Severe Hail Prediction within a Spatiotemporal Relational Data Mining Framework

David John Gagne II*, Amy McGovern†, Jerald Brotzge‡*, Ming Xue‡*
*School of Meteorology, University of Oklahoma, Norman, Oklahoma
†School of Computer Science, University of Oklahoma, Norman, Oklahoma
‡Center for the Analysis and Prediction of Storms, Norman, Oklahoma
Email: djgagne@ou.edu

***Abstract*—Severe hail, or spherical ice precipitation over 1 inch in diameter, has caused billions of dollars in damage to crops, buildings, automobiles, and aircraft. Accurate predictions of severe hail with enough lead time can allow people to mitigate some hail damage by sheltering themselves and their vehicles and by rerouting their aircraft. Current pinpoint forecasts of severe hail rely on detection of hail in existing storms with radar-based methods. Predictions beyond an hour are limited to probabilistic predictions over larger areas based on expected environmental conditions. This paper describes a technique that could increase the accuracy of severe hail forecasts by incorporating output from an ensemble of storm scale numerical weather prediction models into a spatiotemporal relational data mining model that would produce probabilistic predictions of severe hail. The spatiotemporal relational framework represents the ensemble output as a network of storm objects connected by spatial relationships. Composites of the ensemble data show spatial biases in the placement of severe and non severe hail storms. The spatiotemporal relational model performs significantly better at discriminating between severe and non-severe hail compared to a traditional data mining model of the same type. Variable importance rankings show results physically consistent with previous studies and highlight the importance of the relational data.**

***Index Terms*—Spatiotemporal data mining, Relational learning, Random forests, Environmental hazards, Hail**

I. Introduction

Severe hail, defined as precipitation in the form of spherical aggregates of ice with a diameter larger than 1 inch [1], is responsible for economic damage that has totaled over $1 billion [2], [3]. Hail most often damages crops, buildings, and vehicles, which has resulted in a continuing interest from the research and insurance community [4] in understanding the ingredients for severe hail and finding ways to anticipate its occurrence. Given enough lead time, people can mitigate damage to vulnerable assets by moving people indoors, sheltering vehicles, and rerouting aircraft.

Any severe hail forecasts beyond the lead time of an hour are limited to probabilities over broad regions due to difficulties in forecasting the track and intensities of the parent storms. Forecasters at the National Oceanic and Atmospheric Administration Storm Prediction Center currently issue probabilistic severe hail convective outlooks 1 day in advance and Severe Thunderstorm Watches within a few hours of an event. National Weather Service Severe Thunderstorm Warnings provide an expected hail size over the area of a warning polygon. In 2012, the warnings provided an average lead time of 18.7 minutes, detected 83% of the severe storm reports, and raised a false alarm 46% of the time [5] (because a severe storm includes several criteria, these statistics include tornado and wind reports). Physics-based [6] and neural network models [7], [8] have also been used to predict the probability of severe hail (POSH) and maximum expected hail size (MEHS) from radar data and observed and forecast proximity soundings. Statistical models that have derived hail size from numerical model output have used coarse resolution models that can resolve the large scale environmental conditions but cannot predict individual storms. Some microphysics parameterization schemes within higher resolution numerical models can explicitly predict the size distributions of graupel, which is a small ice particle, and hail but require some additional assumptions in order to predict a maximum hail size [9], [10].

Ensembles of numerical models with the horizontal grid spacing capable of resolving individual storms have been run experimentally with the NOAA Hazardous Weather Testbed Spring Experiment by the Center for the Analysis and Prediction of Storms (CAPS) [11] since 2007. Each ensemble member outputs the expected position of storms as well as a wide range of variables describing the storm environment and strength. The members produce predictions of hail and graupel mixing ratios and number concentrations, which can be used to derive an estimate of the maximum hail size assuming a parametric bulk distribution of hail diameter [10]. Those values could be derived in an additional procedure involving the quantization of an assumed empirical distribution of sizes [9], but it would require so many extra computations that it would not be operationally feasible. Instead, output from each ensemble member can be used by statistical post-processing algorithms, which are much less computationally expensive and can be run independently of the numerical models, in order to produce probabilistic predictions of high impact weather events. Previous projects have focused on probabilities of heavy rain using grid point [12] or local neighborhood data [13]. There has been no work focusing on deriving severe hail probabilities from these ensembles up to this point.

Neither the grid point nor neighborhood data collection

methods capture the potentially complex relationships among a set of storms in the model. In a traditional supervised classification framework, each case is associated with a set of independent attributes. While they may be related to each other, there is no way to specify connections explicitly. A spatiotemporal relational framework provides a way to organize complex spatially and temporally varying data into a high-level network while still interrogating low-level data values. Discrete areas/regions of interest are classified as objects. In this domain, a single thunderstorm could be considered an object. Objects can exist for periods of time, change shape, and move in space. The data values contained within or derived from those objects are stored as attributes of each object. For example, the thunderstorm object could have attributes describing size of the storm and the intensity of the rainfall within it. Attributes can be static, time series, or fielded in two or three dimensions. The fielded attributes can be either scalar or vector values. The individual objects are then connected through spatial relationships, which can also have associated attributes, such as distance and direction. The framework has successfully captured the complex relationships in the domains of tornado and turbulence prediction [14] in which very short-range forecasts were made from a fusion of multiple data sources. This project extends that framework to severe hail prediction and ensemble forecasting by making day-ahead forecasts that learn patterns from objects and relationships representing ensemble predictions.

The primary goals of this paper are to introduce a probabilistic severe hail prediction algorithm utilizing storm scale ensemble model forecasts, to extend the spatiotemporal relational framework to non-deterministic datasets and predictions at longer time frames, and to compare the performance of a spatiotemporal relational model with traditional machine learning methods, and to compare the relative importance of neighborhood and object-based data representations. The model presented here is the first step in a larger project to improve severe hail prediction by more accurately representing hail in numerical models and by applying data mining techniques to numerical weather prediction ensemble forecasts.

II. Data

A. Verification Data

There is currently no operational automated system to observe hail size directly. Hail pads, which are styrofoam rectangles covered in aluminum foil that are placed outside on the ground and then dented by hail impacts [8], can be used to measure hail size regularly but require a human observer to check the pad regularly. Verification data for this project come from two sources of hail reports. The National Oceanic and Atmospheric Administration (NOAA) Storm Prediction Center (SPC) aggregates local storm reports for tornadoes, hail, and high winds on its website. The NOAA National Severe Storms Laboratory (NSSL) Meteorological Phenomena Identification Near the Ground (mPING) project collects precipitation type and severe weather reports from the general public using a smartphone app. Both the SPC and mPING reports provide direct confirmation of severe hail along with a size estimate, and the mPING dataset also contains reports of non-severe hail and rain, which makes mPING especially valuable as a source for null events.

Both report databases have many limitations. They only provide one hail size per report, so the actual distribution of hail within a storm is unknown. The reports are subject to population biases, so hail near cities and major highways are more likely to be reported. Reports also give no indication of the spatial coverage or duration of the hail. Lack of a hail report does not necessarily mean that hail did not occur at that location.

In future work, we plan to incorporate radar-derived hail estimates [15], [16] into the verification dataset. They would provide much finer spatial and temporal coverage while controlling for population biases in the report datasets. The radar-derived hail sizes are based on radar returns from above the ground, so they can overestimate the size of hail that reaches the ground, and the algorithm was calibrated for the southern Plains, so it tends to perform worse in the eastern US. A combination of the reports and radar-derived data would provide the best estimates of "truth." We will also acquire more reports from the NSSL Severe Hazards Analysis and Verification Experiment [17], which actively collects hail reports through phone interviews in potentially affected areas.

For this paper, we selected SPC and mPING hail reports from between 1800 UTC and 0600 UTC for each model run. The period covers the peak time for severe weather, which runs from noon to midnight CST in the Great Plains states. Reports of hail 1 inch or larger were considered severe, in accordance with the National Weather Service definition [1]; and hail reports smaller than 1 inch were considered non-severe. With this form of report sampling, only locations that actually received hail are included, and since severe hail is more likely to be reported than non-severe hail, the distribution of severe versus non-severe hail may not approximate the true climatological distribution of hail as non-severe hail tends to occur more often [4]. A map of all hail reports in the time period of the study is shown in Fig. 1. Most of the reports occur in the Plains region of the country, which is to be expected for the May-June time period [15]. Other parts of the country do receive inclusion in the dataset, such as the cluster of reports in North Carolina and in Pennsylvania. There are also reports in the climatologically rare areas west of the Rockies in Idaho, California, Montana, and Washington.

B. Ensemble Forecast Data

This project uses the 2013 CAPS Storm Scale Ensemble Forecast (SSEF) system for its ensemble forecast data [18]. The 2013 SSEF consists of 30 numerical weather prediction models run at 4-km horizontal resolution with varied initial and boundary conditions as well as different combinations of physical parameterization schemes. Each model is initialized at 0000 UTC and outputs predictions every hour for 48 hours in a domain covering the contiguous United States. The SSEF was run every weekday from April 24 through June 6 with

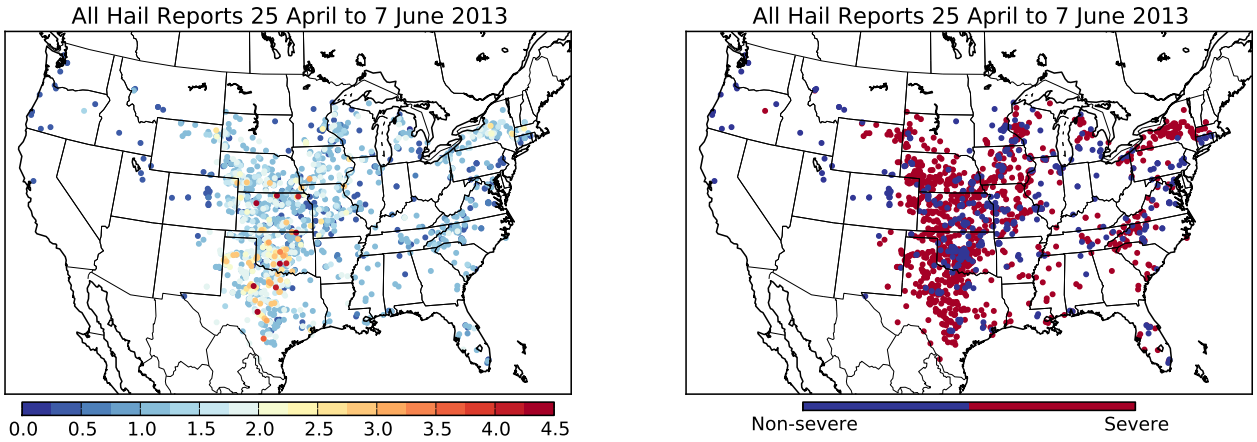


Fig. 1: All hail reports from 18 UTC to 6 UTC for each day of the 2013 SSEF. Reports are colored by size (left) and by whether or not the report is severe (right).

special weekend runs for major severe weather events. The ensemble output consists of a series of two-dimensional grids containing forecast values for a wide range of parameters that describe environmental conditions related to storm potential. Because storms can form and die within a single hour, some of the grids contain the maximum value of a parameter over the previous hour, and others are derived from column sums or maximums. Of the 30 total ensemble members, the 15 that used perturbed initial and boundary conditions as well as varied physical parameterizations were mined. The others were excluded because they used identical initial conditions to the control run and may bias the results of any aggregations if included.

The configurations of the ensemble members are shown in Table I. The control members receive their initial conditions from the 0000 UTC Advanced Regional Prediction System (ARPS) model [19], [20], [21] analysis and boundary conditions from the 0000 UTC North American Mesoscale (NAM) model. The other ensemble members use the initial conditions from the control member plus a perturbation from one of the members of the coarser resolution 2100 UTC Short Range Ensemble Forecast system. All ensemble members except for the ARPS control member use the Advanced Research Weather Research and Forecasting (WRF) model [22]. The microphysics parameterization schemes are responsible for determining the distributions and amounts of each precipitation type in the model. All of the microphysics schemes produce explicit predictions of graupel, but only the Milbrandt and Yau (M-Y) scheme has a hail term. Because of that, we will only be using the graupel mixing ratio fields from the models.

TABLE I: Initial conditions, boundary conditions, and physical parameterization scheme choices for each ensemble member.

Member	IC	BC	Microphysics	LSM	PBL
arw_cn	ARPSa	NAMf	Thompson	Noah	MYJ
arps_cn	ARPSa	NAMf	Lin	–	–
arw_m3	+em-p1	em-p1	Morrison	RUC	YSU
arw_m4	+nmm-n2	nmm-n2	Morrison	Noah	MYJ
arw_m5	+em-n2	em-n2	Thompson	Noah	ACM2
arw_m6	+nmmb-p2	nmmb-p2	M-Y	RUC	ACM2
arw_m7	+nmm-p1	nmm-p1	Morrison	Noah	MYNN
arw_m8	+nmmb-n1	nmmb-n1	WDM6	RUC	MYJ
arw_m9	-nmmb-p1	nmmb-p1	M-Y	Noah	YSU
arw_m10	+em-n1	em-n1	WDM6	Noah	QNSE
arw_m11	-em-p2	em-p2	M-Y	Noah	MYNN
arw_m12	-nmmb-p3	nmmb-p3	WDM6	Noah	YSU
arw_m13	-nmmb-p3	nmmb-p3	Thompson	Noah	YSU
arw_m14	-em-p3	em-p3	Thompson	Noah	MYNN
arw_m15	-nmm-p2	nmm-p2	Morrison	Noah	QNSE

III. Methods

A. Data Sampling

Each instance in the training and testing datasets was centered on either a severe or non-severe hail report. Over the period of the SSEF there were 1257 severe hail reports and 588 non-severe hail reports (Fig. 1). Each report was matched with the nearest and preceding forecast hours. Ensemble data and objects were extracted from a 84×84 km box centered on each report. Neighborhood objects contain fielded attributes consisting of data from the centermost 10×10 grid points for each variable being mined. Storm regions in the surrounding area were extracted and connected to the neighborhood object with spatial relationships (Fig. 2). Each spatial relationship contains time-series attributes describing the distance and direction from the neighborhood object.

B. Object Segmentation and Tracking

Within the box of interest, spatiotemporal objects related to forecasted storms were extracted from each ensemble member. Object segmentation was performed using the enhanced watershed method [23]. The traditional watershed method grows regions from local maxima in a grid until the regions meet or a minimum depth is reached in the data. The enhanced watershed adds minimum size and maximum depth criteria into the region growing step while keeping track of foothills, which are buffer zones around each region. The size and depth

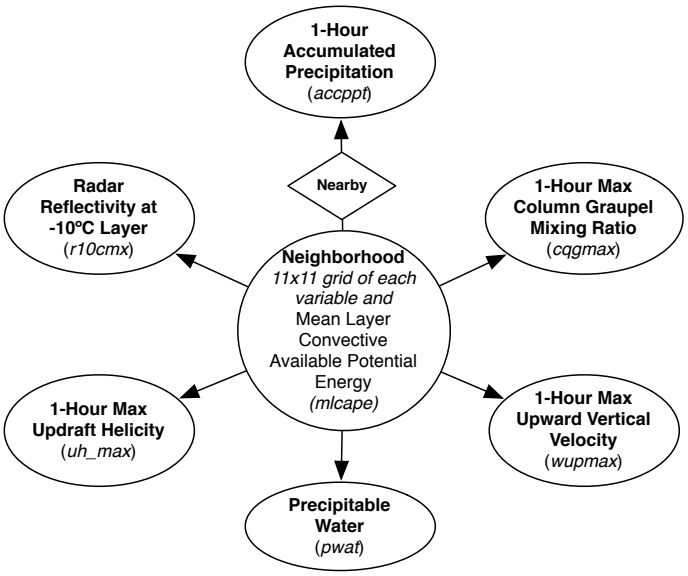


Fig. 2: Schema showing the different types of objects (ovals) in each hail case and how spatial relationships (diamonds and arrows) connect them.

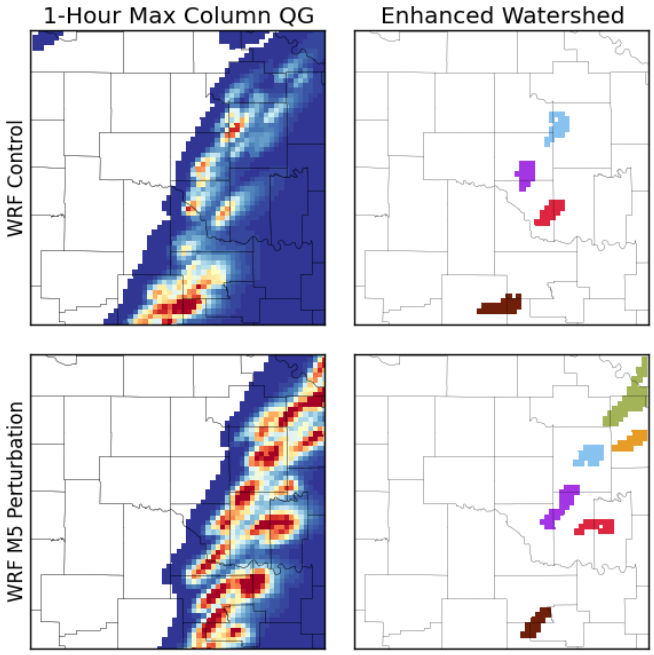


Fig. 3: Example of the enhanced watershed segmentation process applied on 20 May 2013 at 2100 UTC for a severe hail report in Oklahoma City. The left column shows the intensities of the 1-Hour Max Column Graupel Mixing Ratio (WUPMAX) for both ensemble members while the right column shows the objects that the enhanced watershed selected from the grids on the left.

criteria can be used to find objects at varying scales while filtering smaller peaks. It produces more physically accurate objects and does not require any global thresholds. Watershed parameters for each variable were based on their distributions of values and the ranges that were physically significant. The size parameters were set to find objects on the scale of individual storm cells.

The enhanced watershed was run on grids most likely to produce discrete objects associated with storms: 1-hour max column graupel mixing ratio, radar reflectivity in the -10^{o}C layer, 1-hour max updraft helicity, 1-hour accumulated precipitation, precipitable water, and 1-hour max upward vertical velocity. The 1-hour max column graupel mixing ratio measures the amount of ice balls in a column of air, which would be high in grid cells containing hail cores. The radar reflectivity indicates the presence of storms, and hail results in high reflectivity values. The updraft helicity is high in the updrafts of supercell thunderstorms, which are the primary producer of severe hail. High upward vertical velocity indicates the presence of a strong updraft, which produces large hail by lofting hailstones higher in the atmosphere and holding them aloft longer. Accumulated precipitation indicates the extent of the storms as well.

An example case of the enhanced watershed is shown in Fig. 3. The predicted 1-Hour Max Column Graupel Mixing Ratio is shown for the WRF Control and WRF M5 Perturbations. Both show multiple graupel cores over central Oklahoma with some differences in location and size. In both cases, the enhanced watershed is able to identify the major graupel cores while either ignoring objects that do not meet the minimum size criteria or merging them with nearby objects if the two are close enough together and linked with high enough values.

The objects found at each time step were then linked through a customized object tracking technique. Objects from the first time step were associated with objects in the following time step by calculating the Euclidean distance to each object and keeping those within a tracking radius that corresponds to the approximate distance a storm could travel in 1 hour. If only one object is within the tracking radius, then it is selected. For more than one matching object, the percent overlap is used to select the best match. If no overlaps occur, then the object with the closest centroid is selected. This technique can handle storm splits and mergers. It is a variant of the techniques discussed in [24].

C. Spatiotemporal Relational Random Forests

We used the spatiotemporal relational random forest (SRRF) [14] to perform the severe hail predictions. The SRRF is an ensemble of spatiotemporal relational probability trees, which are probability estimation trees that learn from relational data incorporating space and time variations. The SRRF uses the Random Forest [25] procedures of bagging the training set for each tree, selecting a random subset of questions at each node, and not pruning. The trees use a traditional greedy growth algorithm but differ from traditional decision trees in the questions used to split the data. Instead of splitting based

on a category or single value, the tree can ask questions about the objects, attributes, and relationships, such as "Is there a precipitation object with attribute 1 hour accumulated precipitation that exceeds 5 mm?". Additional attributes are precomputed from the provided data in order to examine dynamic statistics and spatiotemporal derivatives, such as gradients. The objects are also decomposed into shapelets [26], which are pieces of time-series derived from the shapes of objects [27]. The shapelets found from each case can be matched against a reference shapelet. Questions can be generated based on the match or based on comparisons of the statistical distributions of the shapes. Spatial relationships can also be discovered between fielded objects with location information.

For the experiments in this paper, we used a 46-tree SRRF that sampled 1000 questions at each node in each tree with a maximum tree depth of 10. These settings balanced predictive and computational performance. These settings have provided strong performance in the past while allowing the forests to be generated in a reasonable amount of computational time. For our baseline comparison, we used a 46-tree random forest from the R *randomForest* library. Since the full spatiotemporal relational dataset would not be translatable to a traditional framework, we extracted the mean and standard deviation of the last time step of the neighborhood objects for each variable and ensemble member for a total of 180 attributes.

Both the SRRF and Random Forest can produce rankings of permutation-based variable importance [25]. For each attribute in the dataset, the values are randomly permuted among the cases that were in the training set but not used for training each tree due to resampling. The permuted cases are then re-evaluated by each tree, and the average decrease in accuracy compared to the original performance is the variable importance score. Unlike procedures used for stepwise selection in regression problems, permutation variable importance can account for nonlinear interactions among variables, but it is unstable due to its random components. For that reason, the variable importance scores are averaged over 30 runs of the SRRF and Random Forest to filter the variability.

D. Verification and Analysis

The models are evaluated with statistics examining aspects of how well they discriminate between severe and non-severe hail. For each run of each model, 1466 cases were uniformly randomly selected for training, and the remaining 367 were used for testing. Some days may be represented in both the training and test sets, but the relative performance should not be affected. The primary evaluation tool is the Receiver Operator Characteristic (ROC) curve [28], a plot of the probability detection (POD) versus the probability of false detection (POFD). The POD and POFD are calculated over all thresholds. A larger area under the ROC curve (AUC) corresponds to larger skill relative to climatology. The Peirce Skill Score (PSS) [29], or POD−POFD, can be used to find the optimal threshold for the classifier at the point on the ROC curve where it is maximized [30]. At that threshold, other two-class verification scores such as False Alarm Ratio (FAR) and frequency bias can be calculated from the binary contingency table. The formulas for each of the scores used in the paper are shown in Table II. Statistical significance of the differences in the distributions of the scores was tested calculating the bootstrap confidence intervals of the ratios of the SRRF and RF scores and determining if

TABLE II: Contingency Table Scores. H is a hit, m is a miss, f is a false alarm, and n is a true negative.

Score	Formula
PSS	$\frac{h}{h+m} - \frac{f}{f+n}$
POD	$\frac{h}{h+m}$
POFD	$\frac{f}{f+n}$
FAR	$\frac{f}{f+h}$
Bias	$\frac{h+f}{h+m}$

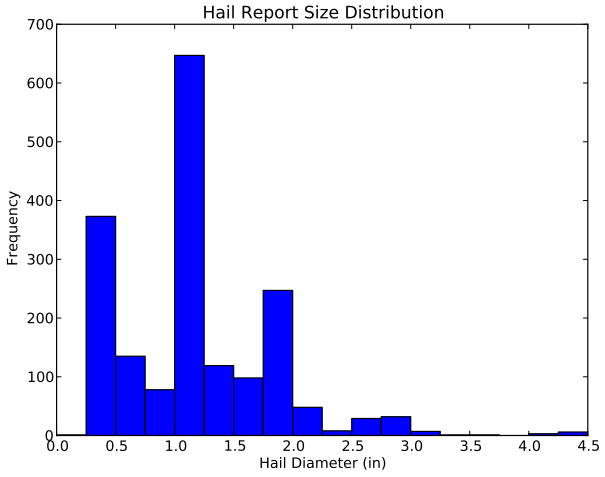


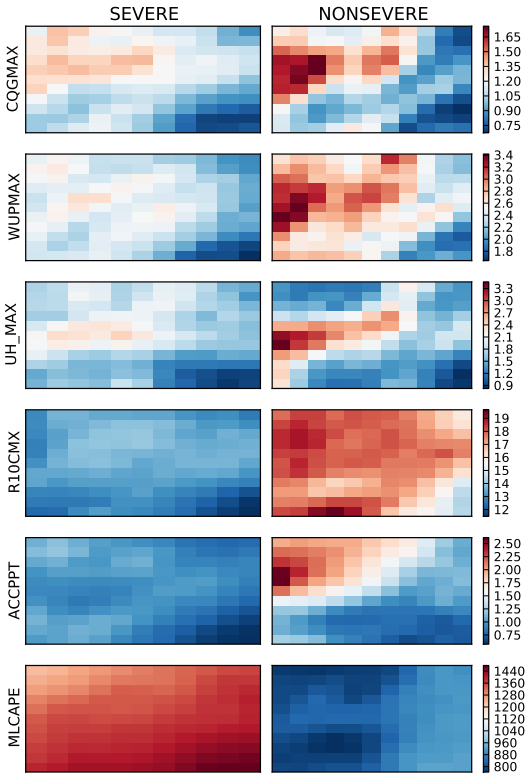
Fig. 4: Frequency of hail reports by size in 0.25 inch increments.

IV. Results

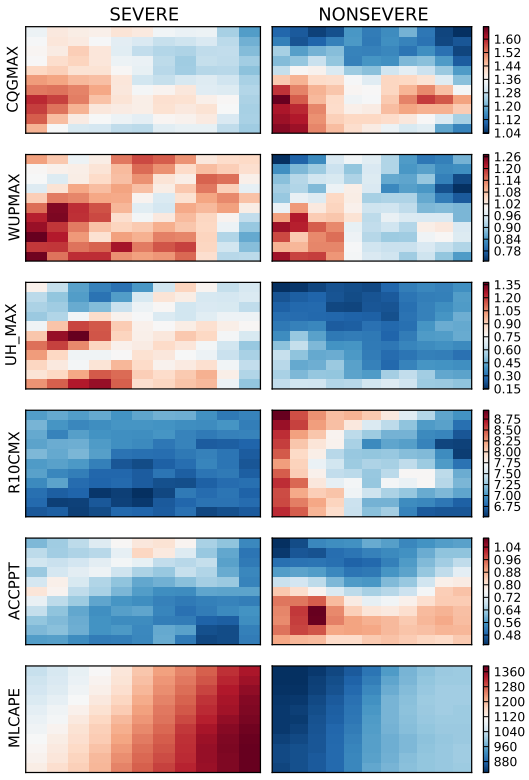
A. Data Distributions

The distribution of hail size in the training data shows shows of the biases of the report dataset. The frequency of reports by size is shown in Fig. 4. Hail reports are binned in 0.25 inch increments, and the smallest reported hail size is 0.25 inches. The spikes at 1 inch and 1.75 inches are due to the public reporting the hail size based on comparisons to quarters and golfballs, respectively instead of directly measuring the hail [17]. Incorporating additional reports and remotely-sensed hail datasets should help to correct those biases.

In order to visually examine the differences among the different SSEF ensemble members, composites were made of the last time step of the neighborhood objects (Fig. 5). The control member (Fig. 5a) shows relatively small differences in the neighborhood values for each of the variables, but the non severe cases do have relatively higher values in the



(a) WRF ARW Control



(b) WRF ARW M6

Fig. 5: Neighborhood composites for the WRF control member and one of the perturbations.

TABLE III: Comparison of the 30-run mean and the 95% bootstrap confidence intervals for various verification scores between the SRRF and random forest.

	SRRF			RF		
Score	2.5	Median	97.5	2.5	Median	97.5
AUC	0.740	0.747	0.755	0.716	0.726	0.736
PSS	0.371	0.385	0.398	0.368	0.386	0.403
POD	0.780	0.813	0.843	0.792	0.822	0.849
POFD	0.390	0.428	0.465	0.406	0.436	0.466
FAR	0.188	0.199	0.210	0.185	0.196	0.207
Bias	0.968	1.019	1.067	0.9810	1.026	1.067

TABLE IV: Top 10 variable importance rankings for 30 SRRFs.

Type	Item	Attribute	Mean	SD
Object	pwat	gridded values	214.75	30.90
Object	wupmax	gridded values	193.29	31.61
Relation	nearby	distance	155.97	41.49
Object	neighborhood	mlcape	131.73	27.15
Object	echotp	gridded values	129.91	21.56
Object	accppt	gridded values	110.85	22.85
Relation	nearby	east-distance	78.31	18.16
Relation	nearby	direction	77.02	17.71
Relation	nearby	north-distance	73.34	16.15
Object	r10cmx	gridded values	71.14	12.99

western half of the neighborhood and stronger gradients. The MLCAPE is consistently higher in the severe cases and has little gradient. In the WRF ARW M6 perturbation, the severe cases show higher values in the neighborhood for 3 of the 6 variables. Some of the variability may be due to storms being outside of the immediate neighborhood of the reports or differing biases in different regimes.

B. SRRF Verification

Fig. 6 shows the ROC curves for all 30 runs of the SRRF and the mean of those runs. All of the SRRF and RF runs show positive skill in terms of Area Under the ROC Curve. The range of ROC curves indicates a smaller variance in the SRRF runs compared to the RF. At the higher probability thresholds the SRRF maintains higher PODs for a given POFD, and at lower probability thresholds the performance is similar.

Table III compares the verification scores of the SRRF and Random Forest. The median and 95% bootstrap confidence intervals are provided. The SRRF has a statistically significantly ($\alpha < 0.01$) larger AUC compared with the Random Forest. The differences in the other scores are not statistically significant. This supports the similar POD and POFDs seen in Fig. 6. Both the SRRF and the RF show little to no frequency bias. The high POD and low FAR for both models is encouraging in that it is successfully picking severe hail cases without causing too many false alarms. Given that it is effectively making an 18 to 30 hour forecast when making this distinction, the scores are particularly impressive.

C. Variable Importance

Variable importance rankings averaged over 30 SRRF runs are shown in Table IV. The objects most prominent in the

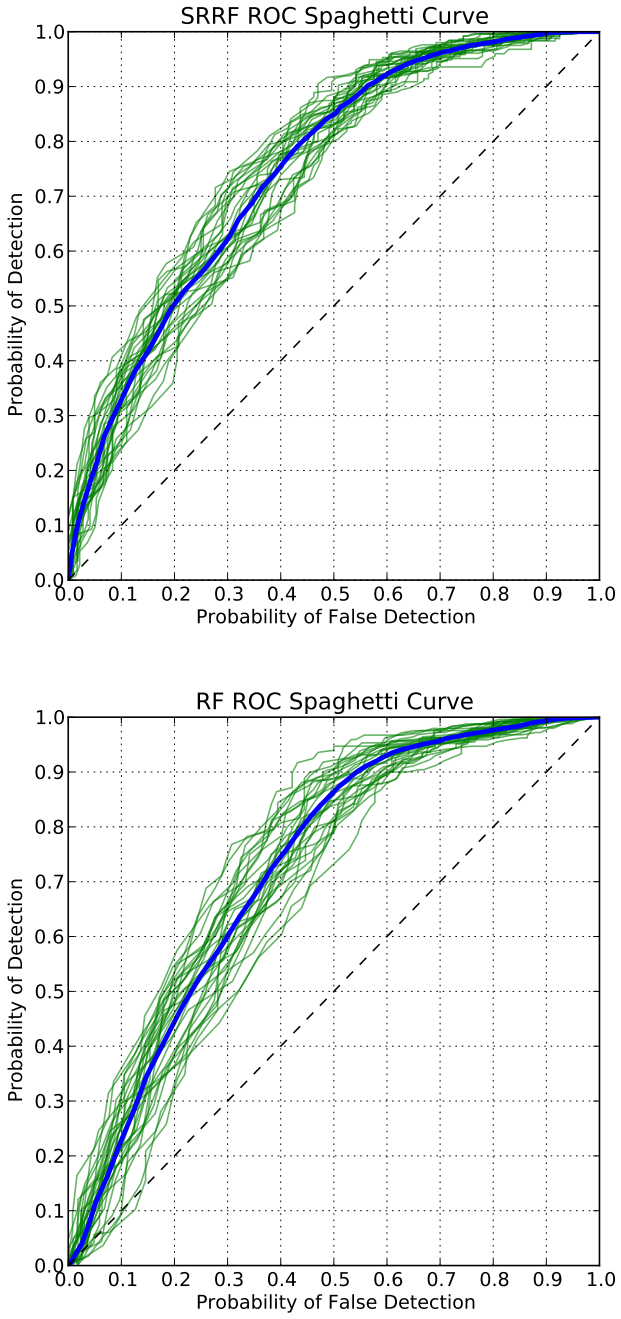


Fig. 6: ROC curves for each of the 30 runs (green) and the mean of the curves (blue). The dashed line indicates the no-skill threshold.

TABLE V: Top 10 variable importance rankings for 30 random forests.

Attribute	Mean Score
ARW M8 MLCAPE Mean	0.148
ARW M13 MLCAPE SD	0.145
ARW M6 MLCAPE SD	0.139
ARW M14 MLCAPE SD	0.137
ARW M11 MLCAPE SD	0.137
ARW M07 Wupmax SD	0.134
ARW M13 Cqgmax SD	0.133
ARW M14 R10cmx SD	0.129
ARW M12 MLCAPE Mean	0.126
ARW M15 MLCAPE SD	0.125

rankings are wupmax, or 1-hour max upward vertical velocity, and pwat, or precipitable water. Precipitable water tends to be correlated with heavy precipitation events and when coupled with other factors correlates with large hail. The wupmax object corresponds to the updrafts of storms in the model. Stronger updrafts tend to be associated with larger hail size because the hailstones can be lofted higher into the storm and stay in the hail growth region of the storm for a longer period of time. All of the attributes associated with the nearby relationship were also very important. The direction, east-distance, and north-distance all provide some information about the position of the storms relative to the hail report. This kind of information is typically not captured in more traditional statistical post-processing methods, so its high importance vindicates the usefulness of the spatiotemporal relational framework. The neighborhood objects did not appear to provide significant information except for MLCAPE, which tends to be fairly constant throughout storm regions. The echotp, or echo top object indicates the height of the storm and tends to be higher in storms with strong updrafts.

The variable importance rankings for the average of 30 Random Forest runs are shown in Table V. Unlike the SRRF, each ensemble member had separate, independent attributes, so it may provide some indication of which model had the most impact on the predictions. Most of the top 10 was dominated by the MLCAPE from various ensemble members. The upward vertical velocity and radar reflectivity were also important for similar reasons as in the SRRF. The column graupel did appear in the top 10 but not in the SRRF top 10. While it is directly connected to the presence of hail, it may occur in large values for cases in which there is a lot of small graupel and hail, so it is not as effective as a discriminator. None of the ensemble members was consistently used more often in the upper section of the variable importance rankings, so no preferences can be ascertained from this dataset.

V. Discussion

The SRRF and the spatiotemporal relational framework show great initial promise for predicting severe hail. The additional spatiotemporal data was shown to have provided additional predictive skill compared to data in a traditional framework with a traditional Random Forest. Variable importance rankings also show high rankings for objects outside of

the grid point neighborhood and for the relationships showing the positions of those more distant objects. Discovering the full significance of those objects and relationships requires more in-depth analysis of their associated tendencies. One possible hypothesis for their significance is that the SRRF is discovering spatial biases in each ensemble member in terms of storm placement.

These initial results lead to many potential avenues for future exploration. Although the data do capture some of the major ingredients of large hail events, the large storage and computational requirements make running the over anything other than a coarse grid or a grid over a small domain infeasible. We plan to explore making model storm-centric hail predictions and on finding ways to characterize the spatial uncertainty of the predictions more accurately and efficiently. We also plan to expand the predictions from severe/non-severe to multi class based on major size categories and also try exact prediction of hail size with regression techniques.

Reproducibility of research: In conjunction with the publication of this paper, we have released the full SRPT/SRRF code and the hail training and testing data in the format used by our algorithm at *http://idea.cs.ou.edu/software.html*.

ACKNOWLEDGEMENTS

CAPS SSEF forecasts were supported by a grant (NWSPO-2010-201696) from the NOAA Collaborative Science, Technology, and Applied Research (CSTAR) Program, and the SSEF and SRRF 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. Nate Snook and Youngsun Jung contributed ideas and feedback to the project. Hail reports were retrieved from the NOAA Storm Prediction Center website. Data from mPING were provided by the NOAA National Severe Storms Laboratory courtesy of Zac Flamig. This study was funded by the NSF Graduate Research Fellowship under Grant 2011099434 and by NSF grants AGS-0802888 and AGS-1261776.

REFERENCES

[1] T. P. Marshall, R. F. Herzog, S. J. Morrison, and S. R. Smith, "Hail damage threshold sizes for common roofing materials," in *Preprints, 21st Conference on Severe Local Storms*. San Antonio, TX: Amer. Meteor. Soc., 2002, p. P3.2.

[2] F. Glass and M. Britt, "The historic missouri-illinois high precipitation supercell of 10 april 2001," in *Preprints, 21st Conf. on Severe Local Storms, San Antonio, TX, Amer. Meteor. Soc*, 2002, pp. 99–104.

[3] S. A. Changnon, "Increasing major hail losses in the u.s." *Climate Change*, vol. 96, pp. 161–166, 2009.

[4] ——, "The scales of hail," *J. Appl. Meteor.*, vol. 16, pp. 626–648, 1977.

[5] NWS. (2013, 5) Gpra metrics national yearly trends. [Online]. Available: http://verification.nws.noaa.gov

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

[7] C. Marzban and A. Witt, "A bayesian neural network for severe-hail size prediction," *Wea. Forecasting*, vol. 16, pp. 600–610, 2001.

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

[9] M. S. Gilmore, J. M. Straka, and E. N. Rasmussen, "Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme," *Mon. Wea. Rev.*, vol. 132, pp. 2610–2627, 2004.

[10] J. A. Milbrandt and M. K. Yau, "A multimoment bulk microphysics parameterization. part iii: Control simulation of a hailstorm," in *J. Atmos. Sci.*, vol. 63, 2006, pp. 3114–3136.

[11] 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.

[12] D. Gagne, A. McGovern, and M. Xue, "Machine learning enhancement of storm scale ensemble precipitation forecasts," in *Intelligent Data Understanding (CIDU), 2012 Conference on*. Boulder, CO: IEEE CIS, 2012, pp. 39–46.

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

[14] A. McGovern, D. J. Gagne II, J. K. Williams, R. A. Brown, and J. B. Basara, "Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning," *Machine Learning*, pp. 1–24, 2013. [Online]. Available: http://dx.doi.org/10.1007/s10994-013-5343-x

[15] J. L. Cintineo, T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, "An objective high-resolution hail climatology of the contiguous united states," *Wea. Forecasting*, vol. 27, pp. 1235–1248, 2012.

[16] A. Witt, M. D. Eilts, G. J. Stumph, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, "An enhanced hail detection algorithm for the wsr-88d," *Wea. Forecasting*, vol. 13, pp. 286–303, 1998.

[17] K. L. Ortega, T. M. Smith, K. L. Manross, K. A. Scharfenberg, A. Witt, A. G. Kolodziej, and J. J. Gourley, "The severe hazards analysis and verification experiment," *Bull. Amer. Meteor. Soc.*, vol. 90, pp. 1519–1530, 2009.

[18] F. Kong, "2013 caps spring forecast experiment program plan," Center for the Analysis and Prediction of Storms, Tech. Rep., 2013.

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

[20] M. Xue, K. K. Droegemeier, V. Wong, A. Shapiro, K. Brewster, F. Carr, D. Weber, Y. Liu, and D. Wang, "The Advanced Regional Prediction System (ARPS) - A multiscale nonhydrostatic atmospheric simulation and prediction model. Part II: Model physics and applications," *Meteor. Atmos. Phys.*, vol. 76, pp. 134–165, 2001.

[21] M. Xue, D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, "The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation," *Meteor. Atmos. Phys.*, vol. 82, pp. 139–170, 2003.

[22] W. C. Skamarock and J. B. Klemp, "A time-split nonhydrostatic atmospheric model for weather research and forecasting applications," *Journal of Computational Physics*, vol. 227, pp. 3465–3485, 2008.

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

[24] V. Lakshmanan and T. Smith, "An objective method of evaluating and devising storm-tracking algorithms," *Wea. Forecasting*, vol. 25, pp. 701–709, 2010.

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

[26] L. Ye and E. Keogh, "Time series shapelets: A new primitive for data mining," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 947–956.

[27] E. Keogh, L. Wei, X. Xi, S.-H. Lee, and M. Vlachos, "Lb_keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures," in *Proceedings of the 32nd international conference on Very large data bases*, ser. VLDB '06. VLDB Endowment, 2006, pp. 882–893.

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

[29] C. S. Peirce, "The numerical measure of the success of predictions," *Science*, vol. 4, pp. 453–454, 1884.

[30] A. Manzato, "A note on the maximum Peirce skill score," *Wea. Forecasting*, vol. 22, pp. 1148–1154, 2007.