Toward Understanding Tornado Formation Through Spatiotemporal Data Mining

Amy McGovern and Derek H. Rosendahl and Rodger A. Brown

1 Motivation

Tornadoes, which are one of the most feared natural phenomena, present a significant challenge to forecasters who strive to provide adequate warnings of the imminent danger. Forecasters recognize the general environmental conditions within which a tornadic thunderstorm, called a supercell thunderstorm, will form. They also recognize a supercell thunderstorm with its rotating updraft, or mesocyclone, when it appears on radar. However, only a minority of supercell storms produce tornadoes. There are no obvious clues within any of the routinely observed data to indicate which supercell storms are going to produce tornadoes and which ones are not. So to be on the safe side, forecasters issue a tornado warning whenever they detect on radar a supercell thunderstorm with a strengthening mesocyclone, which is the parent circulation within which tornadoes form. This approach results in the warning being issued an average 10 to 15 minutes before the appearance of a tornado, but a tornado appears only 20 to 30% of the time that a warning is issued, which results in a large percentage of false alarms (e.g., Simmons and Sutter, 2011).

Surveys conducted by the National Weather Service following devastating U.S. tornadoes reveal that these false alarms are one of the factors contributing to desensitization on the part of the public concerning the need to adhere to warnings (e.g., NWS, 2009, 2011). Having heard many warnings when a tornado did not form, members of the public tend to ignore the warning and only take shelter if they see

Amy McGovern
School of Computer Science, University of Oklahoma, 110 W Boyd St, Norman OK 73019, e-mail: amcgovern@ou.edu

Derek H. Rosendahl
School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd. Suite 5900, Norman, OK 73072, e-mail: drose@ou.edu

Rodger A. Brown
National Severe Storms Laboratory/National Oceanic and Atmospheric Administration, 120 David L Boren Blvd, Norman, OK 73072, e-mail: Rodger.Brown@noaa.gov

a tornado approaching, with the result that some of them do not make it to a safe place in time. Therefore, there is a need to explore ways in which to identify unique factors within storms that lead to tornado formation.

One way to help understand the evolutionary characteristics of supercell thunderstorms, and especially those that produce tornadoes, is to conduct numerical modeling studies of the storms (e.g., Wicker and Wilhelmson, 1995; Noda and Niino, 2005; Xue et al, 2007). These types of studies typically investigate a single storm that develops within a given environment. A more informative approach is to conduct a number of fine-resolution (i.e., horizontal grid spacing less than 100 m) numerical modeling studies of supercell storms under a variety of environmental conditions. Data mining techniques then can be applied to the modeling results in order to discover the differences between supercell storms that produce tornadoes and those that do not.

In this chapter, we discuss the development of novel spatiotemporal data mining techniques that were initially applied to numerical models having coarse 500 meter horizontal grid spacing that did not resolve tornadoes but did resolve the parent mesocyclones. Note that much of this chapter is derived from the following papers (Rosendahl, 2008; Supinie et al, 2009; McGovern et al, 2010, 2011a,b, under review). Although we are in the process of generating 75 to 100 numerically-modeled supercell thunderstorms using tornado-resolving 75 meter horizontal grid spacing, we focus our discussion in this chapter on the 500 meter resolution storms. We anticipate that the mining of our new higher-resolution data set will reveal more clues about the processes that lead to tornado formation within some supercell thunderstorms but not within others. This approach is discussed at the end of the chapter.

2 Natural Hazard Domain: Severe Storm Simulations

Rosendahl (2008) created a set of 261 simulations of supercell thunderstorms, which are the most severe type of thunderstorm and which generate the most violent tornadoes. Each simulation was generated using the Advanced Regional Prediction System (ARPS, Xue et al, 2000, 2001, 2003). The full details on the parameters chosen to create the storms are described in Rosendahl (2008). Each simulation is run for 3 hours of storm time. The simulation saves the state of all relevant meteorological variables every 30 seconds of storm time for every grid point in the domain. With a resolution of 500 m horizontally and a domain size of 100 kilometers by 100 kilometers, a stretched vertical resolution focusing on the lower altitudes (with 50 voxels vertically) , the simulations must save over 100 different variables every 30 seconds for each of 2 million grid squares. In total, each simulation produces over 21 GB of data, which requires us to intelligently process and mine this data.

Although each simulation generates a full gridded field of meteorological variables, the variables near a storm cell will provide the most information. We identify and track storm cells using a modified form of the Storm Cell Identification and Tracking algorithm (Johnson et al, 1998; McGovern et al, 2007) where we track the

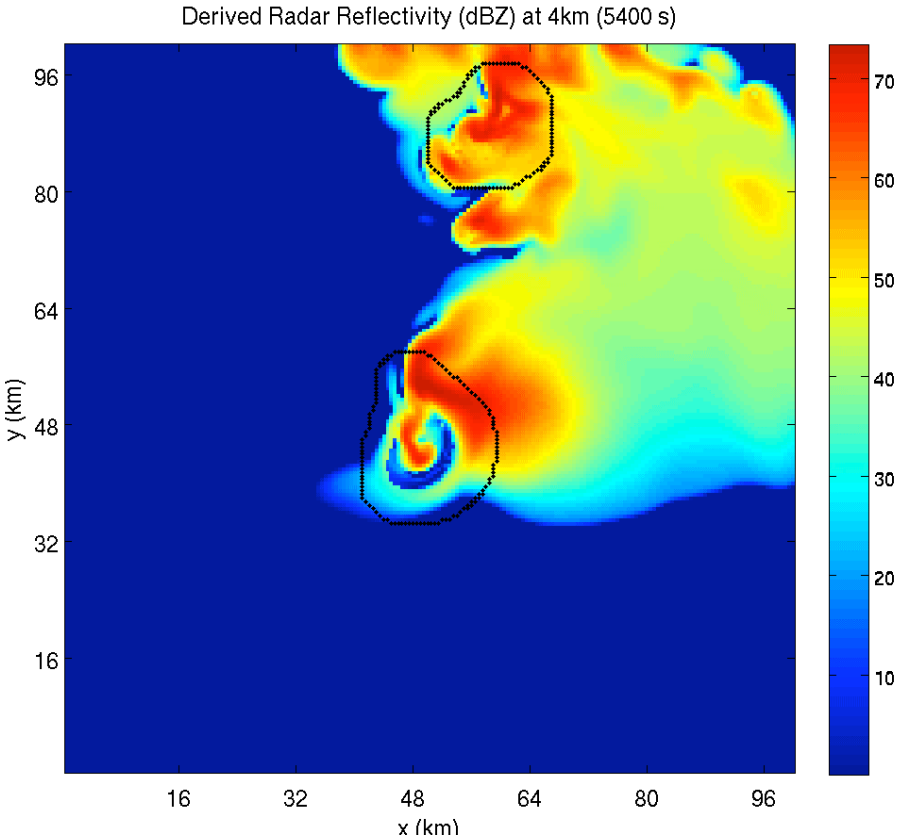


Fig. 1 Reflectivity of an example numerical storm simulation 90 minutes into the storm's lifetime. The scale on the right shows the intensity of the reflectivity in dBZ. Higher reflectivity regions indicate areas where the storm is producing intense precipitation. The black outlines highlight individual storm cells which are used to extract the storm metadata.

cells based on their dominant updraft region (localized area with rising air) because it is the defining feature of a thunderstorm. Figure 1 shows an example of simulated radar reflectivity 90 minutes into a simulation. Reflectivity measures the intensity of the precipitation within the storm which means that regions with more intense rain, snow, ice, or hail have higher reflectivity values. The black outlines in Figure 1 show the two storm regions that are being tracked during that period. Because weak short-lived storms are not of interest in this study, we only track cells that last for at least 30 minutes. Each simulation typically produces 3-4 such cells. The 261 storm simulations generated 1168 separate storm cells that each lasted at least 30 minutes.

3 Spatiotemporal Data Mining

Weather phenomena vary as a function of both time and space. It is difficult to ignore one aspect in favor of the other so our goal has been to develop spatiotemporal data mining algorithms. This chapter first reviews our time series approach and then presents several spatiotemporal algorithms that we have developed. Each storm defines a four dimensional region of interest for data mining. The first three dimensions are the spatial dimensions and the fourth dimension is time. Given the sheer size of the data, all of our approaches reduce the data by extracting high-level metadata and then mining the metadata.

3.1 Multi-variate time series approach

In the first approach to improving our understanding about tornado formation, we seek to identify a series of rules that highlight how environmental characteristics must change to favor the development of tornadoes. We do not know in advance what characteristics are most important so the learning and mining algorithms must identify the most salient variables and discover the temporal motifs most predictive of tornadic rotation. Informally, the goal of our approach is to identify the most relevant dimensions of a multi-dimensional time series, grow a set of predictive rules from motifs discovered in each of those dimensions, and use these to improve our understanding of the data and for prediction. We briefly review definitions necessary for our time series mining algorithm.

3.1.1 Definitions

Definition 1. A time series $T = \langle t_1, t_2, \ldots, t_{n-1}, t_n \rangle$ is an ordered sequence of real-valued observations taken at discrete times: $1, 2, \ldots, n-1, n$. A d-dimensional time series $T^d = \langle T_1, T_2, \ldots, T_d \rangle$ is a set of time series all associated with a single event and correlated in time.

This is the standard definition of time series (e.g., Mueen et al, 2009). Our examples are all temporally ordered sequences but other orderings are possible. Rather than examining only a single attribute as it varies in time, we assume that the event can be measured with a variety of attributes (d in this definition), each of which is measured on the same discrete time interval. These measurements are not required to be independent of one another. Our severe weather simulations have $d = 100$ but the independent dimensionality of the data is much less (approximately $d = 40$).

Definition 2. A labeled multi-dimensional time series is a tuple $E = \{T^d, l\}$ where T^d is a d-dimensional time series and $l \in \mathscr{L}$ where $\mathscr{L}$ is a discrete set of labels (and it is not required to be binary).

Because each of our d-dimensional time series is associated with an event, each example E_i is labeled. In the case of our severe weather data, $\mathscr{L}$ is either binary (positive/negative) or takes on one of three possible values: positive, negative, or intermediate. The approach does not restrict the possible numbers of labels but it does require that the cardinality of $\mathscr{L}$ be finite.

Our data set, $D = \langle E_1, E_2, \ldots, E_n \rangle$, consists of a set of labeled multi-dimensional time series, each of which can last for a variable amount of time but each of which is assumed to have the same dimensionality. That is, all attributes that measure an event are assumed to be present in each labeled example.

Definition 3. A single dimensional time series motif $M_j = \langle t_i, t_{i+1}, \ldots, t_{i+m} \rangle$ consists of a temporally ordered subsequence of a time series where $1 \leq i \leq n$ and $0 < m \leq n$. This motif is of length m and is on dimension j where $1 \leq j \leq d$.

As stated above, our goal is to identify multi-dimensional times series motifs that can be used for prediction. As such, we build on the previous definition.

Definition 4. A multi-dimensional time series motif $P = \langle M_{i_1}, M_{i_2}, \ldots M_{i_m} \rangle$ is a temporally ordered set of single-dimensional time series motifs (see Definition 3). The temporal ordering specifies that the initiation of each single-dimensional time series M_{i_j} must begin after or simultaneously with the previous single-dimensional time series in the set $M_{i_{j-1}}$.

A multi-dimensional time series motif does not specify how many dimensions of the overall available dimensions must be used and it can even repeat dimensions, given that each is temporally ordered. For example, the first motif may be on dimension 1, the second on dimension 3, and the third on dimension 1 again. The temporal ordering is not strict as it requires that M_j begins after M_{j-1} but simultaneous initiations are also acceptable. M_j cannot begin before M_{j-1}. This definition differs slightly from the definition in Mueen et al (2009), where there is a requirement for a temporal gap in between the different subsequences. We were interested in rules that could identify two features both firing at once and did not impose the temporal gap.

3.1.2 Algorithm

A general outline of our approach for identifying multi-dimensional time series motifs is given in Algorithm 1 and we describe each step in detail below. The general idea is to search for the critical dimensions of the data by identifying single dimensional motifs first, narrow down the set of possible single dimensional motifs using user specified minimum performance metrics and then grow the motifs across dimensions using the single dimensional motifs as building blocks for larger motifs.

The rules are built and scored using three standard performance measures. The probability of detection (POD) is the number of times that an event was correctly predicted divided by the total number of observed events. POD ranges from 0 to 1

Algorithm 1: Grow multi-dimensional time series motifs/rules

Input: D: training data, SAX parameters (alphabet size, word size, averaging interval), minimum POD, maximum FAR
Output: A list of rules sorted by CSI
foreach *dimension d* **do**
  Discretize E_i^d for all examples *i* using SAX
  Build trie with pointers to the start and to the end of each word in the sliding window
end
foreach *dimension d* **do**
  Identify all single dimensional words with minimum POD and maximum FAR
  Recursively grow longer rules within dimension *d*
end
for *all rules that meet minimum POD and maximum FAR criteria* **do**
  Grow rules across dimensions
end
return *list of rules sorted by CSI score*

with 1 representing a perfect score. The false alarm ratio (FAR) is the number of times that an event was incorrectly predicted to occur divided by the total number of events predicted to occur. FAR ranges from 0 to 1 with 0 representing a perfect score. The critical success index (CSI, Donaldson Jr et al (1975); Schaefer (1990)) evaluates success as a function of only the events that are predicted to be positive and the ones that were actually positive. Thus, CSI ignores the true negatives and provides an index that incorporates both POD and FAR into one measure. It ranges from 0 to 1 with 1 being a perfect score (perfect POD and perfect FAR). This measure is particularly important for assessing rare events such as tornadoes where POD and FAR alone are inadequate measures of predictability. For instance, a perfect score can be attained for POD by simply predicting an event (e.g., tornado) to occur every time or for FAR by never predicting an event to occur. CSI, however, combines both of these and therefore gives a more robust performance measure.

Because a brute force approach to searching multi-dimensional data grows exponentially in the number of dimensions and the size of the data, it quickly becomes intractable. Some researchers approach this issue using a random type of search Minnen et al (2007); Vahdatpour et al (2009) but our pruning and discretization enables us to search the full space efficiently. Additionally, searching for motifs in real-valued data is difficult, as has been noted by many researchers (for example, Das et al, 1998; Chiu et al, 2003; Mueen et al, 2009). We chose to address this latter problem using the SAX discretization technique (Lin et al, 2003) and we address the computational aspects using a combination of approximate search and intelligent data structures such as the trie described in Keogh et al (2005).

The first step of our approach is to discretize each of the dimensions of data using SAX, which is a standard time series discretization technique (e.g., Lin et al, 2003; Keogh et al, 2005; Lin et al, 2007; Minnen et al, 2007; Shieh and Keogh, 2009; Vahdatpour et al, 2009). To do this, we discretize all examples at once for each dimension. This ensures that the discretizations can be easily compared and that *a* in one time series has a similar meaning to *a* in another example. Given the

number of dimensions and examples, multiple passes through the data would be computationally expensive. To address this, we make use of the trie data structure as discussed in Keogh et al (2005). Since each leaf in the trie has information about exactly which time series that word occurs in, the trie also stores the POD and FAR measures for use in the mining. We have demonstrated in previous work that the results are not sensitive to the choices of parameters to SAX (McGovern et al, 2011b).

Once the tries are built, the data mining algorithm makes use of them to efficiently narrow down the search for more complicated motifs. To do this, we look through each dimension of the data and narrow down the set of basic SAX words using user specified thresholds of the POD and FAR performance measures. By specifying a minimum POD and a maximum FAR, we limit the number of elementary motifs identified. These numbers can come from a user's experience in the domain. This significantly improves the running time of the search since all possible combinations of small motifs are used to grow the larger motfis. Since the search proceeds from general motifs (e.g. short ones) to specific motifs (longer multi-dimensional motifs), the performance of the motifs will only improve POD and FAR. This is similar to the pruning search discussed in Webb (1995), Oates and Cohen (1996), and McGovern and Jensen (2008) where the type of search and evaluation measures can be combined to enable admissible pruning. Thus, we specify minimum levels of performance that would be acceptable and expect that the final numbers will be significantly better with the more specific rules.

Once the basic words are identified, we grow the motifs recursively using the basic words that pass the POD/FAR thresholds. The recursive growth of the motifs within a dimension works by doubling each motif, similar to the method employed by SPADE (Zaki, 2001). Each motif is doubled while maintaining the minimum POD and maximum FAR requirements. For example, $[a,b]$ can be doubled to $[a,b,a,b]$ and it can also be combined with other valid words such as $[a,b,b,a]$. When a doubling fails, the motif is grown linearly by adding words until it is at its maximum length. Since motifs are grown in chunks of word size, not all possible motifs can be detected. For example, the motif $[a,b,a]$ is not discoverable in the example because the minimum word size is 2. However, in all of our experiments, we keep the word size small to minimize this issue.

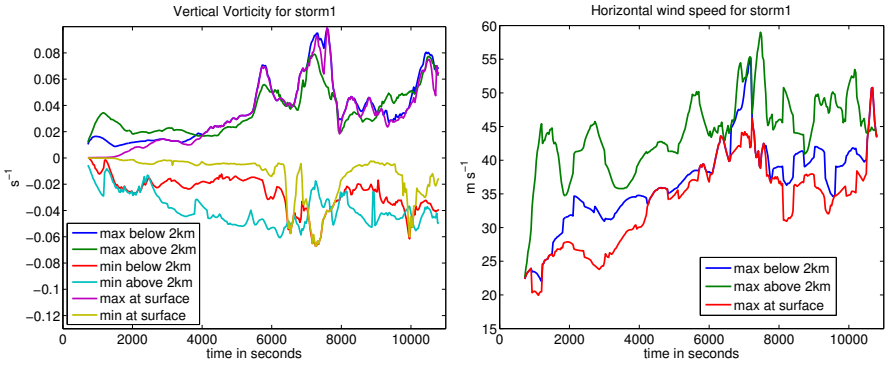
The motif growing within a single dimension is repeated for all dimensions before searching across the dimensions. Although this growing sounds computationally expensive, only one pass through the original data is required. Using the trie data structure, all further motifs can be grown by examining the starting and ending times of each individual motif and ensuring that they follow one another temporally (e.g. they satisfy definition 4 above). The POD and FAR measures continue to be directly computed from the trie by intersecting the positive and negative graphs computed for each individual piece of the motif.

The last step of the algorithm is to repeat the search by combining the different words across the dimensions of the data. Once all of the motifs have been identified for each dimension, an exhaustive search of temporal orderings across dimensions is performed. Although doing an exhaustive search sounds infeasible, it is possi-

ble because of the admissible pruning afforded by the user's minimum POD and maximum FAR measures. In addition, the O(1) access time of the trie facilitates the overall approach. Without the pruning, this approach would be untenable but the pruning significantly improves the running time by enabling the search to ignore large portions of the search space while guaranteeing that the rules that could be identified in that space will never meet the user's specified performance measures. In addition, the trie again can be used to quickly compute the POD/FAR measures across dimensions by continuing to intersect the positive and negative series observed by each piece of the motif.

3.1.3 Empirical results

To create the metadata for the time series data mining, we extract maximum and minimum values of relevant meteorological quantities within each of the simulated storm cells. We measure the maximum and minimum value for each variable from the surface to 2 km in height and then from 2 km to 8 km. For some variables, we also store the maximum and minimum values at the surface. This allows us to identify whether a maximum or minimum value is associated with a surface, low, or mid to upper altitude feature. This yields a 100 dimensional time series for each storm. The full set of quantities is defined in Rosendahl (2008).



(a) Example vertical vorticity data (b) Example horizontal wind speed data

Fig. 2 Meteorological quantities extracted from an example storm. Shown are the maximum and minimum quantities for vertical vorticity (left) and the maximum values for horizontal wind speed (right).

We extract the maximum and minimum values every 30 seconds for the entire three hours of simulation. Figure 2 shows an example of several time series extracted for two of the meteorological quantities. The left panel shows the evolution of the vertical vorticity (an instantaneous measure of spin about a vertical axis) at the surface, low altitudes, and mid- to upper-altitudes for the center storm shown

in Figure 1. Positive (maximum) and negative (minimum) values in the left panel correspond to counterclockwise and clockwise rotation respectively. The right panel shows the maximum horizontal wind speed values.

Given the sheer number of storm cells, we developed an automated labeling approach based on the key characteristics of tornadic storms. Because the horizontal grid spacing of the simulations is too coarse to detect rotation on the scale of a tornado and creating higher resolution simulations requires exponentially more computational time and space, we labeled each storm as to whether it produced strong low-altitude rotation ("positive"), produced either no or very weak rotation ("negative") or was in between these two categories ("intermediate" or "maybe"). Strong low-altitude rotation was defined as a storm where there was a decrease in the surface pressure perturbation of at least -900 Pa in 1000 seconds and either an increase in the horizontal wind speed at the surface of at least 5 m s^{-1} within 750 seconds or an increase in the absolute value of the vertical vorticity of at least 0.03 s^{-1} within 500 seconds. These features had to overlap within a 600 second window to ensure that they were correlated. Storms where the pressure drop fell within the range of -900 Pa to -300 Pa and met the vertical vorticity and wind speed criteria or that had a pressure drop but no corresponding increase in vertical vorticity or wind speed were labeled as intermediate storms. The remainder of the storms were labeled as negative storms. This yielded 58 positive storms, 373 intermediate storms, and 737 negatives.

Given the labeled data, we further processed it in two ways. First, if we were to feed all of the time series information to the data mining algorithm, it would identify the approach that we took to label the data. To avoid rules that simply state that pressure perturbations or vertical vorticity are critical, we remove from consideration each of the features used to label the data from consideration. Further, to ensure that identified precursors were associated with the developing strong low-altitude rotation in positive storms, we saved data for 30 minutes immediately prior to the beginning of the corresponding pressure drop, which was defined using a Gaussian derivative filter on the pressure perturbation time series. For non-positive cases, we randomly sampled 30 minutes to avoid obvious labeling based on the length of the time series given to the data mining algorithm.

Figure 3 shows an example rule identified by the time series data mining algorithm. We explored a wide variety of parameter variations and show only a single example rule, due to space considerations. This rule identifies four salient environmental characteristics associated with rotation in our simulations and also shows the behaviors of each measurement. In general, the rule indicates that strong low-altitude rotation has a greater probability of developing when the four identified meteorological quantities reach relatively extreme values in a short time span. Hundreds of such rules were identified by the data mining algorithm and were ranked according to their performance measure scores.

This information can be highly useful to those in the meteorological community because each rule is simply a sequence of events taking place within a storm and therefore identifying the most important of these sequences may offer insight into how tornadoes form and how one can better predict their occurrence. Also, the

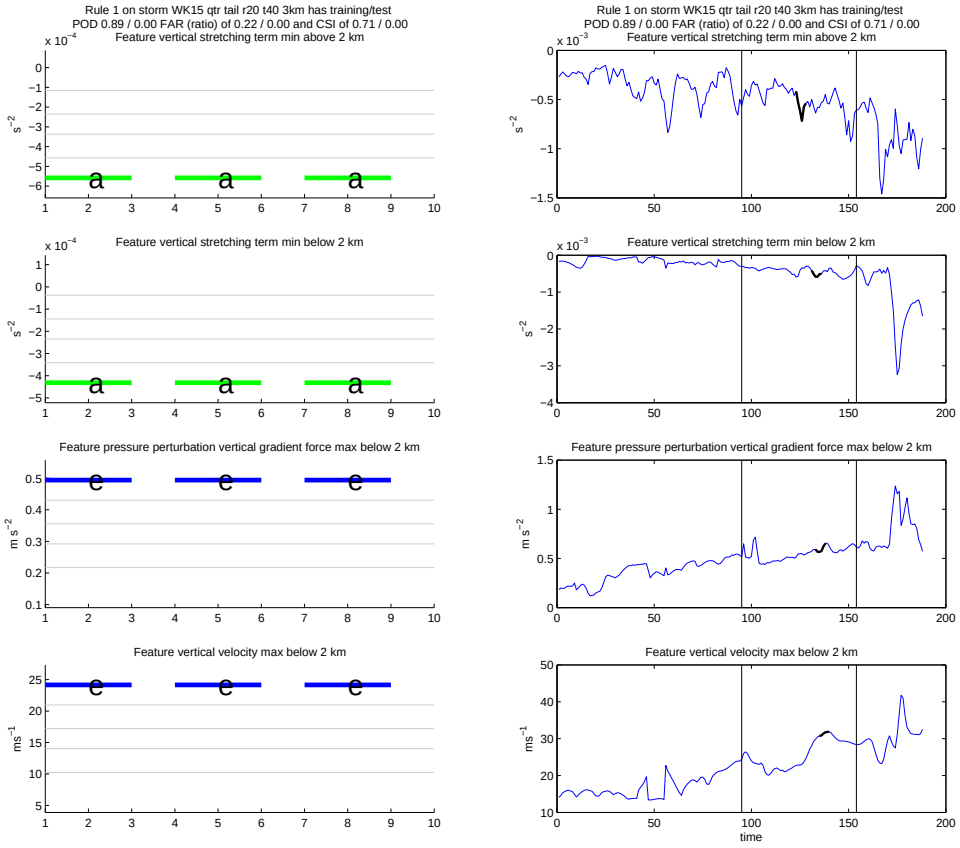


Fig. 3 Example rule identified by time series mining algorithm. The left column identifies 3-letter words comprising the rule in sequential order (baroclinic generation term (vertical) min below 2 km, vertical stretching term max below 2 km, pressure perturbation vertical gradient force max above 2 km and vertical stretching term min below 2 km). Time step within each word plot in left column is arbitrary but colored line letter segments correspond to a 30 s time period (one output interval). The five equiprobable Gaussian regions associated with each word are demarcated by light gray horizontal lines. Right column provides meteorological quantity metadata from an example storm that contains the rule. Each word from the left column is identified by a black line segment in the metadata. The 30 minute window prior to the development of strong low-level rotation is contained within the two vertical gray lines. Corresponding performance measures are listed at the top of the figure.

automated method we used provides an innovative alternative for assessing storm feature evolution which otherwise would require a brute force approach of evaluating individual sequences of model output across hundreds of simulated storms.

3.2 Spatiotemporal Relational Probability Trees

The multi-variate time series analysis was promising and demonstrated that mining of the storm data would likely yield useful information. However, our long-term goal was to develop an approach that would reason with both the spatial and the temporal nature of the data. As such, we developed the Spatiotemporal Relational Probability Tree (SRPT). The SRPT is a probability estimation tree that learns with spatiotemporally varying relational data. For example, a SRPT can predict new class labels based on questions such as "Has a downdraft lasted for at least 5 minutes?" or "Did a region of strong tilting of horizontal vorticity appear before the downdraft doubled in intensity?" While the SRPT was inspired by the the relational probability tree (RPT) (Neville et al, 2003), the SRPT represents both the data and the decision tree distinctions in a very different manner.

3.2.1 Data representation

By moving to a relational representation of our weather data, we gain the ability to reason about the high-level objects already identified by meteorologists. Such objects can represent concepts that they believe are associated with rotations and tornadoes or even different regions of a storm cell. The relational representation enables us to reason about the spatial or spatiotemporal relationships between the objects. Our data are represented as spatiotemporal attributed relational graphs, as we first presented in McGovern et al (2008). This representation is an extension of the attributed graph approach to handle spatiotemporally varying data (Neville et al, 2003; Neville and Jensen, 2004; Jensen, 2005) . All *objects*, such as updrafts or hail cores, are represented by vertices in the graph. *Relationships* between the objects are represented using edges. With the severe weather data, the majority of the relationships are spatial. Both objects and relationships can have *attributes* associated with them and these attributes can vary both spatially and temporally. In the case of a spatially or spatiotemporally varying attribute, the data are represented as either a scalar or a vector field, depending on the nature of the data. The ability to represent spatial fields of objects is a key addition to the SRRF results presented here. This field can be two or three dimensional for space and can also vary as a function of time. In addition to attributes varying over space and time, the existence of objects and relationships can also vary as a function of time. If an object or a relationship is *dynamic*, it has a starting and an ending time associated with it.

Figure 4 shows a schema of our spatiotemporal relational data for the severe storms domain. Note that this shows the possible objects and relationships within

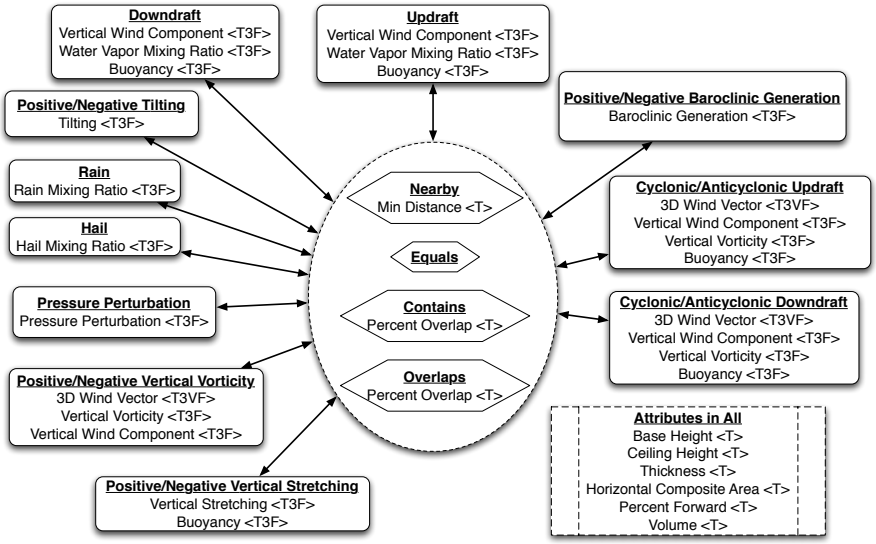


Fig. 4 Schema for the simulated supercell storms domain. The rounded rectangles represent objects and the hexagons show the relations. The type of each object or relation is bolded and the arrows show the directionality of the relationships. Attributes are listed inside both objects and relations. T denotes a temporally varying attribute; T3F denotes a three-dimensional spatiotemporal field.

a single example graph but it does not show the specific instantiation of a graph as that will differ for each storm. Both objects and relations are required to be assigned a type and graphs are required to be labeled (the label does not have to be binary). Objects and relations can be either static (existing for the duration of the graph) or dynamic. Temporal consistency is enforced for all dynamic objects or relations. Attributes can be associated with both objects and relations. Attributes can also be static or dynamic. If an attribute is static, its value stays the same throughout the lifetime of the object or relation. If it is dynamic, it can vary as a function of time or with space and time.

3.2.2 Algorithm

The SRPT is a decision-type tree with the ability to differentiate the data based on spatial, temporal, and spatiotemporal questions at each node. The tree is learned in the standard greedy manner and the differences lie in the questions the SRPT can ask at each node. In a standard decision tree such as C4.5 (Quinlan, 1993), there are a finite number of possible questions about the data. Because there are a very large number of possible splits or questions for spatiotemporal data, we sample the specific splits using a user specified sampling rate. For each sample, a split template

is selected randomly and the pieces of the template are filled in using randomly chosen examples in the training data. These templates are described below.

The non-temporal splits are:

- Exists: Does an object or relation of a particular type exist?
- Attribute: Does an object or a relation with attribute a have a [MAX, MIN, AVG, ANY] value $\geq$ than a particular value v?
- Count Conjugate: Are there at least n yes answers to distinction d? Distinction d can be any distinction other than Count Conjugate.
- Structural Conjugate: Is the answer to distinction d related to an object of type t through a relation of type r? Distinction d can be any distinction other than Structural Conjugate.

The temporal splits are:

- Temporal Exists: Does an object or a relation of a particular type exist for time period t?
- Temporal Ordering: Do the matching items from basic distinction a occur in a temporal relationship with the matching items from basic distinction b? The seven types of temporal ordering are: *before*, *meets*, *overlaps*, *equals*, *starts*, *finishes*, and *during* (Allen, 1991).
- Temporal Partial Derivative: Is the partial derivative with respect to time on attribute a on an object or relation of type $t \geq v$?

The spatial and spatiotemporal splits are:

- Spatial Partial Derivative: Is the partial derivative with respect to space of attribute a on an object or relation of type $t \geq v$?
- Spatial Curl: Is the curl of fielded attribute $a \geq v$?
- Spatial Gradient: Is the magnitude of the gradient of fielded attribute $a \geq v$?
- Shape: Is the primary 3D shape of a fielded object a cube, sphere, cylinder, or cone? This question also works for 2D objects and uses the corresponding 2D shapes.
- Shape Change: Has the shape of an object changed from one of the primary shapes over to a new shape over the course of t steps?

3.2.3 Empirical results

Figure 5 shows an example SRPT learned from the severe storm simulation data. Note that the high probabilities of the formation of strong low-altitude rotation occur only along the left portion of the tree. The low probabilities of low-altitude rotation occur along the right portion of the tree. The category “maybe” is found throughout the tree.

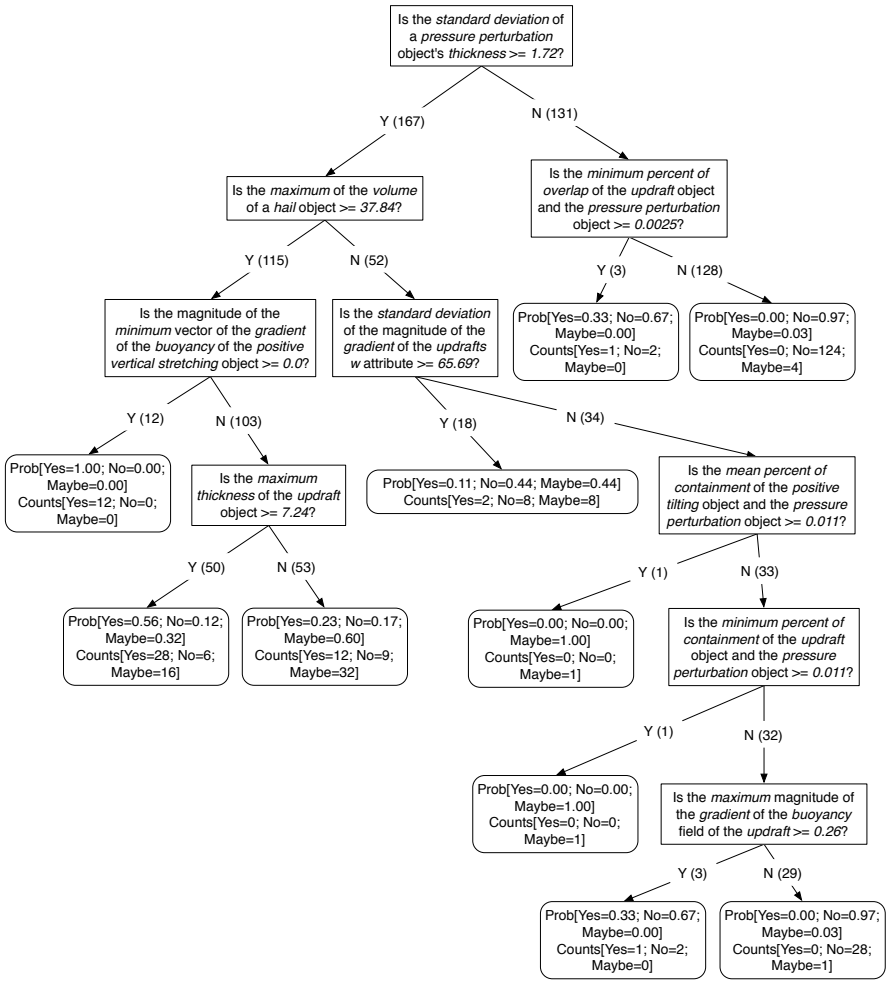


Fig. 5 Highest scoring single tree in the simulated storms domain using the SRPT.

3.3 Spatiotemporal Relational Random Forests

Although one reason for developing a tree-based model was for its human readability by domain scientists, a single decision-tree type model is known to be brittle (Pérez et al, 2005; Dwyer and Holte, 2007). This brittleness makes it harder for a domain scientist to trust the results from a single tree. As such, we developed an ensemble approach of SRPTs, following the Random Forest paradigm(Breiman, 2001). The Spatiotemporal Relational Random Forest (SRRF) enables a domain scientist to grow a robust predictive model of spatiotemporally varying relational data. Similar to Random Forests, the SRRF provides a method for variable impor-

tance, which measures the importance of attributes on objects or relationships to the predictive power of the forest.

3.3.1 Method

Growing a SRRF is very similar to the approach used to grow a Random Forest with the only differences occurring in the individual tree growing algorithm (described above) and from the nature of the spatiotemporal relational data. The training data for each tree in the forest is created using a bootstrap resampling of the original training data. The difference in the learning methods arises from the nature of the spatiotemporal relational data and the SRPTs versus C4.5 trees. In the RF algorithm, each node of each tree in the forest was trained on a different subset of the available attributes. Since the individual trees were standard C4.5 decision trees, this limited the number of possible splits each tree could make. Because each tree was also trained on a different bootstrap resampled set of the original data, the trees were sufficiently different from one another to make a powerful ensemble. Because there are a very large number of possible splits that the SRPTs can choose from, an SRPT finds the best split through sampling, as described above. Like the original RF trees, SRPTs are still built using the best split identified at each level. With fewer samples, these splits may not be the overall best for a single tree, but they will be sufficiently different across the sets of trees to ensure diversity in the forest. However, if the number of samples is too small, the number of trees needed in the ensemble to obtain good results may be prohibitively large.

For a particular attribute a, RFs measure variable importance by querying each tree in the forest for its vote on the out-of-bag data. Then, the attribute values for attribute a are permuted within the out-of-bag instances and each tree is re-queried for its vote on the permuted out-of-bag data. The average difference between the votes on the unpermuted data for the correct class and the votes for the correct class on the permuted data is the raw variable importance score. We have directly converted this approach to the SRRFs and can measure variable importance on any attribute of an object or relation. Spatially and temporally varying attributes are treated as a single entity and permuted across the objects/relations but their spatial and/or temporal ordering is preserved. We examine the variable importance in each of our data sets.

3.3.2 Empirical results

Table 1 shows the top 10 most statistically significant attributes associated with storms that develop strong low-altitude rotation, measured using variable importance on the SRRFs. Because rotation is associated with pressure drops, the inclusion of multiple attributes of the pressure perturbation object seems quite reasonable. Updrafts are the thermodynamic engines that power supercell thunderstorms and they play a major role in the concentration of rotation into tornado-scale vor-

Fields	Mean Variable Importance
PressurePerturbation.PressurePerturbationField	0.251
PressurePerturbation→ Overlap.PercentOverlap→Rain	0.188
Hail.HorizontalCompositeArea	0.114
PressurePerturbation.Volume	0.107
PressurePerturbation.Thickness	0.101
PressurePerturbation.BaseHeight	0.099
Hail.Volume	0.089
Updraft.Volume	0.086
Updraft.Thickness	0.061
Updraft.Buoyancy	0.060

Table 1 Top 10 statistically significant attributes according to variable importance on the simulated storms data.

tices. Owing to the presence of strong updrafts within such storms, conditions also are favorable for the formation of hail. Therefore, it is logical that the SRRF approach should find that pressure, updrafts, and hail attributes are important features of supercell storms that develop significant low-altitude rotation

4 Ongoing Research

The methods described in this chapter have proven both promising and fruitful in mining the 500 m resolution storm simulations. However, these simulations are limited in that they cannot resolve the circulation within a tornado. In our current research, we are developing a set of high resolution simulations, capable of resolving tornadoes. These simulations have horizontal grid spacings of 75 meters. Figure 6 shows the simulated reflectivity from one of these storms. This storm generated a tornado approximately 2 hours into the simulation and the effects of the tornado can be seen on the reflectivity.

We are also developing several new spatiotemporal relational data mining methods. The first focuses on enhancements to the SRPT/SRRF approach that will enable the true discovery of spatial and temporal relationships, not pre-specified by the user. The second focuses on using Bayesian Network structure learning techniques to identify salient relationships in the data.

Acknowledgements This material is based upon work supported by the National Science Foundation under IIS/CAREER/0746816 and corresponding REU Supplements IIS/0840956, 0938138, 1036023, 1129292, and the NSF ERC Center for Collaborative Adaptive Sensing of the Atmosphere (CASA, NSF ERC 0313747).

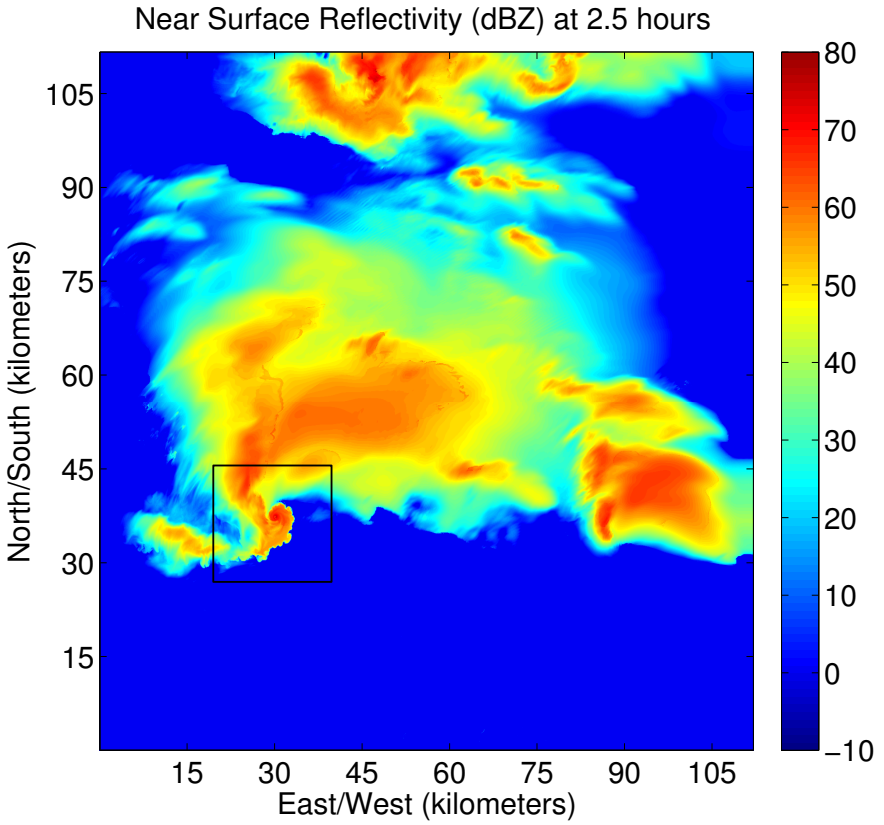


Fig. 6 Simulated near-surface reflectivity of a 75 m horizontal resolution storm. The boxed region highlights a tightly wound up end of the hook echo that indicates the presence of a tornado.

References

Allen JF (1991) Time and time again: The many ways to represent time. International Journal of Intelligent Systems 6(4):341–355

Breiman L (2001) Random forests. Machine Learning 45(1):5–32

Chiu B, Keogh E, Lonardi S (2003) Probabilistic discovery of time series motifs. In: In the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, pp 493–498

Das G, Lin K, Mannila H, Renganathan G, Smyth P (1998) Rule discovery from time series. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, pp 16–22

Donaldson Jr RJ, Dyer RM, Kraus MJ (1975) An objective evaluator of techniques for predicting severe weather events. In: Preprints: Ninth conference on severe local storms, American Meteorological Society, pp 321–326

Dwyer K, Holte R (2007) Decision tree instability and active learning. In: ECML '07: Proceedings of the 18th European conference on Machine Learning, Springer-Verlag, Berlin, Heidelberg, pp 128–139

Jensen D (2005) Proximity knowledge discovery system. kdl.cs.umass.edu/proximity

Johnson JT, MacKeen PL, Witt A, Mitchell ED, Stumpf GJ, Eilts MD, Thomas KW (1998) The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm. Weather and Forecasting 13(2):263–276

Keogh E, Lin J, Fu A (2005) HOT SAX: Efficiently finding the most unusual time series subsequence. In: Proc. of the 5th IEEE International Conference on Data Mining (ICDM 2005), Houston, Texas, pp 226–233

Lin J, Keogh E, Lonardi S, Chiu B (2003) A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp 2–11

Lin J, Keogh E, Li W, Lonardi S (2007) Experiencing SAX: A novel symbolic representation of time series. Data Mining and Knowledge Discovery 15(2):107–144

McGovern A, Jensen D (2008) Optimistic pruning for multiple instance learning. Pattern Recognition Letters 29(9):1252–1260

McGovern A, Rosendahl DH, Kruger A, Beaton MG, Brown RA, Droegemeier KK (2007) Anticipating the formation of tornadoes through data mining. In: Preprints of the Fifth Conference on Artificial Intelligence and its Applications to Environmental Sciences at the American Meteorological Society Annual Meeting, American Meteorological Society, San Antonio, TX, Paper 4.3A

McGovern A, Hiers N, Collier M, Gagne II DJ, Brown RA (2008) Spatiotemporal relational probability trees. In: Proceedings of the 2008 IEEE International Conference on Data Mining, Pisa, Italy, pp 935–940

McGovern A, Supinie T, Gagne II DJ, Troutman N, Collier M, Brown RA, Basara J, Williams J (2010) Understanding severe weather processes through spatiotemporal relational random forests. In: Proceedings of the 2010 NASA Conference on Intelligent Data Understanding, pp 213–227

McGovern A, Gagne II DJ, Troutman N, Brown RA, Basara J, Williams J (2011a) Using spatiotemporal relational random forests to improve our understanding of severe weather processes. Statistical Analysis and Data Mining 4(4):407–429

McGovern A, Rosendahl DH, Brown RA, Droegemeier KK (2011b) Identifying predictive multi-dimensional time series motifs: An application to understanding severe weather. Data Mining and Knowledge Discovery 22(1):232–258

McGovern A, Troutman N, Brown RA, Williams JK, Abernethy J (under review) Enhanced spatiotemporal relational probability trees and forests. Data Mining and Knowledge Discovery

Minnen D, Isbell C, Essa I, Starner T (2007) Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (ICDM '07), pp 601–606

Mueen A, Keogh E, Zhu Q, Cash S, Westover B (2009) Exact discovery of time series motifs. In: Proceedings of the SIAM International Conference on Data Mining, pp 473–484

Neville J, Jensen D (2004) Dependency networks for relational data. In: Proceedings of the Fourth IEEE International Conference on Data Mining, pp 170–177

Neville J, Jensen D, Friedland L, Hay M (2003) Learning relational probability trees. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 625–630

Noda A, Niino H (2005) Genesis and structure of a major tornado in a numerically–simulated supercell storm: Importance of vertical vorticity in a gust front. Science Online Letters on the Atmosphere pp 5–8

NWS (2009) Service assessment, mother's day weekend tornado in oklahoma and missouri, may 10, 2008. http://www.nws.noaa.gov/os/assessments/pdfs/mothers_day09.pdf

NWS (2011) Nws central region service assessment, joplin, missouri, tornado – may 22, 2011. http://www.nws.noaa.gov/os/assessments/pdfs/Joplin_tornado.pdf

Oates T, Cohen PR (1996) Searching for structure in multiple streams of data. In: Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann, pp 346–354

Pérez JM, Muguerza J, Arbelaitz O, Gurrutxaga I, Martìn JI (2005) Consolidated trees: Classifiers with stable explanation. a model to achieve the desired stability in explanation. In: Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp 99–17

Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann

Rosendahl DH (2008) Identifying precursors to strong low-level rotation within numerically simulated supercell thunderstorms: A data mining approach. Master's thesis, School of Meteorology, University of Oklahoma

Schaefer JT (1990) The critical success index as an indicator of warning skill. Weather and Forecasting 5(4):570–575

Shieh J, Keogh E (2009) iSAX: Indexing and mining terabyte sized time series. In: Proceedings of the IEEE International Conference on Data Mining

Simmons KM, Sutter D (2011) Economic and Societal Impacts of Tornadoes. American Meteorological Society

Supinie T, McGovern A, Williams J, Abernethy J (2009) Spatiotemporal relational random forests. In: Proceedings of the IEEE International Conference on Data Mining (ICDM) workshop on Spatiotemporal Data Mining, p electronically published

Vahdatpour A, Amini N, Sarrafzadeh M (2009) Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In: Proceedings of the 21st international joint conference on Artificial intelligence (IJCAI'09), pp 1261–1266

Webb GI (1995) OPUS: An efficient admissible algorithm for unordered search. Journal of Artificial Intelligence Research 3:431–465

Wicker LJ, Wilhelmson RB (1995) Simulation and analysis of tornado development and decay within a three–dimensional supercell thunderstorm. Journal of the Atmospheric Sciences 52(15):2675–2703

Xue M, Droegemeier KK, Wong V (2000) The Advanced Regional Prediction System (ARPS) - a multiscale nonhydrostatic atmospheric simulation and prediction model. Part I: Model dynamics and verification. Meteorology and Atmospheric Physics 75:161–193

Xue M, Droegemeier KK, Wong V, Shapiro A, Brewster K, Carr F, Weber D, Liu Y, Wang D (2001) The Advanced Regional Prediction System (ARPS) - a multiscale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications. Meteorology and Atmospheric Physics 76:143–165

Xue M, Wang D, Gao J, Brewster K, Droegemeier KK (2003) The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. Meteorology and Atmospheric Physics 82:139–170

Xue M, Droegemeier KK, Weber D (2007) Numerical prediction of high-impact local weather:a driver for petascale computing. In: Petascale Computing: Algorithms and Applications, Chapman and Hall/CRC Press, chap 18, pp 103–125

Zaki MJ (2001) Spade: An efficient algorithm for mining frequent sequences. Machine Learning 42(1/2):31–60, special issue on unsupervised learning