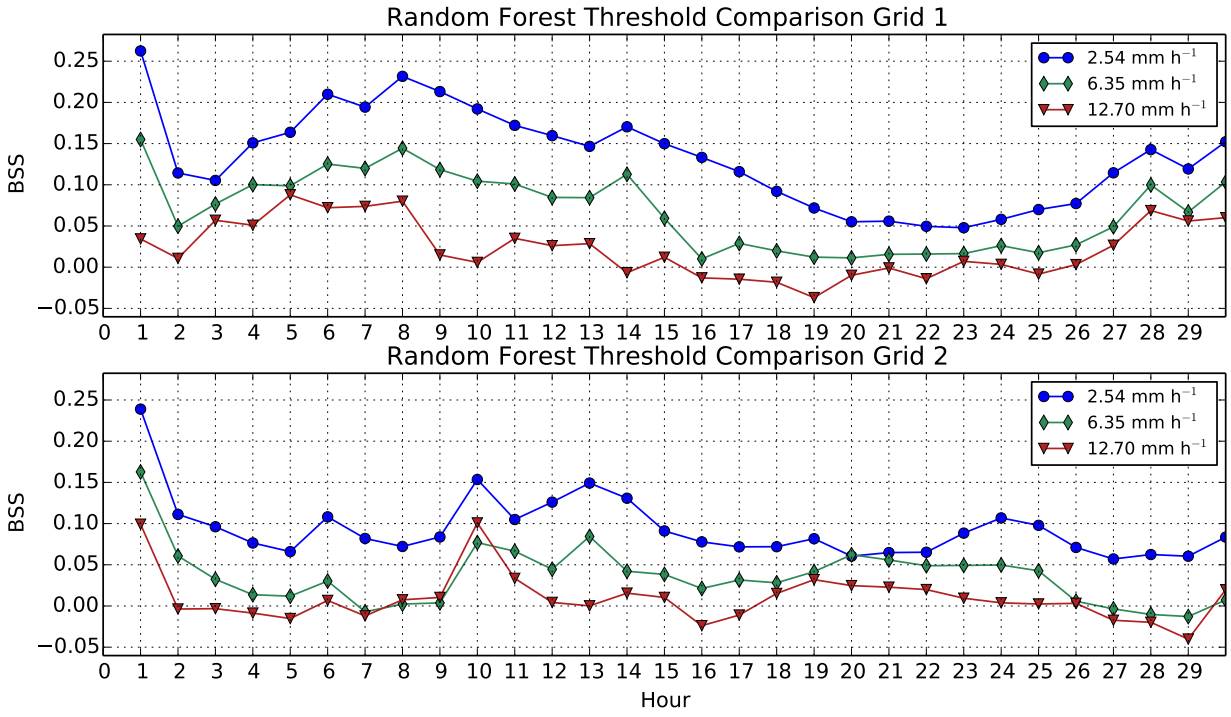


FIG. 8. Evaluation of random forests trained with 3 precipitation thresholds at each forecast hour. Brier Skill Score (BSS) is the evaluation metric.

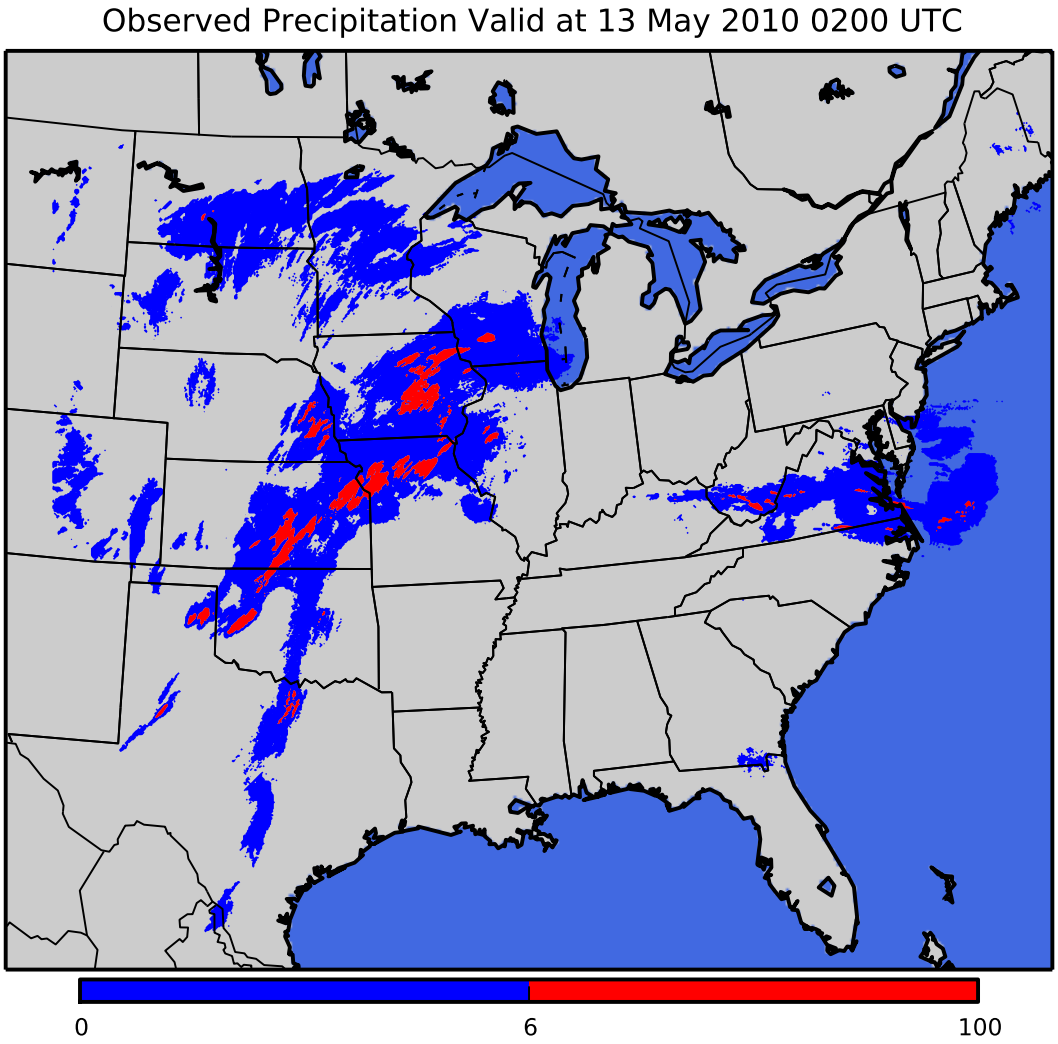


FIG. 9. Observed 1-hour precipitation on 13 May 2010 at 02 UTC.

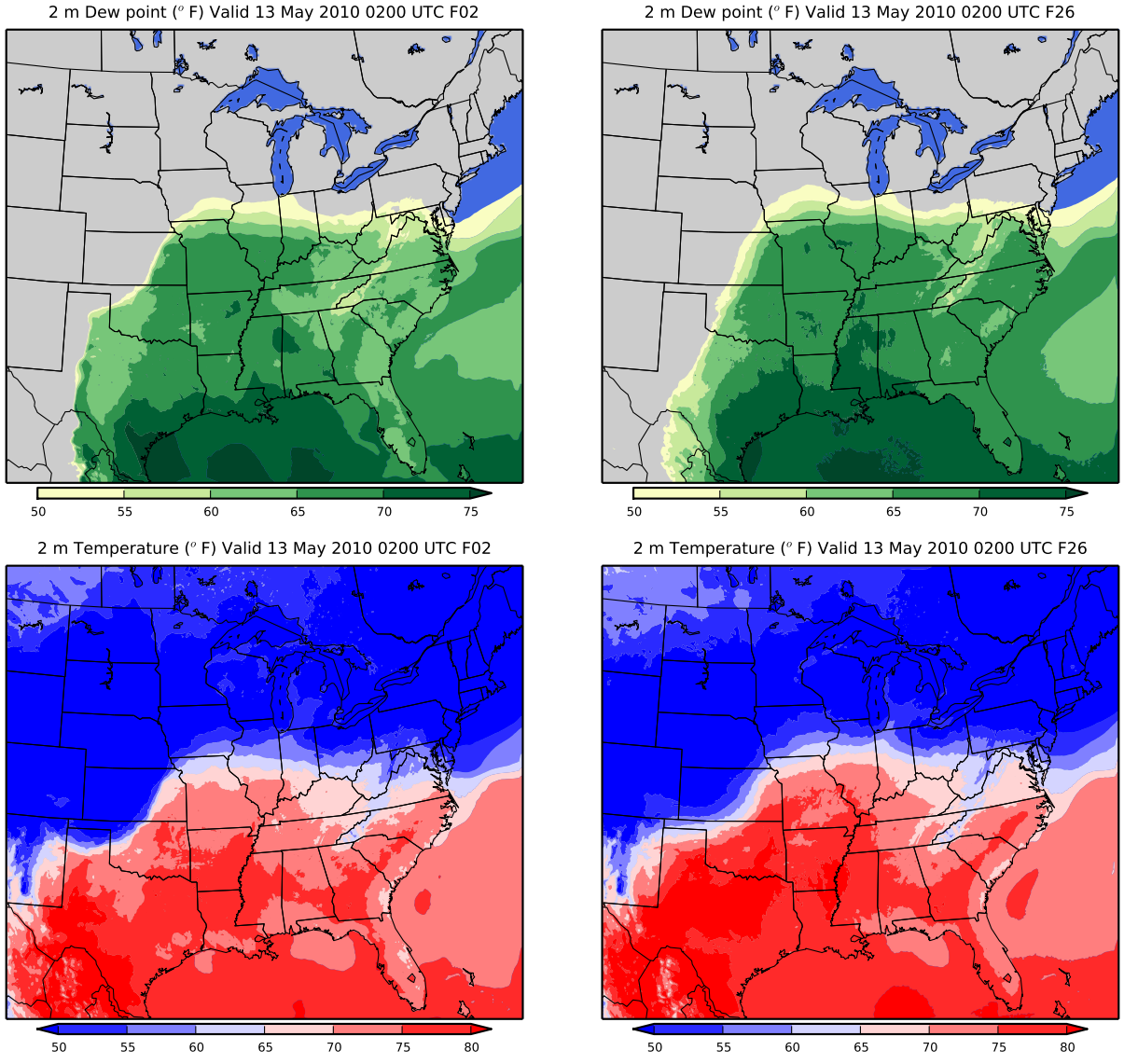


FIG. 10. Filled contours of 2 m dew point and 2 m temperature for the 2-hour and 26-hour SSEF forecasts valid on 13 May 2010 at 0200 UTC.

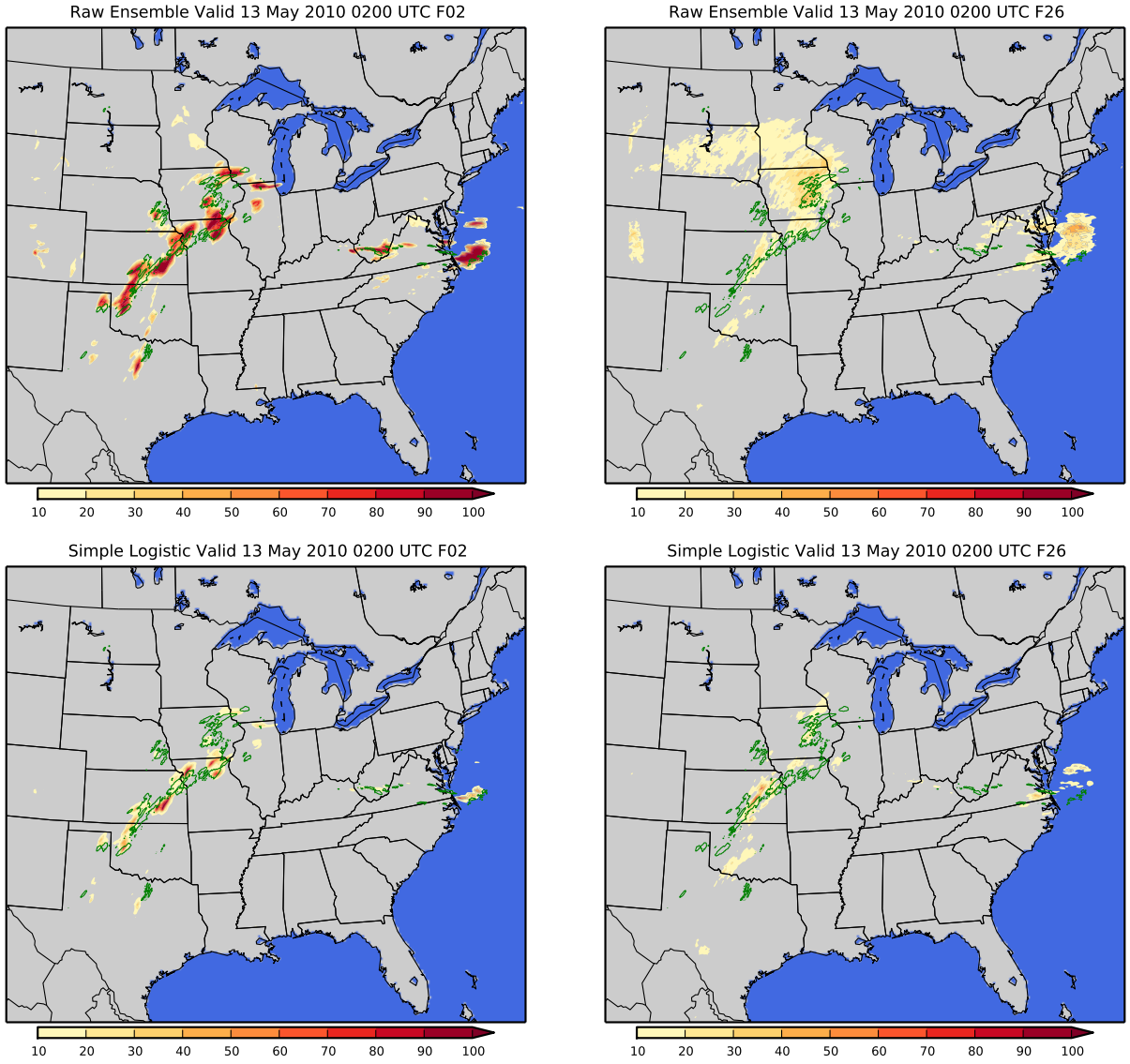


FIG. 11. The 2-hour (left) and 26-hour (right) forecasts of the SSEF ensemble probability (top) and the simple logistic regression (bottom). The green contours indicate the observed areas of 1-hour precipitation greater than 6.35 mm $h^{-1}$.

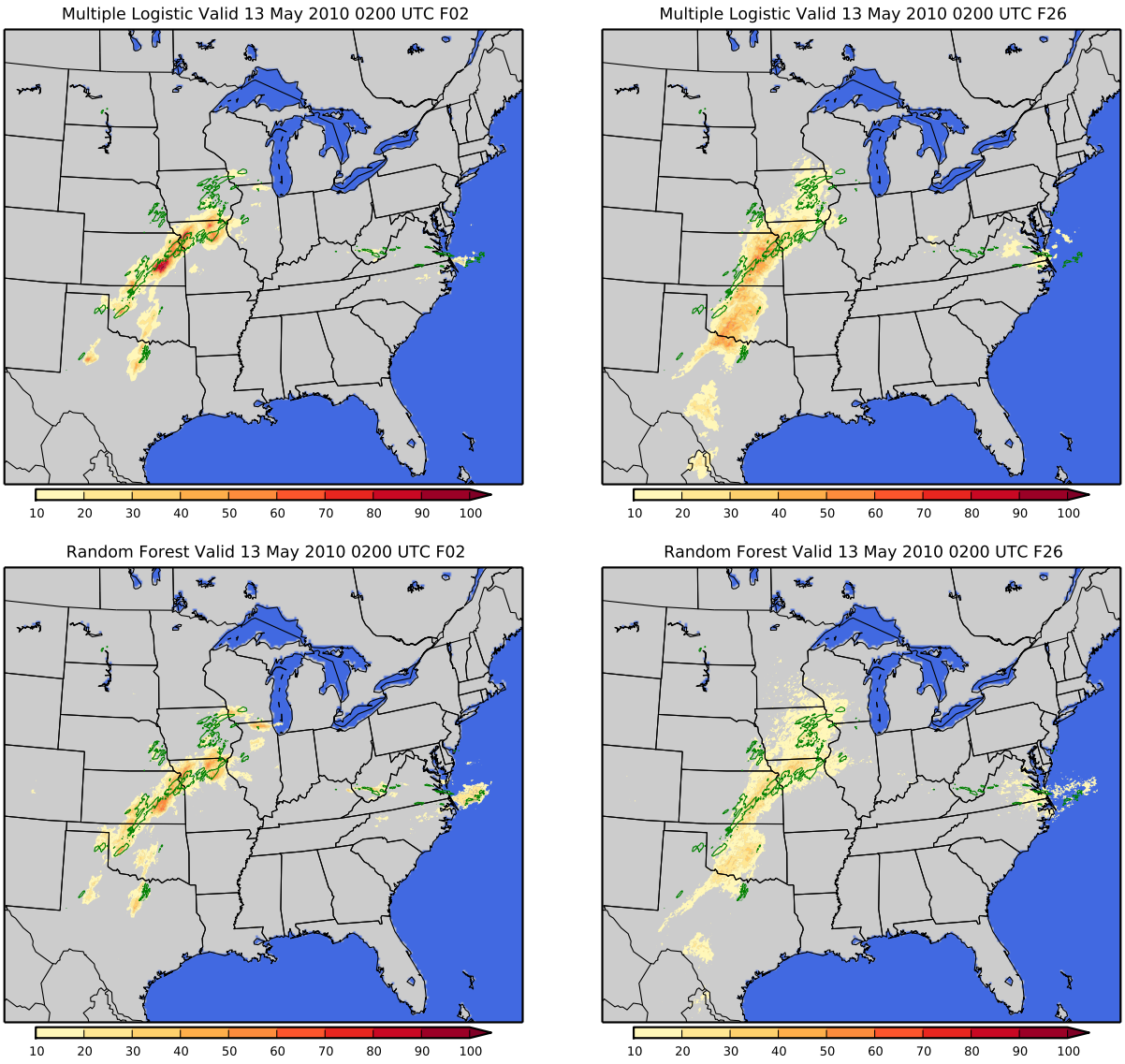


FIG. 12. The 2-hour (left) and 26-hour (right) forecasts of the multiple logistic regression (top) and the random forest (bottom). The green contours indicate the observed areas of 1-hour precipitation greater than 6.35 mm $h^{-1}$.

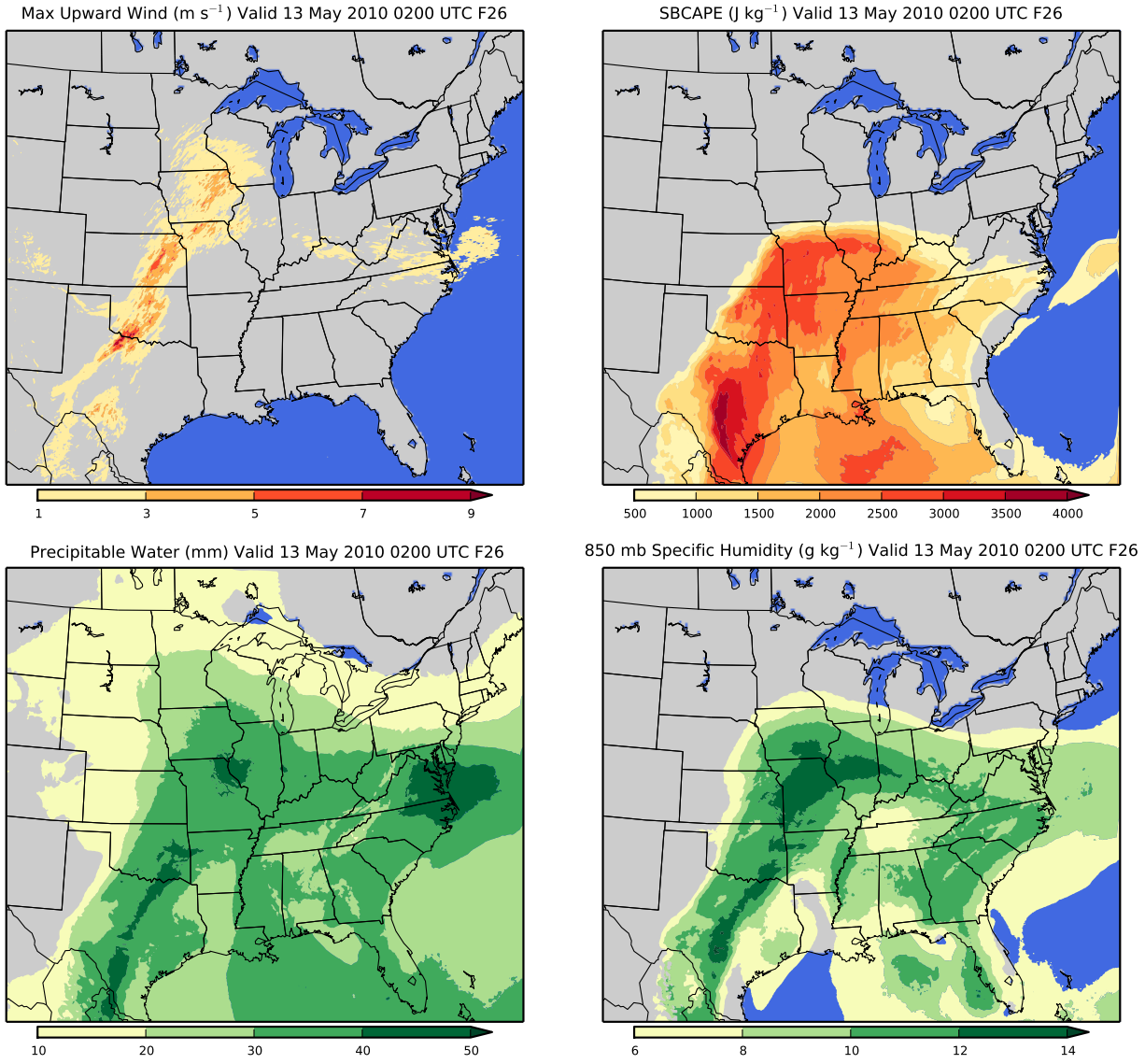


FIG. 13. Maps of the SSEF 26-hour forecasts of predictors considered important by the random forest model.