Discriminating Developing versus Nondeveloping Tropical Disturbances in the Western North Pacific through Decision Tree Analysis

WEI ZHANG

Key Laboratory of Meteorological Disaster, Ministry of Education, and Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, and Climate Dynamics Research Center and Earth System Modeling Center, Nanjing International Academy of Meteorological Sciences (NIAMS), and Nanjing University of Information Science and Technology, Nanjing, China

BING FU

International Pacific Research Center, University of Hawai'i at Mānoa, Honolulu, Hawaii

MELINDA S. PENG

Marine Meteorology Division, Naval Research Laboratory, Monterey, California

TIM LI

International Pacific Research Center, and Department of Atmospheric Science, University of Hawai'i at Mānoa, Honolulu, Hawaii

(Manuscript received 21 February 2014, in final form 6 January 2015)

ABSTRACT

This study investigates the classification of developing and nondeveloping tropical disturbances in the western North Pacific (WNP) through the C4.5 algorithm. A decision tree is built based on this algorithm and can be used as a tool to predict future tropical cyclone (TC) genesis events. The results show that the maximum 800-hPa relative vorticity, SST, precipitation rate, divergence averaged between 1000- and 500-hPa levels, and 300-hPa air temperature anomaly are the five most important variables for separating the developing and nondeveloping tropical disturbances. This algorithm also unravels the thresholds of the five variables (i.e., $4.2 \times 10^{-5}\,s^{-1}$ for maximum 800-hPa relative vorticity, 28.2°C for SST, $0.1\,mm\,h^{-1}$ for precipitation rate, $-0.7 \times 10^{-6}\,s^{-1}$ for vertically averaged convergence, and 0.5°C for 300-hPa air temperature anomaly). Six rules are derived from the decision tree. The classification accuracy of this decision tree is 81.7% for the 2004–10 cases. The hindcast accuracy for the 2011–13 dataset is 84.6%.

1. Introduction

Tropical cyclones (TCs) in the western North Pacific (WNP) bring massive amounts of damage to coastal regions each year (Zhang et al. 2009; Xiao and Xiao 2010). TC genesis has aroused extensive interest within the academic and operational communities (Gray 1968, 1998; Emanuel 1989; Peng et al. 2012). TC genesis is referred to as a process through which a tropical disturbance rapidly develops into a warm-core, cyclonic system with sustainable winds (Gray 1968, 1979, 1998; Fu et al. 2007).

Over the decades, significant advancements have been made in understanding the physical mechanisms and processes involved in TC genesis (Gray 1968; McBride 1981; Craig and Gray 1996; Fu et al. 2007; Wang et al. 2007; Peng et al. 2012; Fu et al. 2012). Gray (1968) suggested several favorable environmental parameters for TC genesis: a sufficiently deep warm ocean layer, conditional instability through a deep atmospheric layer, higher-than-normal midtropospheric relative humidity, above-normal low-level vorticity, weak vertical wind shear over the center of the circulation, and a location far enough from the equator. The first

Corresponding author address: Bing Fu, International Pacific Research Center, 1680 East–West Rd., POST 408, Honolulu, HI 96822. E-mail: bingf@hawaii.edu

DOI: 10.1175/WAF-D-14-00023.1

three parameters are considered to be thermodynamic factors, while the others are dynamic. As those six parameters are usually correlated with each other, Frank (1987) has reduced the six variables to four as follows. Low-level vorticity and the Coriolis parameter are combined into absolute vorticity. In addition, conditional instability is deleted and average vertical motion is added to replace relative humidity. While the thermodynamic conditions and Coriolis parameter are satisfied over a large portion of the tropical ocean for a long period of time, the low-level vorticity and vertical shear vary remarkably on much smaller spatial and temporal scales (Gray 1968; McBride 1981). It has been hypothesized that tropical cyclogenesis occurs when above-normal low-level vorticity and locally weak vertical wind shear are fulfilled within a thermodynamically favorable environment (Gray 1968; McBride 1981). Gray (1998) further reported upon background environmental conditions required for TC genesis and emphasized the important roles of climatology (i.e., region, season, and SST), synoptic flow patterns (monsoon trough or large vorticity with small vertical wind shear), and active mesoscale convective systems (MCSs) within a cloud cluster system.

The processes involved in TC genesis can be subdivided into two consecutive stages (Briegel and Frank 1997; Ritchie and Holland 1999; Nolan 2007). The first stage is known as the transition process from a disturbance to a depression: the initial formation of a rotational circulation with a radius of a few hundred kilometers. The second stage is the transition from a tropical depression to a tropical storm (Briegel and Frank 1997; Ritchie and Holland 1999; Nolan 2007). Two crucial theories, conditional instability of the second kind (CISK; Charney and Eliassen 1964; Ooyama 1969) and wind-induced surface heat exchange (WISHE; Emanuel 1986; Craig and Gray 1996), have been developed to interpret the self-exciting mechanisms that are closely linked to TC formation. CISK and WISHE can largely explain the rapid development of disturbances in the second stage of the two-stage TC formation.

Several indices have been widely used to quantitatively evaluate the potential for TC genesis (e.g., Gray 1979, 1998; Camargo et al. 2007a,b; Nolan 2007). Gray (1979, 1998) developed an index to replicate key features of the seasonal and spatial variability of observed TC genesis using several environmental parameters. A genesis potential index (GPI) has been proposed to represent the potential of TC formation (Emanuel and Nolan 2004) on the basis of Gray's index (Gray 1979). A parameter has also been developed to evaluate the potential for TC formation in the North Atlantic between Africa and the Caribbean Islands (DeMaria et al. 2001). Camargo et al. (2007c) came up with another genesis index constructed from composites of previous GPIs (Gray 1979; Emanuel and Nolan 2004) with respect to both the annual cycle and El Niño–Southern Oscillation (ENSO). More recently, the box difference index (BDI), which accounts for both the mean and variability of an individual parameter, has been introduced to identify controlling parameters measuring the differences between developing and nondeveloping disturbances (Peng et al. 2012; Fu et al. 2012). In their studies, dynamic and thermodynamic variables are highlighted for TC genesis in the WNP and the North Atlantic, respectively (Fu et al. 2012; Peng et al. 2012). BDI has been proved to be useful in examining the future trends of TC genesis (Murakami et al. 2013).

A number of schemes have been developed to predict TC genesis in the WNP and the North Atlantic (e.g., Nicholls 1979; Elsner and Schmertmann 1993; Chia and Ropelewski 2002; Venkatesh and Mathew 2004; Fan 2010). For example, prediction schemes have been implemented through discriminant analysis (Hennon and Hobgood 2003), neural networks (Hennon et al. 2005), and cluster analysis (Hennon et al. 2011).

In spite of those advancements, TC genesis still remains a challenging problem because of the multiscale interactions and a lack of in situ observations over open oceans (Gray 1968, 1998; Emanuel 1989; Peng et al. 2012). The predictions of TC genesis still have large errors in all the ocean basins, though many variables have been suggested to be effective in predicting TC genesis in previous studies.

The C4.5 algorithm is a useful machine learning method and a classic decision tree algorithm, which can deal with inherent nonlinear relationships in variables and missing values (Quinlan 1987, 1993; Fayyad 1997; Fayyad and Stolorz 1997). Moreover, this algorithm enables the quantification of the relative importance of variables and builds decision rules for prediction (Quinlan 1987, 1993). The C4.5 algorithm has aroused tremendous attention because the structure of its classification is visually explicit and readily interpreted (Friedl and Brodley 1997). It is designed for classifying samples into different groups (two classes or more) through information theory (Quinlan 1987, 1993). It has been successfully employed in analyses of the recurvature, landfall, and intensity change of TCs in the WNP (Zhang et al. 2013a–c). In our case, the analysis of TC genesis falls into a binary classification problem, which classifies developing and nondeveloping tropical disturbances using potential factors of TC genesis.

The objective of this study is to quantify the relative importance of those potential factors for TC genesis, to derive rules for predicting TC genesis, and to complement

the existing BDI in analyzing TC genesis. This study also aims at a better understanding of TC genesis and improvements to the techniques of TC genesis prediction.

The remainder of this paper is organized as follows. Section 2 presents a description of the data and methodology. Section 3 is used to present and interpret the research findings as well as compare our results with those of previous studies. A summary of the findings is given in section 4.

2. Data and methodology

a. Data

Daily global analysis data are obtained from the Navy Operational Global Atmospheric Prediction System (NOGAPS). Because synoptic disturbances occur on a scale from hundreds to a thousand kilometers, this $1° \times 1°$ resolution dataset should be sufficient to examine synoptic disturbances in the tropics. Although NOGAPS provides 6-hourly analysis data (i.e., 0000, 0600, 1200, and 1800 UTC), we only utilize the 0000 UTC data. The data used in this study range from 2004 to 2013. SST and the precipitation rate are derived from the Tropical Rainfall Measuring Mission (TRMM) Microwave Imager (TMI). TRMM is a jointly collaborative mission launched by the National Aeronautics and Space Administration (NASA) and the National Space Development Agency of Japan (NASDA). The global tropical oceans (40°S–40°N) are covered by TMI data, which are sufficient for examining tropical synoptic disturbances.

b. Selection of disturbances

The tropical disturbances selected for this study are confined from the equator to 30°N and from 105°E to 180°. We excluded those disturbances north of 30°N because they are closely related to midlatitude weather systems. Also, only tropical disturbances in the WNP from June to September are investigated. The daily 850-hPa relative vorticity fields are used to extract tropical disturbances. The criterion for selecting tropical disturbances is that at least nine grid points ($1° \times 1°$ resolution) whose values of relative vorticity are equal to or greater than $2 \times 10^{-5}\,s^{-1}$ can be found in a $4° \times 4°$ square box centered at the grid with maximum relative vorticity. This criterion is empirically obtained by examining all the tropical disturbances in vorticity fields. It is determined by considering that none of the prestorm disturbances is missed and also there are not too many disturbances.

The samples of developing disturbances are derived from the Joint Typhoon Warning Center (JTWC). We define day −1 as 24 h prior to formation. This study only focuses on 24–48-h TC genesis events. Therefore, we select day −1 samples for developing cases and all other days' samples for nondeveloping ones. As shown in Table 1, there are a total of 5261 nondeveloping cases and 159 developing cases for the period 2004–13 (from June to September). There are 1746 nondeveloping cases and 53 developing ones in the period 2011–13. The decision tree model is trained by data in 2004–10, and data in 2011–13 are used for hindcast validation.

TABLE 1. Number of developing and nondeveloping tropical disturbances during 2004–10, 2011–13, and 2004–13.

Period	Developing	Nondeveloping	Total
2004–10	106	3515	3621
2011–13	53	1746	1799
2004–13	159	5261	5420

After the developing and nondeveloping disturbances are identified, we calculate the values for each variable (Table 2). Table 2 shows the dynamic and thermodynamic variables that potentially influence the development of tropical disturbances into TCs. For example, the 300-hPa temperature anomaly is a thermodynamic factor. Corresponding to each tropical disturbance, this variable is derived by averaging over a square of $10° \times 10°$ centered at the maximum relative vorticity grid (Table 2). In contrast, vertically averaged divergence is a dynamic factor for TC genesis and it is calculated by averaging over a square of $5° \times 5°$ centered at the maximum relative vorticity grid (Table 2). The other variables are calculated likewise. The square box size used for calculating parameters is determined by BDI. We calculated the BDIs of those parameters with different square box sizes. The results show that generally smaller square box sizes for dynamic parameters can yield larger BDIs. This suggests that a smaller average area for dynamic parameters is better for distinguishing developing and nondeveloping disturbances. The size of the averaging area is also an indication of a storm-scale parameter or an environmental large-scale parameter. Previous studies (DeMaria and Kaplan 1994; DeMaria et al. 2005) showed that the Statistical Hurricane Intensity Prediction Scheme (SHIPS) model is also sensitive to the size of the averaging area used to calculate the predictors.

c. The C4.5 algorithm

This study aims to classify developing and nondeveloping tropical disturbances. The C4.5 algorithm uses a decision tree methodology to define rule sets that can be used to assign input data into two or more classes. Therefore, it is used in this study to select variables, thresholds, and rules for the classification of developing and nondeveloping tropical disturbances. A wide variety of decisions or controls come into play when running this

TABLE 2. List of variables for the analysis of TC genesis.

Variable	Name in the algorithm
300-hPa temperature anomaly (10° × 10°)	m_air300
Vertically averaged divergence (5° × 5°)	m_div1000_500
Vertically averaged *du*/*dy* (10° × 5°)	m_dudy1000-400
Precipitation rate (10° × 10°)	m_prate
950-hPa relative humidity (5° × 5°)	m_rhum950
Sea surface temperature (10° × 10°)	m_sst
Max 800-hPa relative vorticity	m_vor800
925–400-hPa water vapor content (5° × 5°)	m_wvc925-400

algorithm. The detailed procedures and parameter settings of this algorithm are provided in the appendix.

d. Cross validation

Cross validation is a common scheme used to verify a learning model. The k-fold cross validation is described as follows. The total dataset is separated into k equal subsets. Training and validation are performed for k iterations. In an individual iteration, one subset is utilized for validation while the remaining subsets (i.e., $k - 1$ subsets) are used for training the model. The prediction accuracy is calculated by dividing the correctly classified samples by the number of samples in the whole dataset. Cross validation assures that each sample of the dataset can be used for training and testing, and in a single iteration the training and testing samples are independent. This thus lends high credence to the generalization capability of the obtained decision tree. This study uses 10-fold cross validation to test the capability of the model built by the C4.5 algorithm.

3. Results and discussion

A decision tree consisting of six rules is built through the C4.5 algorithm to classify developing and nondeveloping tropical disturbances. These rules refer to five meteorological variables. Table 3 shows these variables and their relative importance, which is determined by the sequence in which a variable is selected by the C4.5 algorithm. Among those variables, a dynamic variable (the maximum 800-hPa relative vorticity) is the first selected variable, followed by two thermodynamic variables (i.e., SST and precipitation rate; Table 3). The remaining two variables include one dynamic variable (vertically averaged divergence) and one thermodynamic variable (air temperature anomaly at 300 hPa; Table 3).

TABLE 3. Variables selected to build the decision tree and their relative importance.

Variable	Full name	Order
vor800	Max 800-hPa relative vorticity	1
SST	Sea surface temperature	2
prate	Precipitation rate	3
div1000–500	Vertically averaged divergence	4
air300	Air temperature at 300 hPa	5

Among the six rules included in this decision tree (Table 4), the highest accuracy is 99.5% whereas the lowest is only 52.6%. The classification accuracy of the decision tree is 81.7% by 10-fold cross validation. The confusion matrix shows that 2664 (3081) nondeveloping (developing) observations are correctly classified from 3515 (3515) samples (Table 5).

Rule 1 is stated as follows: if vor800 $\le 4.2 \times 10^{-5}\,s^{-1}$, then the tropical disturbance will not develop into a TC. This rule is formed by one single variable, the maximum 800-hPa relative vorticity, which is the first variable selected to construct the decision tree. Therefore, this rule highlights the importance of maximum 800-hPa relative vorticity in TC genesis. This is consistent with the results from Fu et al. (2012), who suggested that dynamic variables play a more important role than thermodynamic variables in TC genesis in the WNP.

Rule 2 is stated as follows: if vor800 $> 4.2 \times 10^{-5}\,s^{-1}$ and SST ≤ 28.2°C, then the tropical disturbance will not develop into a TC. This rule involves a dynamic variable

TABLE 4. Descriptions and accuracies of the rules derived from the decision tree.

Rule No.	Rule description	Accuracy
1	If vor800 $\le 4.2 \times 10^{-5}\,s^{-1}$, then the tropical disturbance will not develop into a TC	(1949 − 196)/1949 = 84.6%
2	If vor800 $> 4.2 \times 10^{-5}\,s^{-1}$ and SST ≤ 28.2°C, then the tropical disturbance will not develop into a TC	(193 − 1)/193 = 99.5%
3	If vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, and prate $\le 0.1\,mm\,h^{-1}$, then the tropical disturbance will not develop into a TC	(49 − 2)/49 = 95.9%
4	If vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, prate $> 0.1\,mm\,h^{-1}$, and div1000–500 $> -0.7 \times 10^{-6}\,s^{-1}$, then the tropical disturbance will not develop into a TC	(125 − 34)/125 = 72.8%
5	If vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, prate $> 0.1\,mm\,h^{-1}$, div1000–500 $\le -0.7 \times 10^{-6}\,s^{-1}$, and air300 ≤ 0.5°C, then the tropical disturbance will not develop into a TC	(152 − 72)/152 = 52.6%
6	If vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, prate $> 0.1\,mm\,h^{-1}$, div1000–500 $\le -0.7 \times 10^{-6}\,s^{-1}$, and air300 > 0.5°C, then the tropical disturbance will develop into a TC	(2433 − 467)/2433 = 80.8%

TABLE 5. Confusion matrix of the decision tree by 10-fold cross validation.

		Predicted	
		Nondeveloping	Developing
Observed	Nondeveloping	2664	851
	Developing	434	3081

(i.e., maximum 800-hPa relative vorticity) and a thermodynamic variable (i.e., SST), and has the highest classification accuracy (99.5%) among the six rules in this decision tree. This indicates that SST is the second most important variable for TC development. It is noted that SST is ranked as sixth in Fu et al. (2012). SST is a key variable for TC genesis (Gray 1968, 1979; Dare and McBride 2011). Gray (1998) reported that a necessary condition for TC genesis is that SST be greater than 26°C. On a global scale, TCs form over a small temperature range (i.e., 90.4% of TCs formed over the tropical oceans where SST is between 27.5° and 30.5°C; Dare and McBride 2011). This rule suggests that even though the maximum 800-hPa relative vorticity is higher than $4.2 \times 10^{-5}\,s^{-1}$, the tropical disturbance will not develop into a TC if the SST is lower than 28.2°C. This study confirmed previous studies and further unraveled a threshold of 28.2°C, under which TCs have a low chance to form in the WNP.

Rule 3 is stated as follows: if vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, and prate $\leq 0.1\,mm\,h^{-1}$, then the tropical disturbance will not develop into a TC. This rule has a classification accuracy of 95.9%. Precipitation rate is the third variable used to build the decision tree following the maximum 800-hPa relative vorticity and SST. It is noted that precipitation rate is ranked second in Fu et al. (2012). Precipitation rate has also been emphasized to be a key thermodynamic variable for TC development because latent heat release is a crucial factor for the amplification of a tropical disturbance (Kuo 1965; Li et al. 2003). This study corroborates with previous studies that precipitation rate plays a crucial role in developing TCs in the WNP.

Rule 4 is stated as follows: if vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, prate $> 0.1\,mm\,h^{-1}$, and div1000–500 $> -0.7 \times 10^{-6}\,s^{-1}$, then the tropical disturbance will not develop into a TC. This rule consists of four variables: maximum 800-hPa relative vorticity, SST, precipitation rate, and average low-level divergence. This rule includes the information that weak low-level convergence or divergence is not favorable for TC genesis. This rule highlights the key role of strong low-level convergence in TC development, which has been widely mentioned in previous studies (Emanuel 1986; Craig and Gray 1996).

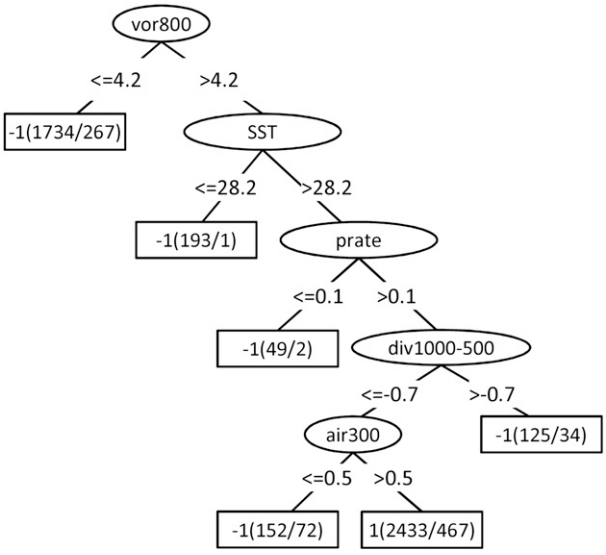

FIG. 1. The decision tree built by the C4.5 algorithm for classifying developing and nondeveloping disturbances. The ellipses contain selected variables and the numbers in the rectangles indicate the predicted class label (1, developing; −1, nondeveloping), the total number of samples from both classes satisfying the conditions, and the number of misclassified samples.

Rule 5 is stated as follows: if vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, prate $> 0.1\,mm\,h^{-1}$, div1000–500 $\leq -0.7 \times 10^{-6}\,s^{-1}$, and air300 ≤ 0.5°C, then the tropical disturbance will not develop into a TC. This rule includes one more variable: an upper-level air temperature anomaly. It has been reported in previous studies that the upper-level warm core plays an important role in the development and intensification of TCs (Zhang and Chen 2012; Chen and Zhang 2013). Although the 300-hPa air temperature anomaly is not selected in Fu et al. (2012), it is ranked fifth in this study, indicating that the formation of an upper-level warm core is necessary for TC development.

Rule 6 is stated as follows: if vor800 $> 4.2 \times 10^{-5}\,s^{-1}$, SST > 28.2°C, prate $> 0.1\,mm\,h^{-1}$, div1000–500 $\leq -0.7 \times 10^{-6}\,s^{-1}$, and air300 > 0.5°C, then the tropical disturbance will develop into a TC. This rule has an accuracy of 80.8% and is the only rule for the developing group of tropical disturbances. This rule reveals the five necessary conditions for a tropical disturbance to be classified as a developing case.

To verify the accuracy of this decision tree in classifying developing and nondeveloping disturbances, both cross validation and hindcast analyses are conducted. This decision tree has been verified via 10-fold cross validation (Fig. 1). It did excellently in classifying developing and nondeveloping cases, with an accuracy of 81.7%. It performed even better in hindcasting 2011–13 cases. For example, 34 (1488) developing (nondeveloping) cases are correctly predicted from 53 (1746) cases during

TABLE 6. Confusion matrix of the hindcast for the period 2011–13.

		Predicted	
		Nondeveloping	Developing
Observed	Nondeveloping	1488	258
	Developing	19	34

2011–13 (see Table 6), which gives a hindcast accuracy of 84.6%.

There are few studies focused on short-range TC genesis forecasting, either from numerical models or from statistical approaches. Recently, Halperin et al. (2013) evaluated the forecast performance of five individual numerical models: the Canadian Meteorological Centre's (CMC) Global Environmental Multiscale (GEM) model, the European Centre for Medium-Range Weather Forecasts (ECMWF) global model, the Global Forecast System (GFS), the Navy Operational Global Atmospheric Prediction System, and the Met Office global model (UKMET), as well as the combination of these models in the Atlantic from 2004 to 2011. For 24–48-h TC genesis forecasts, neither the individual models nor the combination of them can have a hit rate greater than 50%. Also, they all have high false alarm rates, similar to the hit rates. Our results show the decision tree method produces a slightly higher hit rate (64%) and a much better false alarm rate (15%) in the WNP. Another statistical approach (manuscript in preparation) obtained around a 65% hit rate in the North Atlantic for 24–48-h TC genesis forecast. The results suggest that, for short-range (24–48 h) TC genesis forecasts, a statistical approach is better than the use of global model forecasts.

Gray's parameter (Gray 1968, 1979) has been widely accepted to develop forecast indices for TC formation. However, most of the indices are used for evaluating the potential of TC formation. Unlike our approach, which only focuses on the individual tropical disturbances, those indices are calculated at each grid point of all of the tropical oceans. For example, DeMaria et al. (2001) developed a TC genesis index for the tropical Atlantic. They showed the index is useful for identifying periods with above- and below-normal probabilities of TC genesis. Emanuel and Nolan (2004) defined an index for the TC genesis potential of global tropical oceans. The annual cycle of TC occurrences is successfully derived using this index in both the Northern and Southern Hemispheres. For a statistical approach, Perrone and Lowe (1986) utilized discriminant analysis to predict TC formation. The accuracy of 24-h prediction is 72% for developing cloud clusters and 93% for nondeveloping ones in the WNP. Ward (1995) developed a hurricane index for evaluating intensification potential for individual disturbances in the western South Pacific basin. SST, vorticity, and vertical wind shear are included in this index. He verified this index by the use of a few cases and suggested that a threshold value of 24 can be used to discriminate between developing and nondeveloping disturbances. Compared to existing approaches, our findings have shown promising results for forecasting individual TC genesis events in the WNP.

4. Summary

TC genesis has been a mystery in the field of atmospheric sciences for a long time. Numerous studies have been carried out to reveal this mystery from a wide spectrum of perspectives. Based on some important parameters for TC genesis, several TC genesis potential indices have been developed for different ocean basins (Gray 1979, 1998; Camargo et al. 2007a,b; Nolan 2007). Recently, the BDI has been introduced to assess the relative importance of potential variables for TC genesis in the North Atlantic and the WNP (Peng et al. 2012; Fu et al. 2012; Murakami et al. 2013). The BDI has proved to be useful in evaluating the potential of TC genesis for current climate and global warming.

This study used the C4.5 algorithm to obtain the relative importance of potential variables, determine the thresholds (splitting values) of each variable, and create the rules for the classification of the developing and nondeveloping cases. More importantly, the rules obtained by the decision tree can be easily used for the prediction of future TC genesis events.

The research findings are summarized as follow:

1) A decision tree is built by the C4.5 algorithm. Maximum 800-hPa relative vorticity, SST, precipitation rate, divergence averaged between 1000- and 500-hPa levels, and 300-hPa air temperature anomalies are selected to separate the developing and nondeveloping tropical disturbances. This algorithm determines the splitting values of the four variables (e.g., $4.2 \times 10^{-5}\,s^{-1}$ for maximum 800-hPa relative vorticity, 28.2°C for SST, $0.1\,mm\,h^{-1}$ for precipitation rate, $-0.7 \times 10^{-6}\,s^{-1}$ for average convergence, and 0.5°C for 300-hPa air temperature anomaly).
2) Six rules are derived by tracking the root node to each leaf node and combining the splitting values and variables. The decision tree has a classification accuracy of 81.7% for the 2004–10 dataset and hindcast accuracy of 84.6% for 2011–13.

The thresholds obtained in this study are based on NOGAPS data. It should be noted that these thresholds are dataset dependent. This is not surprising since different biases exist in different model analysis datasets. In this study, we only focus on the WNP. But it is obvious

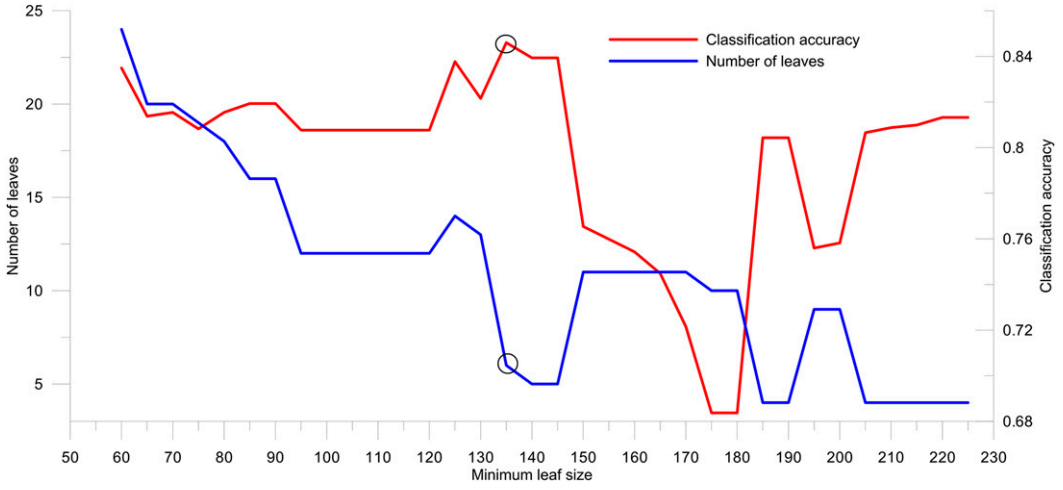

FIG. A1. The classification accuracy of decision trees when the min leaf size varies from 60 to 225. The black circles mark the highest classification accuracy and the corresponding number of leaves on the decision tree.

that this method can be used to build decision trees for all basins. According to previous studies, the characteristics of tropical disturbances, as well as the environments in which they are embedded, are different between the WNP and the North Atlantic (Fu et al. 2012; Peng et al. 2012). The decision trees for different basins are expected to be quite different. The application of this method to the other basins belongs to our future studies and will be later reported.

Acknowledgments. We appreciate the comments and suggestions from anonymous reviewers. This research was jointly supported by the National 973 Fundamental Research Program of the Ministry of Science and Technology of China (2013CB430102), National Natural Science Foundation of China (Grant No.: 41430427; 41201045), Open Fund of Key Laboratory of Geographic Information Science (MOE), East China Normal University (Grant No. KLGIS2012A04), Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140047), the Startup Foundation for Introducing Talent of NUIST, and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

APPENDIX

The C4.5 Algorithm for Building Decision Trees

A decision tree consists of a root node, a leaf node, and a branch. A leaf node is a terminal of the decision tree, indicating the value of the target attribute (class). A root node is defined as the topmost decision node in a tree that corresponds to the best predictor. Each branch of the decision tree represents a possible decision or occurrence. A decision tree traverses a tree of questions depending on the answer of each question until a leaf node is reached, at which point the leaf node states the classification of the input. In the processes of classification, a dataset is sequentially separated according to the decision framework, and a class label (e.g., 1 and −1 for binary classification) is assigned to each sample according to the label of the leaf node to which this sample belongs. The C4.5 algorithm, based on information theory, is a key member of the decision tree family (Quinlan 1987, 1993).

a. Parameter settings

The C4.5 algorithm is implemented in Weka 3.6.2, which is a collection of machine learning algorithms for data mining tasks and an open-source software (available online at http://www.cs.waikato.ac.nz/ml/weka/index.html). The parameters are set as follows: minimum leaf size, 135; confidence factor, 0.25; binary split, true; debug, false; unpruned, false; reducedErrorPruning, true; numFolds, 3; uselaplace, false; seed, 1; subtreeRaising, true; and saveInstanceData, false.

b. Setting minimum leaf size

The parameter minimum leaf size indicates the minimum number of samples a leaf node should hold in building a decision tree. The smaller the minimum leaf size is, the larger the tree size tends to be. However, a larger tree is more likely a result of overfitting of the training samples and usually results in unsatisfactory prediction accuracy for future events. On the other hand, a smaller tree cannot grab sufficient information from the training samples (Hastie et al. 2001). Therefore, it is always a challenging issue to decide the optimal minimum leaf size. A common strategy is to prune nodes that convey little information from the training samples.

Reduced error pruning aims at making a better generalization (prediction capability) of the decision tree (Quinlan 1987). Because higher accuracy represents a better generalization of new samples, this procedure prunes nodes and branches with lower prediction accuracy. Of the total sample size, 5% is usually used as the minimum leaf size (DeLisle and Dixon 2004).

In this study, the minimum leaf size varies from 60 to 225. The prediction accuracy peaks at 84.6% when the minimum leaf size is 135 (Fig. A1). Therefore, the minimum leaf size is set as 135. It is noted that there are two nodes whose numbers of samples are less than 135. Such nodes should be caused by the reduced error pruning applied to the decision tree because some branches of these nodes with low generalization have been pruned.

c. *Resampling*

A dataset is said to be imbalanced if the sample from one class is a higher number than the other (Longadge et al. 2013). Imbalanced samples tend to bias the classification results produced by the C4.5 algorithm (Chawla 2003; Estabrooks et al. 2004; Han et al. 2005). Therefore, resampling is adopted to avoid biased results. This study employs the synthetic minority oversampling technique (SMOTE; Chawla 2003) to oversample the cases in the ''developing'' group (i.e., the minority class).

REFERENCES

Briegel, L. M., and W. M. Frank, 1997: Large-scale influences on tropical cyclogenesis in the western North Pacific. *Mon. Wea. Rev.,* **125,** 1397–1413, doi:10.1175/1520-0493(1997)125<1397:LSIOTC>2.0.CO;2.

Camargo, S. J., K. A. Emanuel, and A. H. Sobel, 2007a: Use of a genesis potential index to diagnose ENSO effects on tropical cyclone genesis. *J. Climate,* **20,** 4819–4834, doi:10.1175/JCLI4282.1.

——, H. L. Li, and L. Q. Sun, 2007b: Feasibility study for downscaling seasonal tropical cyclone activity using the NCEP Regional Spectral Model. *Int. J. Climatol.,* **27,** 311–325, doi:10.1002/joc.1400.

——, A. H. Sobel, A. G. Barnston, and K. A. Emanuel, 2007c: Tropical cyclone genesis potential index in climate models. *Tellus,* **59A,** 428–443, doi:10.1111/j.1600-0870.2007.00238.x.

Charney, J. G., and A. Eliassen, 1964: On the growth of the hurricane depression. *J. Atmos. Sci.,* **21,** 68–75, doi:10.1175/1520-0469(1964)021<0068:OTGOTH>2.0.CO;2.

Chawla, N. V., 2003: C4.5 and imbalanced data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure. *Proc. Int. Conf. on Machine Learning,* Washington, DC, International Machine Learning Society. [Available online at https://www3.nd.edu/~dial/papers/ICML03.pdf.]

Chen, H., and D.-L. Zhang, 2013: On the rapid intensification of Hurricane Wilma (2005). Part II: Convective bursts and the upper-level warm core. *J. Atmos. Sci.,* **70,** 146–172, doi:10.1175/JAS-D-12-062.1.

Chia, H. H., and C. F. Ropelewski, 2002: The interannual variability in the genesis location of tropical cyclones in the northwest Pacific. *J. Climate,* **15,** 2934–2944, doi:10.1175/1520-0442(2002)015<2934:TIVITG>2.0.CO;2.

Craig, G., and S. Gray, 1996: CISK or WISHE as the mechanism for tropical cyclone intensification. *J. Atmos. Sci.,* **53,** 3528–3540, doi:10.1175/1520-0469(1996)053<3528:COWATM>2.0.CO;2.

Dare, R. A., and J. L. McBride, 2011: The threshold sea surface temperature condition for tropical cyclogenesis. *J. Climate,* **24,** 4570–4576, doi:10.1175/JCLI-D-10-05006.1.

DeLisle, R. K., and S. L. Dixon, 2004: Induction of decision trees via evolutionary programming. *J. Chem. Inf. Comput. Sci.,* **44,** 862–870, doi:10.1021/ci034188s.

DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting,* **9,** 209–220, doi:10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2.

——, J. A. Knaff, and B. H. Connell, 2001: A tropical cyclone genesis parameter for the tropical Atlantic. *Wea. Forecasting,* **16,** 219–233, doi:10.1175/1520-0434(2001)016<0219:ATCGPF>2.0.CO;2.

——, M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting,* **20,** 531–543, doi:10.1175/WAF862.1.

Elsner, J. B., and C. Schmertmann, 1993: Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Wea. Forecasting,* **8,** 345–351, doi:10.1175/1520-0434(1993)008<0345:IERSPO>2.0.CO;2.

Emanuel, K. A., 1986: An air–sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. *J. Atmos. Sci.,* **43,** 585–604, doi:10.1175/1520-0469(1986)043<0585:AASITF>2.0.CO;2.

——, 1989: The finite-amplitude nature of tropical cyclogenesis. *J. Atmos. Sci.,* **46,** 3431–3456, doi:10.1175/1520-0469(1989)046<3431:TFANOT>2.0.CO;2.

——, and D. S. Nolan, 2004: Tropical cyclone activity and the global climate system. Preprints, *26th Conf. on Hurricanes and Tropical Meteorology,* Miami, FL, Amer. Meteor. Soc., 10A.2. [Available online at https://ams.confex.com/ams/pdfpapers/75463.pdf.]

Estabrooks, A., T. Jo, and N. Japkowicz, 2004: A multiple resampling method for learning from imbalanced data sets. *Comput. Intell.,* **20,** 18–36, doi:10.1111/j.0824-7935.2004.t01-1-00228.x.

Fan, K., 2010: A prediction model for Atlantic named storm frequency using a year-by-year increment approach. *Wea. Forecasting,* **25,** 1842–1851, doi:10.1175/2010WAF2222406.1.

Fayyad, U., 1997: Knowledge discovery in databases: An overview. *Inductive Logic Programming,* N. Lavrač and S. Džeroski, Eds., Springer, 1–16.

——, and P. Stolorz, 1997: Data mining and KDD: Promise and challenges. *Future Gener. Comput. Syst.,* **13,** 99–115, doi:10.1016/S0167-739X(97)00015-0.

Frank, W. M., 1987: Tropical cyclone formation. *A Global View of Tropical Cyclones,* R. L. Elsberry, Ed., Office of Naval Research, 53–90.

Friedl, M., and C. Brodley, 1997: Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.,* **61,** 399–409, doi:10.1016/S0034-4257(97)00049-7.

Fu, B., T. Li, M. S. Peng, and F. Weng, 2007: Analysis of tropical cyclogenesis in the western North Pacific for 2000 and 2001. *Wea. Forecasting,* **22,** 763–780, doi:10.1175/WAF1013.1.

——, M. S. Peng, T. Li, and D. E. Stevens, 2012: Developing versus nondeveloping disturbances for tropical cyclone formation. Part II: Western North Pacific. *Mon. Wea. Rev.,* **140,** 1067–1080, doi:10.1175/2011MWR3618.1.

Gray, W. M., 1968: Global view of origin of tropical disturbances and storms. *Mon. Wea. Rev.,* **96,** 669–699, doi:10.1175/1520-0493(1968)096<0669:GVOTOO>2.0.CO;2.

——, 1979: Hurricanes: Their formation, structure and likely role in the tropical circulation. *Meteorology over the Tropical Oceans,* D. B. Shaw, Ed., Royal Meteorological Society, 155–218.

——, 1998: The formation of tropical cyclones. *Meteor. Atmos. Phys.,* **67,** 37–69, doi:10.1007/BF01277501.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of tropical cyclone genesis forecasts from global numerical models. *Wea. Forecasting,* **28,** 1423–1445, doi:10.1175/WAF-D-13-00008.1.

Han, H., W.-Y. Wang, and B.-H. Mao, 2005: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Adv. Intell. Comput.,* **3644,** 878–887, doi:10.1007/11538059_91.

Hastie, T., R. Tibshirani, and J. J. H. Friedman, 2001: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics, Springer, 552 pp.

Hennon, C. C., and J. S. Hobgood, 2003: Forecasting tropical cyclogenesis over the Atlantic basin using large-scale data. *Mon. Wea. Rev.,* **131,** 2927–2940, doi:10.1175/1520-0493(2003)131<2927:FTCOTA>2.0.CO;2.

——, C. Marzban, and J. S. Hobgood, 2005: Improving tropical cyclogenesis statistical model forecasts through the application of a neural network classifier. *Wea. Forecasting,* **20,** 1073–1083, doi:10.1175/WAF890.1.

——, C. N. Helms, K. R. Knapp, and A. R. Bowen, 2011: An objective algorithm for detecting and tracking tropical cloud clusters: Implications for tropical cyclogenesis prediction. *J. Atmos. Oceanic Technol.,* **28,** 1007–1018, doi:10.1175/2010JTECHA1522.1.

Kuo, H.-L., 1965: On formation and intensification of tropical cyclones through latent heat release by cumulus convection. *J. Atmos. Sci.,* **22,** 40–63, doi:10.1175/1520-0469(1965)022<0040:OFAIOT>2.0.CO;2.

Li, T., B. Fu, X. Ge, B. Wang, and M. Peng, 2003: Satellite data analysis and numerical simulation of tropical cyclone formation. *Geophys. Res. Lett.,* **30,** 2122, doi:10.1029/2003GL018556.

Longadge, R., S. Dongre, and L. Malik, 2013: Class imbalance problem in data mining. *Int. J. Comput. Sci. Network,* **2,** 83–87. [Available online at http://ijcsn.org/IJCSN-2013/2-1/IJCSN-2013-2-1-58.pdf.]

McBride, J. L., 1981: Observational analysis of tropical cyclone formation. Part I: Basic description of data sets. *J. Atmos. Sci.,* **38,** 1117–1131, doi:10.1175/1520-0469(1981)038<1117:OAOTCF>2.0.CO;2.

Murakami, H., T. Li, and M. Peng, 2013: Changes to environmental parameters that control tropical cyclone genesis under global warming. *Geophys. Res. Lett.,* **40,** 2265–2270, doi:10.1002/grl.50393.

Nicholls, N., 1979: A possible method for predicting seasonal tropical cyclone activity in the Australian region. *Mon. Wea. Rev.,* **107,** 1221–1224, doi:10.1175/1520-0493(1979)107<1221:APMFPS>2.0.CO;2.

Nolan, D. S., 2007: What is the trigger for tropical cyclogenesis? *Aust. Meteor. Mag.,* **56,** 241–266.

Ooyama, K., 1969: Numerical simulation of life cycle of tropical cyclones. *J. Atmos. Sci.,* **26,** 3–40, doi:10.1175/1520-0469(1969)026<0003:NSOTLC>2.0.CO;2.

Peng, M. S., B. Fu, T. Li, and D. E. Stevens, 2012: Developing versus nondeveloping disturbances for tropical cyclone formation. Part I: North Atlantic. *Mon. Wea. Rev.,* **140,** 1047–1066, doi:10.1175/2011MWR3617.1.

Perrone, T. J., and P. R. Lowe, 1986: A statistically derived prediction procedure for tropical storm formation. *Mon. Wea. Rev.,* **114,** 165–177, doi:10.1175/1520-0493(1986)114<0165:ASDPPF>2.0.CO;2.

Quinlan, J., 1987: Decision trees as probabilistic classifiers. *Proc. Fourth Int. Workshop on Machine Learning,* Irvine, CA, International Machine Learning Society, 31–37.

——, 1993: *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 302 pp.

Ritchie, E. A., and G. J. Holland, 1999: Large-scale patterns associated with tropical cyclogenesis in the western Pacific. *Mon. Wea. Rev.,* **127,** 2027–2043, doi:10.1175/1520-0493(1999)127<2027:LSPAWT>2.0.CO;2.

Venkatesh, T. N., and J. Mathew, 2004: Prediction of tropical cyclone genesis using a vortex merger index. *Geophys. Res. Lett.,* **31,** L04105, doi:10.1029/2003GL019005.

Wang, G. H., J. L. Su, Y. H. Ding, and D. Chen, 2007: Tropical cyclone genesis over the South China Sea. *J. Mar. Syst.,* **68,** 318–326, doi:10.1016/j.jmarsys.2006.12.002.

Ward, G. F. A., 1995: Prediction of tropical cyclone formation in terms of sea-surface temperature, vorticity and vertical wind shear. *Aust. Meteor. Mag.,* **44,** 61–70.

Xiao, F., and Z. Xiao, 2010: Characteristics of tropical cyclones in China and their impacts analysis. *Nat. Hazards,* **54,** 827–837, doi:10.1007/s11069-010-9508-7.

Zhang, D.-L., and H. Chen, 2012: Importance of the upper-level warm core in the rapid intensification of a tropical cyclone. *Geophys. Res. Lett.,* **39,** L02806, doi:10.1029/2011GL050578.

Zhang, Q., Q. Liu, and L. Wu, 2009: Tropical cyclone damages in China: 1983–2006. *Bull. Amer. Meteor. Soc.,* **90,** 489–495, doi:10.1175/2008BAMS2631.1.

Zhang, W., S. Gao, B. Chen, and K. Cao, 2013a: The application of decision tree to intensity change classification of tropical cyclones in western North Pacific. *Geophys. Res. Lett.,* **40,** 1883–1887, doi:10.1002/grl.50280.

——, Y. Leung, and J. C. L. Chan, 2013b: The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part I: Tropical cyclone recurvature. *J. Appl. Meteor. Climatol.,* **52,** 1394–1416, doi:10.1175/JAMC-D-12-045.1.

——, ——, and ——, 2013c: The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part II: Tropical cyclone landfall. *J. Appl. Meteor. Climatol.,* **52,** 1417–1432, doi:10.1175/JAMC-D-12-046.1.