Tornadic Supercell Environments Analyzed Using Surface and Reanalysis Data: A Spatiotemporal Relational Data-Mining Approach

DAVID JOHN GAGNE II
School of Meteorology, University of Oklahoma, Norman, Oklahoma

AMY MCGOVERN
School of Computer Science, University of Oklahoma, Norman, Oklahoma

JEFFREY B. BASARA
Oklahoma Climatological Survey and School of Meteorology, University of Oklahoma, Norman, Oklahoma

RODGER A. BROWN
NOAA/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 22 March 2011, in final form 22 May 2012)

ABSTRACT

Oklahoma Mesonet surface data and North American Regional Reanalysis data were integrated with the tracks of over 900 tornadic and nontornadic supercell thunderstorms in Oklahoma from 1994 to 2003 to observe the evolution of near-storm environments with data currently available to operational forecasters. These data are used to train a complex data-mining algorithm that can analyze the variability of meteorological data in both space and time and produce a probabilistic prediction of tornadogenesis given variables describing the near-storm environment. The algorithm was assessed for utility in four ways. First, its probability forecasts were scored. The algorithm did produce some useful skill in discriminating between tornadic and nontornadic supercells as well as in producing reliable probabilities. Second, its selection of relevant attributes was assessed for physical significance. Surface thermodynamic parameters, instability, and bulk wind shear were among the most significant attributes. Third, the algorithm’s skill was compared with the skill of single variables commonly used for tornado prediction. The algorithm did noticeably outperform all of the single variables, including composite parameters. Fourth, the situational variations of the predictions from the algorithm were shown in case studies. They revealed instances both in which the algorithm excelled and in which the algorithm was limited.

1. Introduction

Most observational studies of tornadic supercell thunderstorms focus on either gathering many details about a few storms or gathering a few general characteristics of many storms. Highly detailed studies of a small number of storms (e.g., Markowski et al. 2002; Shabbott and Markowski 2006; Grzych et al. 2007) have found connections between equivalent and virtual potential temperature deficits in the rear-flank and forward-flank downdrafts and resulting tornadogenesis for those particular storms, but their selection of storms is likely not representative of the full spectrum of possible tornadic and nontornadic supercells. Their observations are also collected during field projects using specialized equipment not routinely available to operational forecasters, so additional signals found would require a significant overhaul of the current observing networks for the advances in knowledge to be useful operationally. On the other end of the spectrum, tornadic supercell climatology studies examine the general characteristics of a large number of supercells through the use of “representative” proximity soundings (e.g., Brooks et al. 1994; Rasmussen and Blanchard 1998; Bunkers et al. 2006b) or model analyses (e.g., Thompson et al. 2003; Markowski et al. 2003) and reanalyses (e.g., Brooks et al. 2003b, 2007;
Grünwald and Brooks 2011). These studies generally look at a single snapshot in time of the near-storm environment and have provided distributions of severe weather parameters useful to forecasters, but they have not accounted for the evolution of the environment as storms evolve.

This study takes an approach closer to the middle of the spectrum. The environments of a large number of tornadic and nontornadic supercells in Oklahoma are examined by combining a dense operational surface observing network with regional upper-air reanalysis data to track how the environment changes as each supercell evolves in levels of detail currently available to operational forecasters in that area. The resulting data are used to train a complex data-mining algorithm that can assign each storm a probability of tornadogenesis based on what the algorithm deems to be the most important environmental characteristics. The algorithm will be assessed for utility in four ways. First, it will be scored on the skill of its probability forecasts. Second, its selection of relevant attributes will be assessed for physical significance. Third, its skill will be compared with the skill of single attributes that are commonly used for tornado prediction. Fourth, the spatial variations of its predictions will be shown in case studies.

2. Background

a. Distinguishing tornadic and nontornadic supercells

Previous studies have established climatologies of severe weather forecast parameters and examined storm environment variability through the use of proximity soundings. Rasmussen and Blanchard (1998) produced a baseline climatology of many severe weather parameters for ordinary thunderstorms, nontornadic and weakly tornadic [Fujita scale 0 (F0)—F1] supercells, and strongly tornadic (F2—F5) supercells. The best discriminators between strongly tornadic and other supercells were the lifted condensation level (LCL); mean shear; and a composite of mixing ratio, storm-relative helicity (SRH), and minimum midlevel wind speed (Brooks et al. 1994). Studies of long duration supercells (e.g., Bunkers et al. 2006a,b) found that long-lived supercells (lasting longer than 4 h) produced significantly more strong tornadoes than short-lived supercells, and that long-lived storms tended to be isolated. Long-lived supercells also formed in environments with low LCL heights and high 0–1-km SRH, which supported the link between long-lived supercells and strong tornadoes.

Severe weather climatological studies have also been conducted with model analysis and reanalysis data. Thompson et al. (2003) and Markowski et al. (2003) analyzed proximity soundings derived from the Rapid Update Cycle (RUC) model for differences in the environments between significantly tornadic, weakly tornadic, and nontornadic supercells from 1999 to 2001. Significantly tornadic storms were found to have higher mean-layer convective available potential energy (MLCAPE) and 0–1-km SRH and lower LCL heights. The best discriminators were found to be the supercell composite parameter (SCP) and significant tornado parameter (STP; Thompson et al. 2002). Global reanalysis data have been used to map the spatial distribution of severe weather parameters (Brooks et al. 2003b), to chart the monthly variation in those parameters (Brooks et al. 2007), and to compare weakly tornadic and strongly tornadic environments (Grünwald and Brooks 2011). Regional reanalysis data have been used to correlate seasonal tornado activity with variations in atmospheric forcing (Trapp 2010).

The position of mesoscale boundaries and storm motion relative to them have also been shown to impact the possibility of tornadogenesis in a supercell. Markowski et al. (1998) proposed that mesoscale boundaries can create regions of enhanced horizontal vorticity that can be tilted into the vertical and subsequently stretched, increasing the likelihood of tornadogenesis. Nearly 70% of the tornadoes in their study occurred within 30 km of a boundary. They also hypothesized that storms moving along boundaries, rather than across them, stay in a favorable environment for a longer period and are more likely to be tornadic. Bunkers et al. (2006b) supports this hypothesis with statistics indicating that storms moving along boundaries within a large warm sector tend to be long-lived, and that long-lived supercells tend to be strongly tornadic.

Thermodynamic variations at the surface also have shown some potential as predictors of tornadogenesis. Markowski et al. (2002) analyzed mobile mesonet observations of the outflow within the rear-flank downdrafts of 12 tornadic and nontornadic supercells and found that the significantly tornadic (F2 and greater) supercells had smaller equivalent and virtual potential temperature deficits compared to weakly tornadic and nontornadic supercells. Further research (e.g., Shabbott and Markowski 2006; Grzych et al. 2007) supported those conclusions, but none of the studies explored the phenomenon in large numbers of supercells.

b. Spatiotemporal relational framework

The spatiotemporal relational data framework is designed to handle large amounts of data from multiple sources in an organized and descriptive way. Traditional statistical and data-mining techniques use a flat data representation (e.g., Quinlan 1993). The flat data representation consists of a series of cases that each
have a set of attributes, or descriptive values, and a label, or the value being predicted. Attributes could include temperature, wind speed, and dewpoint for a particular storm environment. A model in this framework picks and/or weights the attributes to make a prediction of the label, which in this dataset is whether a supercell is tornadoic or nontornadoic. By representing each case as a collection of independent attributes, the flat approach ensures a small search space that can be explored by a wide range of models, but it also ignores connections and influences among attributes that could have a major effect on predictions.

Relational data frameworks expand the flat list of cases into a network representation in which nodes (or objects) in the network are connected by relationships based on spatial, temporal, or hierarchical connections. Objects, in this case, are aggregations of data describing a particular physical construct, such as a storm or boundary. Relationships are descriptive connections, such as nearby or overlapping. Each object and relationship contain descriptive attributes like the ones in the flat framework. Relational networks can support multiple objects of the same type as well as different types of objects. In addition to finding patterns in the attributes, relational data-mining algorithms can examine patterns in the existence of and the relationships among different objects. The relational probability tree (RPT) of Neville et al. (2003) extended the traditional decision-tree machine-learning algorithm (Quinlan 1986) to these relational data frameworks.

The spatiotemporal relational data framework extends the relational framework into domains with spatially and temporally varying objects, attributes, and relations. Within the spatiotemporal framework, the network tracks the existence and motion of objects and relationships. Attributes can represent as time series of either scalar values or two- or three-dimensional grids of scalar or vector data. With the flat data framework, a storm could just have one temperature, but with the spatiotemporal relational framework, the storm can have an initiation and termination time and a grid of temperature values from the near-storm environment that can vary as the storm matures. Spatiotemporal networks can accommodate a large number of varied object and relationship types, which allows for complex domains to be modeled. The abundance of organized data gives spatiotemporal relational data-mining algorithms an opportunity to explore and to find the most appropriate aggregations of the data from which to make predictions.

c. Spatiotemporal relational random forests

Creating a model from data in a spatiotemporal relational framework requires an algorithm designed specifically to interpret that framework. The first model created for this framework was the spatiotemporal RPT (SRPT) by McGovern et al. (2008), an extension of the RPT (Neville et al. 2003) to spatiotemporal relational datasets. SRPTs grow a binary decision-tree structure. Decision trees are flowchart-like models that contain branching paths between nodes eventually leading to a predicted value. Nodes, in the decision-tree context, are points in the tree structure where a function is evaluated on a given case. The top and interior nodes of a tree are evaluation nodes, which contain a yes-or-no question concerning an aspect of the data. From each evaluation node extend “yes” and “no” branches that connect to either other evaluation nodes or to leaf nodes that contain probabilities for each possible label. The types of questions that can be asked at each node range from whether or not an object exists to if the spatial gradient of a gridded attribute exceeds a certain value. The full list of possible questions can be found in McGovern et al. (2011). The questions are structured as templates focusing on a particular aspect of the dataset and can have any given combination of objects, attributes, or relationships included, such as, “Is the storm surface maximum temperature gradient ever greater than 2°C km⁻¹?” Random sampling of the questions, instead of a full search of question types as done in other decision-tree implementations, is performed owing to the large search space created by the number of question types, objects, attributes, and relationships. The question with the highest chi-square goodness-of-fit value (Wilks 2011), based on the similarity of the distributions of cases that answered yes and no to the question, is selected, and the cases are sorted by whether they answered yes or no to that question. The question then becomes a node of the tree. The node-building process continues recursively until one of the stopping criteria has been met. At that point, probabilities based on the relative frequency of the training labels are calculated and are stored in a leaf node.

The spatiotemporal relational random forest (SRRF; Supinie et al. 2009; McGovern et al. 2011) combines the SRPT decision tree with the random forest (Breiman 2001) approach for growing a decision-tree type ensemble. For each tree in the forest, the original training set is replaced with a bootstrap resampled set of the same size, so some cases appear multiple times and others do not appear at all. The question sampling feature of the SRPT is used in place of the random selection of attributes at each node in traditional random forests. To calculate the overall label for the forest, the means of the probabilities for each label type from each tree are calculated, and the label with the highest mean probability is selected as the label. By
incorporating multiple trees trained on resampled data, the random forest explores a larger portion of a search space and reaches a level of high performance more consistently than a single decision tree. With a large decision-tree ensemble, the impact of a particular attribute on the dataset can no longer be shown by visual inspection of the decision tree. Instead, a statistical estimate of the impact of an attribute can be made by aggregating statistics measuring the impact within all trees.

3. Data and methods

a. Data sources

The Hocker and Basara (2008) Oklahoma supercell thunderstorm dataset was used for this study. The dataset contains 926 supercell tracks from 1994 to 2003 in Oklahoma that were derived from Next Generation Weather Radar (NEXRAD) radar data in a GIS framework. Information about the location, movement, and timing of each storm was included in those data. Storms that formed just outside the state borders and moved into Oklahoma were included as well as storms that formed in Oklahoma and exited the state, although only the Oklahoma portion of their track was included. Tornado tracks were obtained from the Storm Prediction Center (SPC) Tornado Tracks database (Schaefer and Edwards 1999), which included location, timing, and Fujita scale information for tornadoes from 1950 to the present. Tornadoes were individually matched in space and time with supercell tracks using ArcGIS and the SPC’s Online SeverePlot. For each supercell, metadata were recorded; metadata included whether or not the storm produced a tornado, the Fujita ranking of the strongest tornado, and the touchdown time of the strongest tornado.

Surface observations were obtained for each storm from the Oklahoma Mesonet (McPherson et al. 2007), a dense network of 120 surface observation sites across Oklahoma, with an average spacing of approximately 30 km (Fiebrich and Crawford 2001). Included among the data parameters collected at 5-min intervals at each site were temperature, moisture, pressure, wind, and rainfall. Because the mesonet has been operating continuously since 1 January 1994, it provided a detailed array of surface thermodynamic and wind data for every storm in the climatology. In addition to the parameters directly retrieved from observations, dewpoint temperature, potential temperature $\theta$, equivalent potential temperature $\theta_e$, and virtual potential temperature $\theta_v$ were derived from the temperature, relative humidity, and pressure variables. LCL height in meters was approximated using Espy’s equation (Espy 1841).

Environmental data were retrieved from the North American Regional Reanalysis (NARR) (Mesinger et al. 2006). The NARR provides a wide array of meteorological parameters over North America with a horizontal grid spacing of 32 km and a temporal resolution of 3 h. The grid spacing is comparable to the average distance between Oklahoma Mesonet stations although numerical filtering limits its resolution of atmospheric phenomena to those that are hundreds of kilometers in size. When compared with observed proximity soundings, the NARR provides closer representations of the environment for each storm both spatially and temporally and has been assessed for data quality (Mesinger et al. 2006). Surface-based (SB) and mean-layer (ML) CAPE and convective inhibition (CIN); 0–1-, 0–2-, and 0–3-km SRH; 0–1-, 0–3-, and 0–6-km bulk wind shear; the energy helicity index (EHI) and STP (fixed layer); and tropopause potential temperature were obtained from the NARR.

Because of constraints posed by the main data sources for this project, the spatial and temporal extents of the domain have been necessarily limited. The Oklahoma Mesonet and national NEXRAD network became operational in 1994, so storms occurring before then could not be included with the same detail as the rest of the set. No other state has a surface observing network of the same density as the Oklahoma Mesonet. Resources were not available to extend the supercell track data later than 2003 for this project, so more recent storms could not be included. Although RUC model output has been used in previous tornado environment studies, RUC model runs are not available for the time period of this study.

b. Preprocessing

The various data sources were combined in a spatiotemporal relational framework. Figure 1 displays the schema used to organize the data. The schema details the possible set of objects, relationships, and attributes available in any single case. There were three object types: storm surface, boundary, and near-storm environment. Storm surface objects contained attributes that either described the movement of the storm or the surrounding environment at the surface level. Boundaries in this context referred to any boundaries analyzed from the equivalent potential temperature field, including fronts and drylines. Near-storm environment objects contained attributes that are derived from the NARR and described conditions occurring either at a level above the surface or throughout a vertical column. Attribute data could be composed of up to three data types: time series (T), which contains a series of values at discrete time steps; grid (G), which has the attribute
values spread across a 2D grid; and vector (V), which has the attribute values split into vector components (Fig. 1). The wind attribute in Fig. 1 has a TGV structure, so it is a time series of \( u \) and \( v \) grids. Grid points containing missing or erroneous data can be masked and thus can be ignored in any grid-based calculations.

To create each network representation, surface and reanalysis data were processed to record the environmental conditions around each supercell from initiation through either tornadogenesis or the supercell’s termination. First, the Oklahoma Mesonet surface observations were objectively analyzed to a 0.1° latitude by 0.1° longitude grid over Oklahoma using a bilinear interpolation scheme. The grids were transformed into anomalies from the mean of the observations at each time to control for diurnal and seasonal variations. Then areas of the grid corresponding to objectively analyzed boundary regions and within a 50-km radius of each storm were extracted as gridded field attributes for the boundaries and storms, respectively.

An objective frontal analysis located any air mass boundaries in each supercell environment. The thermal frontal parameter (TFP), developed by Renard and Clarke (1965) and shown in Eq. (1), is calculated from the equivalent potential temperature field:

\[
\text{TFP}(\theta_e) = \frac{\mathbf{V} \cdot \mathbf{\theta_e}}{|\mathbf{\theta_e}|}.
\]  

(1)

Since the maximum in TFP occurred along the warm side of the frontal boundary, the maxima in TFP were used as the designated boundary zones. An additional filter to remove areas with TFP gradients below 0.5 (Jenkner et al. 2010) was applied to the TFP field to eliminate noise at the edges of the domain and boundaries that only lasted for one time step. Temporal front tracking was performed by placing an elliptical buffer zone around each boundary and checking for overlaps in the buffers between time steps. In addition to mesonet data fields, attributes describing the orientation (angle relative to the east), major axis length, minor axis length, and eccentricity (ratio of major axis length to minor axis length) of the boundary and surrounding buffer zone were calculated and listed in Fig. 1. These attributes were generated by applying an ellipsoidal regression-fitting routine over the grid points determined to be within the boundary and buffer zone. Although boundaries are often represented as strict discontinuities, the associated thermodynamic gradient can occur over a significant horizontal distance. When rendered as a discrete grid, the boundaries have an elliptical shape. Spatial

---

**Fig. 1.** Schema for each supercell environment in the dataset. Boxes are objects in the network, and the hexagon represents a relationship. Inside the objects and relationships are lists of attributes with letters in parentheses indicating how they are structured. The quantity TG would describe a time series of gridded data, for example. The SD is standard deviation.
relationships between the storms and boundaries were determined by calculating the distance between the storm’s centroid and the nearest point on the boundary. A “nearby” relationship was created if the storm was less than 200 km from the boundary. For each relationship, the boundary relative angle, or the angle between the boundary major axis and the vector connecting the storm and boundary centroids, was also calculated, which could help indicate how the location of the storm relative to the boundary influenced the near-storm environment.

At each point along the path of the storm, a $3 \times 3$ box surrounding the nearest NARR grid point in space and time was extracted and stored. The different depths of bulk wind shear and storm-relative helicity were calculated by interpolating the NARR grid to every 100 m from 0 to 6 km then performing the relevant calculations. EHI was calculated from the MLCAPE and storm relative helicity grids.

c. SRRF training and calibration

Multiple SRRFs with various parameterizations were generated to test the robustness of the approach to variations in the model and the dataset. The number of trees in each forest varied among 1, 10, 50, and 100. The number of questions sampled at each node ranged among 100, 500, and 1000. Evaluation was performed on a test set composed by randomly separating 20% of the cases from the full dataset for each forest. For each set of parameters 30 forests were generated, and their skill scores were averaged to account for any selection bias in the test sets.

The SRRF training process required one additional step to produce reliable probabilities. Since the SRRF probabilities are calculated from the mean of the tree probabilities, the resulting probabilities tend to cluster near 50% and are very rarely near 0% or 100% because as ensemble size increases, the likelihood that all members will be in agreement decreases considerably. To address this issue, 25% of the training set was withheld from the SRRF and was used to fit an isotonic regression (Friedman and Tibshirani 1984) between the SRRF-generated probabilities and the relative frequency of tornadic supercells at each of those probabilities. Isotonic regression produces a stepwise-increasing function that maps the probabilities from an algorithm to the relative frequencies from the calibration set without forcing the calibrated probabilities to fit to a predefined distribution. The function has been shown to improve probabilistic skill in many popular data-mining algorithms, including random forests (Zadrozny and Elkan 2002; Niculescu-mizil and Caruana 2005).

d. Evaluation

Since SRRFs produce a probabilistic forecast, two standard probabilistic forecast evaluation metrics are used for verification. The Brier score (BS; Brier 1950) is a standard measure for probabilistic forecast verification. It can be decomposed into three components (Murphy 1973), as shown in Eq. (2):

$$
BS = \frac{1}{N} \sum_{k=1}^{K} n_k (p_k - \bar{\sigma}_k)^2 - \frac{1}{N} \sum_{k=1}^{K} n_k (\bar{\sigma}_k - \bar{\sigma})^2 + \bar{\sigma}(1 - \bar{\sigma}).
$$

The first term evaluates the reliability, which is the difference between the forecast probability $p_k$ and the observed frequency of yes forecasts for that probability $\bar{\sigma}_k$ weighted by the number of forecasts with that probability value $n_k$. Better reliability occurs when, for instance, a forecast of 40% probability verifies 40% of the time. The second term evaluates the resolution, which is the difference between the observed frequency of yes forecasts and the climatological probability $\bar{\sigma}$. Better resolution is achieved by associating probabilities above the climatological probability with mainly tornadic cases and vice versa. The third term is the uncertainty, which is indicative of the inherent difficulty of the forecast problem. For ease of comparison, the Brier skill score (BSS) is used, which performs a normalized comparison with a reference forecast, taken to be climatology in this case. BSS is shown in its original and component form in Eq. (3):

$$
BSS = 1 - \frac{BS_{\text{ref}}}{BS} = \frac{\text{resolution} - \text{reliability}}{\text{uncertainty}}.
$$

BSS reaches a maximum of 1 when resolution is maximized and reliability is minimized, and it is 0 when resolution and reliability are equal. The components of the Brier score and the Brier skill score can be visualized with an attributes diagram (Wilks 2011). The observed frequencies of the yes forecasts at each probability threshold are plotted against the probability thresholds. That line is compared with the perfect reliability line, which has a slope of 1 from 0% to 100%; the no-resolution lines, which are horizontal and vertical lines at the climatological probability value; and the no-skill line, which is halfway between the perfect reliability and no-resolution lines and indicates where the reliability and resolution are equal.

Probabilistic forecasts can also be evaluated based on how well they discriminate between yes and no forecasts over the full range of probabilities. Relative operating characteristic (ROC) curves (Mason 1982) show this
performance by splitting forecasts into two categories through a series of probability thresholds and then calculating the probability of detection (POD) and the probability of false detection (POFD) at each threshold. The points are plotted on a grid of POD versus POFD and connected to form the ROC curve. Skill can be derived from the area under the ROC curve (AUC) because better forecasts have higher POD and lower POFD for a larger range of probability thresholds. An AUC of 1 indicates a perfect forecast, and an AUC of 0.5 is equivalent to a random forecast.

e. Attribute evaluation

Individual attributes were also evaluated for their contribution to SRRF performance. Breiman (2001) described a technique for evaluating random forest attributes called variable or attribute importance that can incorporate the relative placement and use of an attribute across multiple trees and forests into one score. During the bootstrap resampling process, some cases, called the out-of-bag cases, were not selected and were independent of that particular tree. To calculate attribute importance, first each tree evaluated the out-of-bag cases to establish a baseline percent correct. Then the values of each attribute were shuffled among the out-of-bag examples, and those cases were reevaluated to find the shuffled percent correct. The difference between the baseline and shuffled percent correct was the tree’s importance score. The raw importance scores for the forest were the means of all the tree importance scores. The raw importance score was then normalized by dividing it by the standard deviation of the tree importance scores. The attribute importance score distribution from the 30 SRRFs generated for each parameterization was used to determine which attributes had a statistically significant importance, assuming a normal distribution. Attributes had higher importance scores if they were selected more often, were used in nodes closer to the top evaluation node, and were traversed by more examples.

Attributes could also be used as individual models. To set a baseline for SRRF performance, the AUC was calculated by varying the value of the decision threshold among all possible values for each statistical aggregation of the attributes in the dataset, similar to a method proposed by Doswell and Schultz (2006). At each threshold, a contingency table was constructed by assuming a predicted label of ‘tornadic’ for values above the threshold and ‘nontornadic’ otherwise. Calculating the POD and POFD at each threshold allowed for a ROC curve to be plotted, as described in the previous section. From the curve of POD versus POFD, the Peirce skill score (PSS), equivalent to POD – POFD, could be calculated. Because of this relationship, the maximum PSS was at the threshold that was farthest from the no-skill line (Manzato 2007). In effect, the threshold where the maximum PSS occurred was the best discrimination threshold for the data. The false-alarm ratio (FAR), which was the ratio of tornadic storms misclassified to all storms labeled tornadic, could also be calculated at the point of maximum PSS.

4. Results

a. Storm distributions

The distribution of supercells favored nontornadic and weakly tornadic storms. Of the 926 storms in the dataset, 215 (23.5%) were tornadic and 711 (76.5%) were nontornadic. The distribution of tornadic storms showed a sharp decrease in frequency as the Fujita scale rating (Fujita 1981) increased (Fig. 2). Of the tornadic storms, 79.5% were weakly tornadic (F0–F1).

b. Skill evaluation

The SRRFs showed increasingly positive skill with larger numbers of trees and more questions sampled per node in both probabilistic verification metrics. Although single trees performed with little to no skill in most cases, forests of even 10 trees had significant skill (Fig. 3). Increasing both number of samples and number of trees increased the AUC, although there is no significant difference between 50 and 100 trees for 500 and 1000 samples. Forest size had the largest impact on BSS differences at 500 samples. At 1000 samples, the forests of at least 10 trees produce very similar BSS values likely because the trees are more similar and the calibration function ensures that the range of probabilities
predicted does not decrease with additional trees. With these BSS and AUC values, the SRRFs show useful skill for discriminating between tornadic and nontornadic supercells given the environmental characteristics.

c. Attribute importance

Attribute importance calculations were performed on SRRFs using different attribute sets to determine which attributes had the largest impact on SRRF predictive performance. For example, if the randomization of an attribute’s values caused a significant decrease in performance across multiple runs of the SRRF, then there was likely a strong correlation between that attribute and tornadogenesis in supercell thunderstorms. Although the importance of an attribute alone cannot explain or discern the physical connection between attribute and tornadogenesis, it could highlight areas for further investigation. Table 1 shows the attribute importance scores for a set of thirty 100-tree 1000-sample SRRFs ranked by \( t \)-test value that determined if the mean importance values were significantly different from 0. Attributes with low standard deviations had similar impacts on performance no matter the situation while attributes with higher standard deviations were only important in select situations.

Among the top variables are multiple values that describe the surface thermodynamic and moisture fields surrounding the storms. Box-and-whisker plots in particular can help illuminate why these attributes received higher importance scores. The top and bottom of each box indicate the upper and lower quartiles of the distribution, the line across the middle indicates the median, the indentations show the 95% confidence interval generated from a bootstrap resampling of the median, and the whiskers indicate the full range of the distribution. Analysis of the distributions of \( u_9 \), \( u_y \), and \( u_e \) at the surface surrounding each storm, shown in Fig. 4, reveals that the medians and quartiles of the quantities are slightly higher for tornadic storms, and that there are some statistically significant differences between the medians. Higher values of \( \theta_e \), \( \theta_v \), and \( \theta' \) anomalies are correlated with increased instability and low level moisture. Because of the coarseness of the

---

### Table 1. The statistically significant important attributes with normalized attribute importance scores ranked by \( t \)-test value.

<table>
<thead>
<tr>
<th>Object</th>
<th>Attribute</th>
<th>Mean</th>
<th>SD</th>
<th>( t )-test value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storm surface</td>
<td>( \theta_v )</td>
<td>0.152</td>
<td>0.085</td>
<td>9.79</td>
</tr>
<tr>
<td>Environment</td>
<td>MLCAPE</td>
<td>0.155</td>
<td>0.09</td>
<td>9.43</td>
</tr>
<tr>
<td>Environment</td>
<td>MLCIN</td>
<td>0.246</td>
<td>0.145</td>
<td>9.29</td>
</tr>
<tr>
<td>Storm surface</td>
<td>( \theta_v )</td>
<td>0.131</td>
<td>0.083</td>
<td>8.64</td>
</tr>
<tr>
<td>Environment</td>
<td>SBCIN</td>
<td>0.203</td>
<td>0.132</td>
<td>8.42</td>
</tr>
<tr>
<td>Storm surface</td>
<td>( \theta_e )</td>
<td>0.142</td>
<td>0.119</td>
<td>6.54</td>
</tr>
<tr>
<td>Environment</td>
<td>0–3-km bulk wind shear</td>
<td>0.134</td>
<td>0.113</td>
<td>6.50</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Dewpoint’</td>
<td>0.119</td>
<td>0.112</td>
<td>5.82</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Relative humidity’</td>
<td>0.087</td>
<td>0.085</td>
<td>5.61</td>
</tr>
<tr>
<td>Environment</td>
<td>0–6-km wind shear</td>
<td>0.091</td>
<td>0.093</td>
<td>5.36</td>
</tr>
<tr>
<td>Environment</td>
<td>0–1-km EHI</td>
<td>0.092</td>
<td>0.096</td>
<td>5.25</td>
</tr>
<tr>
<td>Environment</td>
<td>0–3-km EHI</td>
<td>0.090</td>
<td>0.106</td>
<td>4.65</td>
</tr>
<tr>
<td>Environment</td>
<td>SBCAPE</td>
<td>0.103</td>
<td>0.122</td>
<td>4.62</td>
</tr>
<tr>
<td>Environment</td>
<td>STP (fixed layer)</td>
<td>0.078</td>
<td>0.095</td>
<td>4.50</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Wind</td>
<td>0.096</td>
<td>0.118</td>
<td>4.50</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Radar cloud height</td>
<td>0.092</td>
<td>0.122</td>
<td>4.13</td>
</tr>
<tr>
<td>Storm surface</td>
<td>WDSD</td>
<td>0.081</td>
<td>0.108</td>
<td>4.11</td>
</tr>
<tr>
<td>Environment</td>
<td>0–1-km SRH</td>
<td>0.089</td>
<td>0.121</td>
<td>3.93</td>
</tr>
<tr>
<td>Environment</td>
<td>0–1-km bulk wind shear</td>
<td>0.099</td>
<td>0.121</td>
<td>4.03</td>
</tr>
<tr>
<td>Boundary</td>
<td>Dewpoint’</td>
<td>0.063</td>
<td>0.089</td>
<td>3.88</td>
</tr>
</tbody>
</table>
mesonet observations, direct comparisons cannot be made between the \( u_9 \) analyses here and those observed in Markowski et al. (2002) and Shabbott and Markowski (2006). The overall environment is well sampled, but the changes within a storm are not resolved.

The most important near-storm environment variables (Table 1) included individual instability and shear parameters, which generally agrees with other studies of tornadic storm environments. MLCAPE and MLCIN were the two most important near-storm environment attributes with SBCAPE and SBCIN also appearing in the top 15. The distributions of MLCAPE and MLCIN are shown in Fig. 5. In the case of MLCAPE, although the medians are very similar, the upper range and upper quartiles are consistently higher for MLCAPE in tornadic storm environments. In contrast, both the medians and quartiles of the MLCIN distributions are significantly less negative for tornadic storms. The relatively large range of the MLCIN grid minimums (largest CIN magnitudes) stems from a combination of storms forming along a boundary where more negative MLCIN occurs and the MLCIN eroding between NARR time steps at a particular location with the values from the previous time step being retained. This temporal resolution issue could potentially bias other attributes derived from NARR data, especially in areas near sharp, moving boundaries.

Bulk wind shear parameters and related composite indices also showed importance, but their skill may have been hampered by the vertical resolution of the NARR. Of the shear and helicity parameters evaluated, 0–3-km bulk wind shear had the highest importance. Relative to 0–1-km and 0–6-km bulk wind shear (Fig. 6), 0–3-km bulk wind shear had greater differences within the distributions, although none of the distributions were significantly different. As expected, the storm-relative helicity distributions for tornadic and nontornadic near-storm environments also showed few differences (not shown). These similarities led to similar distributions for composite parameters, including STP and EHI (Fig. 7). The lack of significant differences in the bulk wind shear distribution likely originates in the 25-hPa spacing of the vertical levels in the NARR data. With fewer data points to draw from, complexities in the vertical wind profile were smoothed over, resulting in smaller calculated bulk wind shear and SRH.

The other important attributes addressed aspects of the storm environment not covered by the other
categories (Table 1). The surface wind field could provide a faster updating estimate of the bulk wind shear than available from the NARR. The wind direction standard deviation (WDSD) was higher on average for tornadic supercells than nontornadic supercells, which may be a weak signal of changes in winds at shorter intervals than resolvable by the mesonet. Radar cloud height is also higher for tornadic supercells than nontornadic supercells, but that may be an artifact of the radar beam attaining a higher altitude with increasing distance from the radar site.

d. Single attribute performance

The SRRF so far has shown positive skill in relation to random and climatological forecasts, but it requires an increase in the complexity of the data framework and the interpreting model. Most current methods of diagnosing the tornadic supercell potential focus on how a particular environmental parameter, such as CAPE, helicity, or some multiplicative combination of the two (EHI), varies. Calculating the AUC for each attribute statistic allowed for the SRRFs to be compared against a simple approach and allowed for attribute importance results to be compared against a method that looks at an attribute’s raw performance as opposed to its performance and strength of interactions. The top 20 spatial and temporal means of the attributes ranked by their AUC are shown in Table 2.

The rankings in Table 2 show that single attribute variation has worse skill than most SRRFs (Fig. 3a), and that they provide similar results to attribute importance. SBCIN and MLCIN appear at the top of the AUC rankings; $\theta_r^c$, $\theta_v^c$, and $\theta^a$ all appear in the top 10, and all of them have among the highest PODs but also fairly high POFDs and FARs. As with the attribute importance rankings, 0–3-km bulk wind shear had the highest AUC of the bulk wind shear parameters. The 0–3-km EHI had the highest POD of the near-storm environment attributes but also had high POFD and FAR. Storm speed has the highest POD of any attribute but also has the highest POFD and FAR, meaning that although almost all tornadic storms travel over 10 m s$^{-1}$, many of the nontornadic storms do so as well. The relatively low AUCs for the full range of attributes indicate that, individually, none of the parameters
associated with a tornadic environment can be trusted to discriminate between tornadic and nontornadic supercells. The low skill is not surprising given the similarity of the severe weather parameter distributions for weakly tornadic and nontornadic supercells in various proximity sounding studies (e.g., Rasmussen and Blanchard 1998; Thompson et al. 2003) and the large proportion of weakly tornadic supercells in this dataset.

e. Individual event analysis

A set of SRRFs trained on data from 1994 to 2002 was used to evaluate cases in 2003 to test the SRRF’s predictive abilities. The year 2003 was chosen because it had multiple events with large numbers of supercells and included both strong and weak tornadoes. Storms were then grouped by day and evaluated for spatial patterns in the probabilities of the predictions. A 100-tree, 500-sample SRRF was used because it maximizes performance while ensuring more diverse trees than those trained at larger sample sizes.

Analyzing how the probabilistic forecasts from the SRRFs verify illustrates the strengths and weaknesses of this approach. Figure 8 shows an attributes diagram and ROC curve of the performance of the best SRRF on the 2003 cases. The attributes diagram shows good reliability and resolution at low and high probabilities with a slight overforecasting bias near climatology. The magnitude of the biases may be affected by the small sample size in that range. The wide range of probabilities and BSS of 0.33 show that the SRRF is making good predictions and that the calibration function is ensuring that the SRRF produces probabilities spanning from 0 to 1. The ROC curve shows useful skill with an AUC of 0.76. The low POFD when the POD is between 0 and 0.5 indicates that half of the tornadic storms are easily distinguishable from the nontornadic storms, while the rest are more similar to the nontornadic storms since the POD and POFD increase linearly at that point.

Table 2. The top 20 attributes ranked by AUC.

<table>
<thead>
<tr>
<th>Object</th>
<th>Attribute</th>
<th>AUC</th>
<th>PSS</th>
<th>Threshold</th>
<th>POD</th>
<th>POFD</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment</td>
<td>SBCIN</td>
<td>0.602</td>
<td>0.16</td>
<td>−45.60</td>
<td>0.51</td>
<td>0.35</td>
<td>0.69</td>
</tr>
<tr>
<td>Environment</td>
<td>MLCIN</td>
<td>0.595</td>
<td>0.17</td>
<td>−13.79</td>
<td>0.33</td>
<td>0.16</td>
<td>0.62</td>
</tr>
<tr>
<td>Storm surface</td>
<td>( \theta_e )</td>
<td>0.579</td>
<td>0.18</td>
<td>2.91</td>
<td>0.68</td>
<td>0.50</td>
<td>0.71</td>
</tr>
<tr>
<td>Boundary</td>
<td>Pressure</td>
<td>0.559</td>
<td>0.14</td>
<td>140.88</td>
<td>0.63</td>
<td>0.49</td>
<td>0.71</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Wind speed SD</td>
<td>0.549</td>
<td>0.10</td>
<td>1.30</td>
<td>0.40</td>
<td>0.30</td>
<td>0.72</td>
</tr>
<tr>
<td>Storm surface</td>
<td>( \theta_e )</td>
<td>0.548</td>
<td>0.14</td>
<td>−1.81</td>
<td>0.77</td>
<td>0.63</td>
<td>0.73</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Direction change</td>
<td>0.544</td>
<td>0.12</td>
<td>1.01</td>
<td>0.47</td>
<td>0.35</td>
<td>0.71</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Wind</td>
<td>0.543</td>
<td>0.10</td>
<td>−3.85</td>
<td>0.57</td>
<td>0.48</td>
<td>0.73</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Radar cloud height</td>
<td>0.541</td>
<td>0.09</td>
<td>12074.73</td>
<td>0.47</td>
<td>0.38</td>
<td>0.73</td>
</tr>
<tr>
<td>Storm surface</td>
<td>( \theta_e )</td>
<td>0.540</td>
<td>0.13</td>
<td>−8.80</td>
<td>0.79</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Dewpoint</td>
<td>0.537</td>
<td>0.11</td>
<td>6.61</td>
<td>0.67</td>
<td>0.56</td>
<td>0.73</td>
</tr>
<tr>
<td>Storm surface</td>
<td>LCL height</td>
<td>0.536</td>
<td>0.09</td>
<td>1100.91</td>
<td>0.70</td>
<td>0.61</td>
<td>0.74</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Direction</td>
<td>0.535</td>
<td>0.09</td>
<td>304.00</td>
<td>0.52</td>
<td>0.43</td>
<td>0.73</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Temperature</td>
<td>0.529</td>
<td>0.08</td>
<td>0.26</td>
<td>0.72</td>
<td>0.64</td>
<td>0.75</td>
</tr>
<tr>
<td>Environment</td>
<td>0–3-km bulk wind shear</td>
<td>0.529</td>
<td>0.08</td>
<td>19.10</td>
<td>0.39</td>
<td>0.31</td>
<td>0.72</td>
</tr>
<tr>
<td>Environment</td>
<td>0–3-km EHI</td>
<td>0.528</td>
<td>0.11</td>
<td>2.89</td>
<td>0.53</td>
<td>0.41</td>
<td>0.72</td>
</tr>
<tr>
<td>Environment</td>
<td>SBCAPE</td>
<td>0.528</td>
<td>0.07</td>
<td>4098.33</td>
<td>0.29</td>
<td>0.22</td>
<td>0.71</td>
</tr>
<tr>
<td>Environment</td>
<td>MLCAPE</td>
<td>0.520</td>
<td>0.09</td>
<td>1949.19</td>
<td>0.28</td>
<td>0.20</td>
<td>0.70</td>
</tr>
<tr>
<td>Environment</td>
<td>0–1-km EHI</td>
<td>0.514</td>
<td>0.09</td>
<td>2.47</td>
<td>0.37</td>
<td>0.28</td>
<td>0.72</td>
</tr>
<tr>
<td>Storm surface</td>
<td>Speed</td>
<td>0.505</td>
<td>0.05</td>
<td>10.63</td>
<td>0.93</td>
<td>0.88</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Figure 9 shows the paths of storms and the SRRF predictions for each storm on 19 April 2003. The wide range of probabilities even within small geographic areas indicated that the SRRFs are incorporating the variations in the surface and reanalysis data into their analysis. Tornadic storms are generally receiving much higher probabilities than nontornadic storms. One notable exception is in east-central Oklahoma, where two tornadic storms are nearly adjacent to each other, but one has a probability of 30% while the other has a probability of 60%. Although the two storms had similar ML CAPE ranges, the 60% storm’s surrounding 0–1-km SRH had a maximum of 129 m² s⁻² while the 30% storm had a maximum of 62 m² s⁻². This disparity likely contributed to the vast probability difference over such a small area.

The predictions for the 8 May 2003 tornadoes can be seen in Fig. 10. The primary features are a sharp gradient in MLCAPE through central Oklahoma associated with a dryline, and moderate to very strong bulk wind shear from central to eastern Oklahoma. The storm in central Oklahoma became supercellular at 2126 UTC and
produced an F4 tornado near Moore, Oklahoma, at 2210 UTC. Although this storm was notably strong, the SRRF only provided a 15% chance of tornadogenesis. The likely reason behind this low probability is the wide range of values in the severe weather parameters over the area surrounding the storm. Within the 96 km × 96 km box following the storm from initiation to tornadogenesis, the MLCAPE ranged from 640 to 4510 J kg⁻¹, the MLCIN ranged from −142 to −0.5, and the 0–1-km SRH ranged from 22 to 44 m² s⁻². The wide range of CAPE and CIN values combined with low SRH most likely led to the low probabilities from the SRRF. The 0000 UTC Norman, Oklahoma, sounding did record MLCAPE of 5005 J kg⁻¹ and 0–1-km SRH of 317 m² s⁻², but it was taken 2 h after the tornado formed. The 0000 UTC NARR output also underestimates the SRH at that time in Norman, likely because of insufficient resolution to capture the strong thermodynamic gradients along the dryline. The clearest evidence for negative effects of the MLCAPE gradient on the predictions is the nontornadic supercell in the middle of the maximum MLCAPE area that received a 60% probability from the SRRF.

5. Summary

This study examined a radar-derived dataset of supercell thunderstorms in Oklahoma from 1994 to 2003 and collected associated Oklahoma Mesonet and North American Regional Reanalysis data for each storm. Storm reports were used to determine whether each storm produced a tornado. The surface and reanalysis data were organized into a spatiotemporal relational framework consisting of storm surface, near-storm environment, and boundary objects linked by spatial relationships. Each object contained attributes that could vary spatially and temporally as the storms moved and evolved. The spatiotemporal relational random forest, an ensemble of spatiotemporal relational probability trees, was trained with a variety of parameter variations to find the range of optimal predictive conditions. Individual attributes were evaluated using both the attribute importance method derived from the SRRFs and by calculating the area under the relative operating characteristic curve for the full distribution of each attribute. SRRFs were also evaluated on individual cases to show how their probabilistic forecasts vary in different situations.

Skill evaluation reveals that SRRFs can produce skilled probabilistic forecasts with calibration and that larger forests and question sample sizes produce more skilled forests. Single trees perform significantly worse because they can only address a limited range of possibilities. Both attribute importance measures and single attribute evaluation focused heavily on the surface thermodynamic and moisture fields along with CAPE, CIN, and bulk wind shear. The boundary attributes did not significantly impact the SRRF predictions. However, many of the near-storm environment parameters are sensitive to the presence of boundaries, so boundaries can have an indirect effect on the predictions.

FIG. 8. (a) Attributes diagram and (b) ROC curve showing the overall performance of the best SRRF run on the 2003 cases. Areas in the diagrams shaded gray contain positive skill. No skill in the attributes diagram refers to when the difference of reliability and resolution has a BSS of 0. No skill in the ROC curve refers to an AUC of 0.5.
On individual events, the SRRF showed that its probabilistic forecasts could handle complexities in the environment but were limited by the constraints of the given data sources. Tornadic storms generally received higher probabilities compared to nontornadic storms, and storms in less favorable environments received lower probabilities compared to those that were in more favorable environments. For a given environment, the SRRF performs fairly well in assigning relatively higher probabilities to the tornadic storms, so the SRRF could have use in prioritizing what storms are of most interest at a given time.

These findings do have some caveats. Owing to the limited area and 10-yr time period, the sample size is too small to get the full range of tornadic supercell behavior, especially for strongly tornadic storms. The coarseness of mesonet station spacing and NARR spatiotemporal limitations can result in the observed environment not matching what was occurring at any specific place or time but is as close as can be determined from the data available. The magnitudes of gradients are underestimated because of the coarse resolution, which has sometimes resulted in lower CAPE and bulk wind shear than observed in soundings. The storm reports database has uncertainties in time and location of reports as well as population and awareness biases over time (Brooks et al. 2003a), so some storms may have been misclassified. As evidenced by the scores and diagrams shown in Figs. 3 and 8, the SRRF has handled these data limitations and has still provided useful skill.

This study has shown the value of the SRRF in analyzing a large dataset of tornadic supercells and the abilities and limitations of operationally available datasets. However, the technique can be applied to any complex spatiotemporal dataset. Current work is being conducted to apply the SRRFs to large numbers of high-resolution supercell simulations to analyze physical processes leading to tornadogenesis that cannot be resolved by current observing networks. Although single-storm simulations can be analyzed with more traditional techniques, generalizing the variations from hundreds of models is made easier by the use of a data-mining technique like the SRRF. When computing power reaches the point where storm-scale models can be run in real time, the SRRF or another spatiotemporal relational data-mining algorithm could be used as a postprocessing tool to derive tornadic probabilities from the model ensemble based on how the simulated storm is forecast to evolve. Until that advance in computing power is reached, spatiotemporal relational machine-learning algorithms could provide added value from

---

![Figure 9: Tracks of storms on 19 Apr 2003, with predicted tornadic probabilities superimposed on the tracks. Solid black and gray lines are tornadic storm tracks, and dashed gray lines are nontornadic storm tracks. Black lines correspond to probabilities ≥50%, and gray lines indicate <50%. The black or gray triangle markers indicate where the strongest tornado occurred for each tornadic storm. Solid contours are MLCAPE (J kg⁻¹), and dash–dot contours are 0–3-km bulk wind shear (kt; 1 kt ≈ 0.5 m s⁻¹), analyzed at 0000 UTC 20 Apr.](image-url)
current observational systems and could be used as an additional tool in the forecaster’s toolbox.

Acknowledgments. Special thanks go to Nathaniel Troutman for maintaining and supporting the SRRF codebase. The anonymous reviewers provided much helpful feedback that strengthened the quality of this paper. Some of the computing for this project was performed at the University of Oklahoma Supercomputing Center for Education & Research (OSCER). This study was funded by the National Science Foundation under Grant NSF/IIS/0746816 and related REU Supplements NSF/IIS/0840956, NSF/IIS/0938138, and NSF/IIS/1036023 as well as the NSF Graduate Research Fellowship under Grant 2011099434. The Oklahoma Mesonet is funded by the taxpayers of Oklahoma through the Oklahoma State Regents for Higher Education and the Oklahoma Department of Public Safety. NARR data were retrieved from the NOAA National Climatic Data Center.

REFERENCES


