2007

Improved Associated Conditions in Rapid Intensifications of Tropical Cyclones

Ruixin Yang
George Mason University

J. Tang
Natl Taipei Univ Business

Menas Kafatos
Chapman University, kafatos@chapman.edu

Follow this and additional works at: http://digitalcommons.chapman.edu/scs_articles

Part of the Atmospheric Sciences Commons, and the Meteorology Commons

Recommended Citation

This Article is brought to you for free and open access by the Science and Technology Faculty Articles and Research at Chapman University Digital Commons. It has been accepted for inclusion in Mathematics, Physics, and Computer Science Faculty Articles and Research by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.
Improved Associated Conditions in Rapid Intensifications of Tropical Cyclones

Comments
This article was originally published in Geophysical Research Letters, volume 34, in 2007. DOI:10.1029/2007GL031241

Copyright
Wiley
Improved associated conditions in rapid intensifications of tropical cyclones

Ruixin Yang,¹ Jiang Tang,¹ and Menas Kafatos¹

Received 5 July 2007; revised 29 August 2007; accepted 5 October 2007; published 31 October 2007.

¹Center for Earth Observing and Space Research, College of Science, George Mason University, Fairfax, Virginia, USA.

Copyright 2007 by the American Geophysical Union.

1. Introduction

Rapid intensification (RI) of tropical cyclones (TC) is a major error source in TC intensity forecasting. In order to improve the estimates of RI probability, association rules are used to facilitate the process of mining for candidate sets of conditions. Compared to the relation analysis method, the technique of association rules can simply explore associations among multiple conditions. Our mining results identified a reduced predictor set with fewer factors identified in previous studies but improved RI probabilities. That is, the RI probability with three conditions satisfied: low vertical shear, high humidity, and the TC being in an intensification phase is higher than that with five satisfied conditions including high sea surface temperature and an intensity far away from the maximum potential intensity in addition to the above three.


[1] Rapid intensification (RI) of tropical cyclones (TC) is a major error source in TC intensity forecasting. In order to improve the estimates of RI probability, association rules are used to facilitate the process of mining for candidate sets of conditions. Compared to the relation analysis method, the technique of association rules can simply explore associations among multiple conditions. Our mining results identified a reduced predictor set with fewer factors identified in previous studies but improved RI probabilities. That is, the RI probability with three conditions satisfied: low vertical shear, high humidity, and the TC being in an intensification phase is higher than that with five satisfied conditions including high sea surface temperature and an intensity far away from the maximum potential intensity in addition to the above three.

[2] Forecasting tropical cyclone (TC) intensity changes, rapid intensification (RI) in particular, is a challenge. As Kaplan and DeMaria [2003] defined, a TC undergoes RI if its intensity (defined by the maximum wind) has increased at least 30 knots (15.4 m/s) over a 24-hour period. The favorable factors for TC intensification have been broadly studied. Those factors include warm ocean eddies [Shay et al., 2000; Hong et al., 2000], the contraction of an outer eyewall [Willoughby et al., 1982; Willoughby and Black, 1996; Lee and Bell, 2007], an environment with low vertical shear [Gray, 1968; Merrill, 1988; DeMaria and Kaplan, 1994; DeMaria, 1996; Frank and Ritchie, 1999, 2001], interactions between the upper-level trough and a TC [Molinari and Vollaro, 1989, 1990; DeMaria et al., 1993], and even cloud microphysics [Wang, 2002] and isotopic concentrations [Gedzelman et al., 2003].

[3] Most of the previous studies were largely focused on only one of three categories of factors: ocean characteristics, inner-core processes, and environmental interactions, and it is well known that intensity changes depend on a combination of those factors [Gray, 1968; Zhu et al., 2004]. Holliday and Thompson [1979] examined the rapidly intensifying northwest Pacific typhoons and observed that a sufficiently deep layer of warm water, the development at night time, and a smaller eye size were favorable for those RI typhoons. DeMaria and Kaplan [1994] studied Atlantic TCs and found that the TCs with a smaller size, with a greater potential to reach their maximum potential intensity, with a faster intensification history, and in an environment with low vertical shear and weak upper-level forcing exhibited the largest 48-hour intensification rates. In a study of the RI process of Hurricane Opal (1985), Bosart et al. [2000] concluded that its RI was a result of a combination of several factors: enhanced upper-level divergence, low vertical shear and the enhanced heat and moisture from a warm Gulf eddy. More recently, Kaplan and DeMaria [2003] examined the large-scale characteristics of rapidly intensifying Atlantic TCs from 1989–2000 using the NHC HURDAT file and the SHIPS database. Their results confirmed the aforementioned studies. Furthermore a scheme to estimate the RI probability was developed in their study by combining the thresholds of the five persistence and synoptic predictors: the persistence of intensity change, the vertical shear, the sea surface temperature, the potential to reach the maximum potential intensity, and the moisture content in the lower atmosphere.

[4] Traditional statistical analysis, used extensively in the aforementioned studies to find the associations between rapidly intensifying TCs and their environmental properties, can be viewed as a type of “one-to-one” relation analysis technique. Compared to the “one-to-one” relation analysis, the technique of association rules from the data mining community can explore associations among multiple conditions without extra effort because it examines all possible combinations of frequent condition sets automatically in a large complex dataset [Agrawal et al., 1993]. This data mining technique provides an as complete as possible picture of the dataset to scientists so that the connections among multiple conditions will not be overlooked by a theory-driven analysis approach. The goal of this study is to introduce the technique of association rules from the data mining research area as an “unsupervised,” “automatic” data exploration method to discover “multiple-to-one” associations among a large number of environmental characteristics that are responsible for rapidly intensifying TCs. The results from this technique can then be used to shed light on the hypothesis generation regarding the underlying physical mechanisms, which can be used as guidance in the traditional statistical analysis.

[5] To meet the goal, a dataset consisting of all Atlantic basin TCs from 1982 to 2003 is used here to examine the ability of the association rule algorithm to discover the associated environmental conditions in rapid TC development. The dataset construction, the RI definition, RI thresholds and RI probability definitions are following those in Kaplan and DeMaria [2003, hereafter KD03] to the greatest possible extent. The purpose is to create a context similar to
Table 1. Case Numbers in This Study

<table>
<thead>
<tr>
<th>Time Coverage</th>
<th>Total</th>
<th>RI</th>
<th>Non-RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982–1988</td>
<td>1170</td>
<td>60</td>
<td>1110</td>
</tr>
<tr>
<td>1989–2000</td>
<td>3306</td>
<td>169</td>
<td>3137</td>
</tr>
<tr>
<td>2001–2003</td>
<td>1029</td>
<td>36</td>
<td>993</td>
</tr>
<tr>
<td>1982–2003</td>
<td>5505</td>
<td>265</td>
<td>5240</td>
</tr>
</tbody>
</table>

KD03 so that the results from the association rules can be easily mapped to and verified by the traditional statistical terms and methods.

2. Data and Methods

2.1. Datasets

[6] The datasets for this study are the NHC HURDAT file [Jarvinen et al., 1984] and the SHIPS 1982–2003 database [DeMaria and Kaplan, 1994, 1999; DeMaria et al., 2005]. A detailed description of the variables in the HURDAT and SHIPS datasets can be found in KD03. The HURDAT file consists of 6-hr estimates of position and maximum sustained surface wind speeds for all named Atlantic TCs from 1851 to the present. The SHIPS database contains synoptic information for every 12-hr for all Atlantic TCs from 1982 to the present. In this study, the time period is limited to 1982–2003 due to the initial data availability. One should note that the SHIPS database was recently updated to include information at 6-hr intervals [DeMaria et al., 2005]. The 12-hr interval was retained in this study to allow comparison of the results with KD03.

[7] The two datasets are merged based on a methodology which is identical to that described in KD03 except for the following conditions. First, KD03 considered the systems remaining over both water and tropical regions during the period from t = 12hr to t + 24hr, but we did not consider the locations in the samples. Moreover, in this study we did not include the non-developing tropical depressions, and excluding such data should not change the validation of the method and the general results although it will cause differences on detailed comparison with the results of KD03. In total, there are 34 values (variables) for each case, and each variable was evaluated at the beginning (t = 0h) of each 24-h period. The 0000 and 1200 UTC synoptic predictor values in the SHIPS database were averaged to estimate the magnitude of the corresponding values at 0600 and 1800 UTC. KD03 uses only SHIPS data from 1989 to 2000. Since we have more data from 1982 to 2003, we divided the data into 3 subsets: 1989–2000, 1982–1988, and 2001–2003. In other words, we used the first subset as a training set to repeat the results of KD03, and used the last two subsets as testing sets for the conclusion.

[8] Furthermore, we divided the data into rapidly intensifying cases (RI) and non-rapidly intensifying cases (non-RI) based on the RI definition proposed in KD03; at least 30 knots of intensity increase in the next 24 hours. After the records with missing values were removed from the dataset, we were left with a total of 5505 valid records with 265 RI cases. The numbers of cases after those divisions are listed in Table 1. It should be noted that the total case number of 3306 for the 1989–2000 period is higher than that in KD03 (2621) because we did not restrict samples based on locations.

2.2. Association Rule Algorithm

[9] To discover “multiple-to-one” associations among a large number of factors favoring rapidly intensifying TCs, we will “mine” the above cleaned data by using an association rule algorithm. Association rule induction [Agrawal et al., 1993] is a powerful method for market basket analysis, which aims at finding regularities in the shopping behavior of customers. An association rule is a rule like “Z <= X, Y.” The items X and Y are called antecedents in the rule and Z is the consequent. This rule expresses an association between items X, Y, and Z. It states that if a customer is picked randomly and the customer selected items X and Y, it is likely that the customer also selected item Z. The number of antecedents can range from one to the total number of items in a database.

[10] Three parameters are reported for association rules, namely the support estimating the probability P({X, Y, Z}), the confidence defined as the probability P(Z|{X, Y}), and the lift measuring the strength of a rule “Z <= X, Y,” by the ratio between the actual probability of the item set containing both antecedent and consequent divided by the product of the individual probabilities of the antecedent set and the consequent, P({X, Y, Z})/P({X, Y})P(Z). An association rule Z <= X, Y is strong if it has a large support, a high confidence and a large lift.

[11] The version of the association rule algorithm we used in this study is implemented by C. Borgelt (Apriori-association rule induction/frequent item set mining, available at http://www.borgelt.net/apriori.html). The support value in this implementation is defined as P({X, Y}) instead of P({X, Y, Z}), and this definition P({X, Y}) is the same as those used in KD03 for description. Since no previous experience exists on this specific data mining task for RI hurricanes, no specific control parameters (predefined support, confidence, and lift) are chosen. Instead, results are compared to the results of KD03, and the rules given significant results are discussed.

2.3. Data Discretization

[12] For the mining task, the original 34 attributes are reduced to 11 independent predictors as listed in Table 2. These predictors are chosen because KD03 found that the mean initial conditions of these predictors for RI cases and non-RI cases are statistically different at least at the 95% significance level based on an unequal-variance, two-sided t test. To mine those attributes with the association rule algorithm, we should convert continuous values of the attributes into disjoint conditions. Here, we divide the values into “High” or “Low” conditions based on the threshold values provided in KD03, which was derived from the mean values of the RI samples, as listed in Table 2.

3. Results and Discussion

[13] When observing that the RI probability for any individual predictor was not particularly high, Kaplan and DeMaria (KD03) tried to find a set of predictors to provide improved RI probability estimates. They found that “employing the set of [seven] predictors that were signif-
The predictors DVMX, SHR, and SLYR in KD03 are renamed as PD12, SHRD, and PSLV here. A single, double, or triple asterisk was placed after the predictor names if the results of a two-sided t test indicated the mean difference of the predictor was statistically significant at either the 95% (single asterisk), 99% (double asterisk), or 99.9% (triple asterisk) level.

During the pruning process, the original rule with the five KD03 constraints satisfied, “RI <= SHRD = L, PD12 = H, SST = H, POT = H, RHLO = H (supp = 0.7%, conf = 43.5%, lift = 850.5%),” is removed as a redundant rule. The above cited result can be easily translated into an association rule as “RI <= SHRD = L, PD12 = H, SST = H, POT = H, RHLO = H” with a support of the antecedent as 15% (J. Kaplan, personal communication, 2005) and a confidence of the rule as 41%. Indeed, the exact associate rule with a support = 0.7%, a confidence = 43.5% and a lift = 850.5% is found by using the association rule technique. The lower support is likely due to the larger portion of non-RI samples in our dataset. However the higher confidence is likely due to the excluded non-developing depressions. We suspect that the non-developing depressions (1989–2000) may contain a higher proportion of cases satisfying the five conditions at the same time without undergoing RI, which resulted in the confidence of 43.5% in our study, higher than the reported 41% from KD03.

The pruning process of association rules can help us to easily identify conditions for improved RI probabilities. The pruning process would first identify redundant rules, those with confidence values being equal to or smaller than that of a more general rule. A redundant rule cannot provide improved RI probabilities and will be removed, while the ones with improved RI probabilities over other rules will survive the pruning process.

To compare the results with those in KD03, the composite RI probability defined in KD03 is computed with statistical analysis. The results for the time periods 1982–2003, 1982–1988, 1989–2000, and 2001–2003 are displayed in Figure 1. The plots for the whole dataset from 1982 to 2003 and from 1989 to 2000 are similar to the results of KD03. When the number of satisfied constraints increases, the RI probability increases. However, there are no RI cases which satisfied all 5 constraints at the same time in the datasets for the 1982–1988 and the 2001–2003 periods, thus the probabilities decrease to zeros. The RI probabilities for any 4 satisfied constraints also decrease in these two datasets. This observation indicates that not all 5 conditions are required for a TC to undergo RI.

The zero RI probability with all five constraints satisfied is an interesting result. More detailed investigation indicates that not only was no case found satisfying all five constraints at the same time for RI cases, but the number of records with all five identified conditions for both RI and non-RI cases are much lower for the 1982–1988 and 2001–2003 time periods. In the 22 years from 1982 to 2003, there are a total of 31 cases with all five RI favoring conditions. Twenty-three of them fall in the 12-year base period, 1989–2000; six cases are in the last three years, 2001–2003. For the seven years from 1982 to 1988, only two cases satisfying all five constraints are found. This result may suggest that the five constraints are not the fundamental factors for RI.

Table 2. Eleven Statistically Significant Predictors (From KD03, Table 4)

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD12 ***</td>
<td>Intensity change during the previous 12 hours.</td>
<td>4.6 m/s</td>
</tr>
<tr>
<td>SHRD ***</td>
<td>850–200 hPa vertical shear.</td>
<td>4.9 m/s</td>
</tr>
<tr>
<td>SST ***</td>
<td>Sea surface temperature.</td>
<td>28.4°C</td>
</tr>
<tr>
<td>POT ***</td>
<td>Maximum potential intensity (MPI) – initial intensity.</td>
<td>47.6 m/s</td>
</tr>
<tr>
<td>RHLO ***</td>
<td>850–700 hPa relative humidity.</td>
<td>69.7%</td>
</tr>
<tr>
<td>LAT ***</td>
<td>Latitude</td>
<td>19.7°N</td>
</tr>
<tr>
<td>LON *</td>
<td>Longitude</td>
<td>63.2°W</td>
</tr>
<tr>
<td>USTM **</td>
<td>Zonal (u) component of storm motion.</td>
<td>−3.1 m/s</td>
</tr>
<tr>
<td>U200 ***</td>
<td>200 hPa zonal (u) component of wind.</td>
<td>−0.6 m/s</td>
</tr>
<tr>
<td>REFEC **</td>
<td>200 hPa relative eddy angular momentum flux convergence</td>
<td>0.9 m/s/day</td>
</tr>
<tr>
<td>PSLV **</td>
<td>Pressure of the center of mass of layer for which the environment winds best match the current storm motion.</td>
<td>583.4 hPa</td>
</tr>
</tbody>
</table>

*The predictors DVMX, SHR, and SLYR in KD03 are renamed as PD12, SHRD, and PSLV here. A single, double, or triple asterisk was placed after the predictor names if the results of a two-sided t test indicated the mean difference of the predictor was statistically significant at either the 95% (single asterisk), 99% (double asterisk), or 99.9% (triple asterisk) level.**
least one constraint satisfied for the cases of the five identified constraints and of the three reduced constraints. However, when at least two constraints are satisfied, the probabilities with three identified constraints are significantly higher than the corresponding probabilities in the five constraint cases. The most significant result of this work is that when all three identified constraints are satisfied, the RI probabilities are higher than the probabilities with all five satisfied constraints identified in KD03. The RI probabilities for the whole data and the base 1989–2000 period with all five KD03 conditions satisfied are 32.3% and 43.5%, respectively. The corresponding numbers with the three conditions mined out rises to 33.0% and 47.6%. This result is significant not only to the rapid intensification of hurricanes but also to the data mining procedures in identifying meaningful scientific results.

It is worth noting that the new reduced predictor list does not include “$SST = H$,” high sea surface temperature, which is commonly considered as the main factor influencing the TC intensity and its change. Nevertheless, it was noticed that although SST is an important factor in hurricane intensity, it is not a controlling factor [Evans, 1993; Paterson et al., 2005]. Even still, as pointed by DeMaria and Kaplan [1994], the role of SST can be reflected through the differences between the maximum potential intensity and initial intensity, the potential (POT). In this work, the threshold of potential is 47.6 m/s based on KD03. However, since RI is defined by a 30 knot (about 15 m/s) wind increase in next 24 hours, a more logical threshold for POT should be 15 m/s. We will explore the impact of this threshold value in future research. Moreover, Emanuel et al. [2004] explored the sensitivities of TC intensity changes to initial conditions and environmental factors by using a simple coupled model with wind shear parameterization. Their results demonstrated that the intensity changes are very sensitive to wind shear and humidity in middle troposphere air, which is very consistent with the short list of RI favoring factors, “$SHRD = L$, $RHLO = H$.”

4. Concluding Remarks

Compared to statistical analysis, the technique of association rules can explore associations among multiple conditions without extra effort because it examines all possible combinations of frequent condition sets automatically. It provides an as complete as possible picture of the dataset to scientists so that the connections among multiple conditions will not be overlooked by a theory-driven analysis approach. Compared to the statistical analysis of KD03, the data mining technique used in this work not only identified the predictors giving an improved RI probability but also obtained this result with fewer predictors through a pruning process of association rules. That is, the RI probability with three conditions satisfied: low vertical shear of horizontal wind ($SHRD = L$), high humidity in the 850–700 hPa level ($RHLO = H$), and the TC being in an intensification phase ($PD12 = H$) is higher than that with five satisfied conditions including high sea surface temperature ($SST = H$) and an intensity far away from the maximum potential intensity (POT = H) in addition to the above three. This work demonstrates that the association rule data mining technique can be used as an exploration tool to explore the rapid intensification of hurricanes.
method to generate hypotheses, and the statistical analysis should be performed as confirmation of the hypotheses, as generally expected for data mining applications.

[22] Acknowledgments. The authors would like to thank Dr. Mark DeMaria for providing the 2003 SHIPS data and Dr. Christian Borgelt for making his implementation of the Apriori association rule algorithm available. The authors also thank anonymous reviewers for their comments and suggestions which help to improve this manuscript and especially one of them for pointing out the physical POT threshold value. This work was partially supported by NASA Grant NNX06AF30G.

References


M. Kafatos, J. Tang, and R. Yang, Center for Earth Observing and Space Research, College of Science, George Mason University, Fairfax, VA 22030-4444, USA, (ryang@gmu.edu)