J2.3 USING SPATIOTEMPORAL RELATIONAL DATA MINING TO IDENTIFY THE KEY PARAMETERS FOR ANTICIPATING ROTATION INITIATION IN SIMULATED SUPERCELL THUNDERSTORMS

Nathan C. Hiers
School of Meteorology
University of Oklahoma
Norman, OK
Nathan.C.Hiers-1@ou.edu

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

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

This project expands upon the work outlined in McGovern et al (2007) and Rosendahl (2008) in two important ways: changing the definition of storm cells and extending the data mining techniques. The overall goal of each of these projects is to identify key parameters for tornadogenesis by applying data mining techniques to a full set of meteorological fields. This significantly differs from the traditional approach of analyzing radar data containing only reflectivity and the radial wind component.

Tornadoes kill an average of 33.7 people annually[1] and cause, in part, the $13B (Pielke and Carbone, 2002) of economic impact due to mesoscale storms annually in the United States. Although tornado forecasting has improved significantly with the introduction of Doppler radars, there is still room for improvement. The probability of detection (POD) is currently about 75%, and the false alarm ratio (FAR) is currently around 78%. Improving these numbers would help to mitigate the loss of life and property.

The current primary instrumentation for observing mesoscale convection is the WSR-88D NEXRAD radar system. However, due to the earth's curvature, current WSR-88D radars are unable to see 76% of the volume between the ground and two kilometers (Brotzge et al., 2006), which is where a tornado forms. Another constraint of radars is that the current WSR-88Ds can only observe reflectivity and radial velocity. In order to gain an understanding of the dynamic and physical properties that are causing tornadogenesis, a rich dataset that includes many derived and observable meteorological quantities is needed. To address this issue, we use simulated data from the Advanced Regional Prediction System (ARPS), which is a top weather forecasting system for mesoscale data (Xue et al. 2000, 2001, 2003). When assimilated data is available in larger quantities, we can study that data as well.

We introduce a spatiotemporal model that will produce human readable results. The spatiotemporal relational probability tree (SRPT) is an extension of the relational probability tree (Neville et al, 2003) to spatiotemporal data. Both of these models

[1] http://www.nssl.noaa.gov/users/brooks/public_html/deathtrivia/

```
Identify-and-track-storms(time t)

    $w_{max}$ ← domain-wide maximum vertical wind velocity (w) at 4-8 km height
    $refl_{max}$ ← domain-wide maximum reflectivity at 4-8km height
    If $w_{max} < 15$ m $s^{-1}$ or $refl_{max} < 55$ dbZ
        return
    Otherwise
        Measure Euclidean distance from location of $w_{max}$ or $refl_{max}$ to the storms identified at time t-1
        If the closest previous storm shares the most common points, 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 storm attributes.
            If $w_{max}$ within a storm drops below 10 m $s^{-1}$, stop tracking this storm
```

Table 1: Storm identification and tracking algorithm.

are built on the successful decision tree model (Rusell & Norvig, 2002). Like all weather phenomenon, tornadoes are the result of many processes that change in both space and time, and the only way to understand the processes is to be able to model and analyze them in both space and time. The SRPTs are able to analyze both spatial and temporal data, making them an ideal candidate to use for exploring tornadogenesis. Using the SRPTs should give meteorologists a better understanding of the important dynamic processes that result in tornadogenesis, leading to better understanding and prediction of tornadoes as well as mitigation of the loss of life and property.

2. Meteorological Data

Our data was created using over 250 ARPS simulations, each lasting 3 hours. The simulations use supercell-favorable conditions (Rosendahl, 2008). The horizontal grid spacing is 500m with a stretched hyperbolic tangent in the vertical. Each grid point contains derived and fundamental quantities that are important for understanding storm morphology and tornadogenesis. Each simulation can contain multiple storms, and each storm can be abstractly broken down into high level features that are dynamically important to its morphology. Unfortunately, the grid spacing currently being used is too large for a definitive resolution of a tornado, so the storms exhibiting attributes commonly associated with a tornado will be called 'strong low level rotations' in this paper.

2a. Data Extraction

Any given storm simulation may generate several separate storms. We define a storm based on a combination of the maximum updraft and reflectivity in the 4km through 8km section of the simulation. The algorithm for identifying and tracking individual storms is given in Table 1.

Once a single storm is being tracked, new storms can be identified if their maximum reflectivity or updraft goes beyond a determined threshold. Using a weighted Voronoi diagram (Aurenhammer, 1991), the area a storm receives is based upon the maximum vertical wind speed of the competing storms. For example, if there is a storm with a maximum updraft of 30 m/s and another storm with a maximum updraft of 15 m/s and each storm begins with roughly the same area, as they expand, the 30 m/s storm would get roughly twice the area that the 15 m/s storm would get. An example of this is shown in Figure 1.

2b. Identifying storm features

Our eventual goal is to examine the data using a large array of high level features, such as those shown in Figure 2c, which more accurately describe current important features in supercell and tornado development. Automated identification and tracking of high level features is difficult and 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

High Level Features	Relations	Attributes
Updraft (w > 15 m/s)	Contains	Volume
Downdraft (w < -5 m/s)	Equals	Max/Min of variable
Baroclinic generation (bg > 1e-6 s^{-2})	Contained By	Start time
Hail (qh > 0.01 kg/kg)	Overlaps	End time
Pressure Perturbation (pp < -900 pa)		Standard deviation
Rain (qr > 0.0025 kg/kg)		Median
Tilting (tilt > 0.0001 s^{-2})		Mean
Vertical stretching (vs > 0.0001 s^{-2})		Volume
Vertical vorticity (vv > 0.02 s^{-1})		Base height
		Ceiling height
		Thickness
		Horizontal composite area

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 fundamental and derived meteorological quantities using thresholds derived from the literature. These quantities are listed in Table 2. These represent some of 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.

We take a relational view of the data where we identify discrete regions, which we call high-level features, of interest based on current theories of tornadogenesis. In addition, we identify possibly relationships between each high level feature and we measure attributes on each feature. Attributes are a part of the high level features, and describe meteorological quantities and statistical distributions of quantities that exist within that feature. Relations describe how multiple high level features interact with one another. All of this information is summarized in Table 2. In this experiment, six relations, nine high level features, and an average of thirteen attributes per high level features are measured.

Figure 2a shows a top down view of the reflectivity at 4km in the simulation. Figure 2b is an example of the same storm at the same time step using high level features such as updrafts and downdrafts. Figure 2c was the inspiration for choosing many of the high level features. This figure is one of the canonical figures in current theories of tornado development.

3. SPATIOTEMPORAL RELATIONAL PROBABILITY TREES

A decision tree can be viewed as a series of questions asked about data with respect to a classification task. For example, our task is predicting strong low-level rotations and a possible question may be “is there an updraft?” The data is split based on the answer to this question and a new tree is built recursively using the subsets of data. The tree growth process continues until either all the data agrees at the leaf or there is no statistically significant question remaining to split the data. A standard decision tree contains a single answer at a leaf node, usually the modal answer of the data at that leaf. A probability tree instead yields a probability of each class (in our case, probability of a strong low level rotation occurring or of it not occurring).

SRPTs come from relational data, which are comprised of high level features, relations, and attributes. The tree can differentiate the data using the objects, relationships, attributes, and temporal questions based on the objects or relationships. For example, the tree can ask if an updraft exists or how long it has existed. The questions are chosen based on their χ^2 value. Only questions that have statistically significant values of χ^2 are chosen.

To facilitate better understanding of the algorithm, an example follows: Assume there is a single positive storm that is being run through the SRPT algorithm. Assume that in this positive storm, the high level features that have caused a rotation initiation is an updraft that has lasted longer than 23 minutes and a downdraft that has lasted longer than 10 minutes. The SRPT algorithm would look through a user-defined number of possible inquires to see if the storm had that particular feature being sought by the inquiry and if that storm was positive. For this example, assume the first question that has both a positive inquiry and a positive label would be "Does the storm have an updraft lasting longer than 23 minutes?" If no other question produces a better χ^2 value, then it will be the top question, or 'root' of the SRPT. The SRPT algorithm would then search for a new question using only the subset of storms that answered the last question positive or 'yes'. The positive storm also had an attribute involving the downdraft lasting for longer than 10 minutes, so it would be the next question chosen in the SRPT and would be positioned below the root. As there are no other attributes for this storm that are important to rotation initiation, the SRPT would be complete.

4. PRELIMINARY RESULTS

The SRPT generated during this experiment, as seen in Figure 3, shows that the most significant question is the existence of stretching vertical vorticity for at least 15 minutes. The yes branch then uses the existence of baroclinic generation of vertical vorticity as its most significant question. After baroclinic generation of vertical vorticity, the existence of hail for various times is used for the subsequent branches. Going down the 'no' branch for the root node, the existence of stretching of vertical vorticity for various times is seen in the subsequent child and grandchild nodes. The probabilities range from a maximum of 1 to a minimum of 0. The p-value threshold for significance was set to 3e-03.

The six-fold cross validation produced an Area Under the Curve (AUC) of 0.77 with a standard deviation of 0.14. AUC is a measure of performance that does not rely on the underlying class distributions (Bradley, 1997 and Foster and Provost, 1997), something that is critical for understanding rare events such as thunderstorms that develop strong low level rotations. AUC is calculated as the area under the Receiver Operating Curve, a plot that shows how the false positive ratio and true positive ratio varies as the class label cutoff varies. An AUC of 1 means the algorithm is performing perfectly while an AUC of 0.5 means the algorithm is performing as well as randomly guessing. Anything below 0.5 is worse than random. A total of 1023 storms were used in this experiment, with 38 of them being classified as having a strong low level rotation.

Our hypothesis, that using the SRPT would help gain valuable insights into possible causes for rotation initiation, was supported by the results in that both stretching of vertical vorticity was found to be the most significant question, and that this type of data was able to be properly investigated using a SRPT. Those familiar with the current theories of supercell storm evolution may be surprised that neither updraft nor downdraft were found in the SRPT. This can be explained for two reasons; one being that the limited set of questions being asked in this experiment had no inquiry of the attributes of the high level features except for time in existence. The other explanation is that supercell thunderstorms are defined by having a rotating updraft, and are expected to have at least one downdraft. Since all storms used in this experiment were of a supercellular nature, all would be expected to have a sustained updraft and downdraft, meaning that the presence or absence of strong low-level rotation couldn't be differentiated using these types of questions. The AUC value being above 0.50 shows that the SRPT was able to perform better than randomly guessing. The model itself shows that SRPT's can be successfully used to find meaningful knowledge about spatiotemporal data sets that potentially have higher order objects related to strong low-level rotation that are currently not well understood.

One note of particular interest in the SRPT was the 'no' branch beginning with the root node. All subsequent significant questions along the 'no' branch dealt with the stretching of vertical vorticity for varying time intervals, but each time interval became smaller the further down the tree it was. This could be signs of an implicit cyclical pattern in the relative strength of a supercell thunderstorm. A supercell thunderstorm could begin as strong, weaken and reorganize, and strengthen again in order to produce a strong low-level rotation.

5. CURRENT AND FUTURE WORK

This work is preliminary and we are expanding on it in several directions. Meteorologically, the number of high level features and relations that we used in this work is limited. We are actively identifying potential new features from the literature. Second, the resolution of our simulations, 500m, is too coarse to resolve a tornado. Making the resolution smaller requires an exponential increase in both computing time and storage space. In the near future, we hope to develop a set of simulations at 250m resolution but 25m resolution will wait until technology is capable of generating such a simulation in a reasonable amount of time (Xue et al., 2007).

The SRPT algorithm presented here is also preliminary and we are actively expanding the approach. We are working to address questions of bias in the algorithm when spatial and temporal autocorrelation exist, something known to be common in meteorology. Although autocorrelation can be used to improve predictions by knowing that the temperature at one location is autocorrelated to a nearby location, it can introduce bias into the decision tree model where the model chooses inappropriate features in the tree (Neville et al., 2003). In addition to handling autocorrelation, we are also expanding the number of distinctions that the tree can ask, both spatially and temporally.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. NSF/CISE/REU 0453545 and a seed grant from the University of Oklahoma's College of Engineering.

REFERENCES

Aurenhammer, F., 1991: Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys*, **23**(3),345-405, 1991

Bluestein, H. B., 1993: *Synoptic-Dynamic Meteorology in Mid latitudes: Volume II: Observations and Theory of Weather Systems*. Oxford University Press.

Brotzge, J., Droegemeier, K. K., and McLaughlin, D. J., 2006: Collaborative adaptive sensing of the atmosphere (CASA): New radar system for improving analysis and forecasting of surface weather conditions. *Journal of the Transportation Research Board*, **1948**,145–151.

Bradley, A. P., 1997: The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**, 1145–1159.

Davies-Jones, R. P., 1986: Tornado dynamics. In Kessler, E., editor, *Thunderstorm Morphology and Dynamics*, chapter 10, pages 197–236. University of Oklahoma Press.

Lemon, L. R. and Doswell, III, C. A., 1979: Severe thunderstorm evolution and mesocyclone structure as related to tornadogenesis. *Monthly Weather Review*, 107:1184–1197.

McGovern, A., Rosendahl, D. H., Kruger, A., Beaton, M. G., Brown, R. A., and Droegemeier, K. K., 2007: Anticipating the formation of tornadoes through data mining. *Preprints, Fifth Conference on Artificial Intelligence and its Applications to Environmental Sciences, American Meteorological Society*, San Antonio, TX, CD-ROM, 4.1.

Neville, J. and Jensen, D. and Friedland, L. and Hay, M., 2003: Learning Relational Probability Trees. Proceedings, *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 625–630.

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

Provost, F. and Fawcett, T., 1997: Analysis and Visualization of Classifier Performance:

Comparison under Imprecise Class and Cost Distributions. Proceedings, *Third International Conference on Knowledge Discovery and Data Mining*, 43-48.

Rosendahl, D. H., 2008: Identifying Precursors to Strong Low-Level Rotation within Numerically Simulated Supercell Thunderstorms: A Data Mining Approach. Master's thesis, University of Oklahoma.

Russell, S. and Norvig, P., 2002: *Artificial Intelligence: A Modern Approach.* Prentice-Hall, Second Edition.

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

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

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

Xue, M., K.K. Droegemeier, and D. Weber, 2007: Numerical Prediction of High-Impact Local Weather: A driver for Petascale Computing. Chapter 18 in *Petascale Computing: Algorithms and Applications*, Chapman and Hall/CRC Press. In Press.

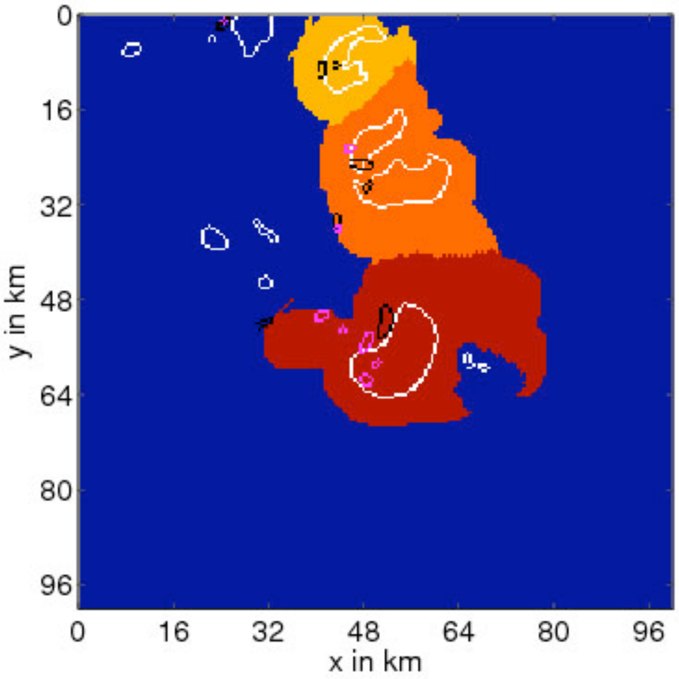
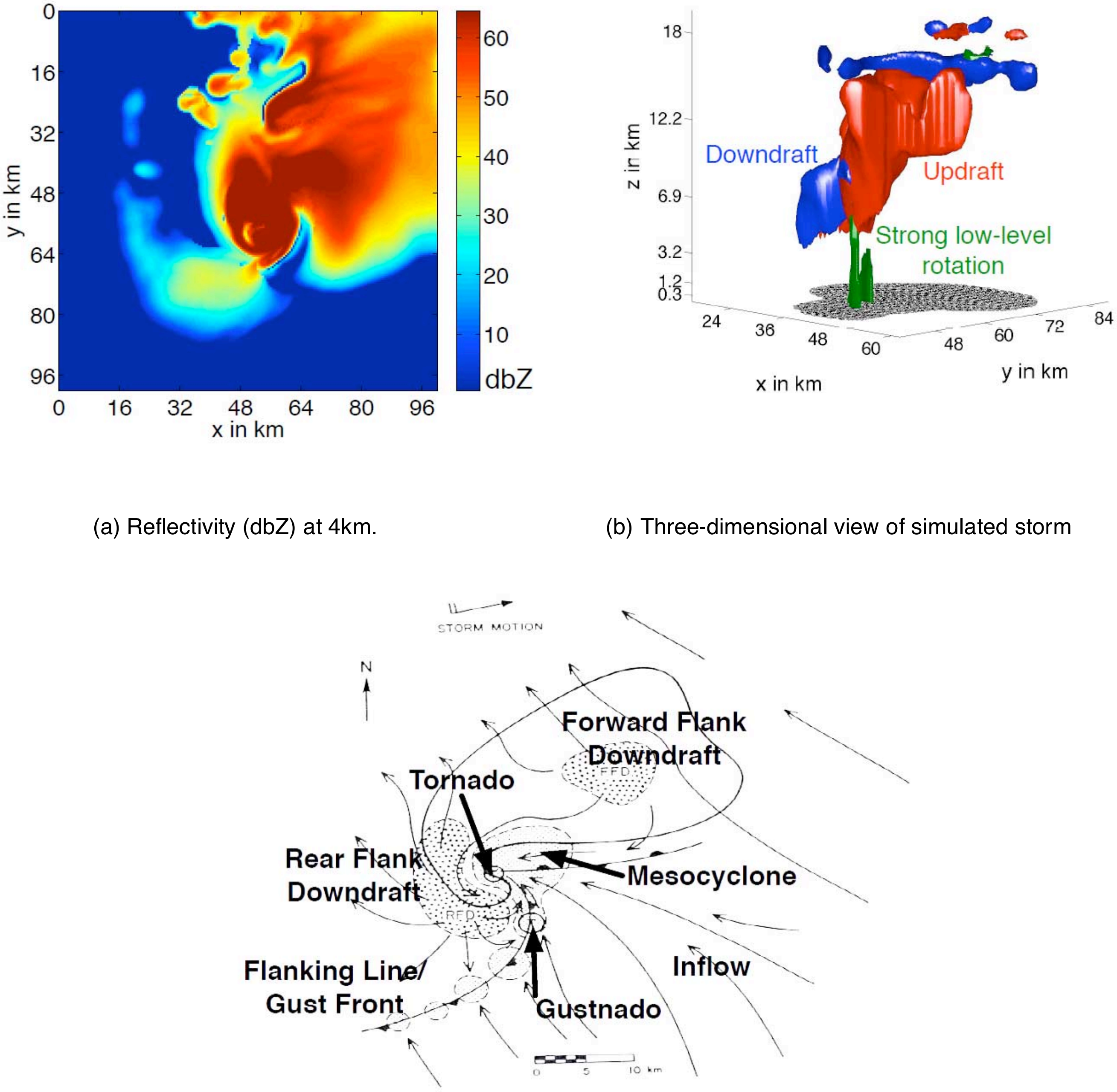


Figure 1: An example of how the storm identification algorithm tracks and splits storms. Individual storms are colored in shades of red in this image. The central updraft is shown with a white outline.



(a) Reflectivity (dbZ) at 4km.

(b) Three-dimensional view of simulated storm

(c) Structure of a supercell

Figure 2: These figures are best viewed in color. A: Reflectivity at 4km altitude B: High-level three-dimensional features based on updrafts, downdrafts, and vertical vorticity. C: Structure of a classic supercell (adapted from Lemon and Doswell, III, 1979; Davies-Jones, 1986; Bluestein, 1993).

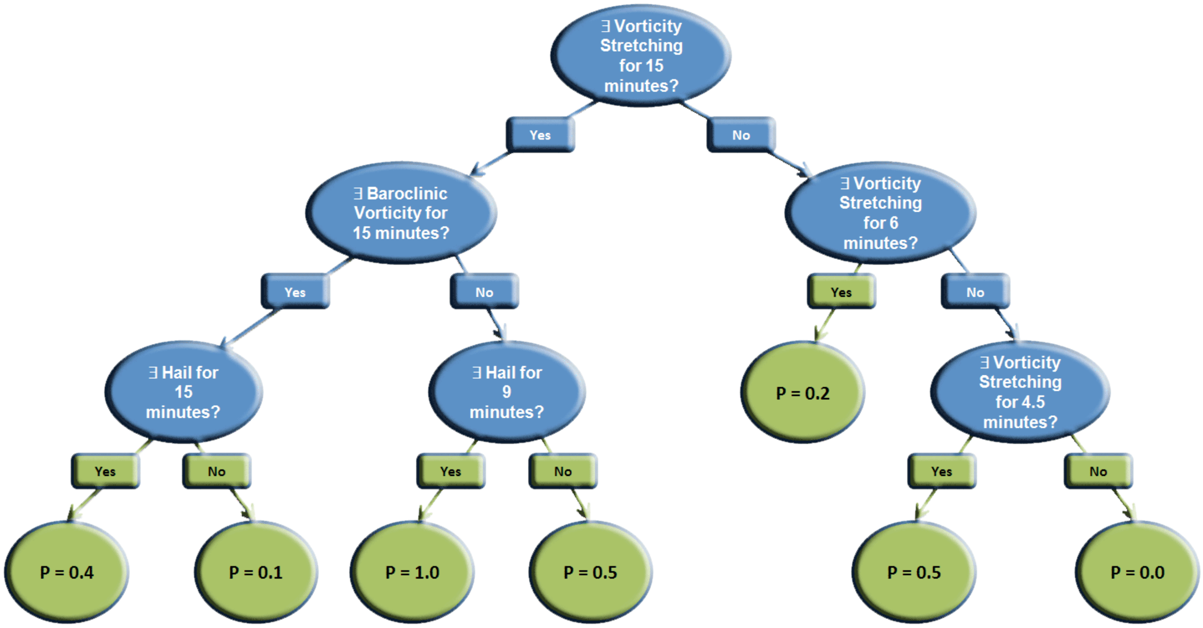

∃ Vorticity Stretching for 15 minutes?
Yes
No
∃ Baroclinic Vorticity for 15 minutes?
∃ Vorticity Stretching for 6 minutes?
Yes
No
Yes
No
∃ Hail for 15 minutes?
∃ Hail for 9 minutes?
P = 0.2
∃ Vorticity Stretching for 4.5 minutes?
Yes
No
Yes
No
Yes
No
P = 0.4
P = 0.1
P = 1.0
P = 0.5
P = 0.5
P = 0.0

Figure 3: The SRPT generated during our preliminary experiment.