Proceedings of the Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence

Day-Ahead Hail Prediction Integrating Machine Learning with Storm-Scale Numerical Weather Models

David John Gagne II
Center for Analysis and Prediction of Storms
University of Oklahoma

Amy McGovern
School of Computer Science
University of Oklahoma

Jerald Brotzge
Dept. of Atmospheric and Environmental
Sciences, University of Albany

Michael Coniglio
NOAA National Severe Storms Laboratory

James Correia, Jr.
NOAA Storm Prediction Center
NOAA/OU Cooperative Institute for
Mesoscale Meteorological Studies

Ming Xue
Center for Analysis and Prediction of Storms
University of Oklahoma

Abstract

Hail causes billions of dollars in losses by damaging buildings, vehicles, and crops. Improving the spatial and temporal accuracy of hail forecasts would allow people to mitigate hail damage. We have developed an approach to forecasting hail that identifies potential hail storms in storm-scale numerical weather prediction models and matches them with observed hailstorms. Machine learning models, including random forests, gradient boosting trees, and linear regression, are used to predict the expected hail size from each forecast storm. The individual hail size forecasts are merged with a spatial neighborhood ensemble probability technique to produce a consensus probability of hail at least 25.4 mm in diameter. The system was evaluated during the 2014 National Oceanic and Atmospheric Administration Hazardous Weather Testbed Experimental Forecast Program and compared with a physics-based hail size model. The machine-learning-based technique shows advantages in producing smaller size errors and more reliable probability forecasts. The machine learning approaches correctly predicted the location and extent of a significant hail event in eastern Nebraska and a marginal severe hail event in Colorado.

Introduction

Hail, or large spherical ice precipitation produced by thunderstorms, has caused billions of dollars in losses by damaging buildings, vehicles, and crops (Changnon 2009). Economic losses from hail have been increasing over the past two decades as populations have increased and cities have expanded in the hail-prone regions of the central United States (Changnon et al. 2000). Some losses from hail could be mitigated with accurate forecasts of severe hail potential that give people and companies time to protect vehicles and property from an incoming hailstorm.

Forecasting hail size and location is a challenging problem for meteorologists due to major uncertainties in both the forecasting and observing processes. Unlike more traditional meteorological conditions such as temperature and rainfall, hail size is not measured directly by automated instruments. The primary source of empirical observations comes from humans estimating the largest size found at their location, and hail size estimated from radar is calibrated on those imperfect human observations. Within a storm, hail size can vary dramatically and is generally not spatially contiguous. Accurate hail forecasts require predictions about the characteristics of potential hail-producing storms and the environmental conditions surrounding them. Ensembles of numerical weather prediction models can estimate the range of possible atmospheric conditions and can partially resolve the individual storm cells that produce hail up to a day in advance (Clark et al. 2012). Current numerical models do not produce explicit hail size forecasts. Hail potential can be inferred indirectly through proxy variables related to storm intensity (Clark et al. 2013) or more directly through a physical (Brimelow et al. 2006) or machine learning model (Manzato 2013) approach linking atmospheric conditions to the largest possible hail size in a given area and time period. While previous studies have focused on predicting hail sizes over large areas and time period, this study investigates how the latest high-resolution numerical weather prediction model output can be integrated with machine learning models to predict hail potential over more specific areas and times. Because of the much larger data volumes associated with these models, this study adapted advanced techniques from the image processing and machine learning fields to make hail predictions in an operational setting.

The purpose of this paper is to describe and evaluate techniques for producing day-ahead, hourly forecasts of hail diameter using storm-scale numerical weather prediction models, image processing, and machine learning. Forecasts are produced for whether or not any hail will occur, the maximum hail diameter produced from a particular storm, and the probability of hail at least 25.4 mm (1 inch) in diameter within 40 km of a point, which are the size criteria for severe hail and the spatial verification threshold used by the National Weather Service. The goal is for the machine-learning-based techniques to equal or exceed the performance of a physics-based hail size model. Forecasts from both machine-learning and physics-based techniques were generated during the 2014 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed Experimental Forecast Program (EFP) and were evaluated statistically and subjectively by teams of research and operational meteorologists (Clark et al. 2012).

Hail Observations

Developing a machine learning model to predict hail requires a reliable estimate of hail spatial coverage and diameter. No automated network exists to detect hail at the ground, so hail size observations come from either storm spotter reports or estimates derived from NEXRAD radar. Reports of hail at least 1 inch in diameter are collected by the NOAA National Weather Service Storm Prediction Center (SPC). The database is extensive and publicly available, but it suffers from many limitations. The recorded hail diameters are often estimated by comparing the stone to analog objects, such as golf balls. This estimation technique results in unnatural peaks in the hail size distribution (Jewell and Brimelow 2009). The locations of hail in the dataset are also biased toward population centers and major highways.

Radar-estimated hail size offers a solution to the population bias issue plaguing hail reports. This project uses the NOAA NSSL Multiradar Multisensor (MRMS) gridded Maximum Estimated Size of Hail (MESH), which derives a maximum hailsize from gridded 3D radar reflectivity (Witt et al. 1998). A multi-year comparison of MRMS MESH to storm reports found that MESH was unbiased and had superior spatial coverage to hail reports (Cintineo et al. 2012). The native MESH data were interpolated to the model domain using cubic spline interpolation.

Storm-Scale Ensemble

This project uses output from the Center for Analysis and Prediction of Storms (CAPS) Storm-Storm Scale Ensemble Forecast (SSEF) system (Kong 2014), which was run in conjunction with the NOAA Hazardous Weather Testbed Experimental Forecast Program. The SSEF consists of an ensemble of Weather Research and Forecasting (WRF) Advanced Research WRF models with randomly perturbed initial and boundary conditions. In addition, each ensemble member used a different combination of microphysics (physics describing how water changes phase and grows into precipitation), planetary boundary layer (atmosphere near the surface), and land surface model (vegetation and soil processes) parameterization schemes in order to increase the diversity of model solutions. Each SSEF run was initialized at 00 UTC and produced hourly output during the period from late April to early June. The 2013 SSEF was used to train and validate the machine learning models while the 2014 SSEF was used for testing. The 2013 SSEF consisted of 30 model runs from 26 April to 7 June 2013, and the 2014 SSEF consisted of 12 model runs between 15 May and 6 June 2014. The 18 to 30 hour forecasts valid from 18 to 6 UTC are evaluated as they cover the time frame when hailstorms are most likely and contain storms that were not present when the SSEF initiated.

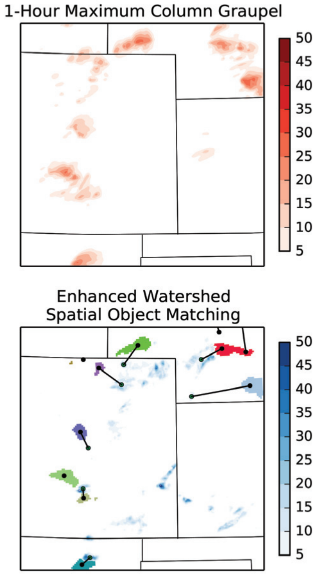


Figure 1: In the first panel, the 1-hour maximum column-summed graupel from a member of the SSEF at 22 UTC on 6 June 2014 is shown. The second panel shows the hailstorm objects extracted from the column graupel grid by the enhanced watershed technique in solid colors. The connecting lines indicate the matches between the forecasted hailstorms and observed MESH (blue contours).

Machine Learning Framework

Hail size forecasts are derived from each ensemble member by identifying forecast hailstorms, matching the forecast storms with observed hailstorms, extracting data within the storm areas, and then fitting a machine learning model between the atmospheric variables and the observed hail size.

Hailstorm Identification and Matching

Hail size prediction first requires determining the areas in which hail is likely to occur. Model atmospheric conditions related to hail should only occur in the areas where the model produces ice-containing storms, so identifying likely storm areas in the model both reduces the noise in the training data and greatly reduces the required computational power. To find ice-containing storms, we examine the 1-hour maximum total column graupel field, which indicates the maximum value over the previous hour of the total mass of spherical ice particles in a column of air. For object identification, we use the enhanced watershed technique (Lakshmanan, Hondl, and Rabin 2009). As with the traditional watershed, local maxima in the field are first identified, and then objects are grown from the maxima in discrete steps until stopping criteria are met. While the traditional watershed uses a global lower threshold or maximum number of steps as its stopping criteria, the enhanced watershed also includes

Table 1: Input variables for the machine learning models from the SSEF ensemble members. Storm variables (S) describe conditions within the storm and environment variables (E) describe the surrounding atmosphere.

Variable	Description	Units
Max Updraft Speed (S)	Upward vertical wind speed	m s^{-1}
Max Downdraft Speed (S)	Downward vertical wind speed	m s^{-1}
Max Updraft Helicity (S)	Proxy for storm intensity	m^2 s^{-2}
Radar Reflectivity (S)	Simulated view of storm structure	dBZ
Max Column Graupel (S)	Total mass of ice particles	kg m^{-2}
0-5 km Total Graupel (S)	Mass of ice particles in lower levels	kg m^{-2}
Storm Height (S)	Height of the top of storm	m
Bunker's Storm Motion (S)	Estimated storm speed and direction	m s^{-1}
Mean Layer CAPE (E)	Mean instability	J kg^{-1}
Most Unstable CAPE (E)	Highest possible instability	J kg^{-1}
Mean Layer CIN (E)	Mean Inhibition	J kg^{-1}
Most Unstable CIN (E)	Lowest possible inhibition	J kg^{-1}
Lifted Condensation Level (E)	Estimated distance from ground to clouds	m
Precipitable Water (E)	Amount of water contained in column of air	mm
0-6 km Wind Shear (E)	Difference in winds at 6 km and surface	m s^{-1}
0-3 km Storm-Rel. Helicity (E)	Estimate of horizontal localized rotation	m^2 s^{-2}
0-3 km Lapse Rate (E)	Change in temperature with height	K km^{-1}
850 mb Specific Humidity (E)	Ratio of water vapor mass to total air mass	g kg^{-1}

a size criteria and buffer zones around local maxima. Prior to applying the enhanced watershed to the data, a Gaussian filter was applied to each grid in order to increase spatial correlations and generate smoother objects.

The enhanced watershed is applied to both the model column graupel fields and the observed MESH field. The enhanced watershed was manually tuned to capture a wide range of hail swath intensities while keeping neighboring swaths as separate objects. The object-based verification approach matches forecast and observed objects iteratively based on spatial distance from closest to farthest away within a 200 km radius. An example of the enhanced watershed and object matching being applied is shown in Fig. 1. Since each observed hail object can only be matched with one forecast hail object, some storms near isolated hail observations do not get matched.

Once storms are identified and matched, statistics describing different properties of the storm and atmosphere are extracted from each hailstorm object. These statistics include the mean, standard deviation, minimum, and maximum of WRF output variables describing the strength of the storm as well as the conditions of the storm environment (Table 1). The forecast label is the maximum hail size within the matched MESH object, or 0 if no match was found.

Hail Classification and Size Regression

Machine learning models first determine if a specific forecast storm will produce any hail, and given that the storm does produce hail, what size the hail will be. A classification model was trained on all cases to produce a binary prediction of whether or not the storm would produce hail, and a regression model was trained on only the storms that were matched with an observed hail event. Three machine learning models are tested: random forest, gradient boosting regression trees, and a combination of a logistic classification model and ridge regression. Random forests (Breiman 2001) are ensembles of decision trees that use resampling of the training data and random subsampling of attributes to increase the diversity of the individual members and improve predictive performance. This project used a 100 tree random forest with default parameters. Gradient boosting trees (Friedman 1999) are an additive ensemble of decision trees that are iteratively trained and weighted based on their error characteristics. We used 1000 trees, a learning rate of 5%, and a max depth of 5. Both methods have produced strong predictive performance in many domains and both can be analyzed using variable importance measures and partial dependence plots. A logistic model is a linear classifier that translates input parameters into a probability through a logit transformation, and ridge regression is linear regression with a penalty term to restrict the size of the coefficients and make the regression more robust. All methods were implemented using the Python *scikit-learn* library (Pedregosa et al. 2011). When the predicted hail sizes were applied to the original forecast grid, the storms producing no hail were removed from the grid, and the predicted size values were applied to the grid points within the area covered by each forecast hail storm.

HAILCAST

HAILCAST is a one-dimensional, physics-based coupled cloud and hail model. HAILCAST grows a set of simulated hail embryos based on the instability, wind shear, and moisture in the local atmosphere. It has shown skill when run with input from forecast and observed vertical atmospheric profiles (Brimelow et al. 2006) in a wide range of storm environments (Jewell and Brimelow 2009). The technique has been further refined to run during the integration of a storm-scale numerical model (Adams-Selin, Ziegler, and Clark 2014) and has been released publicly in WRF version 3.6. In addition to being run during the 2014 EFP, HAILCAST has been incorporated into the operational Air Force Weather Agency storm-scale ensemble. HAILCAST is run at each SSEF member grid point with an updraft speed at least 10 m s^{-1}. The maximum HAILCAST hail size within each forecast hailstorm object was used as the comparison prediction with the machine learning methods because it provided the most analogous estimate to the observed maximum hail size.

Neighborhood Ensemble Probability

The machine learning methods produce a calibrated hail size forecast for each ensemble member and each time step. These machine learning forecasts do not cover the full range of possible hail sizes at every grid point because the SSEF contains spatial and temporal errors in storm placement and intensity and does not fully approximate internal storm dynamics as well as the processes that govern precipitation formation and thermodynamic changes associated with them. These physics errors results in modeled storms that do not form, move, and intensify at the same rate as the real ones. One approach commonly used to account for this spatial error is the neighborhood ensemble probability method (Schwartz et al. 2010). Conditional probabilities of severe hail are calculated by counting the number of grid points in a local, circular neighborhood in which severe hail occurs and dividing by the number of grid points in which any hail

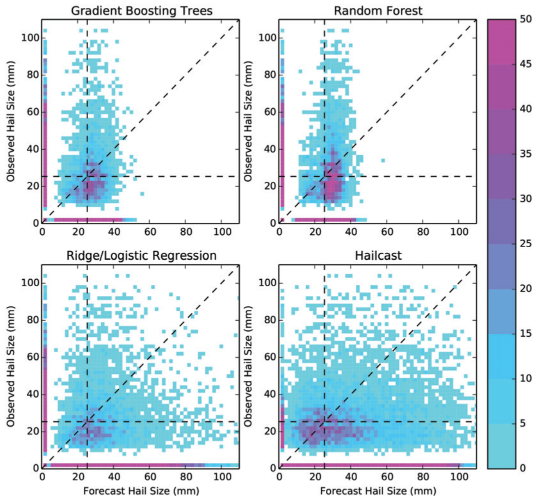


Figure 2: Heatmaps of the distributions of forecast errors for each hail size model.

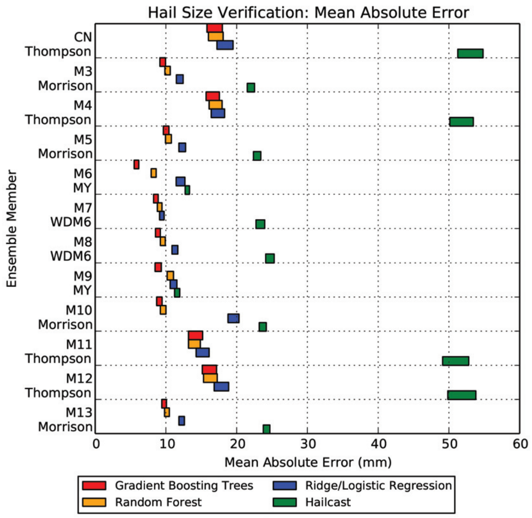


Figure 3: Comparison of the bootstrap 95% confidence intervals by model and ensemble member. The microphysics scheme used in each ensemble member is indicated below the name of the member.

occurs. The probability from all ensemble members are averaged together, and a Gaussian filter is applied to smooth the edges of the non-zero probabilities. Since each model forecast has been bias-corrected by the machine learning regressions, the resulting probabilities should also be unbiased. The size of the neighborhood can be adjusted to capture uncertainties at varying scales. Weather forecasters prefer spatially smooth probabilities as they more closely match human forecasts. The drawbacks of neighborhood ensemble probabilities are that they weaken probability gradients and can understate the threat of single isolated storms while highlighting clusters of more widespread marginal storms.

Results

We statistically validated the hail size and probability forecasts based on 12 hail days from 15 May to 6 June 2014. The predicted hail sizes were compared with the maximum hail sizes within each matched observed hailstorm object. The probability forecasts were compared at each grid point with whether or not hail at least 25.4 mm in diameter was observed within 40 km of that point, which are the evaluation criteria used by the SPC.

Hail Size Forecasts

The machine learning and HAILCAST size forecasts showed skill in predicting hail sizes up to 60 mm in diameter, which account for the bulk of all hail events. Both tree-based methods predicted that most severe hail would be between 25 and 60 mm, and most of their predictions were close to those values. Observed hail over 60 mm was also predicted to be within the 25 to 60 mm range (Fig. 2). While ridge regression and HAILCAST predicted hail sizes over the full range of observed values, both methods tended to overpredict the maximum hail diameter, especially HAILCAST.

Examining the errors for each ensemble member reveals some links between the error characteristics and the microphysics parameterization scheme used by each member (Fig. 3). HAILCAST performed statistically significantly (bootstrap 95% confidence intervals) worse than any of the machine learning methods, and the error was greatest in ensemble members using the Thompson microphysics scheme. The Thompson scheme assumes a relatively larger graupel density compared to the other schemes, which HAILCAST used as the basis for growing its hailstones. The Milbrandt and Yau (MY) scheme has separate graupel and hail densities, and HAILCAST performed best in the members using that scheme. The machine learning models performed similarly across most ensemble members, and gradient boosting trees performed statistically significantly better than the other models for most members. The hail occurrence predictions also showed similar skill among all machine learning methods and ensemble members (Fig. 4). A performance diagram (Roebber 2009) displays the relationships among four binary contingency table scores: probability of detection (y-axis), false alarm ratio (x-axis), frequency bias (dotted diagonal lines), and critical success index (solid curved lines). Performance is best in the upper right corner of the diagram and along the diagonal where the frequency bias is 1. The machine learning methods had similar success ratios, but there was a wider range in the percentage of hailstorms detected. HAILCAST was the best at distinguishing which storms produced hail. Some of the performance issues stem from the enhanced watershed parameters fitting storms from some models better than others due to differences in microphysics.

Neighborhood Probability Forecasts

Since the machine learning approaches produced hail forecasts with little bias, the resulting neighborhood probabilities tended to be more reliable, or occurring at the frequency

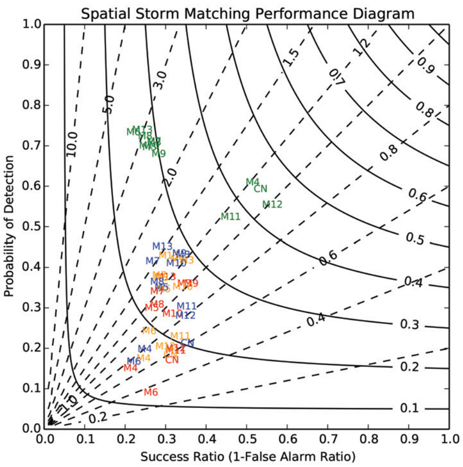


Figure 4: A performance diagram measures the ability of each method and ensemble member to match forecast and observed hail storms spatially.

given by the probability, than the corresponding HAILCAST forecasts. For probabilities ranging from 0 to 20%, gradient boosting trees are nearly reliable and the other methods are slightly overconfident while HAILCAST is more overconfident (Fig. 5). At higher probabilities, there was overconfidence from all methods. From subjective verification of the different hail forecasts, this overconfidence is linked to a spatial displacement of the highest neighborhood probabilities away from where severe hail fell at a particular time.

Case Studies

The worst hail event during the experiment occurred on 3 June 2014 in Nebraska. Multiple rounds of storms produced wind-driven baseball to softball sized hail that left large dents and holes in cars, crops, and the sides and roofs of houses. Each model generated a neighborhood probability prediction for each hour from 18 to 00 UTC. The maximum 1-hour probabilities during that time period are displayed in Fig. 6. All models encompassed the full observed area of 25 mm or greater hail with nonzero probabilities and have their highest confidence in eastern Nebraska where the largest hail was observed. All models also displayed enhanced probabilities in western Nebraska where isolated storms also produced severe hail. Random forest produced the subjectively best forecast of the machine learning methods because its maximum overlapped the 75 mm hail most closely and because it had relatively lower probabilities for the western Nebraska storms. HAILCAST produced the most confident forecast, but it had high probabilities well outside the area where 25 mm hail was observed.

A more marginal but widespread hail event occurred on 21 May 2014 in Colorado, Kansas, Oklahoma, and Texas. An isolated hailstorm dropped severe hail and caused flooding in downtown Denver, and additional storms dropped hail across eastern Colorado. The ensemble means of the hail size forecasts are shown in Fig. 7. HAILCAST and ridge regression generally overestimated the maximum hail sizes

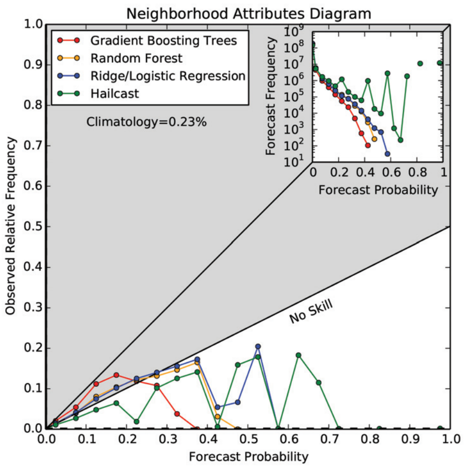
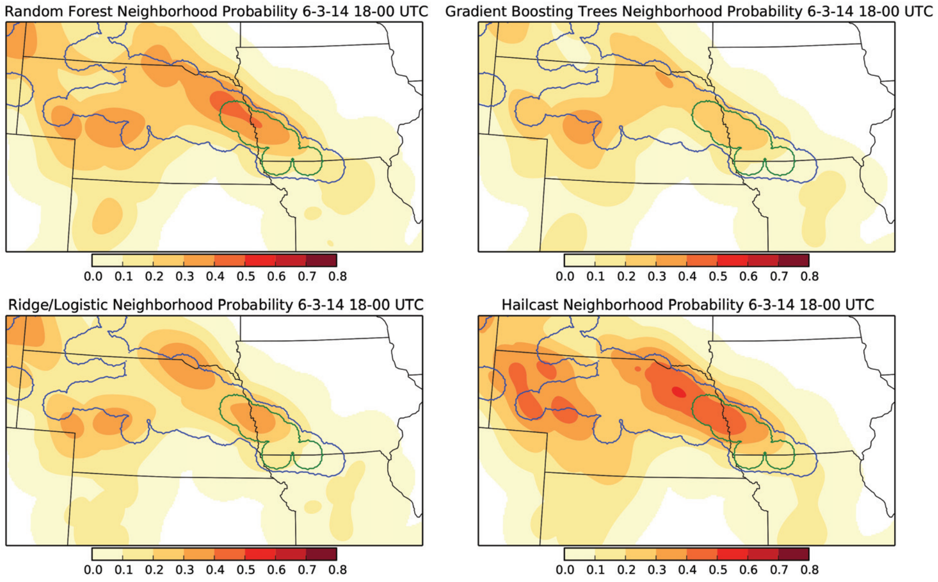


Figure 5: Attributes diagram that compares the forecast probabilities of each model with their corresponding observed relative frequencies. Points in the gray area have positive skill, and points outside the gray area have negative skill. The inset indicates the observed frequency of each probability forecast.

for the day with widespread areas of over 50 mm hail. Random forest and gradient boosting produced hail sizes closer to what occurred, and gradient boosting also had a wider range of hail sizes than random forest. The most intense portions of the forecast hail swaths were shifted northeast of the observed hail swaths, so while the general character of the event is correctly forecast, downtown Denver was forecast to receive no hail in 3 of the 4 models. The neighborhood probabilities in Fig. 8 account for this spatial error and show non-zero probabilities over Denver. The random forest neighborhood probabilities capture the Colorado hail the best by showing two areas of high hail potential and by having non-zero probabilities of hail over Denver.

Discussion

Generating and validating daily hail forecasts with a group of experienced meteorologists provided insights about the good qualities of the forecasts and what needed improvements. The machine-learning neighborhood probabilities were useful because the bias-correction reduced the false alarm area compared to HAILCAST. The probability forecasts were closer to the best forecast from a trained meteorologist given the same information. Further improvement to machine learning model performance is constrained by the model storm information. The storm representation can be improved with better resolution, model physics, and initial conditions, but it will always contain uncertainties and errors because we cannot fully observe the atmosphere, the physical models contain approximations, and computational power is limited. While the different machine learning models were not able to predict hail above 60 mm in diameter, this was largely because there was very little training data at these sizes. HAILCAST, on the other hand, predicted hail



Random Forest Neighborhood Probability 6-3-14 18-00 UTC
Gradient Boosting Trees Neighborhood Probability 6-3-14 18-00 UTC
Ridge/Logistic Neighborhood Probability 6-3-14 18-00 UTC
Hailcast Neighborhood Probability 6-3-14 18-00 UTC
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8

Figure 6: Maximum neighborhood ensemble probabilities between 18 and 00 UTC on 3 June 2014. The blue contour indicates the areas that were within 40 km of 25 mm diameter hail, and the green contour indicates the same distance from 75 mm diameter hail.

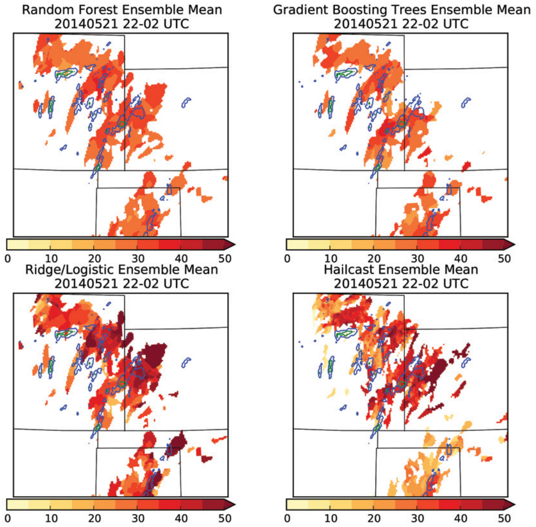


Figure 7: Ensemble mean hail sizes from each model for 21 May 2014 from 22 to 02 UTC. Blue contours indicate observed hail sizes of at least 5 mm and green contours indicate hail sizes of at least 25 mm.

over 60 mm almost every day during the experiment and was not trusted by the meteorologists because of that issue. This was also the first operational test for both models, and the forecaster feedback has been valuable for introducing improvements to both systems.

Additional modifications will be applied to provide better performance in an operational setting. Because the choice of numerical weather prediction model parameters affects the characteristics of the storms, the storm-finding system should be tuned to each SSEF ensemble member. Predicted hail size should be associated with storm intensity, and the weaker connections in this dataset may be due to not including intensity in the matching criteria between forecast and observed hail areas. Forecast and observed storms have also shown tendencies to be offset in time as well as space, and this time offset may also be responsible for the intensity mismatches. The approach will also be applied to other storm-scale models that are being run operationally, such as the High Resolution Rapid Refresh, the NSSL WRF, and the Air Force Weather Agency storm-scale ensemble. The updated modeling approaches will be evaluated by forecasters in the Spring 2015 Hazardous Weather Testbed Experimental Forecast Program. Past and future hail forecasts are viewable at *http://cs.ou.edu/∼djgagne*.

Conclusions

Hail is a dangerous severe weather phenomenon that causes increasingly extensive economic damage each year. Improving hail prediction with more accurate information about expected hail locations and intensity will allow people to mitigate some of the potential impact of hail. We have demonstrated in an operational setting a hail prediction system that applies machine learning and image processing techniques to storm-scale numerical model ensembles. The approach shows accuracy in predicting hail location and in discriminating its severity with lead times of up to a day in advance of a hailstorm. The machine learning approaches demonstrated some advantages over physics-based hail size calculations. Improvements to the numerical models and machine learning approaches should lead to increasingly accurate hail size and location forecasts.

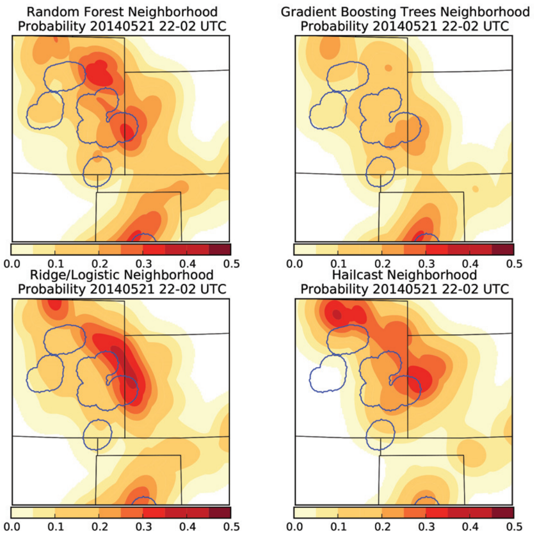

Random Forest Neighborhood
Probability 20140521 22-02 UTC
Gradient Boosting Trees Neighborhood
Probability 20140521 22-02 UTC
Ridge/Logistic Neighborhood
Probability 20140521 22-02 UTC
Hailcast Neighborhood
Probability 20140521 22-02 UTC
0.0
0.1
0.2
0.3
0.4
0.5

Figure 8: Neighborhood ensemble probability of severe hail from each model for 21 May 2014 from 22 to 02 UTC. Blue contours indicate hail sizes of at least 25 mm and green contours indicate hail sizes of at least 75 mm.

Acknowledgments

Special thanks go to the fellow meteorologists who participated in the 2014 EFP. MESH data were provided by NSSL with the help of Kiel Ortega. The SSEF was run by CAPS on the Darter supercomputer by the National Institute for Computational Sciences at the University of Tennessee with the help of Fanyou Kong, Kevin Thomas, Yunheng Wang, and Keith Brewster. This research was funded under the Severe Hail Analysis, Representation, and Prediction project under NSF Grant AGS-0802888 and NSF Graduate Research Fellowship Grant 2011099434.

References

Adams-Selin, R.; Ziegler, C.; and Clark, A. J. 2014. Forecasting hail using a one-dimensional hail growth model inline within WRF. In *Proceedings, 27th Conference on Severe Local Storms*, 11B.2. Madison, WI: Amer. Meteor. Soc.

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

Brimelow, J. C.; Reuter, G. W.; Goodson, R.; and Krauss, T. W. 2006. Spatial forecasts of maximum hail size using prognostic model soundings and HAILCAST. *Wea. Forecasting* 21:206–219.

Changnon, S. A.; Pielke Jr., R. A.; Changnon, D.; Sylves, R. T.; and Pulwarty, R. 2000. Human factors explain the increased losses from weather and climate extremes. *Bull. Amer. Meteor. Soc.* 81:437–442.

Changnon, S. A. 2009. Increasing major hail losses in the u.s. *Climate Change* 96:161–166.

Cintineo, J. L.; Smith, T. M.; Lakshmanan, V.; Brooks, H. E.; and Ortega, K. L. 2012. An objective high-resolution hail climatology of the contiguous United States. *Wea. Forecasting* 27:1235–1248.

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

Clark, A. J.; Gao, J.; Marsh, P. T.; Smith, T.; Kain, J. S.; Correia, Jr., J.; Xue, M.; and Kong, F. 2013. Tornado path length forecasts from 2010-2011 using ensemble updraft helicity. *Wea. Forecasting* 28:387–407.

Friedman, J. 1999. Greedy function approximation: A gradient boosting machine. Technical report, Stanford University.

Jewell, R., and Brimelow, J. C. 2009. Evaluation of an Alberta hail growth model using severe hail proximity soundings from the united states. *Wea. Forecasting* 24:1592–1609.

Kong, F. 2014. 2014 CAPS spring forecast experiment program plan. Technical report, Center for Analysis and Prediction of Storms.

Lakshmanan, V.; Hondl, K.; and Rabin, R. 2009. An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Atmos. Oceanic Technol.* 26:523–537.

Manzato, A. 2013. Hail in northeast italy: A neural network ensemble forecast using sounding-derived indices. *Wea. Forecasting* 28:3–28.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; and Cournapeau, D. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.

Roebber, P. J. 2009. Visualizing multiple measures of forecast quality. *Wea. Forecasting* 24:601–608.

Schwartz, C. S.; Kain, J. S.; Weiss, S. J.; Xue, M.; Bright, D. R.; Kong, F.; Thomas, K. W.; J., L. J.; Coniglio, M. C.; and Wandishin, M. S. 2010. Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting* 25:263–280.

Witt, A.; Eilts, M. D.; Stumph, G. J.; Johnson, J. T.; Mitchell, E. D.; and Thomas, K. W. 1998. An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting* 13:286–303.