Classification of Convective Areas Using Decision Trees

DAVID JOHN GAGNE II

School of Meteorology, The University of Oklahoma, Norman, Oklahoma

AMY MCGOVERN

School of Computer Science, The University of Oklahoma, Norman, Oklahoma

JERRY BROTZGE

Center for Analysis and Prediction of Storms, The University of Oklahoma, Norman, Oklahoma

(Manuscript received 12 August 2008, in final form 22 December 2008)

ABSTRACT

This paper presents an automated approach for classifying storm type from weather radar reflectivity using decision trees. Recent research indicates a strong relationship between storm type (morphology) and severe weather, and such information can aid in the warning process. Furthermore, new adaptive sensing tools, such as the Center for Collaborative Adaptive Sensing of the Atmosphere's (CASA's) weather radar, can make use of storm-type information in real time. Given the volume of weather radar data from those tools, manual classification of storms is not possible when dealing with real-time data streams. An automated system can more quickly and efficiently sort through real-time data streams and return value-added output in a form that can be more easily manipulated and understood. The method of storm classification in this paper combines two machine learning techniques: *K*-means clustering and decision trees. *K*-means segments the reflectivity data into clusters, and decision trees classify each cluster. The *K* means was used to separate isolated cells from linear systems. Each cell received labels such as ''isolated pulse,'' ''isolated strong,'' or ''multicellular.'' Linear systems were labeled as ''trailing stratiform,'' ''leading stratiform,'' and ''parallel stratiform.'' The classification scheme was tested using both simulated and observed storms. The simulated training and test datasets came from the Advanced Regional Prediction System (ARPS) simulated reflectivity data, and observed data were collected from composite reflectivity mosaics from the CASA Integrative Project One (IP1) network. The observations from the CASA network showed that the classification scheme is now ready for operational use.

1. Introduction

We introduce an automated human readable approach for classifying storm type based on weather radar reflectivity and velocity, with the primary goal of creating a technique that can be used to improve radar scanning strategies. The technique is trained on simulated storms and then tested and validated on both simulated model output and observed storm data. Automated techniques for classification based on storm morphology have been previously investigated by other researchers (e.g., Schiesser et al. 1995; Alexiuk et al. 1999; Anagnostou 2004; Baldwin et al. 2005). The principal contribution of this approach is the development of a technique that is more easily understood by human weather forecasters, enabling it to be more widely integrated into operations.

There are two reasons why automated storm classification is needed. First, the type and extent of severe weather produced by a storm is shown to be at least moderately correlated with specific storm morphology. Gallus et al. (2008) classified nine distinct classes of storms based upon their radar reflectivity and identified significant differences in the number and types of severe weather produced by each storm class. Guillot et al. (2008) found that tornado and severe thunderstorm warning lead times are greater for strong supercell and linear storms than for weaker pulse and less organized

Corresponding author address: David John Gagne II, The University of Oklahoma, 120 David L. Boren Blvd., Suite 5900, Norman, OK 73072.
E-mail: djgagne@ou.edu

DOI: 10.1175/2008JTECHA1205.1

systems. A model output statistics (MOS)-type climatology can be developed and applied in real time to severe weather warning operations if the relationships between storm type and morphology and the severe weather produced from them are known.

Second, automated storm classification is needed because of the increasing development of dynamic, data-driven application systems (Darema 2004). These observing networks are driven automatically by the data they observe, often in real time. One such example is the Center for the Collaborative Adaptive Sensing of the Atmosphere (CASA; Brotzge et al. 2006). CASA operates a network of radars that scan collaboratively and *adaptively* as a function of end-user needs. This automated radar scanning strategy, known as distributed collaborative adaptive sensing (DCAS), adapts in real time (within one minute) to what is being observed. DCAS operates automatically and interdependently, based upon a complex set of quality control functions and end-user requirements. One quality control function requires that neighboring radars operate in tandem to minimize the effect of attenuation; a second function defines certain integrated scanning for maximizing dual-Doppler scanning. Experience with this system has shown that the type and orientation of storms can have an important effect on the ability of the system to scan storms properly (Brotzge et al. 2006). For example, isolated cells require narrow sector scanning by multiple radars, whereas linear squall line events require broad scans along the major axis of the feature. Thus, increased knowledge of a storm's morphology could greatly enhance our ability to observe it.

Decision trees (Quinlan 1986) were chosen as the primary approach for classification for two reasons. One of the biggest factors is the intuitive understanding of the model. Many other machine learning models, such as neural networks, are difficult to interpret, particularly by noncomputer scientists. As Fig. 1 illustrates with its sample decision tree, the relationships between attributes can be easily shown and converted into other forms, such as rules. The other reason is that decision trees can identify the most important attributes from a dataset and ignore the less important ones. This ability to be selective adds to their human readability and can yield improved understanding about what is most important in a dataset. However, because decision trees make their decisions based on one attribute at every point in their structure, they are limited to identifying only linear relationships within a dataset. Other approaches may be difficult to interpret but can possibly lead to improved results. With this in mind, the tree-based results were compared to several other standard machine learning approaches.

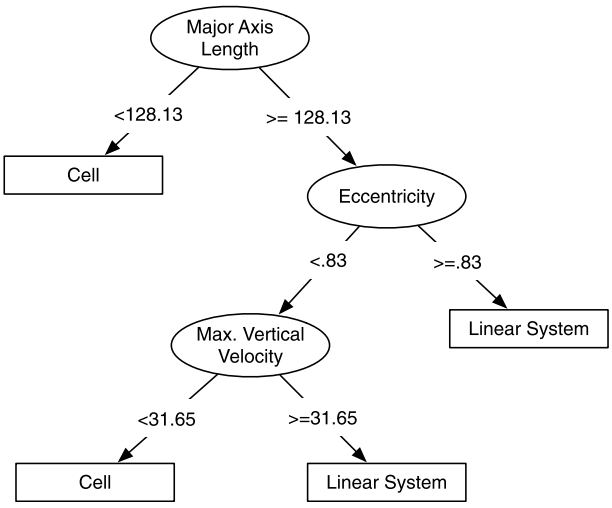


FIG. 1. An example of the general-type decision tree.

The decision-tree approach in this study is most similar to that of Rigo and Llasat (2004). They combined aspects of the storm cell identification and tracking (SCIT) algorithm (Johnson et al. 1998), which tracks the location of the storm center, and the convective–stratiform differentiation algorithm (Steiner et al. 1995) to serve as the basis for a structural classification system. Rigo and Llasat (2004) used the storm areas and statistics about them as guides when classifying each image. This study instead used those storm areas to train an algorithm to automatically classify the storms in its dataset. They also assigned only one storm type per image, even if more than one storm appeared. With the algorithm in this paper, multiple storm types were often found and labeled within the dataset. This study also incorporates two machine learning techniques—*K*-means clustering and decision trees—to identify and classify storm areas. The *K*-means clustering section of the project is derived from a technique used for image segmentation that had previously been applied to radar reflectivity (Lakshmanan 2001; McQueen 1967).

2. Data and methodology

a. Data sources

The data used for this study came from both simulation and observed radar reflectivity. The simulations were generated by the Advanced Regional Prediction System (ARPS) (Xue et al. 2000, 2001, 2003), a storm-scale model with numerical weather prediction and data assimilation features. This model provided more than 250 simulations of mesoscale storms generated in an environment favorable for supercell thunderstorm development (Rosendahl 2008). For this project, simulated

radar reflectivity was derived from model parameters using equations described in Rosendahl (2008). These are the same equations used in ARPS. Those equations converted the mixing ratio values from rain, snow, and hail into radar reflectivity and added them together to produce the full radar reflectivity value. The radar reflectivity was examined at a height of 4 km above the ground of the model. Four kilometers was chosen as the level to be analyzed, because it incorporates a balance of low and midlevel storm features. The simulations used a 100 km $\times$ 100 km grid with 500-m spacing in the horizontal and 50 height levels having increased spacing with increasing height. The center of the grid was adjusted following each time step to keep the moving storms at the center of the grid. The reflectivity values of the simulated data tended to be higher than the observed values of reflectivity because there was no attenuation.

The observed weather radar data came from the CASA Integrative Project One (IP1) test bed, a group of four small X-band Doppler radars located in southwest Oklahoma (Brotzge et al. 2006). The reflectivity measurements from all of the radars were mapped to a single 120 km $\times$ 120 km Cartesian grid with 500-m grid spacing to fit the image as closely as possible to the ARPS simulated data.

b. K-*means clustering*

To identify individual storm regions, each reflectivity image was first divided into similar, congruous regions using the *K*-means clustering algorithm. Doing this required the minimization of the weighted Euclidean distance, as derived from Lakshmanan's (2001) image segmentation algorithm:

$$d_e = \lambda |r_m - r_p| + (1 - \lambda)\sqrt{(x_m - x_p)^2 + (y_m - y_p)^2}, \quad (1)$$

where λ weights the differences in reflectivity versus Cartesian coordinates, r represents the reflectivity value (dBZ) at a certain point, x and y are the coordinates of that point, m designates variables derived from the cluster means, and p designates variables derived from a point in the reflectivity image. The first term of the equation finds the distance of each point from the reflectivity means, while the second term finds the distance between the selected point and the coordinates of the reflectivity means in the image. For this work, a λ of 0.6 was chosen through empirical testing. *K*-means clustering uses this similarity metric to find geographically similar areas with similar reflectivity values. The output of *K*-means clustering is processed by (i) breaking those that are not contiguous into separate clusters and (ii) removing clusters whose area is less than 4 km^2, thus removing regions too small to be really considered a storm.

After clustering, the clusters are divided into *convective*, *stratiform*, and *low reflectivity areas*. Since each cluster contains a range of reflectivity values, simple thresholding could not be used to differentiate the clusters effectively. Instead, a metric that takes into account what range of values are found in the cluster is used. If at least 70% of the cluster contained reflectivity between 20 and 40 dBZ, then it was considered to be stratiform. Forty dBZ was chosen as the upper limit of the threshold for the stratiform area because it was used as the threshold in the convective–stratiform separation algorithm from Steiner et al. (1995). The 70% area threshold was used because not all of the radar reflectivity values in a cluster that could be considered stratiform would be between 20 and 40 dBZ, so it allows for a margin of error with the thresholds. Otherwise, if less than 10% of the cluster contained reflectivity greater than 80% of the maximum reflectivity, then the cluster was considered to be a low reflectivity area. If the cluster fit neither of those categories, then it was considered to be convective. Although this system does not necessarily detect every area of convection in a particular radar reflectivity image, it does successfully remove nonconvective areas and does locate the vast majority of the remaining convective areas.

c. *Attribute calculation*

Given reflectivity data at every grid point in the domain, the number of possible ways to examine the data is large. Humans examine the data using a variety of visual features and combine these with their experiences. Our decision-tree attributes are shown in Table 1. The morphological attributes come from fitting the storm region to an ellipse, which can be fit to areas that have the more circular profile of the cells and the more elongated profile of the linear systems. These attributes include the eccentricity of the ellipse, the length of the major and minor axes, the orientation of the major axis relative to the positive x axis, and the area of the cluster. The reflectivity attributes include the maximum, minimum, mean, standard deviation, and the range of the reflectivity values within each storm region.

The wind attributes sample the wind field from the simulation runs within the storm regions and calculate statistics similar to those for reflectivity for all of the fields within the regions. Wind speed was the magnitude of the horizontal wind vector at 4 km, and vertical velocity was the magnitude of the vertical wind vector at

TABLE 1. Attribute sets used in the decision trees.

Morphological	Reflectivity	Wind	Control
Eccentricity	Max	Max speed	Mean stratiform distance
Major axis length	Min	Min speed	Storm direction
Minor axis length	Mean	Mean speed	
Orientation	Std dev	Max vertical velocity	
Equivalent diameter	Range	Min vertical velocity	
Area		Mean vertical velocity	

4 km from the model output. Maximum, minimum, and mean were calculated for both fields.

One attribute designed to differentiate between the leading, trailing, and parallel stratiform rain areas is the *mean stratiform distance*. To calculate this value, the equation of the major axis line was found and solved to find the offset value d_s from the line through the midpoint of the cluster centroids:

$$d_s = |m_l|(x_s - x_c) + (y_c - y_s), \quad (2)$$

where m_l represents the slope of the line, the s values represent centroid coordinates of the stratiform cluster and the c values represent the same for the convective clusters. A diagram of how it is calculated is shown (Fig. 2). High positive values correlate more with the mean stratiform area lying in front of the convective area, values near zero indicate that the mean stratiform center lies almost along the line, and high negative values indicate the mean stratiform lies behind the line. To calculate the direction of storm motion, the location of the storm area's centroid was compared with the centroid from the previous time step and the center of the ARPS grid between time steps. The angle of the resulting vector was the storm's direction. For cases when a new storm appeared, a storm motion of 0 was calculated. Storm mergers and splits were handled by finding the storm area in the previous time step that most closely overlapped the current storm area. The procedure matched up the storms between time steps correctly in most cases.

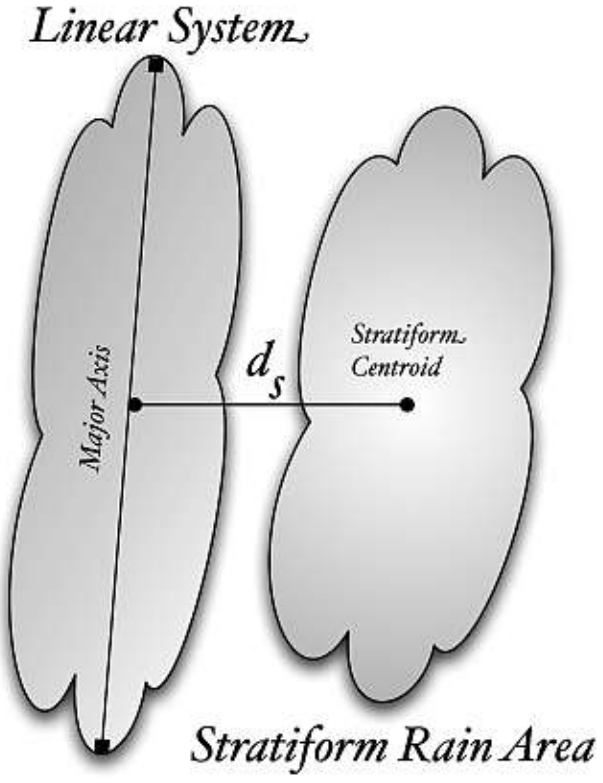


FIG. 2. Mean stratiform distance is an attribute that represents the distance been the centroid of the stratiform rain area and the major axis of the convective area.

d. Hand classification

Identified ''clusters'' are assumed to represent storm regions, and they are labeled with a hierarchal classification system that combines types developed by Parker and Johnson (2000) and Rigo and Llasat (2004). At the highest level, convective areas are divided into *cells* and *linear systems* (Fig. 3), which are referred to as the *general-type* classification scheme. Within the cell category, the storms are divided into ''isolated pulse cells,'' ''isolated strong cells,'' and ''multicells.'' In the linear system category, they are divided into ''leading stratiform,'' ''trailing stratiform,'' and ''parallel stratiform'' depending on the relative location of the stratiform rain area associated with the linear system (Parker and Johnson 2000). The six types of storms are grouped into the *specific-type* classification scheme.

To train the decision tree, a subset of the data was hand labeled. The simulated training set, called ARPS-1, contained 505 storms, and the simulated test set, called ARPS-2, contained 378 storms. A hand-labeling interface (Fig. 4) provided the basic information required to visually classify storms. When a clustered storm appears on the screen, an individual cluster is selected and then the appropriate classification for it is chosen from a pull-down menu. Isolated pulse cells tended to be small areas of light-to-moderate reflectivity with little organization and weak updrafts and winds. Isolated strong cells tended to be isolated areas of high reflectivity combined with other features such as a hook echo, strong winds, and strong vertical velocity. Multicells were generally clusters that contained multiple areas of high reflectivity intermixed with weaker reflectivity. The stratiform area was

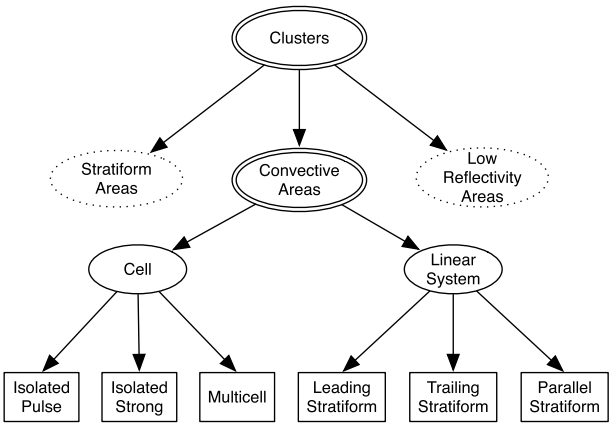


FIG. 3. Hierarchy of the different storm types used in the classification process.

labeled as leading or trailing depending on the indicated storm direction. Five random time steps from each of the ARPS simulations were hand labeled.

e. Classifier training and statistical evaluation

The decision trees were trained using the Waikato Environment for Knowledge Analysis (WEKA), version 3.5.7, developed by the University of Waikato in New Zealand (Witten and Frank 2005). WEKA is a suite of various machine learning and data mining algorithms. For the purposes of this project, the data were used in WEKA to generate decision trees based on different combinations of statistical values from the storm data. Within WEKA, the J48 decision tree was used as the primary decision-tree algorithm, because it is an implementation of the commonly used C4.5 algorithm (Quinlan 1986). Decision trees were generated for the general and specific storm types based on permutations of attribute types: the morphological, reflectivity, and wind attributes; morphological and reflectivity attributes; morphological and wind attributes; reflectivity and wind attributes; morphological attributes; wind attributes; and reflectivity attributes. Figure 1 shows an example of one of the general decision trees.

To compare the different attribute sets and algorithms, the classification accuracy and the Hanssen and Kuipers discriminant, also known as the true skill statistic (TSS) or the Peirce skill score (Woodcock 1976), was used. TSS is a measure for comparing algorithms that are not sensitive to the underlying distribution of the data. In addition, TSS can handle data with more than two labels. The classification performance was compared against the performance of a random classifier. Possible TSS scores range from 1 to −1 with 0 being the score of a random classifier and 1 being the performance of a perfect classifier. Equation 3 shows how TSS is calculated for a classifier with K labels and N instances, where n is the number of instances at a given cell of a confusion matrix, $N(P)$ is the total number of instances along a prediction row, and $N(O)$ is the number of instances along an observed column:

$$\mathrm{TSS} = \frac{\frac{1}{N}\sum_{i=1}^{K} n(P_i, O_i) - \frac{1}{N^2}\sum_{i=1}^{K} N(P_i)N(O_i)}{1 - \frac{1}{N^2}\sum_{i=1}^{K} N(P_o)^2}. \quad (3)$$

Significance testing was performed on these datasets by finding the distribution of correct and incorrect predictions across the ARPS-2 dataset for each attribute set. The mean of the prediction distribution was taken to find the accuracy, and the attribute set with the best accuracy was compared against the other attribute sets using a paired t test at a significance of 0.01. The best attribute set would be used for all algorithms in the algorithm comparison test.

Comparisons against other machine learning algorithms were done to show if the human-readability of the decision trees is outweighed by significantly better classification performance from other algorithms. For our experiments, WEKA machine learning algorithms of a variety of types were used. The *J48 tree* algorithm was compared against other decision-tree algorithms. The *REP tree* is a rapid decision-tree learning algorithm that builds with information gain and prunes with reduced-error pruning (Witten and Frank 2005). The *random tree* is simply a decision tree generated by selecting a random attribute at each node and not pruning; this was used as a control tree. In addition, the decision tree was compared against more function-based algorithms. In this study, *logistic models* and the *multilayer perceptron*, a neural network implementation, were used. Hybrids between the two types were also compared. The *random forest* generates a weighted function from a set of random trees (Brieman 2001), and the *logistic model tree* (LMT) generates a decision tree with logistic models at its leaves (Landwehr et al. 2005).

Because *boosting* (Schapire 1990, 1999) is guaranteed to not hurt the performance and is very likely to improve it, boosting of multiple algorithms was selected as another comparison algorithm. Boosted decision trees still satisfy the goal of being able to share results easily with noncomputer scientists. Boosting takes a single classification algorithm and turns it into an ensemble algorithm. It creates multiple models of the same type

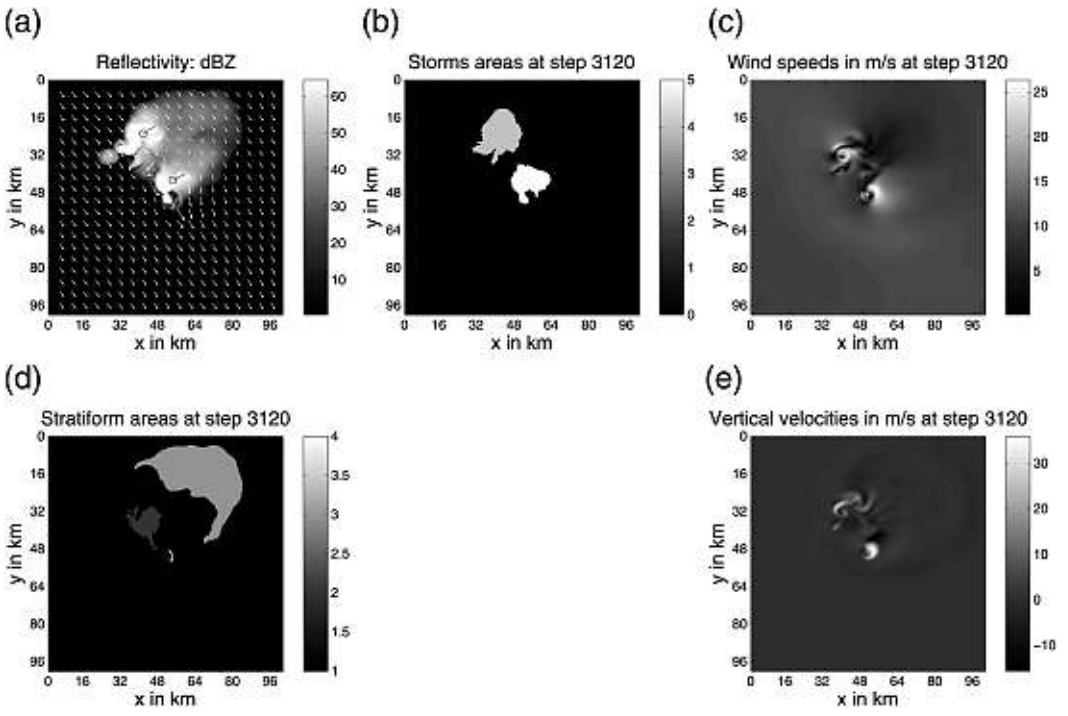


FIG. 4. The interface for hand-labeling storms with their type. A diagram of the convective cluster locations. (b) A storm area is selected and labeled based on the information provided by the other panels. (a) An image of the reflectivity with wind vectors and storm direction markers overlaid. (c) The 4-km wind speed across the domain. (d) Stratiform cluster areas are shown. (e) A display of the 4-km vertical velocity.

and weighs each of them based on their classification accuracy. The resulting weighted model will then be able to make predictions more accurately than the individual models on which it was based. To keep the computation time of the boosting algorithms to within reasonable limits, boosting was expanded on J48 trees, REP trees, and random forests.

Another technique similar to boosting is bootstrapping aggregation, or *bagging* (Brieman 1996). Bagging works by generating multiple classifiers from random subsets of the given training set and taking a plurality vote of their predictions to determine the ultimate prediction. It works best for unstable classifiers, such as decision trees and neural networks. Bagging was also implemented with the same algorithms as boosting.

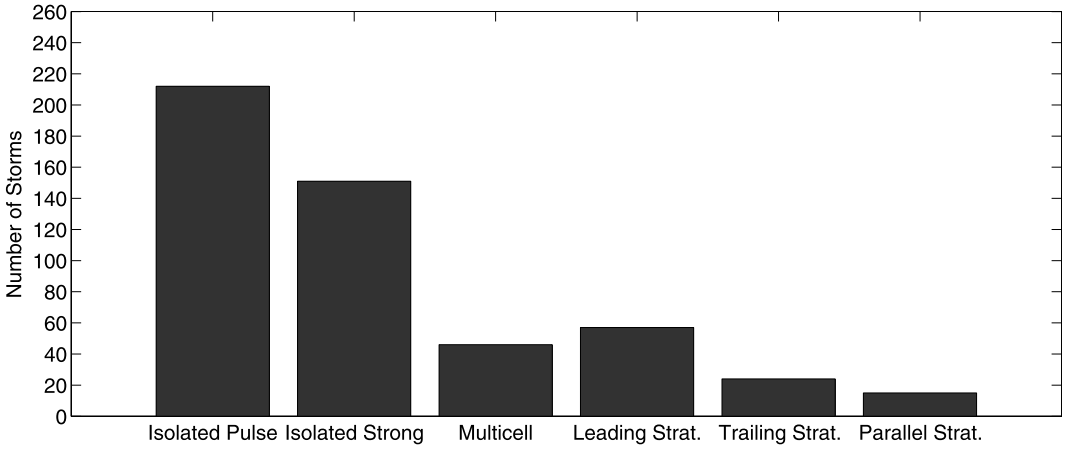
The performance of the decision-tree algorithm was also tested on a set of observed radar data. The radar observations were collected from the CASA IP1 radar network. A smaller number of these storms were hand labeled for the purpose of testing whether an algorithm trained on model storms could still accurately classify real-world storms. Comparisons were made using the datasets with morphological and reflectivity attributes, because wind attributes were not available for the CASA data.

3. Results

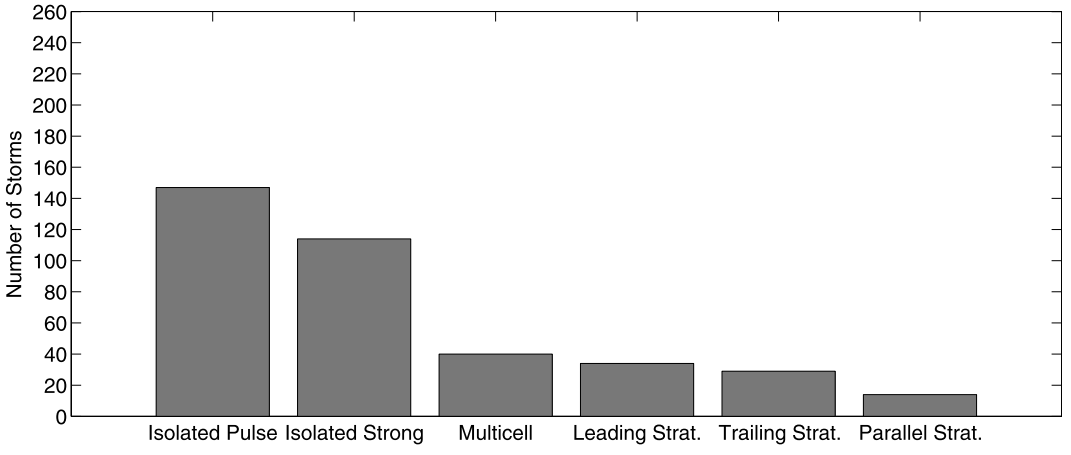
Figure 5 shows the distribution of different storm types in the hand-labeled ARPS training data (ARPS-1), ARPS testing data (ARPS-2), and the CASA observations. Because of the environmental regimes used to generate the simulations, the storm types in ARPS-1 and ARPS-2 data favor cell-type storms.

a. Comparisons among attributes

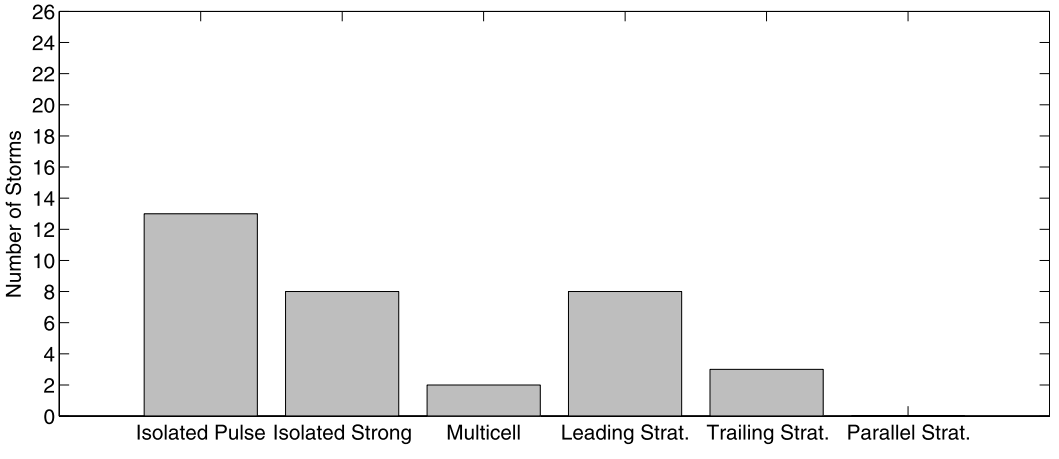
For the first evaluation of the decision-tree classification algorithm, decision trees were generated using different subsets of the attributes. The first decision tree used all of the attribute types in the training set and the J48 tree algorithm (Fig. 6). Each branch along the decision tree forms a set of rules that describes characteristics for each type. For instance, the majority of the isolated strong storms fell into the pattern of an area greater than 222.25 km^2, a major axis length less than 54.5 km, a maximum reflectivity greater than 70.6 dBZ, and a maximum wind speed greater than 22.1 m s^{-1}. Those characteristics correspond strongly to those used for labeling isolated strong storms. It can also find patterns in some of the more uncommon storms, which is evidenced by the multiple branches for other storm types.



(a) ARPS 1 Distribution



(b) ARPS 2 Distribution



(c) CASA Distribution

FIG. 5. The distribution across storm types for the (a) training set and (b),(c) two test sets.

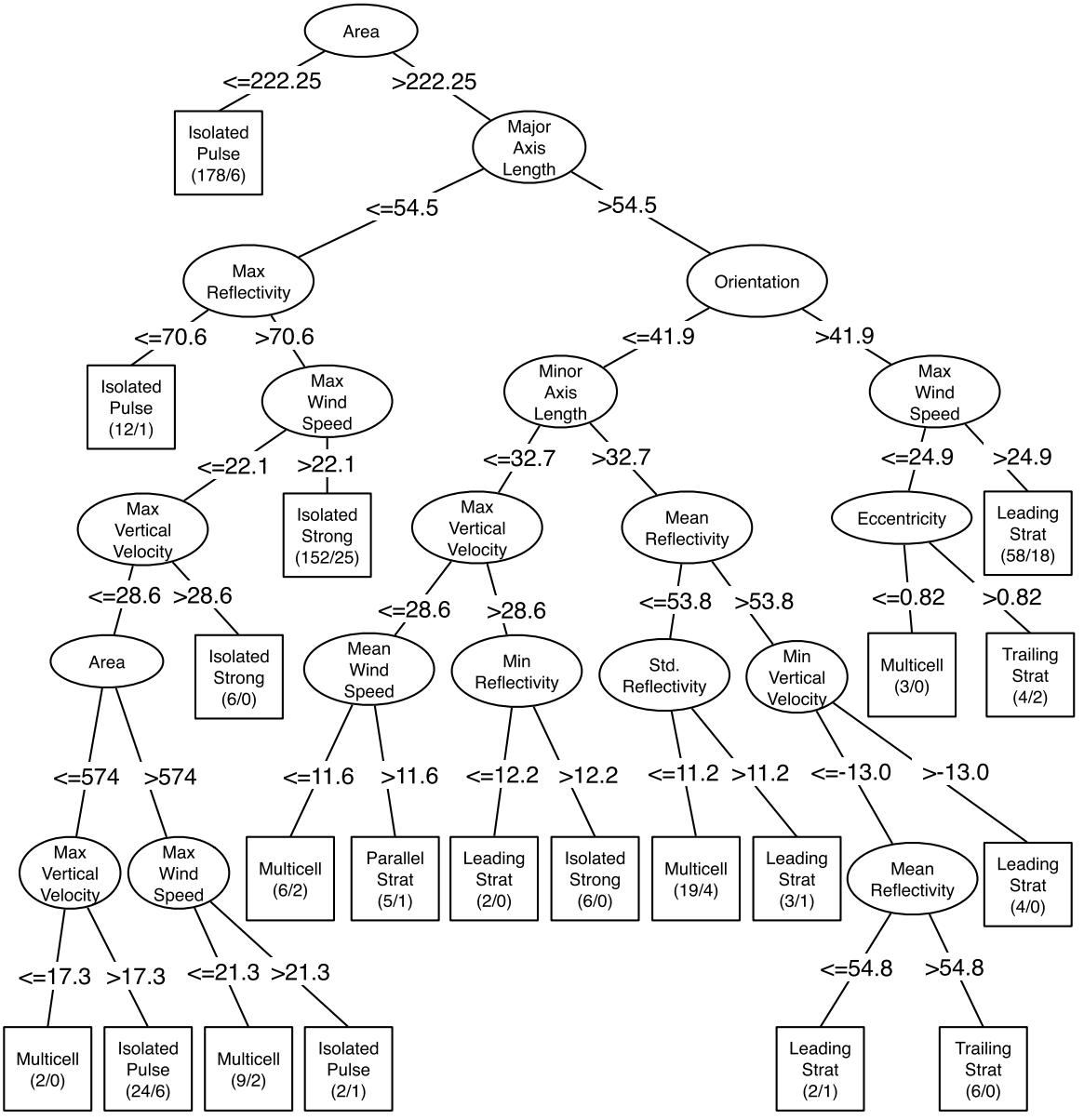


FIG. 6. The J48 tree generated using all types of attributes in the dataset: area (km^2), lengths (km), reflectivity (dBZ), and wind speeds ($m\ s^{-1}$). The numbers in parentheses indicate the number of correctly/incorrectly classified storms that reached that particular node.

Additional trees were created using different subsets of the attributes and tested on the ARPS-2 dataset. Across the different attribute sets, both the morphological, reflectivity, and wind attribute set and the morphological and wind attribute set had the highest accuracy, with the morphological and reflectivity attribute set coming in a close second. Only the reflectivity attribute set and the wind attribute set were significantly worse in accuracy, as shown in Fig. 7. The TSS of most of the trees is relatively high, indicating that strong skill is being found in the classification process. The morphological and reflectivity attribute set had the highest TSS, followed by the morphological, reflectivity, and wind attribute set and the morphological and wind attribute set. The reflectivity and wind attribute set, reflectivity attribute set, and wind attribute sets had the lowest TSS. For the general-type trees, the accuracy remained very high across most attribute sets, with the all-attribute set (includes all attributes from the morphological, reflectivity, and wind attribute set plus simulation time step and centroid coordinates) performing the best (0.886), and the morphological, reflectivity, and wind attribute set was

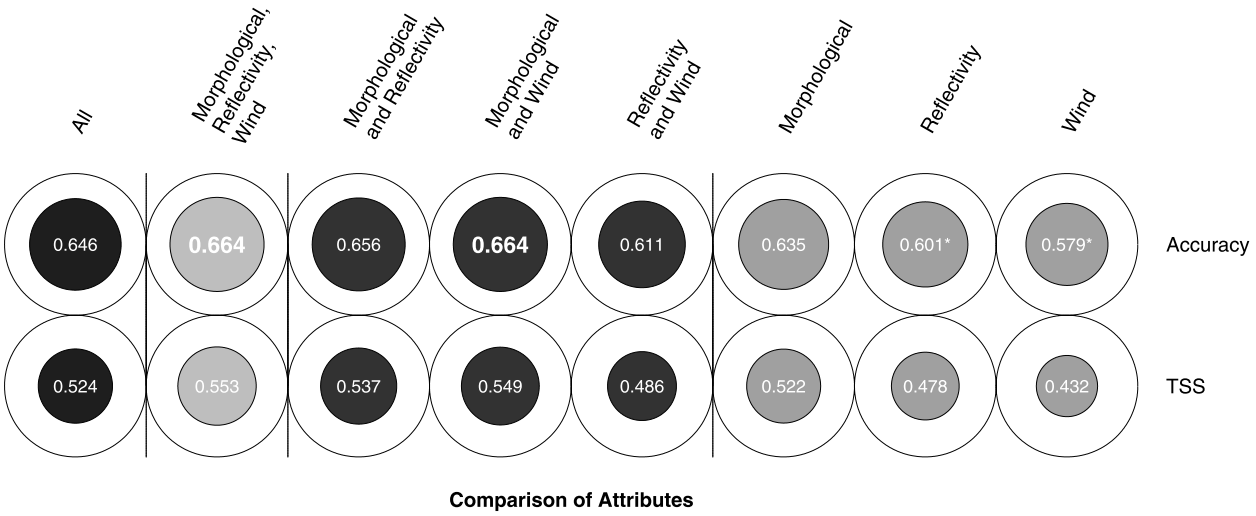


FIG. 7. The accuracy and TSS for J48 trees generated using different combinations of attributes. Bold, larger numbers indicate the highest accuracy, and numbers followed by an asterisk (*) indicate statistically significantly worse accuracy ($p < 0.01$).

a close second (0.881). The reflectivity and wind attribute set and the wind set performed statistically significantly worse ($p < 0.01$).

b. Attributes discussion

The choice of attributes in the tree from Fig. 6 shows the focus of the storm classification process very clearly. Appearing at the root (top) and twice elsewhere is area, making it the most gainful attribute in the tree, indicating that the general size of the storm is a major factor in separating the different types of storms. More types of area attributes would probably further improve the classification ability of the tree. Major axis length is an important attribute because all of the linear storm types are found along one of its branches since the linear storms will have very long, stretched-out, fitted ellipses, so their major axes will in general be much longer than most of the smaller, more circular cell-type storms. Wind speed and vertical velocity were also major differing attributes within the tree, since they were each sampled 4 times and were distributed throughout most levels of the tree. The tree identified the distinction that isolated strong storms generally have stronger winds and updrafts than isolated pulse storms, since most of the isolated strong storms were isolated by maximum wind speed or maximum vertical velocity.

Comparing the performance of decision trees from different attribute sets showed that changing the given attributes can have a significant effect on the decision tree's performance (Fig. 7). It also showed that, of the three types of attribute sets, the morphological attributes set was the strongest, because its performance was not significantly different from the combined sets, whereas reflectivity and wind, both singly and in combination, performed significantly worse. Although attributes from those two sets can be gainful to a tree, by themselves they are a less effective choice for determining storm classification. Because storms are classified based on their general morphology—with a little help from reflectivity strength and wind speed—the result fits with the human classifier's priorities in attributes.

c. Comparisons among algorithms

The performance of the decision tree was compared with the performances of other machine learning algorithms on the ARPS-2 test set using the morphological, reflectivity, and wind attributes set. Comparisons were done against three types of algorithms: standard machine learning algorithms, algorithms with boosting, and algorithms with bagging. The algorithms were first trained on the specific-type classification system, with results shown in Fig. 8. The random forest and bagging with random forests were found to have the best accuracy. The REP tree, random tree, and multilayer perceptron all performed statistically significantly worse ($p < 0.01$) than the random forest. Boosting and bagging did not significantly change the performance of any of the algorithms. TSS stayed well above 0.5 for all but two of the algorithms.

The algorithms were also compared using the general classification scheme. Overall, the algorithms all had very high accuracy and most had high TSS. A paired t test at a 0.01 significance revealed that none of the algorithms is statistically significantly better or worse than the others for classifying cells and lines. Similar

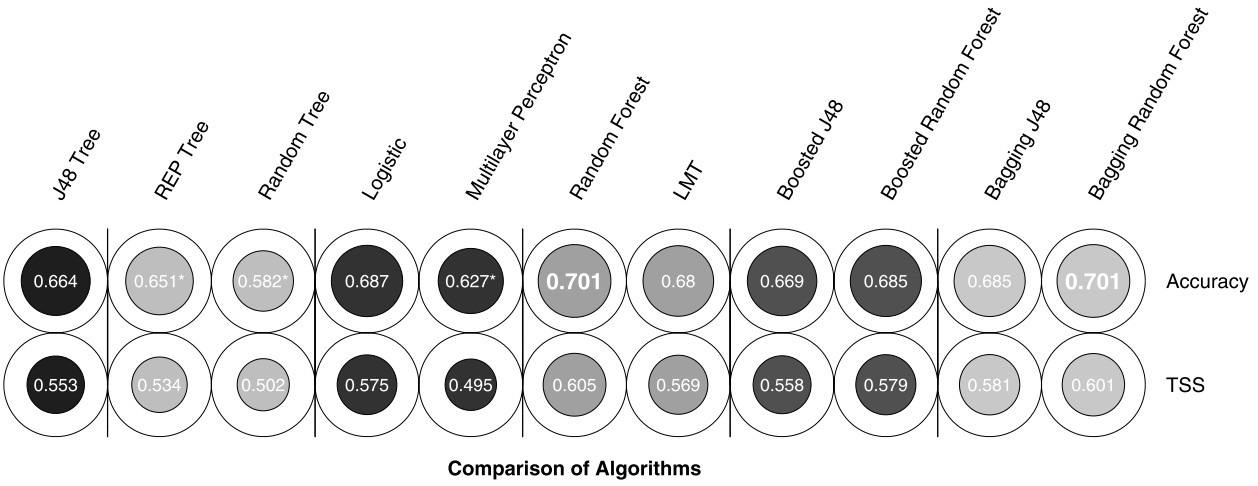


FIG. 8. The accuracy and TSS for each algorithm using the specific-type classification scheme on the ARPS-2 test set. All algorithms were compared against the random forest using a paired *t* test on accuracy, and the J48 tree was not significantly different from the random forest. Same as Fig. 7 for definition of numbers. The boosted J48 tree did not receive a significant increase in performance compared to the original algorithm.

results were found using the general classification scheme. Overall, the algorithms all had very high accuracy and most had high TSS, with the random forest having the highest accuracy (0.918) and TSS (0.727). A paired *t* test ($p < 0.01$) revealed that the REP tree, random tree, and multilayer perceptron were statistically significantly worse than the random forest for classifying cells and lines. The J48 tree was also statistically significantly worse ($p < 0.01$), but the accuracy was still very high at 0.886.

d. Algorithm discussion

The similarity between the accuracy of the decision-tree algorithm and the accuracy of the other common machine learning algorithms indicates that the decision tree's major advantage of human readability is not offset by any loss in classifying ability. Most of the standard algorithms cannot display the relationships between the attributes easily. The decision tree also had the advantage of speedy training. The trees were trained faster than every other algorithm evaluated. Boosting and bagging on this dataset were not shown to be statistically significant ($p < 0.01$) ways for improving the classification ability of the algorithm, so the increase in training time and complexity was not found to be a useful trade-off. Regarding the specific-type classification scheme, random forests had the best performance of any of the algorithms, but since their increase in performance was not statistically significant ($p < 0.01$), the J48 tree still holds a readability advantage over them. Regarding the general classification scheme, random forests did perform statistically significantly better ($p < 0.01$) than J48 trees in terms of accuracy, but the real difference in accuracy was still very small, and the accuracy of the J48 tree was still extremely high.

e. CASA results

Performance was calculated on the CASA test set of radar-observed storms within the CASA IP1 network. Three training sets were used to make J48 trees that were tested on the CASA set. These sets were the ARPS-1 set, which contained 505 storms and was labeled with all three types of attributes; the ARPS-2 set, which contained 378 storms and was labeled with all three types of attributes; and the ARPS-2 without wind set, which contained 519 storms and was labeled with only morphological and reflectivity attributes but used the same storm set as the ARPS-2 data. All three sets used the same attributes to train a J48 tree to classify the CASA data, so even though the sets were hand labeled using wind data, the wind attributes were removed for training. On the CASA set, the ARPS-2 without wind data far outperformed the ARPS-1 and ARPS-2 data, as shown in Fig. 9.

f. CASA discussion

Hand-classifier attribute bias can be seen in the testing data. Since the ARPS-1 data were classified primarily by analyzing the shape of the storms and how strong their winds were, the decision tree successfully identified part of the human decision-making process in its attribute search. On the test set, the morphological, reflectivity, and wind characteristics all played a major role in the

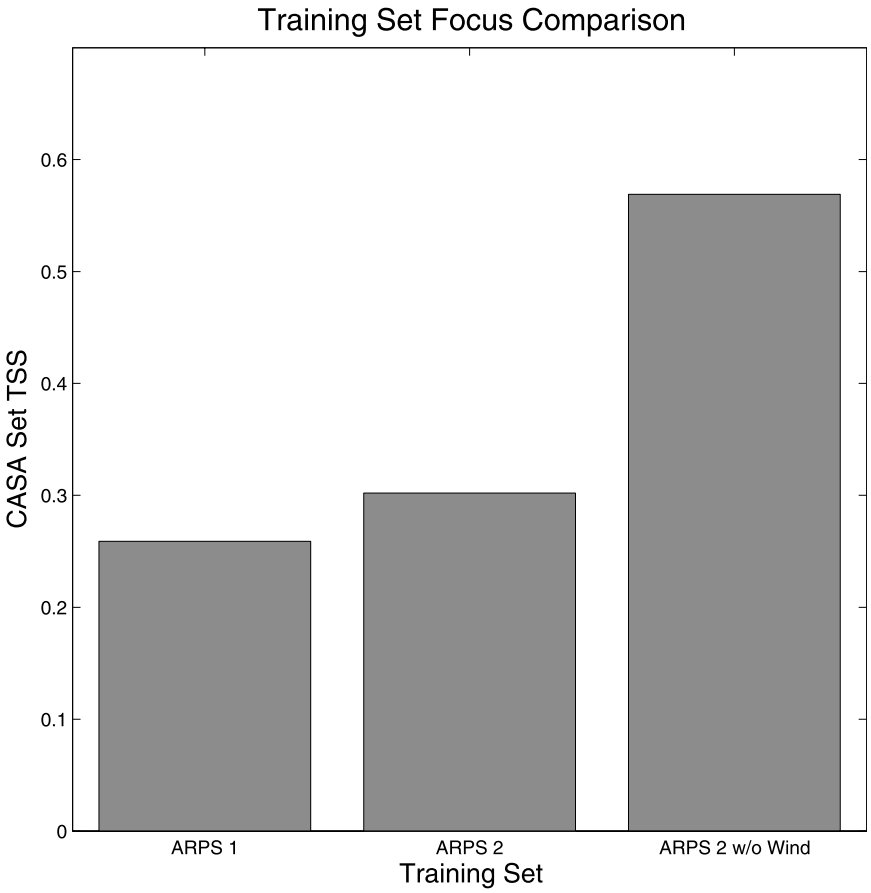


FIG. 9. This is a comparison of classification performance on the CASA test set. Although the ARPS-1 and ARPS-2 datasets used wind attributes for hand labeling, the ARPS-2 without wind set only used morphological and reflectivity attributes. Both datasets used the same attributes to train a decision to classify the CASA data, but the extra data presented during hand labeling changed the patterns in the dataset and caused a major decrease in TSS on the CASA set.

hand-classification process. Including wind variables in the classification process changed how some of the storms were labeled, which changed the underlying trends in the dataset. When the algorithm had those wind attributes removed, those underlying patterns could no longer be found, so performance decreased (Fig. 9). The algorithm based on the ARPS-2 without wind data, which used morphological and reflectivity attributes from the start, achieved a TSS of 0.569. Since the storms in that set were classified using only the morphological and reflectivity data and then trained using the morphological and reflectivity attributes, all the underlying patterns remained intact and were successfully discovered by the decision tree. Even though the three datasets in Fig. 9 trained with the same attribute sets, the trends in them differed significantly. From these tests, it can be seen that matching the training and hand-classification attributes to the attributes used to classify the final product is vital for producing a strong and reliable decision-tree classifier. The tests also show that a classifier trained only on simulated reflectivity data can still classify observed reflectivity data to a high degree of skill and accuracy.

The primary goal of this study was to determine the best methodology for automated classification of storm cells for use in real-time operations. For this purpose, the best methodology was defined as that method which demonstrated the combined greatest accuracy with the greatest ease of interpretation. The ease of interpretation, in turn, allows for new science to be discovered from the rules of the decision tree. From this study, an examination of the rules highlights the critical value of storm size, shape, and orientation, whereas the value of storm reflectivity is of much less importance.

As a next step, severe weather feature information will be included in the data, and the decision tree will be

rerun to identify those storm types that lead to the development of severe weather. In this way, radar scanning can more exclusively focus on only those storms that exhibit severe storm characteristics. A decision tree from this more advanced study would provide valuable insight into those specific attributes that generate (or inhibit) the development of severe weather features, such as tornadoes, hail, severe winds, or flooding.

4. Conclusions

These experiments have found that decision trees are a viable method for automatically determining storm type. The trees had a relatively high TSS and accuracy across all datasets, indicating strong performance overall. Other standard machine learning algorithms and boosting did not perform significantly better than the decision-tree algorithms, so the decision tree's human-readability advantage does not compromise its classification performance. The decision trees could also successfully predict the storm types on an independent model test set with little loss in performance. When the tree was trained within the same attribute set as real-world radar data, it could classify real-world radar data with little loss in performance as well. A decision tree trained on model radar reflectivity data could potentially provide real-time storm classification on any radar system.

Acknowledgments. This work is supported in part by the Engineering Research Centers Program of the National Science Foundation under NSF Award 0313747 and The University of Oklahoma's College of Engineering. In addition, this material is based upon work supported by the NSF Grant IIS/REU/0755462. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. Thanks to Dr. Keith Brewster for providing the programs and assisting with the extraction of the CASA reflectivity data; Dr. Rodger Brown for assisting with the review of this paper; and Derek Rosendahl, who generated the ARPS dataset. Final thanks to the anonymous reviewers for their helpful comments and to Dr. V. Chandrasekar for editing the peer review process.

REFERENCES

Alexiuk, M., N. Pizzi, and W. Pedrycz, 1999: Classification of volumetric storm cell patterns. *Proc. IEEE Canadian Conf. on Electrical and Computer Engineering,* Vol. 2, Edmonton, AB, Canada, IEEE, 1081–1085.

Anagnostou, E., 2004: A convective/stratiform precipitation classification algorithm for volume scanning weather radar observations. *Meteor. Appl.,* **11,** 291–300.

Baldwin, M., J. Kain, and S. Lakshmivarahan, 2005: Development of an automated classification procedure for rainfall systems. *Mon. Wea. Rev.,* **133,** 844–862.

Brieman, L., 1996: Bagging predictors. *Mach. Learn.,* **26,** 123–140.

——, 2001: Random forests. *Mach. Learn.,* **45,** 5–32.

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

Darema, F., 2004: Dynamic data driven application systems: A new paradigm for application simulations and measurements. *Proceedings of Computational Science—ICCS 2004,* M. Bubak et al., Eds. Lecture Notes in Computer Science, Vol. 3038, Springer-Verlag, 662–669.

Gallus, W. A., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting,* **23,** 101–113.

Guillot, E. M., T. M. Smith, V. Lakshmanan, K. L. Elmore, D. W. Burgess, and G. J. Stumpf, 2008: Tornado and severe thunderstorm warning forecast skill and its relationship to storm type. Preprints, *24th Conf. on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology,* New Orleans, LA, Amer. Meteor. Soc., 4A.3. [Available online at http://ams.confex.com/ams/88Annual/techprogram/paper_132244.htm.]

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

Lakshmanan, V., 2001: A hierarchal, multiscale texture segmentation algorithm for real-world scenes. Ph.D. thesis, University of Oklahoma, 108 pp.

Landwehr, N., M. Hall, and E. Frank, 2005: Logistic model trees. *Mach. Learn.,* **59,** 161–205.

McQueen, J. B., 1967: Some methods for classification and analysis of multivariate observations. *Statistics,* L. M. Le Cam and J. Neyman, Eds., Vol. 1, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, 281–296.

Parker, M. D., and R. H. Johnson, 2000: Organizational modes of midlatitude mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 3413–3436.

Quinlan, J., 1986: Induction of decision trees. *Mach. Learn.,* **1,** 81–106.

Rigo, T., and M. C. Llasat, 2004: A methodology for the classification of convective structures using meteorological radar: Application to heavy rainfall events on the Mediterranean coast of the Iberian peninsula. *Nat. Hazards Earth Syst. Sci.,* **4,** 59–68.

Rosendahl, D. H., 2008: Identifying precursors to strong low-level rotation within numerically simulated supercell thunderstorms: A data mining approach. M.S. thesis, School of Meteorology, University of Oklahoma, 200 pp.

Schapire, R. E., 1990: The strength of weak learnability. *Mach. Learn.,* **5,** 197–227.

——, 1999: Theoretical view of boosting and applications. *Algorithmic Learning Theory: Proceedings of the 10th International Conference (ALT '99),* O. Watanabe and T. Yokomori, Eds., Lecture Notes in Computer Science, Vol. 1720, Springer-Verlag, 11–25.

Schiesser, H. H., R. A. Houze Jr., and H. Huntrieser, 1995: The mesoscale structure of severe precipitation systems in Switzerland. *Mon. Wea. Rev.,* **123,** 2070–2097.

Steiner, M., R. A. Houze, and S. E. Yuter, 1995: Climatological characterization of three-dimensional storm structure from operational radar and rain gauge data. *J. Appl. Meteor.,* **34,** 1978–2007.

Witten, I., and E. Frank, 2005: *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. Morgan Kaufmann, 525 pp.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.,* **104,** 1209–1214.

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. *Meteor. Atmos. Phys.,* **75,** 161–193.

——, and Coauthors, 2001: The Advanced Regional Prediction System (ARPS)—A multiscale nonhydrostatic atmospheric simulation and prediction model. Part II: Model physics and applications. *Meteor. Atmos. Phys.,* **76,** 134–165.

——, 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. *Meteor. Atmos. Phys.,* **82,** 139–170.