Li, Yang, and Sun, 2008. Mining Geophysical Parameters through Decision Tree Analysis to Determine Correlation with Tropical Cyclone Development, Computers & Geosciences, (in press)

Mining Geophysical Parameters through Decision Tree Analysis to Determine Correlation with Tropical Cyclone Development

Wenwen LI, Chaowei YANG*, Donglian SUN

Joint Center for Intelligent Spatial Computing and
Earth Systems & GeoInformation Sciences
George Mason University, Fairfax, VA, 22030-4444
{wli6, cyang3, dsun}@gmu.edu

* Corresponding Author: 1-703-993-4742 (phone), 1-703-993-9299 (fax),

Abstract

Correlations between geophysical parameters and tropical cyclones are essential in understanding and predicting the formation of tropical cyclones. Previous studies show that sea surface temperature and vertical wind shear significantly influence the formation and the frequent changes of tropical cyclones. This paper presents the utilization of a new approach, data mining, to discover the collective contributions to tropical cyclones from sea surface temperature, atmospheric water vapor, vertical wind shear, and zonal stretching deformation. A decision tree using the C4.5 algorithm was generated to illustrate the influence of geophysical parameters on the formation of tropical cyclone in weighted correlations. From the decision tree, we also induced decision rules to reveal the quantitative regularities and co-effects of [sea surface temperature, vertical wind shear], [atmospheric water vapor, vertical wind shear], [sea surface temperature, atmospheric water vapor, zonal stretching deformation], [sea surface temperature, vertical wind shear, atmospheric water vapor, zonal stretching deformation], and other combinations to tropical cyclone formation. The research improved previous findings in a) preparing more precise criteria for future tropical cyclone prediction, and 2) applying data mining algorithms in studying tropical cyclones.

Key Words: Hurricane, Natural Disaster, Prediction, Data Mining

1. INTRODUCTION

A Tropical Cyclone (TC) is one of the most devastating natural disasters that frequently cause loss of human lives and serious economic damage through ocean storm surges, destructive winds, and flash flooding (Bengtsson, 2001). For example, the deadly 2005 hurricane, Katrina, is blamed for the deaths of 1,836 people and a reported economic loss of $81.2 billion [1]. Finding out how geophysical parameters contribute to TCs and ideally leading to the prediction of a TC for disaster preparedness is now an urgent research agenda.

Recent research (Bengtsson, 2001; Emanuel, 2005; Hoyos et al., 2006; Knutson and Tuleya, 1999; Latif et al., 2007; Solow and Moore, 2002) has focused on analyzing Sea Surface Temperature (SST) and vertical Wind Shear (WS) with TC characteristics, including frequencies and intensities and the initial formation conditions. Investigations into SST and hurricane frequency and intensity have found that SST higher than 26ºC is the basic requirement for the formation of a TC (Bengtsson, 2001; Webster et al., 2005), and Palme´n (1948) and Gray (1968) revealed that disturbances and storms developed when SST reached 26.5ºC in the Northwest Atlantic, Gulf of Mexico and the Northwest Pacific. Other research (Shapiro and Goldenberg, 1998) found that fluctuations in the

[1] Increased North Atlantic Hurricane Activity: A Summary of Relevant Literature, http://www.rms.com/Publications/TC_Bibliography_Summary.pdf

magnitude of tropospheric WS contribute to changes in seasonal hurricane frequency. Goldenberg (2001) found that the dominant parameter for TC activity was the magnitude of the vertical shear of the horizontal wind between the upper and lower troposphere (a.k.a. WS), and found that WS with a magnitude over 8m/s is generally unfavorable for TC development. Emanuel (2005) found that only part of the observed increase in the TC's power dissipation index (PDI) was caused by increased SSTs, while the rest can only be explained by changes in other geophysical parameters, such as WS. Hoyos et al. (2006) used Mutual Information to estimate the correlation between SST (also WS, Humidity, zonal Stretching Deformation (SD)) and hurricane intensities (Categories 4 and 5, higher levels of TCs) in six ocean basins, and found that SST is the dominant parameter influencing the long-term trend of increasing hurricane intensity, and other parameters influence short-term hurricane intensity, depending on the specific basin.

The frequencies and intensities of TCs change significantly, and researchers haven't found the complete list of geophysical parameters involved and how they influence TCs. Generally, a higher SST, higher atmospheric Water Vapor (WV) content, and minimal WS will increase hurricane activities (Hoyos et al., 2006). Webster et al. (2005) verified that negative zonal deformation values might also result in hurricane development. However, there's no study on how the factors jointly affect the frequencies and intensities of TCs.

This paper reports our efforts in exploring the united effects of these geophysical parameters in the Main Development Region (MDR) of hurricane activities in the North Atlantic Ocean Basin. Data mining algorithms, such as the C4.5 classification algorithm and FCM (Fuzzy C-Means) clustering algorithm, are used to discover more precise criteria.

Section 2 introduces the study area, datasets, and methodology. The decision-tree linking parameters and TCs, decision rules, and rules evaluations are discussed in Section 3. Section 4 draws conclusions and discusses future research.

2. STUDY AREA, DATA, AND PROCEDURE

2.1 Study Area.

The MDR, covering 10º- 20º N, 20º-80º W across the tropical North Atlantic and the Caribbean Sea (Figure 1), is selected as the study area because 1) the abnormality of atmospheric conditions in this area was the main reason for active TCs (Demaria, 1996; Bender, 1997; Gray 1968; Mo et al., 2001); and 2) TCs reaching hurricane intensity in the Atlantic basin have a high probability of hitting a populated area (e.g., only four hurricanes in the Gulf of Mexico failed to make landfall from 1950-1995, Lehmiller et al., 1997; Solow and Moore, 2002) and usually result in large economic losses.

Figure 1 Near Here

2.2 Data used

Because geophysical parameters have interdecadal characteristics due to ENSO and global warming (Vitart and Anderson, 2001), we select the latest nine-year data range for SST, WV, and winds during the North Atlantic hurricane seasons (From June 1st to Nov. 30th, 1998 – 2006). Daily average SST and WV are derived from Tropical Rainfall Measuring Mission (TRMM) Microwave Imager (TMI [2]), which provides daily maps (separated into ascending and descending orbital segments) from December 1997-present.

The data is archived on a 0.25°*0.25° latitude-longitude grid twice per day (day and night).

The WS is defined as the difference between 200- and 850-hpa zonal wind (U_{200}-U_{850}) and the SD is defined as the differential of 850-hpa zonal wind by the adjacent longitude grid (Equation 1). The intermediate parameter of ‘zonal wind’ for computing the above two parameters is retrieved from the National Center for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR) reanalysis (Kalnay et al.1996). Data on a 2.5°*2.5° latitude-longitude grid are available from 1949-present.

$$\frac{\partial U_{850}}{\partial \lambda} = \frac{U_{850}(\lambda_2) - U_{850}(\lambda_1)}{\lambda_2 - \lambda_1} \quad (1)$$

The estimated intensities of all of the North Atlantic TCs from 1998 to 2006 are obtained from the National Hurricane Center’s (NHC) best track data, which provides center locations (latitude and longitude in tenths of degrees) and intensities (maximum 1-minute surface wind speeds in knots and minimum central pressures in millibars) at 6-hour intervals for all tropical storms since 1851 (Jarvinen et al. 1984). The mean of four daily maximum-surface-wind-speed values is used to represent the daily average storm intensity. According to Saffir-Simpson’s Categorization, the TCs with a maximum surface wind speed of over 17m/s are tropical storms and the TCs with a maximum surface wind speed of over 33m/s are hurricanes [3].

[2] The TMI ocean product dataset, http://www.ssmi.com/tmi/tmi_browse.html
[3] Saffir_Simpson Hurricane Scale. http://www.nhc.noaa.gov/HAW2/english/basics/saffir_simpson.shtml

2.3 Methodology

2.3.1 Method selection

Data mining is used for discovering hidden information and knowledge from huge amounts of data (Han and Kamber, 2001). Compared to traditional statistical models, which are often used in related research for measuring correlations between two variables (Hoyos et al., 2006), data mining methods, such as decision tree analysis, can help to find the hidden relationships among multiple effective parameters. A decision tree is a classical prediction model that supports decision-making (Han and Kamber, 2001) by converting complex data into a relatively simple and straightforward structure. Furthermore, it can generate optimized results within a relatively short time period. Also, it has been proven effective in solving scientific problems and supporting decision making in other research areas, such as land cover terrestrial surface type classification and prediction (Colstoun and Yang, 2000). In this paper, we attempt to introduce decision tree analysis into TC data mining.

Among all the algorithms for decision tree generation, the most popular ones are Classification And Regression Trees (CART), CHi-squared Automatic Interaction Detection (CHAID), ID3 and C4.5. CART supports only binary splits, where each parent node can be split into, at most, two child nodes. So the generated decision tree cannot support finer divisions with more than two subdivisions. CHAID and ID3 are limited to classifying variables only, so attributes of continuous variables must be converted to categorical data with loss of accuracy. Previous studies found that SST is the most important factor in TC formation; this research studies the continuous SST for more accurate results. Therefore, CHAID and ID3 cannot fulfill the request. The C4.5

algorithm is chosen because 1) C4.5 is more flexible than ID3, and 2) C4.5 is able to support both multi-split and process continuous parameters.

2.3.2 Algorithm Description

The C4.5 algorithm (Quinlan, 1993) is a supervised learning method based on decision tree induction (Han and Kamber, 2001). The basic strategy is to select an attribute that will best separate samples into individual classes by a measurement, '*Information Gain Ratio*', based on information-theoretic 'entropy'. 'Best' means to find the minimum information needed to keep the least "impurity" of the partitions (Baglioni et al., 2005; Han and Kamber, 2001).

Formally, let S be the training set consisting of s data samples, s (C_i) is the number records in S that belong to class C_i (for i=1, 2… m). The information (entropy) needed to classify S is:

$$Info(S) = -\sum_{i=1}^{m} \frac{s(C_i)}{s} \log_2\left(\frac{s(C_i)}{s}\right) \qquad (2)$$

Hence, the amount of information needed to partition S into {S_1, S_2…S_v} by attribute A (The number of distinct values of attribute A is 'v') is:

$$Info(A|S) = -\sum_{j=1}^{v} \frac{s_j}{s} * Info(S_i) \qquad (3)$$

The gain is computed as:

$$gainRatio(A \mid S) = \frac{gain(A \mid S)}{Info(A \mid S)} \tag{4}$$

where $$gain(A \mid S) = Info(S) - Info(A \mid S) \tag{5}$$

In our study, four attributes/parameters (SST, WV, WS and SD) and three classes (No storm, Non-major storm, and Major-storm) are united for mining and classification.

3. DECISION-TREE BASED CLASSIFICATION ANALYSIS

3.1. Attribute Pre-processing

Attributes (WV, WS, and SD) should be stratified to discrete values for decreasing the abundant branches when building a decision tree with continuous data. A Fuzzy C-Means (FCM, Dunn, 1973; Bezdek, 1981) clustering algorithm is applied on the seasonal mean over the MDR to the three classes. Because a negative value for SD is favorable for hurricane activity (Webster et al., 2005), the data with positive values are recorded in one group and the others are clustered to two other groups by FCM. SST, which acts as the most significant factor in driving a tropical storm (Emanuel, 1987; Holland, 1997), is left for further classification for more accurate results.

Figure 2 displays the statistic charts of the North Atlantic atmospheric conditions from June to November. General variations of SST and WV are similar. They usually reach their peaks around September, and WS reaches its lowest peak around August. The trend of SD doesn't show any regularity but the data mostly concentrate within [-0.54, 0].

Figure 2 Near Here

3.2. C4.5 Classification and Analysis

Tropical storms and hurricanes are collectively called "named storms" (Landsea and Gray, 1992), which have maximum sustained surface winds over 17 m/s. The named storms are calculated and incorporated with daily attribute data, and labeled Class 'Yes' (i.e., there is a tropical storm). Other records with no named storm will be labeled as Class 'No'. If there are two named storms on the same day, there will be two records. The C4.5 classification algorithm (Quinlan, 1993) was run on the training data to build a decision tree. Figure 3 depicts the decision tree generated from 2080 records from 1998 to 2006 with a correctness of 82%. Each node of the tree is associated with an attribute describing a feature of the hurricane data, and each outgoing arrow is labeled with a possible value. For example, we use (low, middle, high) for WV, (mild, middle, strong) for WS, and (neg_s, neg_m, pos_l) for SD. Each leaf node is associated with a certain class of a value. To quantify the relationships of named storms and the four geophysical parameters, the records and their possibilities in each node are counted and presented in the tree. The number in brackets under the possibility value (%) in rectangles at each leaf equals the number of training instances, which belong to that path in the tree. Meanwhile, each number is followed by the number of classification errors encountered in that particular path of the decision tree. For example, the leaf in the path "SST<=27.28°C, WV=low" has the prediction possibility at 14.1%, and there're 256 instances belonging to this path, in which error instances are 36. The final decision tree illustrates that increasing SST, WV and decreasing WS reinforce the formation of tropical storms, which is consistent with Webster et al.'s (2005) finding. Meanwhile, results confirmed that decreasing SD is favorable to the formation of tropical storms.

As discussed in Section 2.3, the node in the upper level of the decision tree has a higher information gain ratio than the lower level node in the classification. Therefore, parameters, such as SST, which appear at the root node of the decision tree, are more important than those in the lower level in triggering tropical storms. When SST is lower than 27.28°C, it has a higher priority than that of WV in the formation of tropical storm. The other two parameters are not shown in this path, because their influences are not obvious in this path. When SST is higher than 27.28°C, the combination based on priority is either (SST>WS>WV) or (SST>WS>SD), depending on the value of WS.

Figure 3 Near Here

3.3. Production rules analysis and evaluation

The complex decision tree can be transformed to that of production rules for both improving classification performance (Quinlan, 1987) and representing knowledge in a well-understood fashion (Winston, 2001). Any path from the root node to a leaf node can be viewed as a rule. But some rules with low accuracy should be eliminated, specifically, if Xi is one of the rules, Xi becomes a candidate for elimination if either its removal will not decrease the certainty of the rule, or if the hypothesis that Xi is irrelevant cannot be rejected at the 1% level or better (Quinlan, 1987). The final production rules (with default class T) are depicted in Figure 4. Among all of these 14 rules, there are nine for predicting tropical storm activity and the other five are for determining that tropical storm will not happen.

Figure 4 Near Here

Rule 51 enhances the empirical threshold (26°C) for hurricane formation at the possibility of 92%. 154 records are counted and correctness for this rule is 92.2% (Table 1). This is consistent with the fact that SST can impact TS frequency in at least two ways: a change in SST may lead to 1) a change in the thermodynamic structure of the atmosphere or 2) a change in the large-scale circulation (Vitart and Anderson, 2001).

Another critical rule, Rule 46, indicates that when WS is over 16.9m/s, the atmospheric environment is very unfavorable (at the possibility of 86.7% with accuracy of 71.5%, Row 9, Table 2) to a developing TS. This gives a more sufficient condition than "when WS is over 8m/s, it's 'generally' unfavorable for a TC", which was derived by (Goldenberg et al., 2001). Other rules demonstrate the united influence of geophysical parameters to the formation of tropical storms, such as combinations [SST, humidity, wind shear], [humidity, zonal deformation], and [humidity, wind shear zonal deformation], which all contribute to hurricane activities.

We evaluated all the production rules and chose those having high correctness (Column 5, Table 1) and high prediction possibilities (Column 3 (1-Error%), Table 1) as the final rules (Table 1). Besides evaluating the error ratio of each rule, an error matrix is given to record the accumulated statistical error of the selected rules, shown in Table 2. Results illustrate that 16.5% of total TC activities have been predicted as non-TC activity and

20% of total None-TC activities have been predicted as TC by the currently decision rules, which reach a satisfying precision for prediction.

Table 1 Near Here

Table 2 Near Here

3.4. Production rules validation

The third most active hurricane season, when nineteen named storms were reported in 1995, is selected for validating the rules. Since there are no daily SST products from TMI for 1995, AVHRR Oceans Pathfinder global 4km data (Vazquez, et al., 1998) are used. WV is from SSM/I and NCEP reanalysis. 263 records are tested and the validation accuracy is 80% (Table 3). The generated rules (Rule 51, 19, 10 and 20) for regulating the occurrence of tropical storms have an accuracy of over 90% (Table 3). Meanwhile, these rules can be applied in prediction. For example, Rule 14, 10 and 35 can be used to predict the occurrence of hurricane Felix (which lasted from August 8th to 23rd).

Table 3 Near Here

4. CONCLUSION AND DISCUSSION

This paper reports our effort to utilize data mining to analyze TC activities in the North Atlantic (NATL) basin. Data from the MDR of the NATL basin during hurricane season (June 1st -Nov.30th) from 1998-present are mined. A decision tree and a set of decision rules are generated to reveal how the four parameters, SST, WS, WV, and SD, relate to TC formation and intensity. The decision tree demonstrates influence of different parameters on hurricane activity. The decision rules reveal the co-effect of multiple parameters and their roles in the formation of tropical storms and major hurricanes. Validation of the decision rules has proven that the generated rules are correct and the potential in predicting the TCs. The results also illustrate potential in utilizing data mining to study other geosciences' phenomena besides TCs.

The knowledge derived is expected to contribute to the study of tropical storm formation and intensification. For example, in the current Statistical Hurricane Intensity Prediction Scheme (SHIPS) (DeMaria and Kaplan, 1999), SST, wind shear, and the flux convergence of eddy angular momentum evaluated at 200 mb were included. We found that the moisture structure in the lower atmosphere and the deforming factor are also very important. Therefore, results derived here could help future research to consider more factors, such as moisture structure, for tropical storm formation and intensification. As a first attempt to introduce decision tree analysis into hurricane prediction, the decision tree method also provides an exemplar to utilizing data mining tools and methods in improving our understanding of tropical storm formation.

However, to reach the maturity of prediction, besides decision tree research introduced in this paper, contributions from a wide range of research areas and scientists are needed. For example, meteorologists may need to discover the complete list of parameters contributing to the formation of tropical storms and to add the parameters in the classification model; computer scientists need to develop more efficient computing platforms for us to make real-time prediction possible. Meanwhile, a more specific classification of tropical storms should be applied into the prediction model, i.e., to divide TCs into tropical storms, hurricane, and major hurricane.

ACKNOWLEDGEMENT

Research reported is supported by a 973 grant # 2006CB701306 and NASA grants # NNG04GH96A and NNX07AD99G. Dr. Eric Grunsky, the Editor-in-Chief of Computers and Geosciences, Dr. Karl Benedict, and the anonymous reviewers provide constructive comments in improving the paper.

References

Baglioni, M., Furletti, B., Turini, F., 2005. C4.5: Improving C4.5 by means of prior knowledge. In: Proceedings of the 2005 ACM Symposium on Applied Computing (SAC), New Mexico, USA, pp. 474-481.

Bender, M. A., 1997. The effect of relative flow on the asymmetric structure in the interior of hurricanes. Journal of Atmospheric Sciences 54, 703-724.

Bengtsson, L., 2001. Enhanced hurricane threats. Science 293, 440-441.

Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, NY, 272 pp.

Colstoun E.C., Yang W., 2000. Surface type visible/infrared imager/radiometer suite algorithm theoretical basis document, V 3.0, 94 pp. http://npoesslib.ipo.noaa.gov/IPOarchive/SCI/atbd/SurfaceType.pdf.

DeMaria, M., 1996. The effect of vertical shear on tropical cyclone intensity change. Journal of Atmospheric Sciences 53, 2076–2087.

DeMaria, M., Kaplan, J., 1999. An updated statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic and Eastern North Pacific basins. Weather Forecasting 14, 326-337.

Dunn, J. C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32-57.

Emanuel, K.A., 1987. The dependence of hurricane intensity on climate. Nature 326, 483-485.

Emanuel, K.A., 2005. Increasing destructiveness of tropical cyclones over the past 30 years. Nature 436, 686-688.

Goldenberg, S.B., Landsea, C.W., Mestas-Nuñez, A.M., Gray, W.M., 2001. The recent increase in Atlantic hurricane activity: causes and implications. Science 293, 474-479.

Gray, W.M., 1968. Global view of the origin of tropical disturbances and storms. Monthly Weather Review 96, 669–700.

Han, J.W., Kamber, M., 2001. Data Mining: Concept and Techniques. Morgan Kaufmann Publishers, San Francisco, CA, 550 pp.

Holland, G.J., 1997. The maximum potential intensity of tropical cyclones. Journal of Atmospheric Sciences 54, 2519-2541.

Hoyos, C.D., Agudelo, P.A., Webster, P. J., Curry, J.A., 2006. Deconvolution of the factors contributing to the increase in global hurricane intensity. Science 312, 94-97.

Jarvinen, B. R., Newmann, C. J., Davis, M. A. S., 1984. A tropical cyclone data tape for the North Atlantic basin, 1886–1983: Contents, limitations and uses. NOAA Technical Memorandum NWS NHC 22, Miami, FL, 21 pp.

Kalnay, E., Kanamitsu, M., Kistler, R. et al., 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society 77, 437-472.

Knutson, T. R., Tuleya, R. E., 1999. Increased hurricane intensities with CO2-induced warming as simulated using the GFDL hurricane prediction system. Climate Dynamics 15, 503-519.

Landsea, C. W., Gray, W. M., 1992. The strong association between Western Sahel monsoon rainfall and intense Atlantic hurricanes. Journal of Climate 5, 435–453.

Latif, M., Keenlyside, N., Bader, J., 2007. Tropical sea surface temperatures, vertical wind shear, and hurricane development. Geophysical Research Letters 34, L01710, DOI: 10.1029/2006GL027969.

Lehmiller, G.S., Kimberlain, T.B., Elsner, J.B., 1997. Seasonal prediction models for North Atlantic basin hurricane location. Monthly Weather Review 125, 1780-1791.

Mo, K., Bell, G. D., Thaiw, W., 2001. Impact of sea surface temperature anomalies on the Atlantic tropical storm activity and West African rainfall. Journal of Atmospheric Sciences 58, 3477-3496.

Palme´n, E., 1948. On the formation and structure of tropical hurricanes. Geophysics 3, 26–38.

Quinlan, J.R., 1987. Generating production rules from decision trees. In: Proceedings of 10th International Joint Conferences on Artificial Intelligence, Milan, Italy, pp.304-307.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA, 316 pp.

Shapiro, L. J., Goldenberg, S. B, 1998. Atlantic sea surface temperatures and tropical cyclone formation. Journal of Climate 11, 578-590.

Solow, A. R., Moore, L. J., 2002. Testing for a trend in North Atlantic hurricane activity, 1900-98. Journal of Climate 15, 3111-3114.

Webster, P. J., Holland, G. J., Curry, J. A., Chang, H.R., 2005. Changes in tropical cyclone number, duration, and intensity in a warming environment. Science 309, 1844-1846.

Winston, P.H., 2001. Artificial Intelligence, 2nd ed., Addison Wesley, Massachusetts, U.S., 527 pp.

Vazquez, J., Perry, K., Kilpatrick, K., 1998. NOAA/NASA AVHRR Oceans Pathfinder Sea Surface Temperature Data Set User's Reference Manual, v 4.0, JPL Publication D-14070, http://www.nodc.noaa.gov/woce_V2/disk13/avhrr/docs/usr_gde4_0_toc.htm.

Vitart, F., Anderson, J. L., 2001. Sensitivity of Atlantic tropical storm frequency to ENSO and interdecadal variability of SSTs in an ensemble of AGCM integrations. Journal of Climate 14, 533-545.

Tables

Table1. Evaluation of generated rules: Field *Size* refers to the condition number of certain rule. If the criterion of SST is a scale with an upper and lower bound, the condition number for that rule will be the arrow number plus one, such as Rules 20 and 36; for the other rules, the condition number is equal to the arrow number. *Error* refers to error margin of the rule, which equals (1-prediction possibility %). *Used* represents the number of times the rule was used, regardless of *Accuracy*. T refers to tropical storms and No refers to no tropical storms in field *Class*.

Rule	Size	Error	Used	Accuracy	Class
51	1	9.8%	154	92.2%	T
22	3	11.0%	136	91.2%	T
19	3	13.3%	230	85.7%	T
14	3	15.3%	54	87%	T
48	2	15.4%	226	78.8%	T
10	3	16.1%	164	75.6%	T
13	2	18.2%	62	79%	T
20	5	19.4%	19	89.5%	T
39	3	21.3%	38	68.4%	T
46	1	23.6%	250	71.6%	No
32	2	16.0%	283	85.5%	No
36	5	22.9%	16	87.5%	No
2	1	26.6%	169	71%	No
18	3	26%	137	78.1%	No

Table 2. Classification Error Matrix: The top heading refers to actual events. The left heading refers to classified events. The cells of 904 and 684 are correct classification. The cells of 171 and 179 are classifications in error.

Triggered Classified As	TC	No TC
TC	904	171
No TC	179	684
Accumulated Statistic Error	16.5%	20%

Table 3. Rules validation.

Year \ Rule		Rule 51	Rule 22	Rule 19	Rule 14	Rule 10	Rule 20	Rule 39	Rule 32	Rule 46	Total
1995	Used	58	17	39	2	41	12	6	3	26	263
	Wrong	3	4	1	1	2	0	3	2	15	53
	veracity	94.8%	76.5%	97.4%	50%	95.1%	100%	50%	33.3 %	42.3 %	80%

Figure 1 Study (Shading) Area covers region of 10°-20°N, 80°-20°W.

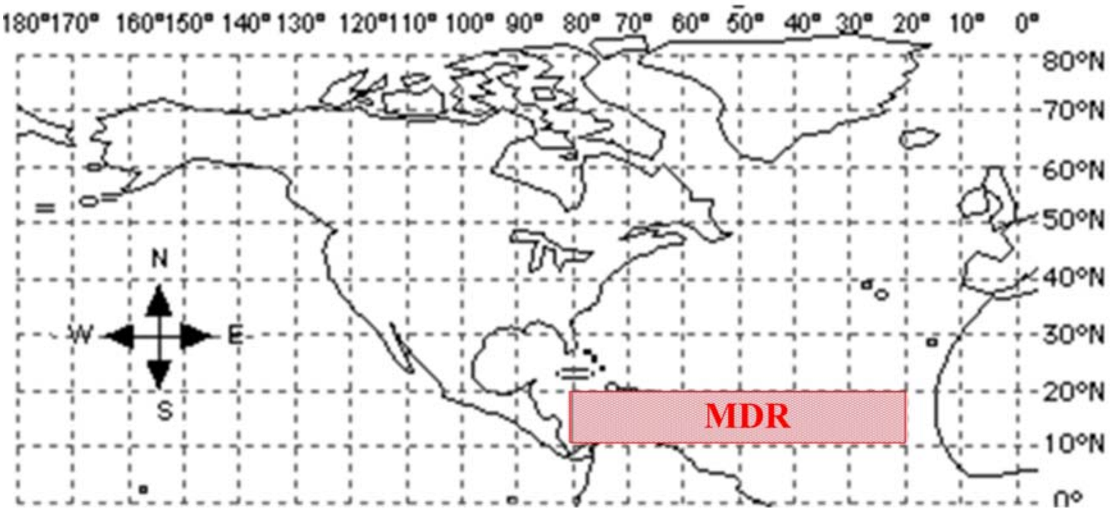


Figure 2 Temporal Distribution and trends of four attributes from 1998 to 2006 in MDR during North Atlantic hurricane season: A) SST, B) WV, C) WS, and D) SD. X axis represents the time from June to November; Y axis is the statistical attribute value. The lines in B), C), and D) represent the partition of attribute datasets.

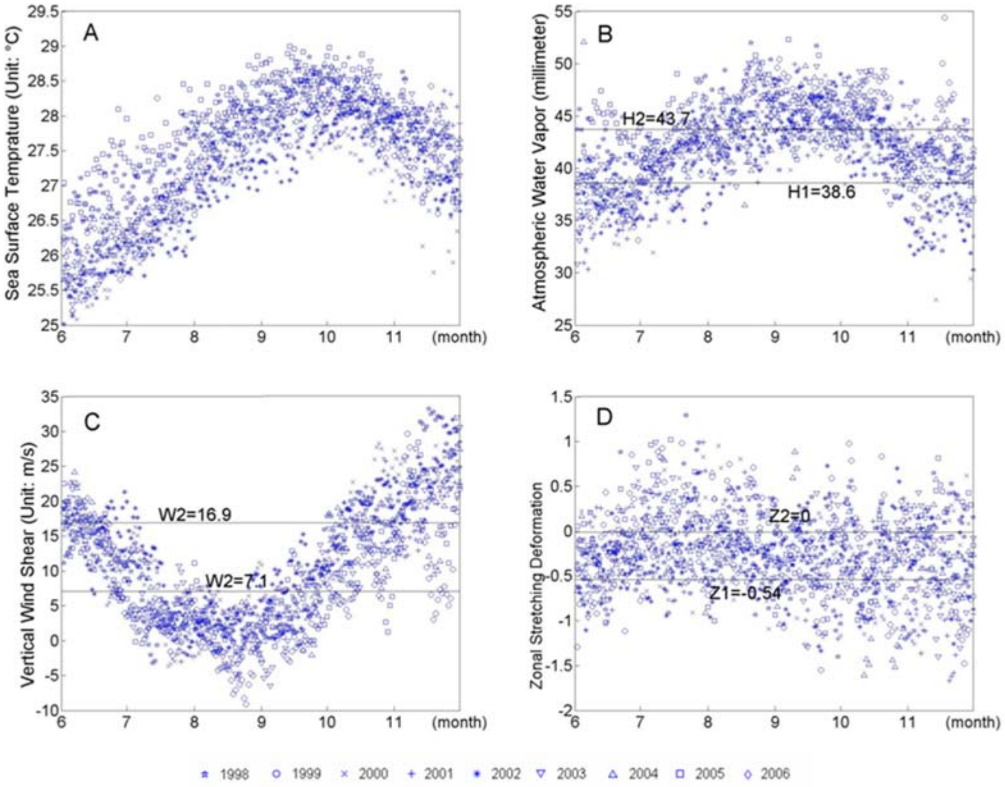


Figure 3 Decision Tree generated from C4.5; invalid leaf nodes have a small number of sample sets.

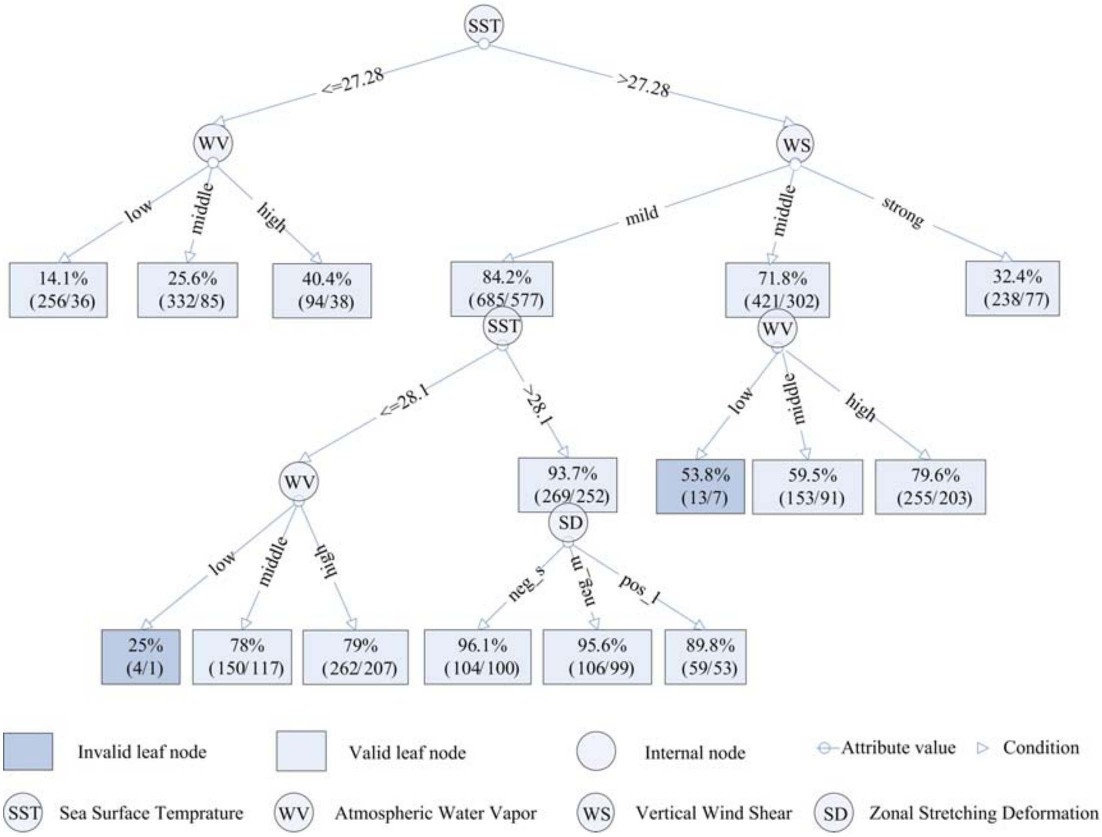
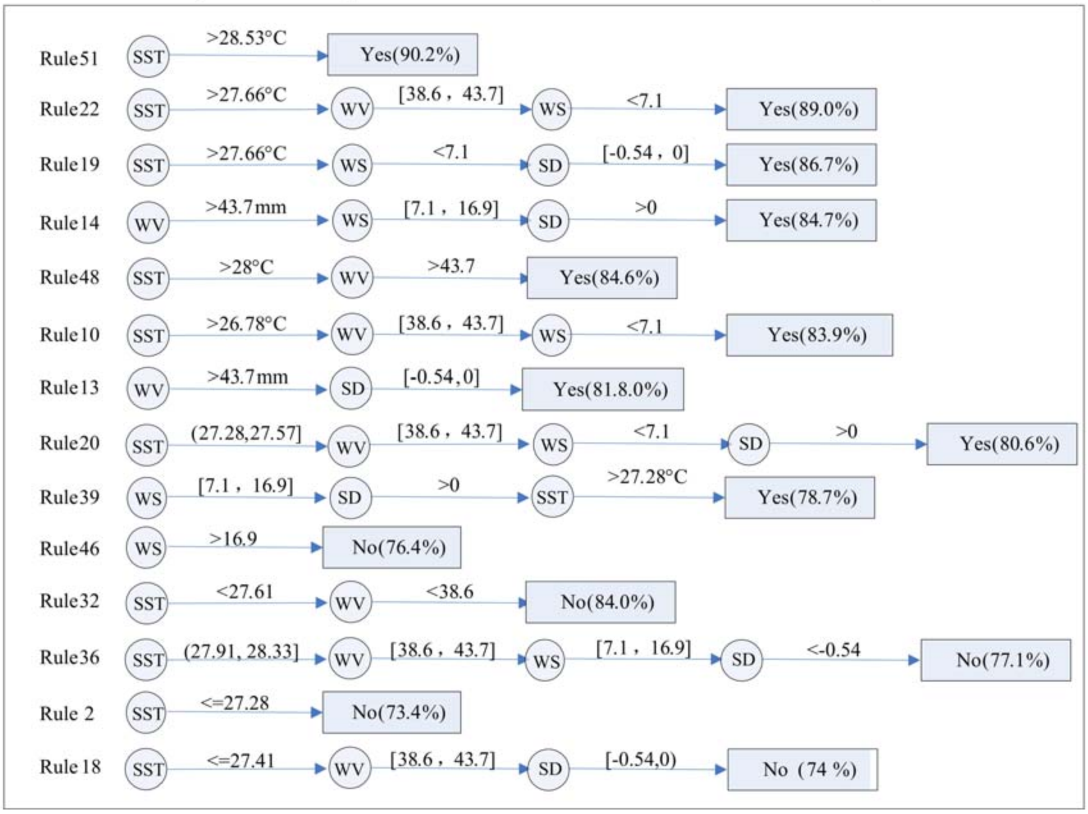


Figure 4 Production Rules generated and selected from decision trees. Class ‘Yes’ means that there is ‘tropical storm’, and class ‘No’ means that there is no ‘tropical storm’.


Rule51 SST >28.53°C Yes(90.2%)
Rule22 SST >27.66°C WV [38.6，43.7] WS <7.1 Yes(89.0%)
Rule19 SST >27.66°C WS <7.1 SD [-0.54，0] Yes(86.7%)
Rule14 WV >43.7mm WS [7.1，16.9] SD >0 Yes(84.7%)
Rule48 SST >28°C WV >43.7 Yes(84.6%)
Rule10 SST >26.78°C WV [38.6，43.7] WS <7.1 Yes(83.9%)
Rule13 WV >43.7mm SD [-0.54,0] Yes(81.8.0%)
Rule20 SST (27.28,27.57] WV [38.6，43.7] WS <7.1 SD >0 Yes(80.6%)
Rule39 WS [7.1，16.9] SD >0 SST >27.28°C Yes(78.7%)
Rule46 WS >16.9 No(76.4%)
Rule32 SST <27.61 WV <38.6 No(84.0%)
Rule36 SST (27.91, 28.33] WV [38.6，43.7] WS [7.1，16.9] SD <-0.54 No(77.1%)
Rule 2 SST <=27.28 No(73.4%)
Rule 18 SST <=27.41 WV [38.6，43.7] SD [-0.54,0) No（74 %)