4.3A Anticipating the formation of tornadoes through data mining

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

Derek H. Rosendahl
School of Meteorology
University of Oklahoma
Norman, OK
drose@ou.edu

Adrianna Kruger
School of Computer Science
University of Oklahoma
Norman, OK
adriannakruger@ou.edu

Meredith G. Beaton
School of Computer Science
University of Oklahoma
Norman, OK
mbeaton@ou.edu

Rodger A. Brown
NOAA
National Severe Storms Laboratory
Norman, OK
Rodger.Brown@noaa.gov

Kelvin K. Droegemeier
School of Meteorology
University of Oklahoma
Norman, OK
kkd@ou.edu

1. Introduction

Severe weather phenomena such as tornados, thunderstorms, hail, and floods, annually cause significant loss of life, property and crop destruction, and disruption of the transportation systems. The annual economic impact of these phenomena is estimated to be greater than 13 billion dollars (Pielke and Carbone 2002). Any mitigation of the effects of these storms would be beneficial.

We propose to enhance our understanding of the formation of severe weather events, specifically focusing on tornadoes, through data mining/knowledge discovery techniques. The process of knowledge discovery is about making sense of data. Generally, the data are too complex for humans to quickly identify and understand the important patterns. Instead, knowledge discovery techniques can be used to highlight salient patterns. We are developing new data mining techniques for use on storm-scale weather data.

The long-term goals of our work are to improve the understanding of tornado formation to the point where refined detection and prediction algorithms can be created. We will do this by engaging in an interdisciplinary knowledge discovery process where we develop new data mining techniques (computer science) to understand the data and analyze the results (meteorology). This becomes a cycle where the results inform new techniques and new techniques produce new results. The results presented in this paper represent the beginning of this research.

2. Meteorological Data

With our goal of improving the detection and anticipation of tornadoes, we are not taking the traditional route of examining radar reflectivity and radial velocity gathered by a specific radar system. These radar systems may be limited in their ability to detect and anticipate tornadoes due to inherent radar characteristics such as a beam increasing in altitude and spreading as it travels away from the radar causing the resolution volume to be too large to accurately identify tornadic circulations (e.g. Donaldson 1970; Brown et al. 1978; Wood and Brown 1997). Additionally, analyzing only radar reflectivity and radial velocity provides a limited number of variables to use in depicting the true state of the atmosphere. We therefore make use of new research

on data assimilation that will eventually enable us to predict and detect severe weather on a real-time three-dimensional gridded data set containing all the fundamental and derived meteorological quantities. Examples of the fundamental quantities include the three components of air motion, temperature, pressure, and precipitation. Examples of the derived variables include divergence, temperature gradient, vertical vorticity, and the pressure gradient force. The ability to have all fundamental and derived meteorological quantities at all grid points results in an improved and significantly broadened representation of the atmosphere. This new representation necessitates the development of more sophisticated detection and anticipation techniques.

Figure 1 gives a sample of the gridded data created using an ensemble Kalman filter assimilation of real observations (Tong and Xue 2005). The top two panels display the observed reflectivity and Doppler wind measurements in the May 29, 2004 tornado in Oklahoma City. The remaining panels show assimilated radar reflectivity and retrieved temperature, pressure, and vertical vorticity.

Because the technology to create real-time three-dimensional gridded data from actual observations is currently under development, we are using simulated storm data produced from the Advanced Regional Prediction System (ARPS), which is a three-dimensional, nonhydrostatic model that is one of the top weather forecasting systems for mesoscale data (Xue et al. 2000, 2001, 2003). The computational grid for our study has a horizontal spacing of 0.5 km within a 100 km by 100 km by 20 km domain. There are 49 levels in the vertical with a stretched grid from 50 m at the ground to 750 m near the top of the domain. The model is run for three hours with history files saved every 30 seconds.

Soundings are used to initialize horizontally homogeneous base state environments wherein a thermal bubble initiates convection. The thermodynamic profiles of the soundings are analytically constructed using the Weisman and Klemp (1982) method with variations in the surface mixing ratios. The hodographs have variations in magnitude and shape similar to Adlerman and Droegemeier (2005) (e.g. half circle, quarter circle, quarter circle with tail, and straight). Only supercell storms are studied in this research and therefore indices such as 0-6 km Bulk Richardson Number (Weisman and Klemp 1982) and storm-relative helicity (Davies-Jones et al. 1990) are used to identify suitable soundings. Our preliminary results used soundings having surface mixing ratios of 13, 14, 15, 16, and 17 g kg^{-1} with a half circle hodograph of radius 10 m s^{-1} turning through 4 km. We also used a sounding from the 20 May 1977 Del City, Oklahoma tornado (Ray et al. 1981)

2a. *Data Extraction*

Our eventual goal is to examine the data using high-level features such as a hook echo, gust front, rear flank downdraft, occlusion downdrafts, etc. However, automated identification and tracking of these high-level features is computationally difficult. Identifying them requires creating a description that a majority of meteorologists will agree on or at least creating a large enough labeled data set such that a machine learning algorithm could learn to extract these features. We will be addressing these issues in future work.

For the results in this paper, we chose to extract a set of 24 fundamental and derived meteorological quantities. These quantities are listed in Table 1. These represent the most important meteorological quantities and enable us to observe each storm with a significant reduction in data size over examining each variable at each grid point.

Any given storm simulation may generate several separate storm cells. We define a storm cell based on the maximum updraft. The algorithm for identifying and tracking individual storms cells is given in Table 2. At each time step, we identify the domain-wide maximum vertical wind speed at 4 km height. Once a single storm is being tracked, new storms will not be identified until their maximum vertical wind speed exceeds that of the main storm. A 20 km by 20 km full height box is drawn around the location of the maximum vertical wind speed at 4 km height for each storm. We then measure the maximum and minimum of each quantity listed in Table 1 within the box. We measure the maximums and minimums for each variable for the surface to 2 km in height and then from 2 km to the top. 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 level feature.

In current work, we are examining the use of a modified Storm Cell Identification and Tracking (SCIT) algorithm (Johnson et al. 1998) for improved storm identification and tracking. The original SCIT algorithm used reflectivity to identify and track each storm. Because we do not have to depend only on reflectivity, we instead use discrete areas of significant updrafts and track around each updraft.

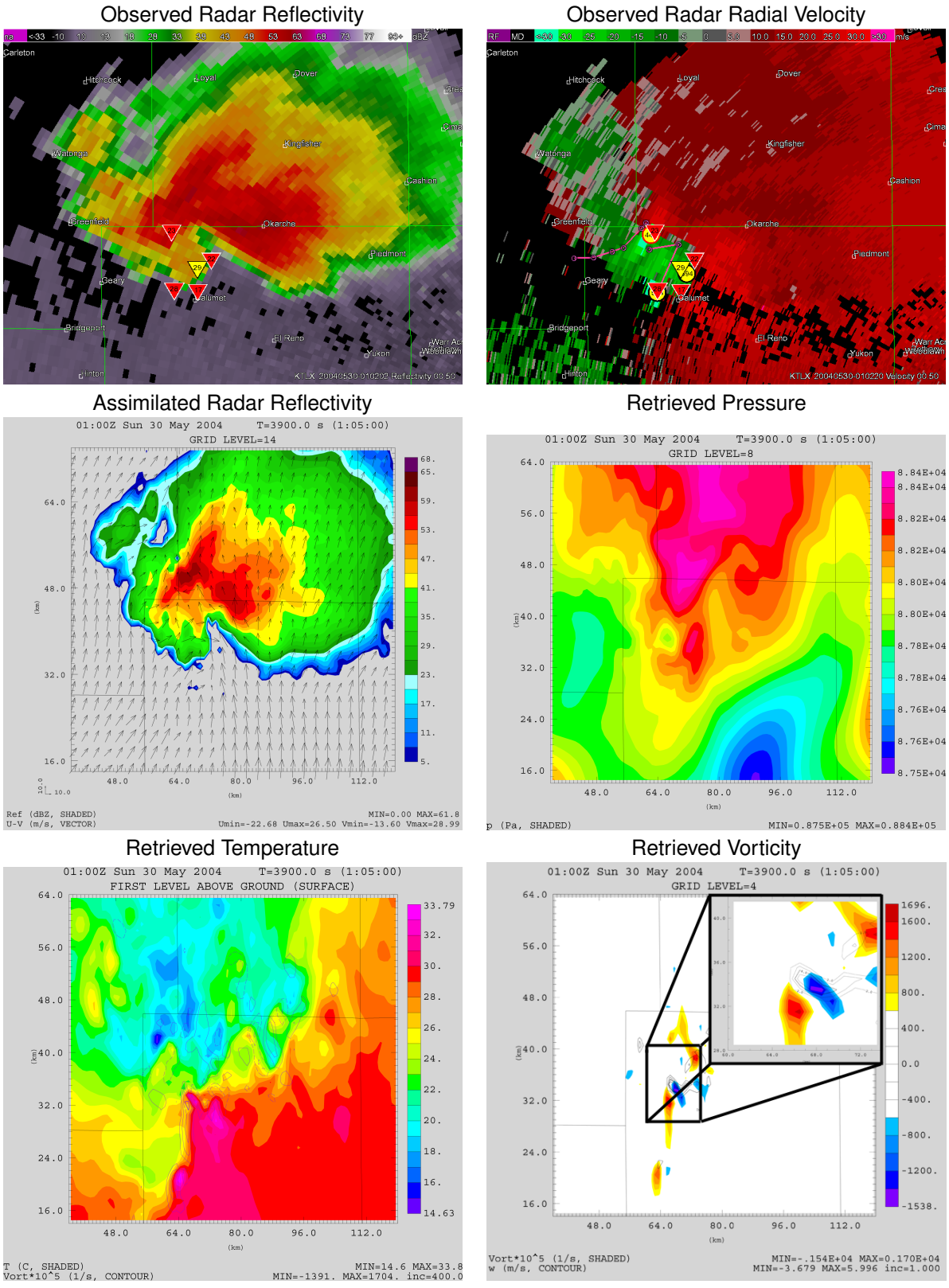

Figure 1: An example of real data (top panels) being used to create gridded data (center and bottom panels). This example is from the May 29, 2004 tornado in Oklahoma City and is courtesy of Fritchie and Droegemeier at the University of Oklahoma.

Table 1: Maximum and minimum quantities extracted for each storm cell. The bars represent averages.

Variable	*Equation*	*Units*
vertical velocity	w	m s^{-1}
vertical velocity horizontal gradient	$\sqrt{\left(\frac{\Delta_h w}{\Delta x}\right)^2 + \left(\frac{\Delta_h w}{\Delta y}\right)^2}$	s^{-1}
vertical vorticity	$\frac{\Delta_h v}{\Delta x} - \frac{\Delta_h u}{\Delta y}$	s^{-1}
rainwater mixing ratio	qr	kg kg^{-1}
rainwater mixing ratio horizontal gradient	$\sqrt{\left(\frac{\Delta_h qr}{\Delta x}\right)^2 + \left(\frac{\Delta_h qr}{\Delta y}\right)^2}$	kg kg^{-1} m^{-1}
rainwater mixing ratio vertical gradient	$\sqrt{\left(\frac{\Delta_v qr}{\Delta z}\right)^2}$	kg kg^{-1} m^{-1}
perturbation potential temperature	pt - ptbar	K
pressure perturbation (p')	p - pbar	Pa
vertical perturbation pressure gradient force	$\frac{-1}{rho} * \frac{\Delta_v p'}{\Delta z}$	m s^2
horizontal divergence	$\frac{\Delta_h u}{\Delta x} + \frac{\Delta_h v}{\Delta y}$	s^-1
hail mixing ratio	qh	kg kg^{-1}
hail mixing ratio horizontal gradient	$\sqrt{\left(\frac{\Delta_h qh}{\Delta x}\right)^2 + \left(\frac{\Delta_h qh}{\Delta y}\right)^2}$	kg kg^{-1} m^{-1}
hail mixing ratio vertical gradient	$\sqrt{\left(\frac{\Delta_h qh}{\Delta z}\right)^2}$	kg kg^{-1} m^{-1}
horizontal wind speed	$\sqrt{u^2 + v^2}$	m s^{-1}
vertical stretching	$-\left(\frac{\Delta_h v}{\Delta x} - \frac{\Delta_h u}{\Delta y}\right)\left(\frac{\Delta_h u}{\Delta x} + \frac{\Delta_h v}{\Delta y}\right)$	s^{-1}
tilting term	$\left(\frac{\Delta_h w}{\Delta y} * \frac{\Delta_h u}{\Delta z}\right) - \left(\frac{\Delta_h w}{\Delta x} * \frac{\Delta_h v}{\Delta z}\right)$	s^{-1}
baroclinic generation	$\left(\frac{1}{(rho)^2} * \frac{\Delta_h (rho)}{\Delta x} * \frac{\Delta_h p}{\Delta y}\right) - \left(\frac{\Delta_h (rho)}{\Delta y} * \frac{\Delta_h p}{\Delta x}\right)$	s^{-1}
vertical velocity (w) and vertical vorticity (ζ) correlation	$\frac{\Sigma(w_{ij} - \overline{w}_{w>1})\left(\zeta_{ij} - \overline{\zeta}_{w>1}\right)}{\sqrt{\Sigma(w_{ij} - \overline{w}_{w>1})}\sqrt{\Sigma\left(\zeta_{ij} - \overline{\zeta}_{w>1}\right)}}$	
horizontal potential temperature gradient	$\sqrt{\left(\frac{\Delta_h pt}{\Delta x}\right)^2 + \left(\frac{\Delta_h pt}{\Delta y}\right)^2}$	K m^{-1}
radar reflectivity	$ref = 10 * LOG_{10}\left(Ze_{rain} + Ze_{snow} + Ze_{hail}\right)$	dBZ
radar reflectivity horizontal gradient	$\sqrt{\left(\frac{\Delta_h (ref)}{\Delta x}\right)^2 + \left(\frac{\Delta_h (ref)}{\Delta y}\right)^2}$	dBZ m^{-1}
radar reflectivity vertical gradient	$\sqrt{\left(\frac{\Delta_v (ref)}{\Delta z}\right)^2}$	dBZ m^{-1}
horizontal Laplacian of radar reflectivity	$\nabla^2(ref)$	dBZ m^{-2}
vertical velocity and horizontal Laplacian of radar reflectivity correlation	$\frac{\Sigma(w_{ij} - \overline{w}_{w>1})\left(\nabla^2(ref)_{ij} - \overline{\nabla^2(ref)}_{w>1}\right)}{\sqrt{\Sigma(w_{ij} - \overline{w}_{w>1})^2}\sqrt{\Sigma\left(\nabla^2(ref)_{ij} - \overline{\nabla^2(ref)}_{w>1}\right)^2}}$	

Table 2: Summary of how we identify individual storm cells and extract the maximum and minimum quantities from each storm.

Identify-storm-cells(time t)
- $w_{max} \leftarrow$ domain-wide maximum vertical wind velocity (w) at 4 km height
- If $w_{max} < 15$ m s^{-1}
  - return
- Otherwise
  - Measure the Euclidean distance from the location of w_{max} to the storms identified at time t-1
  - If the closest previous storm is less than 5 km away, assume it is the same storm
  - Otherwise, mark this as a new storm.
- For all storms identified on the previous step and the new storm if it exists
  - Extract the maximum and minimum quantities in a 20 km by 20 km full height box around w_{max}
  - If w_{max} within a storm drops below 10 m s^{-1}, stop tracking this storm

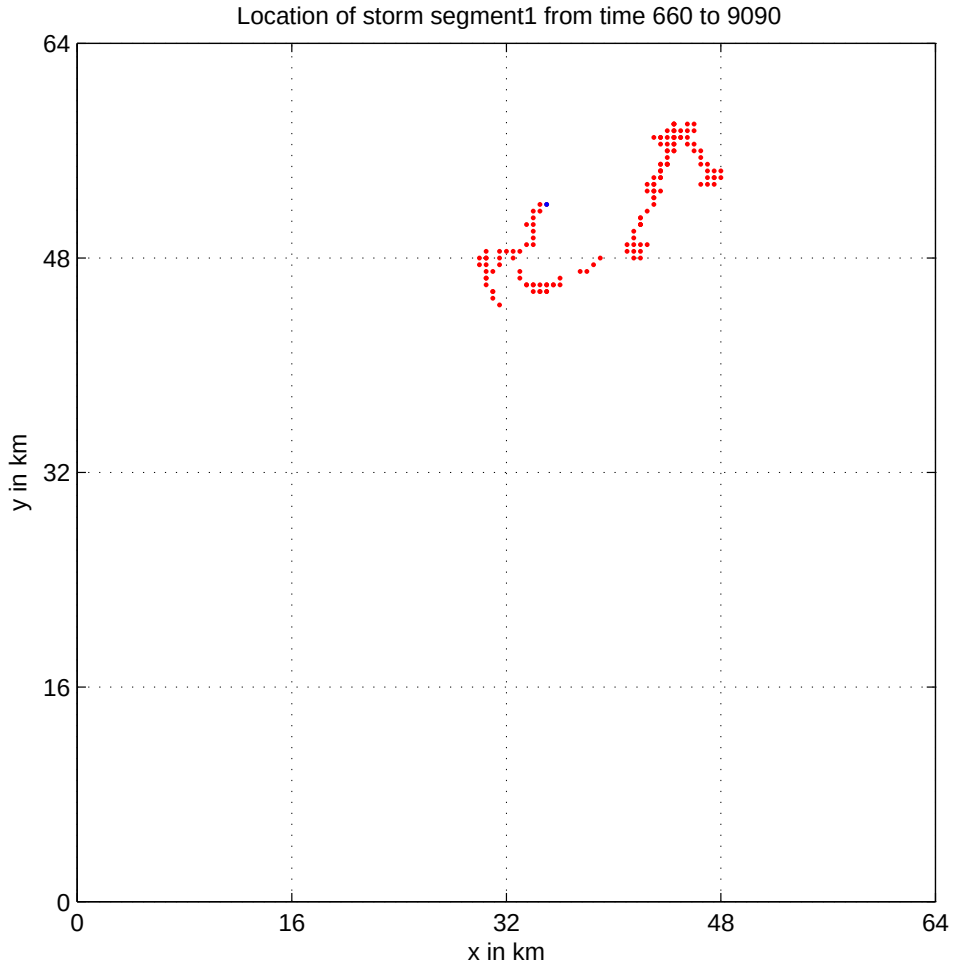

Figure 2: Locations of the maximum vertical velocity (w) at 4km within a storm on the WK14 run. The blue dot shows the ending location. This storm existed from 660 seconds into the simulation until 9090 seconds (nearly the end).

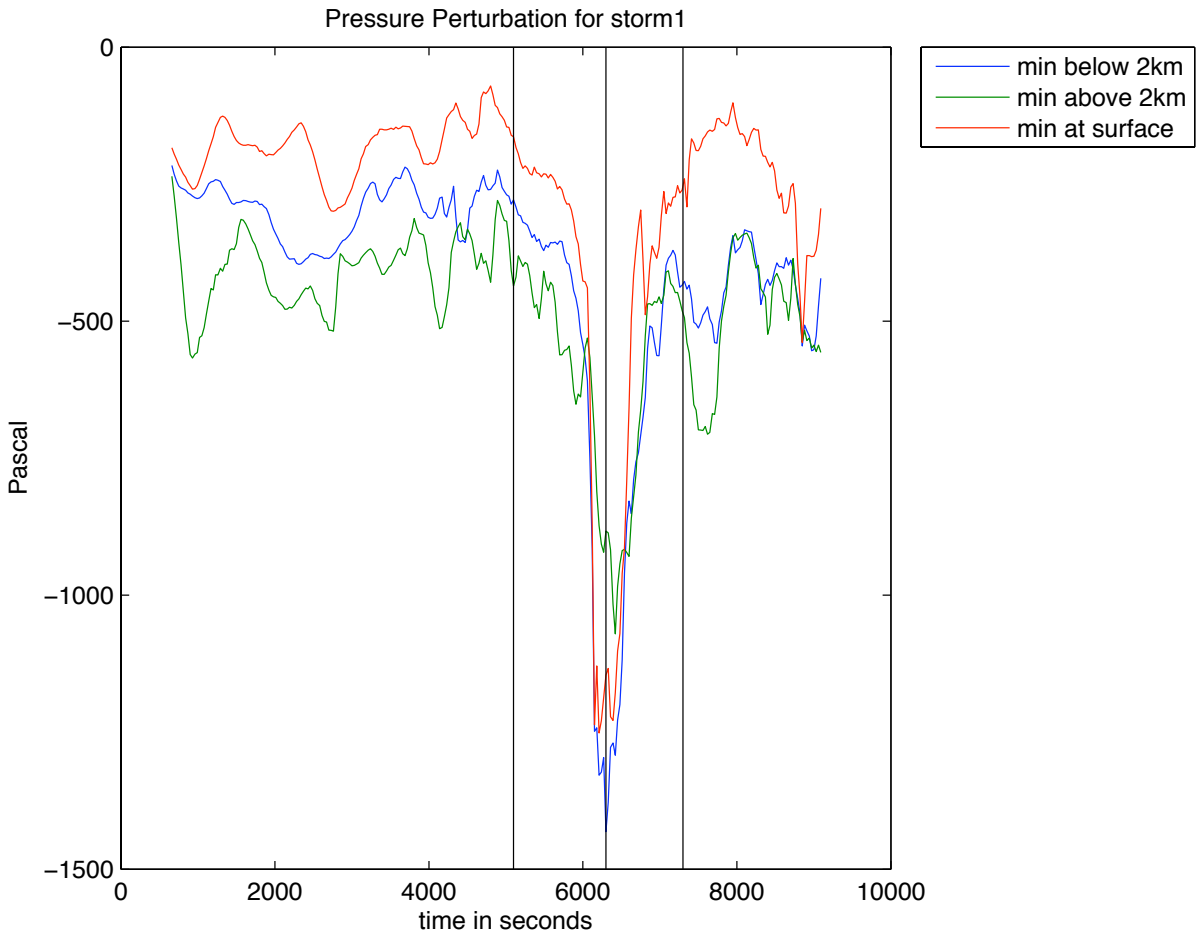

Figure 3: Maximum and minimum pressure pertubations recorded in a 20km by 20km box around the location of the maximum vertical wind speed. Each of the values is recorded for a specific range of height values. The vertical lines highlight several regions of interest for a low-level rotation.

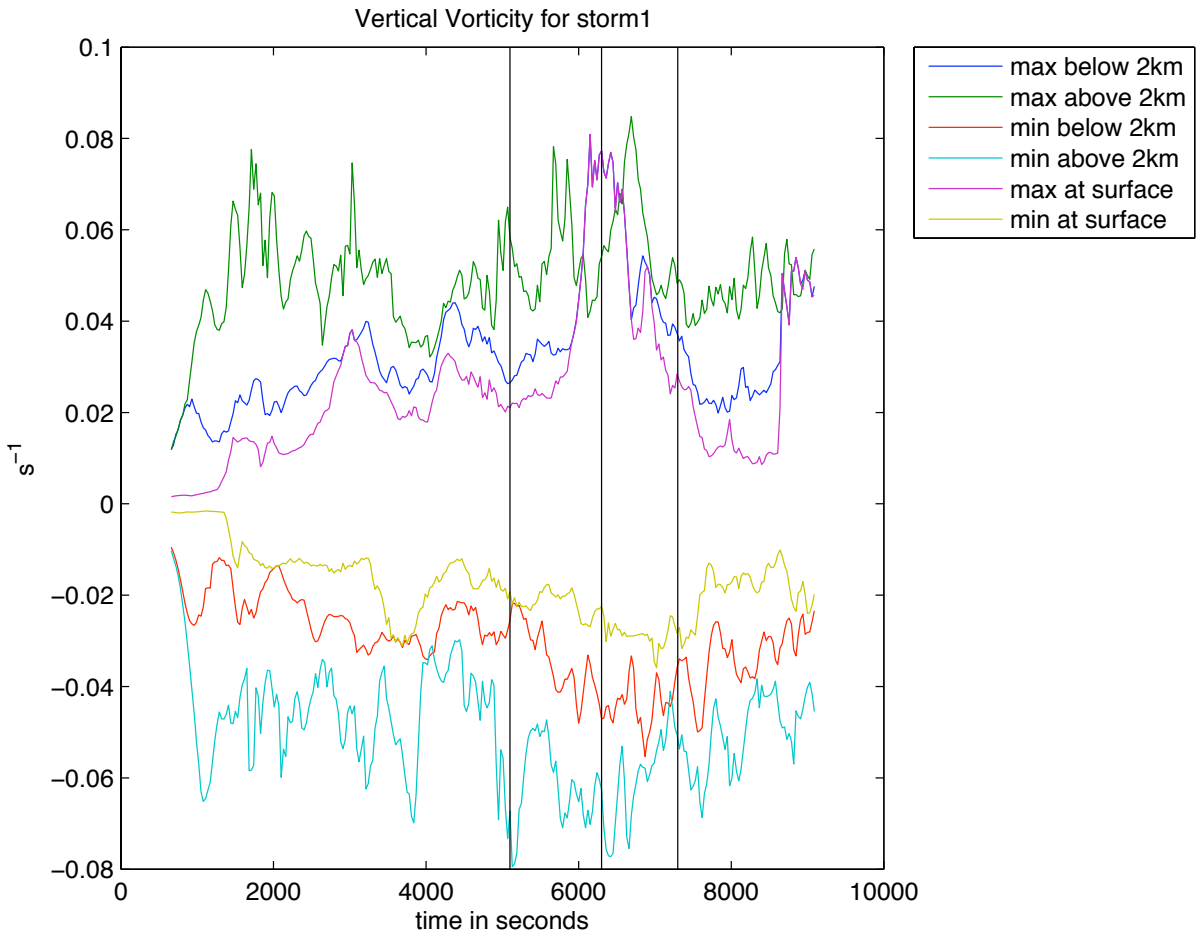

Figure 4: Maximum and minimum vertical vorticities recorded in a 20km by 20km box around the location of the maximum vertical wind speed. Each of the values is recorded for a specific range of height values. The vertical lines highlight several regions of interest for a low-level rotation.

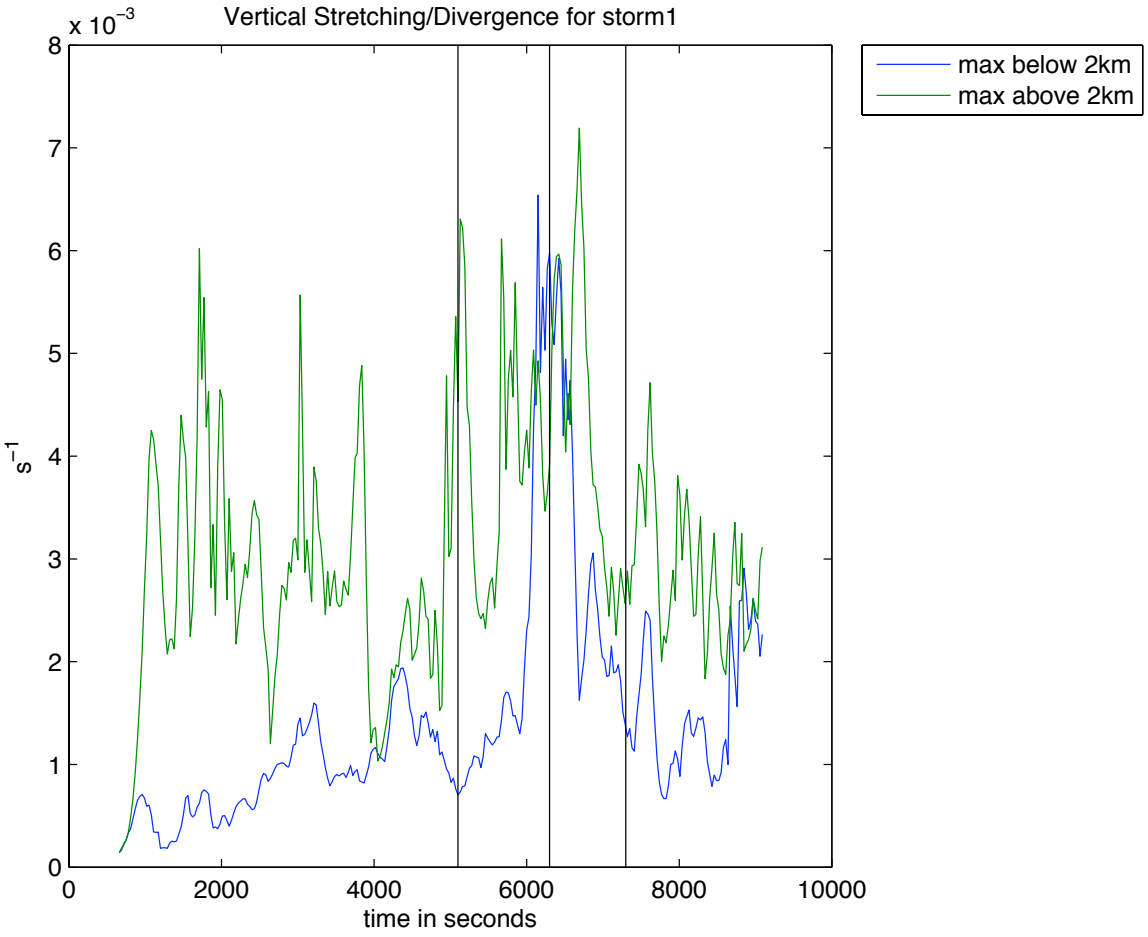

Figure 5: Maximum and minimum vertical stretching recorded in a 20km by 20km box around the location of the maximum vertical wind speed. Each of the values is recorded for a specific range of height values. The vertical lines highlight several regions of interest for a low-level rotation.

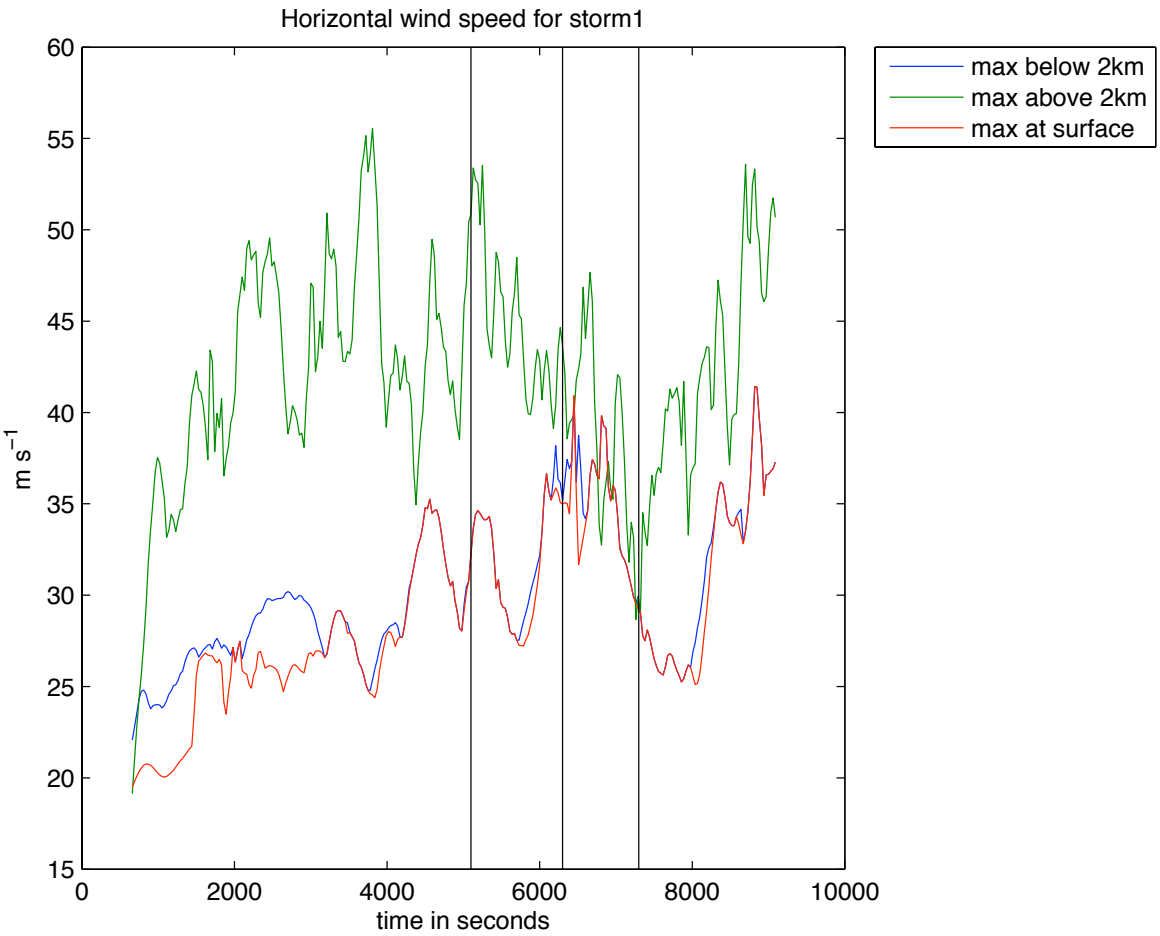

Figure 6: Maximum and minimum horizontal wind speed recorded in a 20km by 20km box around the location of the maximum vertical wind speed. Each of the values is recorded for a specific range of height values. The vertical lines highlight several regions of interest for a low-level rotation.

2b. *Identifying storm features*

The horizontal grid spacing (0.5 km) used by the ARPS model to produce the simulated thunderstorms in this study is of the same order as the diameter of tornadoes. Consequently a tornado cannot be resolved in the model output. However, the model's grid spacing is sufficient to resolve the tornado's parent mesoscale circulation, which is what we use in these results.

Given the inability to resolve a tornado with a 500 m grid spacing, we manually examined the temporal data for evidence of mesoscale circulation which might indicate the presence of a tornado. Current tornado detection algorithms also search for evidence of the mesoscale circulation because the tornadoes themselves are too small to be observed unless they occur near the radar (Lakshmanan et al. 2006). However, radar-based algorithms rely only on reflectivity and radial velocity while we can use any of the retrieved fields. We found that prominent signatures for the following four parameters within the lowest 2 km and at the surface were associated with the development of a low-altitude circulation within a given storm: perturbation pressure, vertical component of vorticity, horizontal wind speed, and the stretching term of the vertical vorticity equation. In particular, we looked for coincident localized pressure minima, vertical vorticity maxima, wind speed maxima, and stretching term maxima. Maximum vertical velocity above 2 km, which is important in the development of a tornado's parent circulation, was not included as a pertinent parameter. We tracked the maximum vertical velocity within a 20 by 20 km box. It is likely that this box contained several updrafts at various stages of evolution which meant it was not possible to isolate the particular updraft that was associated with the circulation. Examination of the data upheld this hypothesis.

Figure 2 shows the location of the maximum vertical wind speed w_{max} at 4 km from the time that it reaches 15 m s^{-1} until the time that the maximum drops below 10 m s^{-1}. The blue dot shows the ending location. Shown in Figures 3-6 are plots of the four parameters for the storm generated using a 14 g kg^{-1} mixing ratio. We refer to this storm as WK14. The portions of the curves that indicate the presence of a low-altitude circulation are located within the band bounded by the solid vertical lines at 5100 and 7300 seconds (37 minutes time interval). The center vertical line at 6300 seconds indicates the approximate time when the circulation was strongest. Perturbation pressure in the developing circulation started to decrease around 5100 seconds (Figure 3). The significant decrease in pressure (about 10 millibars or 1000 pascals) coincided with a marked increase in low-altitude circulation (vertical vorticity maximum in Figure 4). The circulation increase appears to be due primarily to the increase of air converging into the lower portion of the updraft with a resulting increase in vertical stretching of vertical vorticity (Figure 5). As to be expected, low-altitude wind speed increased in conjunction with the increased circulation (Figure 6). Around 6300 seconds the parent mesoscale circulation was most intense and it would be at that time that a tornado would have occurred if we had a finer grid spacing.

About 45 minutes after this circulation began, a weaker secondary circulation developed in the storm (from around 8500 seconds to the end of the data curves). This secondary circulation had a weaker pressure drop associated with it, but there was a significant increase in vertical vorticity (associated with increased vertical stretching of vorticity) and horizontal wind speed. When severe thunderstorms develop a series of mesoscale circulations (each frequently associated with a separate tornado), the circulations develop at roughly 40-minute intervals (e.g., Burgess et al. 1982; Alderman et al. 1999), which is evident in our findings.

Although we labeled strong low-level rotations by hand for these results, we are developing a more automated approach that will require much less human intervention for the full set of 300-500 storm simulations. We cannot remove the human from the process of labeling tornadoes entirely or else tornado detection would already be a solved problem.

3. Data Mining

Mesoscale weather data pose several challenges to current data mining techniques. The sheer size of the data available is more than many current techniques can handle. For example, each of our simulations produces about 20GB of data. The data are dynamic (temporal), continuous, and multidimensional. Even with a propositional representation, identifying patterns in continuous data is difficult. The propositional representation of measuring maximum and minimum quantities gives us a reduced representation for the data that is more manageable. This section presents our algorithm and our approach to dealing with the continuous and multidimensional aspects of the data.

The goal of this project is to identify salient patterns in the data occurring prior to the development of strong low-level rotation that could be used to an-

ticipate tornado formation. That is, we want to identify the conditions in a supercell storm that lead to a tornado in condition A but not in condition A' where A and A' are very similar storm conditions. Intuitively, the goal of our data mining algorithm is to create a set of rules of the form that "if the time series data for feature A fits the characteristic shape X and the time series data for feature B fits the characteristic shape Y within five minutes of the match on feature A, then the probability of a tornado occurring within Z minutes is P." This section builds on this intuition with a formal definition of the patterns and algorithm we use to identify those patterns.

3a. *Definitions*

Using the propositional approach described above, we have continuous multi-dimensional real-valued data. Each dimension of the data is called a *feature*. Each one of the maximum and minimum quantities outlined in Table 1 is a feature. For a single storm, each feature, f, takes the form of $f = \{v_t, v_{t+1}, v_{t+2}, \ldots v_{t+T}\}$ where v is a real number, t is the first time step for the data collected for this storm cell, and T is the number of time steps that this feature was recorded. We call the full set of readings for a single feature in a single storm a *trajectory*, $\mathcal{T}$. It has also been called a stream (Oates and Cohen 1996).

Traditionally in machine learning, features are assumed to be statistically independent. Examination of the table shows that some of the features are clearly related to one another. However, this assumption is often violated in machine learning yet performance is not affected (Holte 1993; Domingos and Pazzani 1997).

Each storm cell, S, has a trajectory per feature and a single binary label $\mathcal{L} \in \{0, 1\}$ associated with it. We refer to $\mathcal{L} = 0$ as negative labels and $\mathcal{L} = 1$ as positive labels. Formally, $S = \{\mathcal{L}, \mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \ldots \mathcal{T}_n\}$ where n is the number of trajectories associated with a single storm. The label applies to the trajectories as a whole and does not point to any specific time within the trajectory that caused the trajectory to be labeled as positive or negative. The data mining algorithm aims to identify the subset of features that are most predictive of the labels.

Given a single feature f, a *pattern*, p, is informally defined as a characteristic shape for that feature. For example, if the feature measures the maximum pressure perturbation at the surface, a characteristic shape could be a sudden drop followed by a rise. These patterns are not predefined but identified automatically by the data mining program. More formally, a pattern is a time series of normalized real-values that the feature should follow, $p = \{f, v_k, v_{k+1}, v_{k+2}, \ldots, v_{k+j}\}$ where $k \geq t$ and $j \leq T$.

A rule, r, is an ordered set of patterns, a time window, a weight, and a probability prediction. Formally, $r = \{p_1, p_2, \ldots p_j, \text{window}, w, \text{prob}\}$. Each of the patterns p_i may come from a different feature. The patterns are assumed to occur in order, meaning that p_2 must occur after p_1, but there is no specific time gap between the patterns so long as all fit within the specified time window. The probability, prob, specifies the probability of seeing a positive label given this rule. The weight, $w \in \mathcal{R}$, is used to determine the relative strengths of the rules in the case that multiple rules are identified.

Although this formulation looks similar to the formulations for multiple-instance learning (MIL) (Dietterich et al. 1997; McGovern and Jensen 2006), there are several big differences. The first is that the data in the trajectories are ordered while MIL data are not. A second difference comes in the formulation of the knowledge representation. Given the unordered nature of the data, MIL algorithms generally look for a subset of predictive features and an associated range on their values but they do not look for patterns, especially those with gaps. This formulation also shares some similarities with association rules (for example, see Zaki 2001). Although association rule algorithms do handle part of the temporal nature of the data, they still assume unordered itemsets.

3b. *Data conversions*

Data mining with continuous data presents a number of challenges. In particular, it is highly unlikely that any two trajectories will ever contain exactly the same pattern. The usual answer to dealing with continuous data is to transform the data to a discrete representation. By mining at the discrete (and often symbolic) level, we significantly cut down the search space while enabling approximate matches between patterns. However, most techniques for creating discrete data from continuous data lose the ability to compare the distance of potential patterns in the original form of the data. This ability is critical because the flip side of creating a discrete representation is that patterns may look interesting when they are not comparable in the original data. We chose to use the SAX (Lin et al. 2003) approach which explicitly addresses these issues. The symbolic data generated using SAX can be compared and there is a lower-bound on the distance of the symbolic data

with respect to the distance of the original continuous data.

Figure 7 and 8 show how SAX converts the vorticity and pressure perturbation graphs shown in Figures 3 and 4 respectively. SAX first normalizes the data in a single trajectory to fit a gaussian with mean of zero and a standard deviation of 1. Symbols (such as a, b, c, d, e, and f) are generated according to the distribution of the normalized data. For example, Figure 7 contains all six symbols while Figure 8 has such a significant drop in pressure that the symbol 'f' is never observed in the normalized data. The lines on the graph show the breakpoints between symbols and the actual symbols themselves can be seen. Each symbol is also colored differently. For these graphs, we averaged together six data points (three minute of simulation time) for each symbol. Each trajectory for each storm is converted (and normalized) separately. The symbolic data are used for data mining but we can refer back to the original data in the final step of choosing the most interesting patterns.

3c. *Algorithm*

This rule-finding algorithm draws from the MSDD algorithm of Oates and Cohen (1996), the multiple-instance learning approach described by McGovern and Jensen (2006), and the data structures for efficient detection of unusual patterns described in Keogh et al. (2005). The parameters are the minimum word size (number of symbols) for each pattern and a p-value used to prune rules that are not statistically significant. Typically the minimum word size should be small. We used three for the results presented in this paper. With each symbol representing three minutes of data, words represent nine minutes of data. The rule finding algorithm works as described below.

1. For each storm cell and each trajectory within a storm cell, create symbolic data using SAX.

2. Generate *tries* using the minimum word size. The tries are a tree-based data structure that allow constant time access to the counts and occurrences for each word. These counts maintain the number of times that a word is seen in a positive or a negative trajectory. Multiple appearances within a single trajectory are only counted once. The tries also store indices into the trajectories for each word.

3. Create an empty allowable word set.

4. Examine all the basic words in the trie. Create a contingency table using the stored positive and negative counts for each word. Calculate χ^2 for each word and obtain the corresponding p-values. Add any words whose p-value is below the user specified threshold to the allowable word set.

   Although this step only identifies small rules of the form *if word A appears, then the storm is positive with probability p*, it is critical for pruning the search space for growing the larger rules.

5. Given a user specified window, recursively search for rule conjunctions using the allowable word list. Stop recursion when no rules with p values less than the user specified threshold have been created.

6. Return all rules with p values that are less than the user specified threshold.

Although generating the tries is computationally expensive, it can be done with a single pass through the data. The complexity of the trie building algorithm is $O(a^w)$ where a is the size of the alphabet and w is the trie word size. With the exponential dependence on word size in the time, we keep word size small (we used three for all results in this paper). Once it is built, the trie stores the critical information for each word, allowing the algorithm to access these counts in constant time. This efficiency allows us to search through all of the data quite quickly (less than a second for the results reported in this paper). If search time becomes an issue as we gather more storms, we can make use of the properties of χ^2 to prune in a guaranteed manner.

With larger data sets, it is unlikely that any single rule can accurately capture the characteristics of the entire data set. We have implemented an extension of the approach described above that uses boosting (Schapire 2003). In this case, the algorithm described above is used to identify the rules but the examples are weighted by their relative importance. On the first step, all examples are weighted equally. As rules are created that correctly classify some examples, the weight of those examples is decreased while the weight of the incorrectly classified examples is increased. We do not report the results of using boosting in this paper because our experimental data set was too small to sustain boosting. However, in preliminary work with simulated data, boosting showed promise and we expect this will be true with the meteorological data as well.

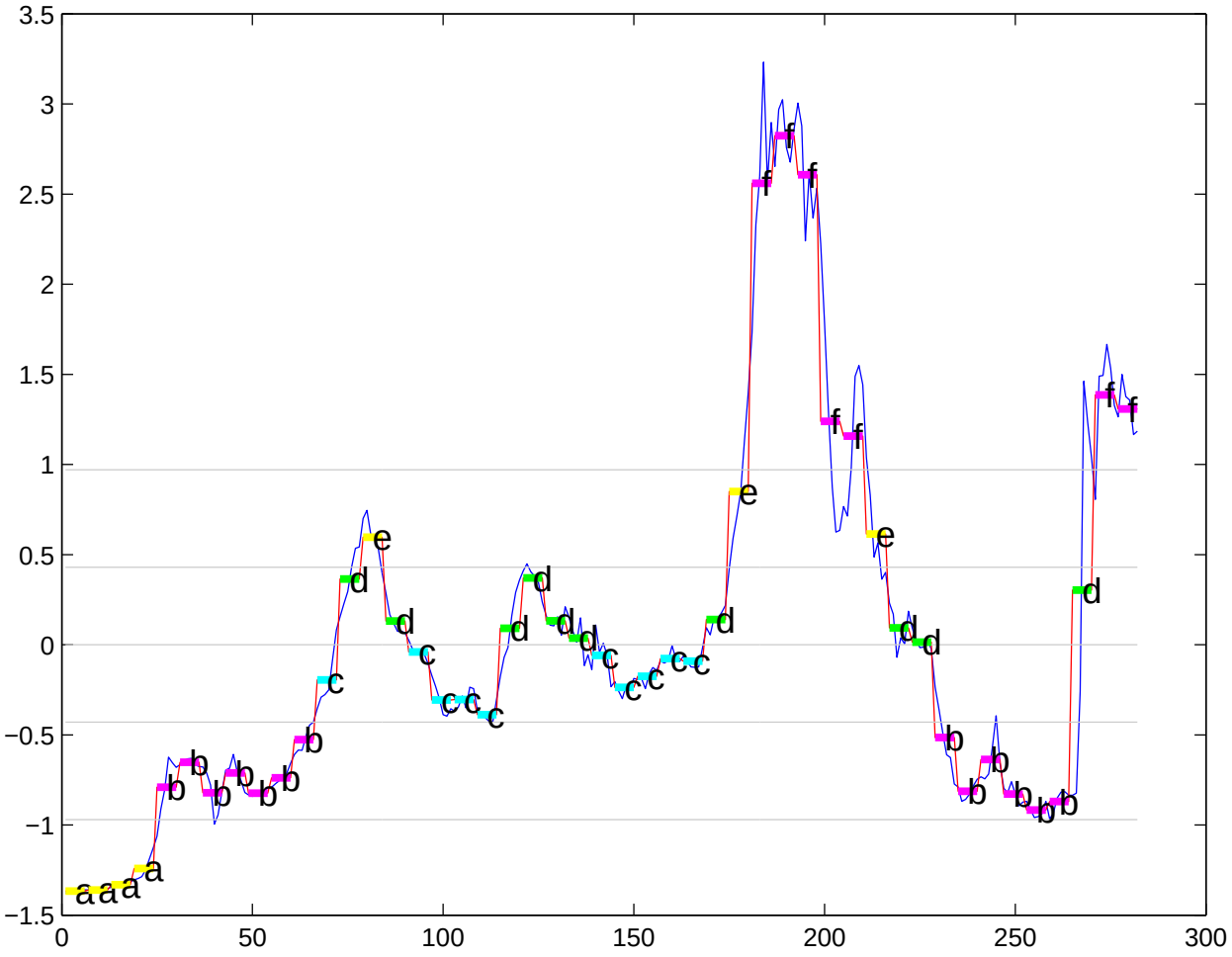

Figure 7: Maximum vertical vorticity at the surface as discretized by the SAX (Lin et al. 2003) algorithm. These values correspond to those shown in Figure 4.

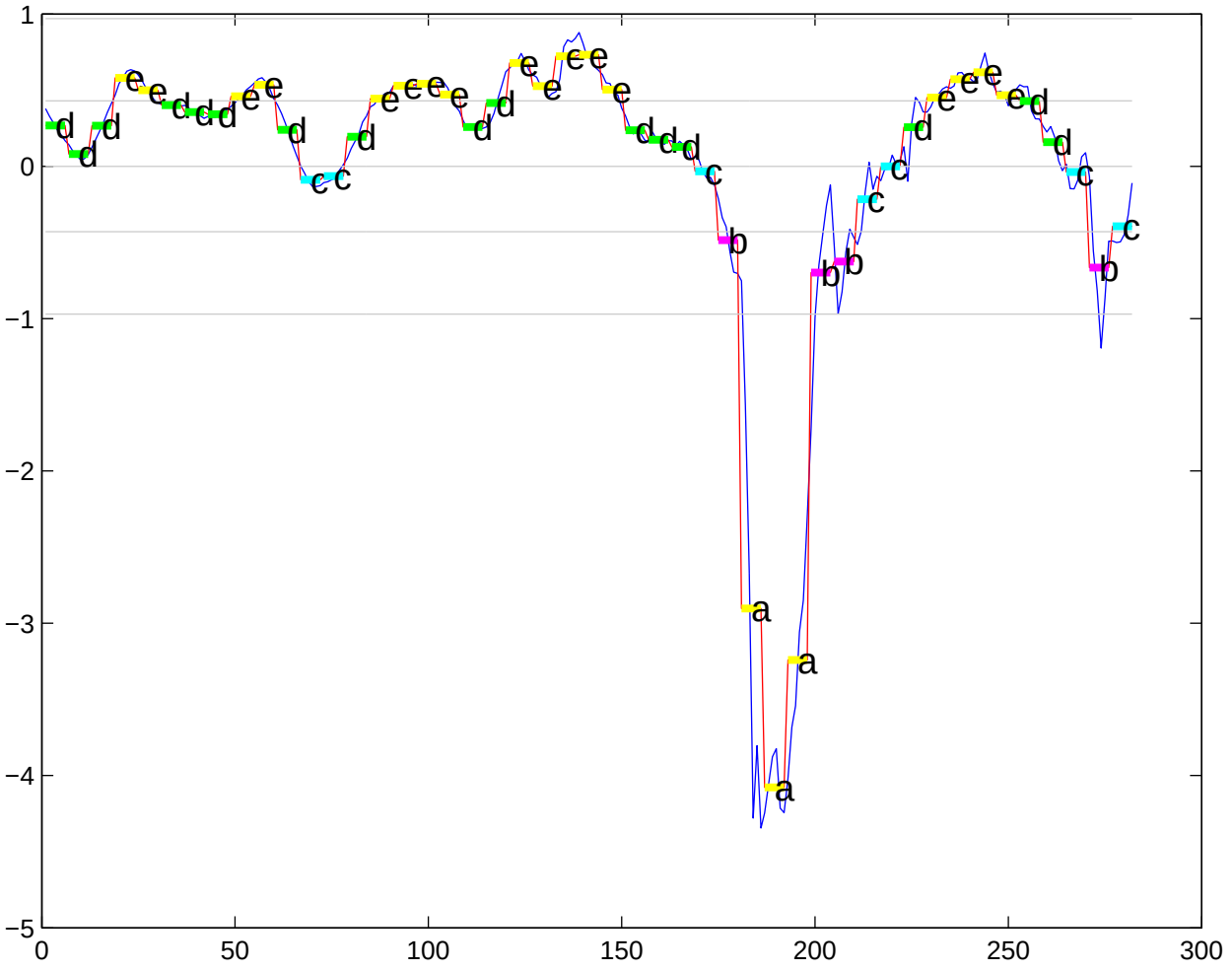

Figure 8: Maximum pressure perturbation at the surface as discretized by the SAX (Lin et al. 2003) algorithm. These values correspond to those shown in Figure 3.

4. Preliminary Results

Our preliminary results all draw from six simulation runs in ARPS, outlined in Section 2. Although our eventual goal is to use 300-500 simulations, this paper only addresses results from the first six simulations due to initial issues with storm tracking and domain translation within the ARPS model. These issues have been resolved and expanded results will be discussed at the conference and in future publications.

Given the lack of data (six simulations yielded 16 labeled storm cells), data mining on all 64 features will likely lead to overfitting. This problem is compounded by the fact that χ^2 works best with a mass of at least 20, while we have only 16. With these issues in mind, we narrowed the set of potential features down to several sets of one, two, and three features.

Figure 9 visually demonstrates the occurrences of the rule identified using the minimum pressure perturbation values at the surface. This rule is conjunctive and specifies that if the minimum pressure perturbation has been steady at a relatively high value and it is followed by a large drop (of the form shown in the graph), there is a 93.75% chance of a tornado occurring. This number comes from the accuracy on the training set. This rule correctly identified three of our four low-level rotations and correctly ruled out the presence of low-level rotations in all 12 of the negative cases. Although a large drop in pressure was identified as one of the primary characteristics in labeling a storm as positive, we did not tell the data mining algorithm what part of the pressure perturbation readings was salient. It was able to identify the drop automatically. In addition, it identified that the drop was not the only critical pattern. Instead, having a higher value for some time followed by the lower value was more predictive.

Figure 10 shows an example of a rule that is clearly overfitting on the data. In this case, the rule achieves a 75% accuracy on the training set by specifying that low-level rotations are likely to form when there is a growth in vorticity at the surface. However, this occurs in a number of cases where no low-level rotations actually occur. With more data, we believe this rule will be refined to use the pressure perturbation and will be able to correctly identify the low-level rotations.

To verify our hypothesis that we have too little data and are overfitting, we used leave-one-out cross validation. In these experiments, we trained on all but one storm and tested on the held-out storm. We repeat this cycle for each storm and report the average accuracy across all tests. If the average training set accuracy differs significantly from the average test set accuracy, then overfitting is likely. Using pressure perturbation, we see an average accuracy of 90 percent on the training set while the test set accuracy is only at 69 percent. All but one of the incorrect classifications occurred in case where a positive storm was pulled out to the test set. Since there are only four positive storms in the data set, pulling one out for cross validation leaves too few to train on. Using vorticity, we see even more dramatic evidence of overfitting with an average accuracy of 83 percent on the training set and only 44 percent on the test set. Since the default accuracy is 75 percent, the data mining is identifying noise that makes the classifier perform worse than just guessing no tornado. Given the more promising results with pressure perturbation, we are sure these results will improve with more storms.

5. Current and Future Work

Given the small number of simulations, the results reported in this paper are only the beginning of what will be a very exciting collaboration between meteorology and computer science. Our next step is to run the data mining algorithm on the full set of simulated storms, which will enable us to examine all of the maximum and minimum features and not just the few we hand picked here. In addition, we expect that the boosting version of the algorithm will perform better once we have the full set of storms.

The six simulations used for this paper generated 16 separate storm cells. With this ratio, we expect that the full 300-500 simulations may generate more than 1000 separate storm cells. Although we hand labeled the 16 storm cells used in this paper, we are developing an automated labeling technique as it is not feasible to hand label 1000 storm cells.

The longer term research on this project will focus on using a relational approach. This will enable us to make use of the high-level features, such as hook echos and gust fronts, that meteorologists can identify. Challenges that we will need to address include identifying and tracking these features, developing a dynamic relational data representation, and developing new statistical relational models for use on dynamic data. We discuss these challenges in depth in McGovern et al. (2006).

This research is a part of the Collaborative Adaptive Sensing of the Atmosphere (CASA) Engineering Research Center. This center is developing new low-powered X-band radars that will sense the low-

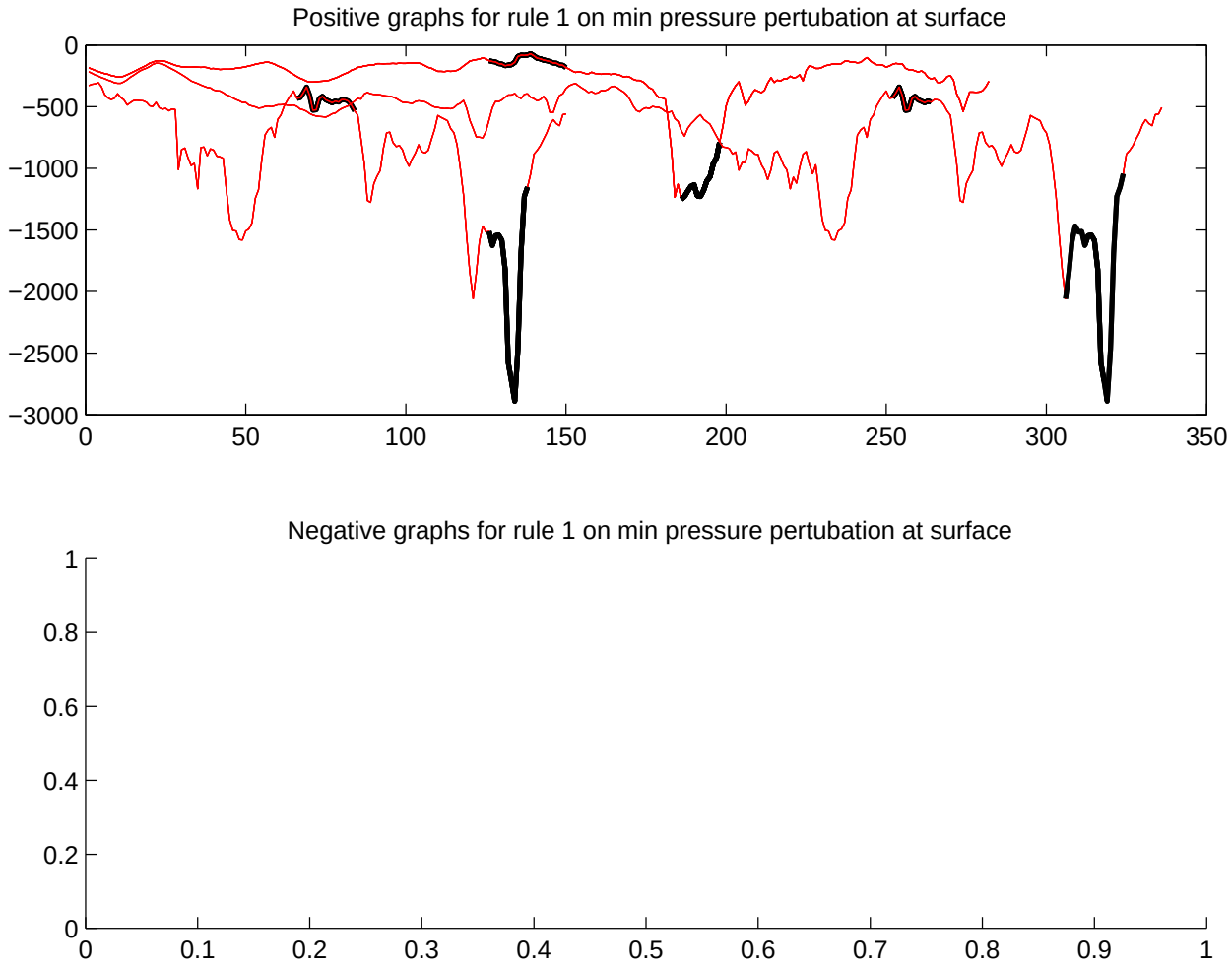

Figure 9: Occurrences of the rule on the minimum pressure perturbation at the surface. This rule is conjunctive and both halves of the rule are shown on the graph. The y-axis is the pressure perturbation and the x-axis is time.

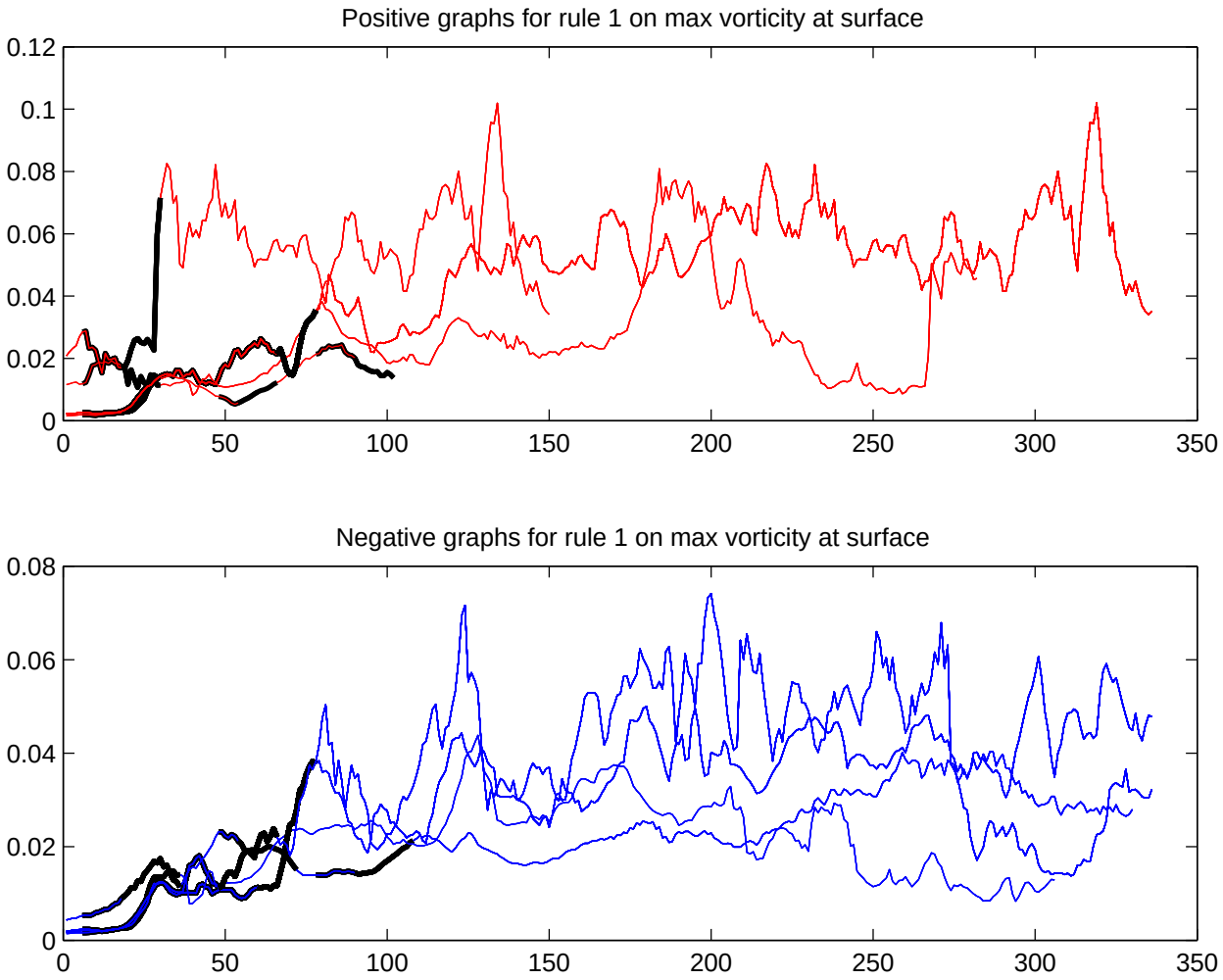

Figure 10: Occurrences of the rule on the maximum vorticity at the surface. This rule has a number of false positives and those are shown in the lower portion of the graph.

est 3 km of the atmosphere better than the current NEXRAD radars (McLaughlin et al. 2005). These radars will dynamically adjust their scanning strategies to the current weather situation. The fact that the radars will be able to scan the lower regions of the atmosphere and that they can change the scanning region every 30 seconds will create valuable data capable of observing previously undetected storm structure. This new data necessitates the development of new detection/prediction techniques.

Doppler radars currently produce reflectivity and Doppler velocity measurements in storms. The combination of successful techniques for identifying tornadoes (Mitchell et al. 1998; Stumpf et al. 1998; Lakshmanan et al. 2006) and Doppler radars have increased the true positive rate for detecting tornadoes (Simmons and Sutter 2005). Even as the true positive rate increased, the false positive rate remained about constant at about 75%. Current techniques for predicting severe weather are tied to specific characteristics of the radar systems. Each new sensing system requires the development of new algorithms for detecting hazardous events. By studying an assimilated data set, where we can measure all the fundamental meteorological quantities, we expect to significantly improve our understanding of tornado formation and to be able to improve tornado detection and prediction techniques in general, not just for a single radar.

Acknowledgement This material is based upon work supported by the National Science Foundation under Grant No. NSF/CISE/REU 0453545 and by the National Science Foundation Engineering Research Centers Program Under Cooperative Agreement EEC-0313747 to the University of Massachusetts-Amherst, and by Cooperative Agreement ATM-0331594 to the University of Oklahoma.

References

Adlerman, E. and K. K. Droegemeier, 2005: The dependence of numerically simulated cyclic mesocyclogenesis upon environmental vertical wind shear. *Mon. Wea. Rev*, **133**, 3595–3623.

Alderman, E. J., K. K. Droegemeier, and R. P. Davies-Jones, 1999: A numerical simulation of cyclic mesocyclogenesis. *Journal of Atmospheric Science*, **56**, 2045–2069.

Brown, R. A., L. R. Lemon, and D. W. Burgess, 1978: Tornado detection by pulsed doppler radar. *Monthly Weather Review*, **106**, 29–38.

Burgess, D. W., V. T. Wood, and R. A. Brown, 1982: Mesocyclone evolution statistics. *Preprints, 10th Conf. on Severe Local Storms*, Amer. Meteor. Soc, San Antonio, TX, 84–89.

Davies-Jones, R., D. Burgess, and M. Foster, 1990: Test of helicity as a tornado forecast parameter. *Preprints, 16th Conference on Severe Local Storms*, Amer. Meteor. Soc, Kananaskis Park, AB, Canada, 588–592.

Dietterich, T. G., R. H. Lathrop, and T. Lozano-Perez, 1997: Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, **89**, 31–71.

Domingos, P. and M. Pazzani, 1997: On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, **29**, 103–130.

Donaldson, J. R., Jr., 1970: Vortex signature recognition by a doppler radar. *Journal of Applied Meteorology*, **9**, 661–670.

Holte, R., 1993: Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, **11**, 63–90.

Johnson, J. T., P. L. Mackeen, A. Witt, E. D. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The storm cell identification and tracking algorithm: An enhanced wsr-88d algorithm. *Weather and Forecasting*, **13**, 263–276.

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

Lakshmanan, V., T. Smith, G. J. Stumpf, and K. Hondl, 2006: The Warning Decision Support System - Integrated Information (WDSS-II). *Weather and Forecasting*, in press.

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

McGovern, A. and D. Jensen, 2006: Chi-squared: A simpler evaluation function for multiple-instance learning. *Under Review*.

McGovern, A., A. Kruger, D. Rosendahl, and K. Droegemeier, 2006: Open problem: Dynamic relational models for improved hazardous weather

prediction, Presented at the ICML Workshop on Open Problems in Statistical Relational Learning.

McLaughlin, D. J., V. Chandrasekar, K. Droegemeier, S. Frasier, J. Kurose, F. Junyent, B. Philips, S. Cruz-Pol, and J. Colom, 2005: Distributed collaborative adaptive sensing (DCAS) for improved detection, understanding, and prediction of atmospheric hazards. *9th Symp. Integrated Obs. Assim. Systems - Atmos. Oceans, Land Surface (IOAS-AOLS)*, Amer. Meteor. Soc, San Diego, CA.

Mitchell, E. D., S. V. Vasiloff, G. J. Stumpf, A. Witt, M. D. Eilts, J. Johnson, and K. W. Thomas, 1998: The National Severe Storms Laboratory tornado detection algorithm. *Weather and Forecasting*, **13**, 352–366.

Oates, T. and P. R. Cohen, 1996: Searching for structure in multiple streams of data. *Proceedings of the Thirteenth International Conference on Machine Learning*, Morgan Kauffman, 346–354.

Pielke, R. and R. Carbone, 2002: Weather impacts, forecasts, and policy. *Bulletin of the American Meteorological Society*, **83**, 393–403.

Ray, P., B. Johnson, K. Johnson, J. Bradberry, J. Stephens, K. Wagner, R. Wilhelmson, and J. Klemp, 1981: The morphology of several tornadic storms on 20 May 1977. *J. Atmos. Sci*, **38**, 1643–1663.

Schapire, R. E., 2003: The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*, D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, and B. Yu, eds., Springer.

Simmons, K. M. and D. Sutter, 2005: WSR-88D radar, tornado warnings, and tornado casualties. *Weather and Forecasting*, **20**, 301–310.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory mesocyclone detection algorithm for the WSR-88D. *Weather and Forecasting*, **13**, 304–326.

Tong, M. and M. Xue, 2005: Ensemble kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS experiments. *Mon. Wea. Rev.*, **133**, 1789–1807.

Weisman, M. and J. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Monthly Weather Review*, **110**, 504–520.

Wood, V. T. and R. A. Brown, 1997: Effects of radar sampling on single-Doppler velocity signatures of mesocyclones and tornadoes. *Weather and Forecasting*, **12**, 928–938.

Xue, M., K. K. Droegemeier, and V. Wong, 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., K. K. Droegemeier, V. Wong, A. Shapiro, K. Brewster, F. Carr, D. Weber, Y. Liu, and D. Wang, 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**, 134–165.

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

Zaki, M. J., 2001: Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, **42**, 31–60, special issue on unsupervised learning.