Machine Learning Enhancement of Storm Scale Ensemble Precipitation Forecasts

David John Gagne II
School of Meteorology
University of Oklahoma
Norman, Oklahoma
djgagne@ou.edu

Amy McGovern
School of Computer Science
University of Oklahoma
Norman, Oklahoma
amcgovern@ou.edu

Ming Xue
Center for the Analysis and Prediction of Storms
School of Meteorology
University of Oklahoma
Norman, Oklahoma
mxue@ou.edu

***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.**

I. 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 [1], [2], [3]. 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 [3]. 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.

Lessons from previous studies on ensemble post-processing of precipitation forecasts provided guidance on our approach to post-processing a storm scale ensemble. Model Output Statistics (MOS; [4], [5]) produced probability of precipitation forecasts from deterministic NWP models using multivariate linear regression to select the most relevant variables from the model and to fit a function. Variables selected in the original MOS equations included primarily precipitation and precipitable water. The MOS

approach was extended to ensemble NWP forecasts with the use of a linear regression fit to a Gamma distribution [6]. Extensive work on creating the best ensemble precipitation forecasts in terms of both areal coverage and magnitude was done in [2] and found that using the ensemble mean for areal coverage and probability matching the precipitation magnitudes produced the best forecast out of the techniques tried. Bremnes [1] 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. Further ensemble post-processing work was done in [7] and [8] by combining the logistic regression approach with extensive reforecast runs of the GFS ensemble to produce skilled short term and extended probability of precipitation forecasts. Bayesian Model Averaging (BMA) [9], a weighted average of the probability distribution functions for bias-corrected ensemble member forecasts, has also been applied to precipitation forecasts [10] with a Gamma distribution in place of the Gaussian distribution. BMA can produce probabilistic, deterministic, and interval forecasts. Both BMA and Bremnes also use a separate algorithm to predict probability of precipitation in addition to the precipitation amount. Our approach builds on these previous approaches by incorporating a range of model variables, using ensemble statistics instead of individual members, and utilizing non-parametric algorithms. In addition, our approach is applied to a much higher resolution model than was used in any previous study, incorporates a radar-mosaic for grid- instead of station-based verification, and uses advanced ensemble machine learning techniques.

II. Data and Methods

A. 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 [11], [12]. The 2010 SSEF is composed of 26 separate model runs from the Weather Research and Forecasting (WRF) Advanced Research WRF and Non-hydrostatic 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 [13], which ran from 3 May to 18 June. The SSEF provides hourly model output over the continental United States at 4 km horizontal resolution 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

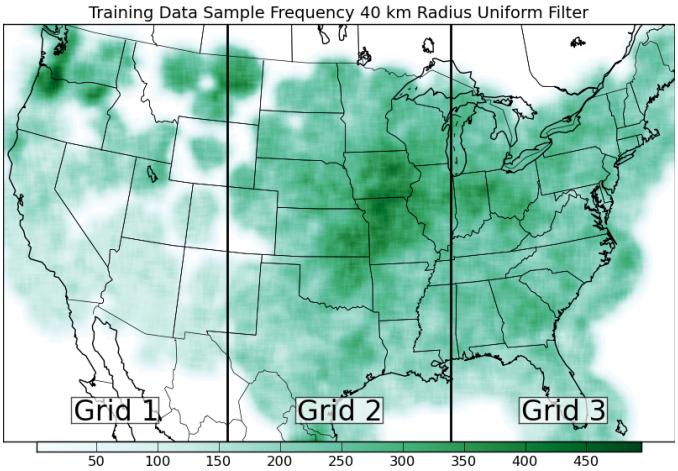

Figure 1. Map of spatial histogram of sampled points within a 20 km radius of each point. The domain sub grids are also shown and labeled.

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 [14] where the Ensembled Continuous Bayesian Networks (ECBN) algorithm is introduced. They compare performance to the random forests presented here.

B. Verification Data

A storm scale verification dataset was paired with the storm scale ensemble forecasts. The National Mosaic Multi-Sensor QPE (NMQ) [15] 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.

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 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 stratified randomly sampled from each subdomain in areas with quality radar coverage. Grid 1 corresponds to the western third of the domain, Grid 2 corresponds to the central third, and Grid 3 corresponds to the eastern third. Fig. 1 shows that most of the continental US was sampled with the exception of areas of the Rocky Mountains with poor radar coverage. The sampling of areas within radar coverage is not uniform

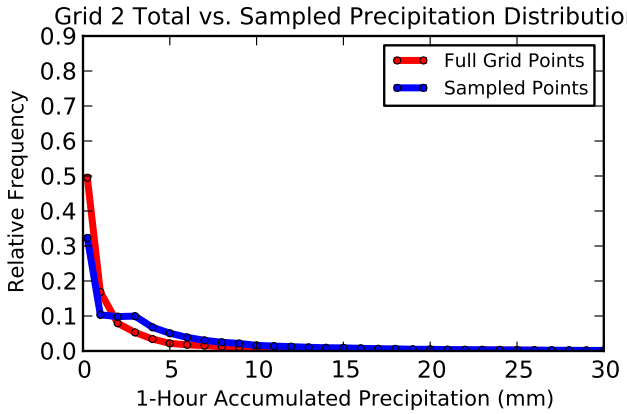

Figure 2. Histogram comparing the relative frequencies of the rainfall distribution for grid 2 to the sampled rainfall distribution.

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 distribution and the full rainfall distribution are shown in Fig. 2. Undersampling of the 0 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 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 [8]. Relevant model variables (Table I) 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 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.

D. Machine Learning Algorithms

Machine learning algorithms utilizing different approaches and assumptions were trained on the probabilistic and quantile datasets. For the probabilistic domain, logistic regressions were used as the baseline algorithm. Logistic

Table I
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

regressions are linear regression models fitted to a logit curve that ranges from 0 to 1. Logistic regressions were trained from the raw ensemble probability and with stepwise selection of variables.

Multivariate adaptive regression splines (MARS; [16]) 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 contains two intersecting linear functions. 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. For computational timing purpose, the maximum number of variables was limited to 15. This project uses the open-source *earth*[1] library as its MARS implementation.

Random forests [17] consist of an ensemble of classification and regression trees 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. Since the tree-building process selects a subset of variables for each tree, a technique called variable importance [17] can be used to rank the relevance of the variables.

In addition to producing a single probability or quantity, two methods can produce fixed probability intervals. Quantile regressions [18] estimate the median and other

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

quantiles by minimizing the absolute error. They also handle non-Gaussian error distributions better. Quantile regression forests [19] 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.

E. 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. The conditional algorithm is trained on only points where some rain occurred, so it implicitly assumes that the rain/no-rain classifier is perfect.

Evaluation techniques depended on the type of forecast. The Brier Skill Score (BSS) [20] is one method used to evaluate probabilistic forecasts. The Brier Skill Score can be decomposed into three terms [21], as shown in Eq. 1:

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

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 [22], 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 Receiver Operating Characteristic (ROC) curve, or AUC [23] was 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.

Deterministic precipitation forecasts are based on the difference from the ensemble mean forecast. This approach produces a distribution closer to Gaussian than predicting rainfall directly and allows the use of Gaussian linear regression. The forecasts are verified with the root mean squared error (RMSE). The RMSE is shown in Eq. 2:

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

where 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 can be evaluated based on the proportion of samples that fall within the quantile intervals as well as the width of the quantiles.

III. Results

A. Probability Forecasts

Probabilistic models were developed for each subdomain to find the probability of 1-hour precipitation exceeding 6.54 mm or 0.25 inches. Due to space limitations, the discussion of the verification will focus on Grid 2. The first part of the process determines the probability of any precipitation occurring. Fig. 3 shows the AUC for each algorithm tested. Random forest consistently has the highest AUC and multi logistic regression and MARS are only slightly worse. The raw ensemble and the calibration logistic regression are significantly worse but still have an AUC that would be considered skilled. Verification scores for the threshold exceedance predictions can be seen in Fig. 4. BSS and AUC both drop severely from hours 1 to 4 before rising again between hours 5 and 10. Slight decreases in performance occur from hours 10 through 25 with an increase in performance again between hours 25 and 30. The decrease in BSS occurs

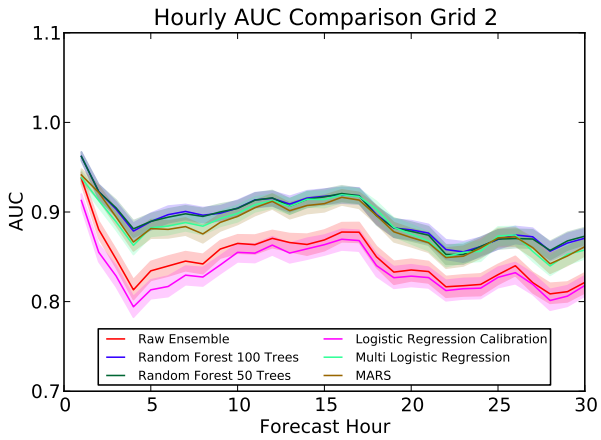

Figure 3. Variation of AUC by hour for probability of precipitation forecast algorithms in Grid 2. The shaded areas indicate the 95% bootstrap confidence intervals.

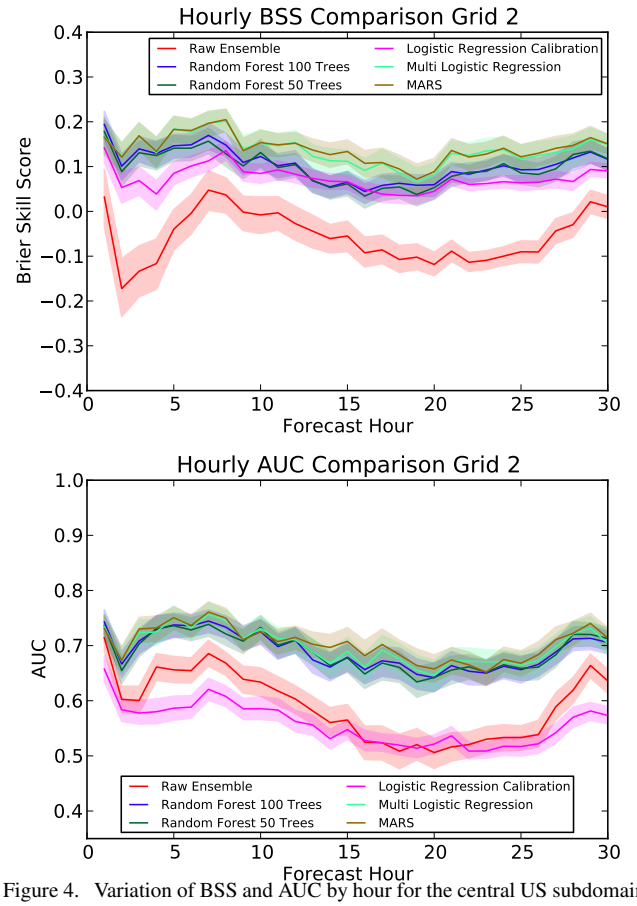

Figure 4. Variation of BSS and AUC by hour for the central US subdomain. The shaded areas indicate the 95% bootstrap confidence intervals.

at a lower rate for the machine learning models compared to the raw ensemble, and all machine learning models stay above 0 for the whole period. MARS and Multi Logistic Regression statistically significantly ($\alpha < 0.05$) outperform random forests and logistic regression in terms of BSS.

AUC shows a slightly different picture than BSS. There is no significant difference in AUC among the multivariate machine learning models. The calibration logistic regression performs worse than the ensemble. The decrease in AUC for the calibration logistic regression is likely due to the logistic regression just increasing the probabilities of the samples but not changing their order relative to each other as well as decreasing the range of probabilities, so the resulting AUC would decrease.

Table II
VARIABLE IMPORTANCE STATISTICS FROM ALL 34 100-TREE RANDOM FORESTS. M IS MEAN AND SD IS STANDARD DEVIATION.

Variable	Z	Mean	SE
V 700 mb M	405.7	0.565	0.0014
V 500 mb M	337.2	0.552	0.0016
Temperature 700 mb M	318.2	0.553	0.0017
Precipitable Water M	304.6	0.553	0.0018
MSLP M	281.5	0.549	0.0020
Max + Vertical Velocity M	270.0	0.519	0.0019
Max - Vertical Velocity M	268.2	0.524	0.0020
U 700 mb M	267.7	0.545	0.0020
Height 700 mb M	266.2	0.546	0.0020
Max - Vertical Velocity SD	236.1	0.519	0.0022

The calibration improvements from each model can be seen in the attributes diagrams from Fig. 5. The raw ensemble is under-predicting the probabilities that are less than 50% and over-predicting those greater than 50%. The calibration logistic regression model adjusts the 0 ensemble probability to correspond with its observed relative frequency near 20%. Since the calibration logistic regression did not have any additional information about the samples beyond the predicted precipitation, this simple transformation was the most it could manage. Random forests, MARS, and multi logistic regression did have additional variables they could select. In many cases, they increased the probability for each sample compared to the raw ensemble, resulting in both higher probabilities for more cases as seen in the bin sizes in Fig. 5 and a wider spatial extent for nonzero probabilities. The BSS was highest for MARS and multiple logistic regression with both random forests having slightly less skill. In this instance, MARS was better at keeping the observed relative frequencies along the perfect reliability line. Since random forests are a non-parametric model, they are more sensitive to the distribution of the training data set.

The top ten most important variables selected by the random forest are shown in Table II. The top two variables correspond with the upper air north-south wind, which is maximized in areas where rain is more likely to occur. The other important variable in the top 5 is mean precipitable water, so where moisture is available for precipitation is more important than where the ensemble actually produces precipitation. The vertical velocity variables are generally tied to where convective updrafts and downdrafts are strongest, which in effect are where the ensemble places convection.

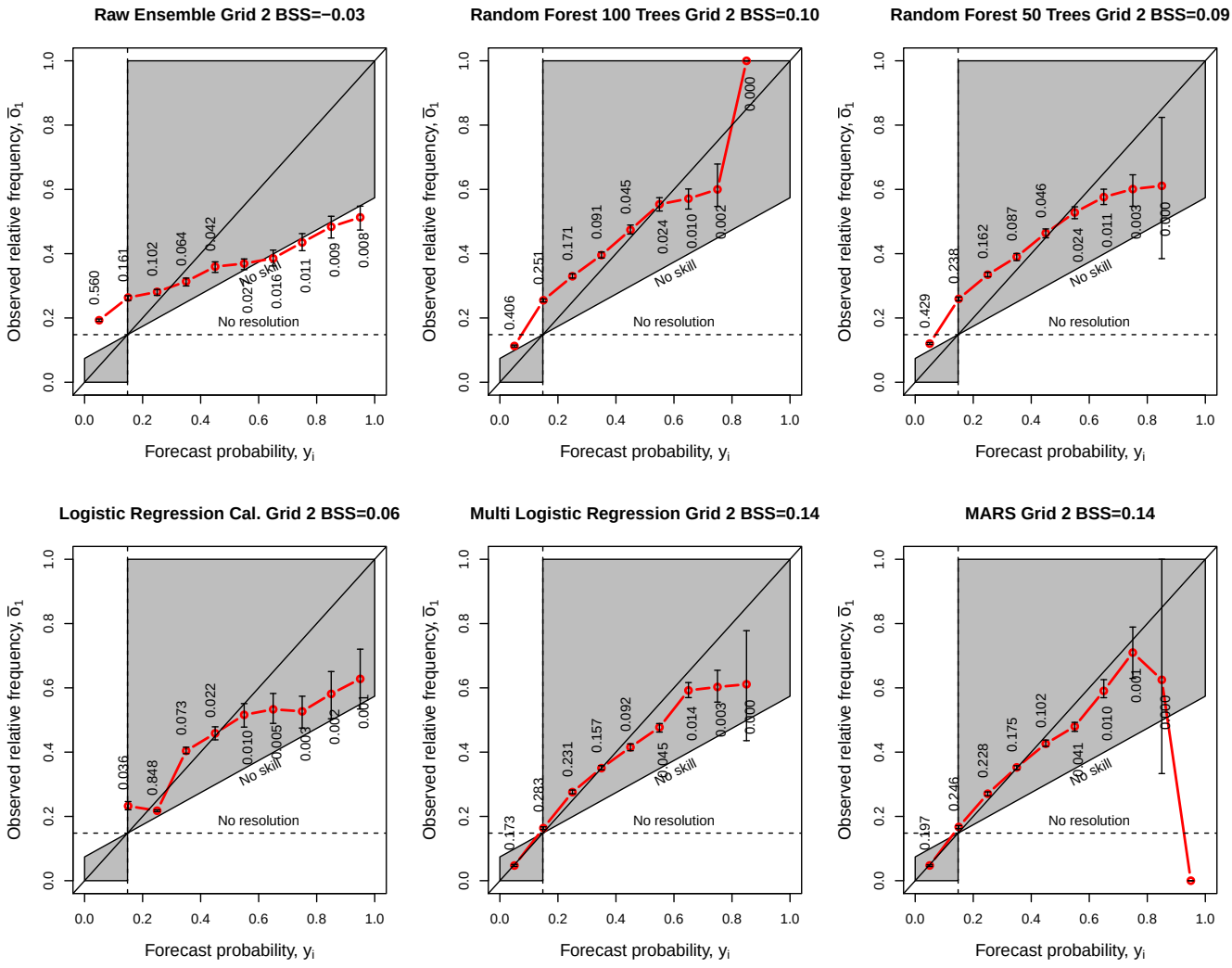

Figure 5. Attributes diagrams for each probabilistic model in grid 2. The data points are at the center of each 10% width bin. The values next to each data point are the proportion of samples within each bin.

The ensemble mean and spread accumulated precipitation forecasts are not among the most important variables because 1-hour precipitation forecasts are heavily dependent on getting the location correct, and the precipitation is often displaced in space and time.

B. Deterministic Forecasts

The ensemble information was also combined to make a single deterministic 1-hour precipitation forecast for each sample. Fig. 6 shows how the RMSEs of the deterministic forecasts vary by hour. As with the probabilistic forecasts, all of the machine learning models decrease the RMSE. The random forest and MARS create the largest decrease in RMSE, although the MARS model is more sensitive to erroneous attribute values, as evidenced by the sharp increases in hourly RMSE. The median prediction from the quantile regression forest performed noticeably worse than the random forest. This difference is likely due to the

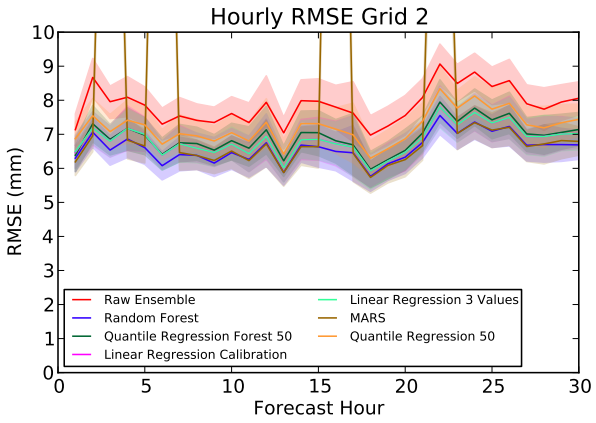

Figure 6. RMSEs by hour for the SSEF median forecast and each machine learning model.

quantile regression forest underestimating the amplitude of

Table III
PERCENTAGE OF SAMPLES IN EACH INTERVAL SECTION.

Model	<5	5-50	50-95	>95
Raw Ensemble	1.2	7.6	34.0	57.3
Quantile Reg. Forest	3.2	47.4	43.7	5.7
Quantile Reg.	5.2	44.7	45.0	5.1

extreme precipitation events since extreme values have less effect on the median than the mean.

C. Interval Forecasts

Evaluation of the interval forecasts from the raw ensemble, quantile regression, and quantile regression forest involves checking that the percentage of samples in each section of the interval match their fixed probability. As seen in Table III, the observed precipitation totals tend to fall above the interval entirely. The quantile regression forest does bring over 80% of the samples within the 5-95 percentile interval, but most of the samples are in the 5-50 percentile interval, so the model exhibits negative skew. The quantile regression, on the other hand, balances the intervals.

D. Case Study: 2010 May 19

Running the post-processing algorithms over the full domain on a particular day revealed how the machine learning techniques improved their verification scores over the raw ensemble. Each model was evaluated at 0000 UTC on 20 May 2010. At that time a line of discrete supercells had formed in central Oklahoma with a widespread area of stratiform rain extending across northern Kansas into Missouri. The probabilistic predictions from the SSEF, multi logistic regression, and 100 tree random forest are shown Fig. 7, along with the observed precipitation. The SSEF placed the highest probabilities of exceeding 6.35 mm of precipitation in southern and central Kansas but gives very low probabilities for the convection in central and southern Oklahoma. No probabilities are given for the precipitation in Missouri and very low probabilities are predicted for the convection in Louisiana. The random forest shifts the higher probabilities southward into Oklahoma and gives some hint that the areas of heavy precipitation would be more isolated. The probabilities in Louisiana are lower and less smooth. The multi logistic regression does the best at capturing the areal coverage of the convection and provides smoother probabilities that still hint at the discrete nature of the convection in Oklahoma. It is the only model that produces any probability of precipitation in Missouri.

While the case study shown is only one time step of one run, some conclusions can be made regarding the machine learning algorithms. The combined probability of precipitation and probability of precipitation exceedance algorithms highlight the areas of greatest threat while constraining the coverage of the probabilities. The multi logistic regression does produce smoother probability contours compared to random forest and maximizes at higher amounts in this instance. While both algorithms handled the area of discrete super cellular convection well, they did not produce high probabilities in the area of stratiform rain. This weakness is likely due to the choice of variables for the post-processing, which are biased toward ingredients for convective precipitation as opposed to stratiform.

IV. 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, multiple logistic regression, and MARS provided significant improvements for probabilistic and continuous forecasts by both increasing the range of precipitation and probabilistic values predicted and by increasing the areal coverage of the precipitation 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 single event to illustrate the geographic variability of the forecasts and tendencies in the predictions.

The verification results and case study also showed that post-processing can be beneficial beyond the larger spatial and temporal scales that had been the focus of previous techniques. By giving more weight to the ingredients for heavy precipitation instead of the model precipitation forecasts, the probabilistic machine learning algorithms accounted for some of the spatial and temporal uncertainty inherent in convective precipitation forecasts. Stratiform rain forecast ingredients can be included in future datasets. Further improvements could be made by incorporating values describing the spatiotemporal variability of the ensemble.

ACKNOWLEDGMENT

Special thanks go to committee members Michael Richman and Fanyou Kong. Zac Flamig provided RQI data and assistance with the NMQ. This study utilized computing resources from the Center for the Analysis and Prediction of Storms. This study was funded by the NSF Graduate Research Fellowship under Grant 2011099434.

REFERENCES

[1] J. B. Bremnes, "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output," *Mon. Wea. Rev.*, vol. 132, pp. 338–347, 2004.

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

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

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

[5] R. J. Bermowitz, "An application of model output statistics to forecasting quantitative precipitation," *Mon. Wea. Rev.*, vol. 103, pp. 149–153, 1975.

[6] T. M. Hamill and S. J. Colucci, "Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts," *Mon. Wea. Rev.*, vol. 126, pp. 711–724, 1998.

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

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

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

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

[11] F. Kong, M. Xue, K. W. Thomas, Y. Wang, K. A. Brewster, X. Wang, J. Gao, S. J. Weiss, A. J. Clark, J. S. Kain, M. C. Coniglio, and J. Du, "Evaluation of CAPS multi-model storm-scale ensemble forecast for the NOAA HWT 2010 spring experiment," in *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.* Seattle, WA: Amer. Meteor. Soc., 2011, p. Paper 457.

[12] M. Xue, F. Kong, K. W. Thomas, Y. Wang, K. Brewster, J. Gao, X. Wang, S. J. Weiss, A. J. Clark, J. S. Kain, M. C. Coniglio, J. Du, T. L. Jensen, and Y. H. Kuo, "CAPS realtime storm scale ensemble and high resolution forecasts for the NOAA hazardous weather testbed 2010 spring experiment," in *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.* Seattle, WA: Amer. Meteor. Soc., 2011, p. PAPER 9A.2.

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

[14] S. Hellman, "Learning ensembles of dynamic continuous bayesian networks," Master's thesis, University of Oklahoma, 2012.

[15] S. Vasiloff, K. W. Howard, J. Zhang, D. H. Kitzmiller, M. G. Mullusky, W. F. Krajewski, E. A. Brandes, R. M. Rabin, D. S. Berkowitz, H. E. Brooks, J. A. McGinley, R. J. Kuligowski, and B. G. Brown, "Improving QPE and very short term QPF: An initiative for a community-wide integrated approach," *Bull. Amer. Meteor. Soc.*, vol. 88, pp. 1899–1911, 2007.

[16] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, pp. 1–67, 1991.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5 – 32, 2001.

[18] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of Economic Perspectives*, vol. 15, pp. 143–156, 2001.

[19] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.

[20] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Wea. Rev.*, vol. 78, pp. 1–3, 1950.

[21] A. H. Murphy, "A new vector partition of the probability score," *J. Appl. Meteor.*, vol. 12, pp. 595–600, 1973.

[22] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Academic Press, 2011.

[23] I. Mason, "A model for assessment of weather forecasts," *Aust. Meteor. Mag.*, vol. 30, pp. 291–303, 1982.

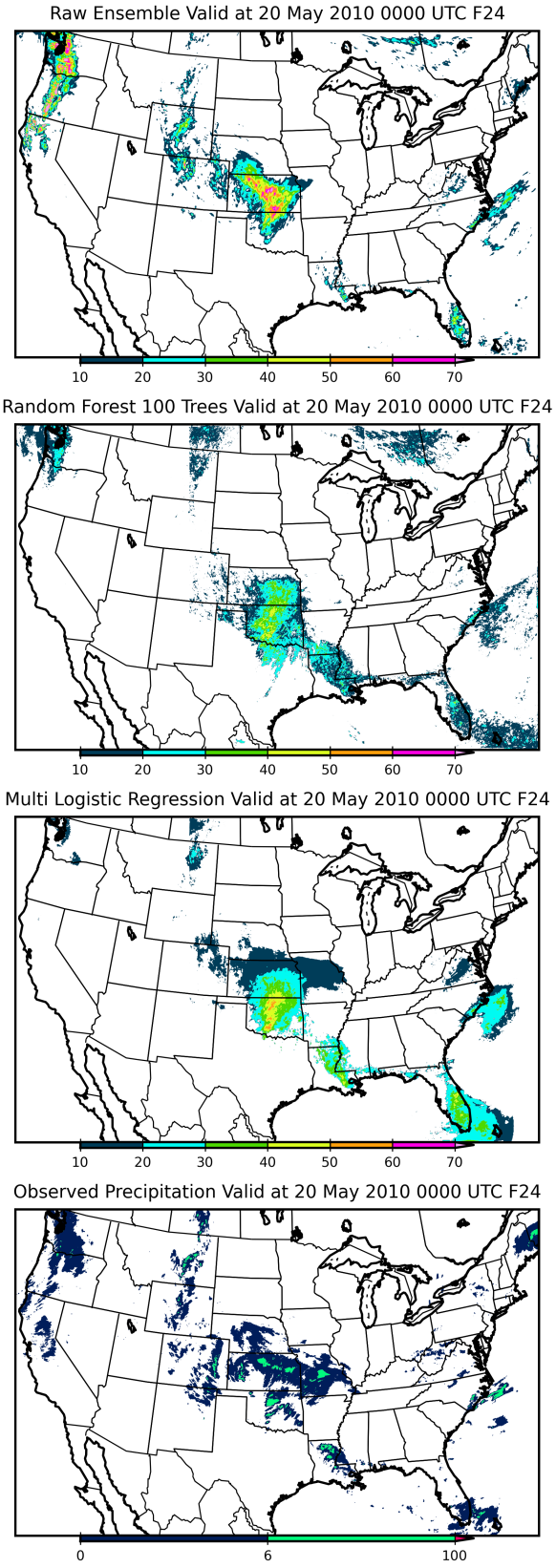

Figure 7. Probabilistic forecasts of 1-hour precipitation exceeding 6.54 mm from the 19 May 2010 0000 UTC run of the SSEF valid at 20 May 2010 0000 UTC.