Prediction of Tropical Cyclone Genesis from Mesoscale Convective Systems Using Machine Learning

TAO ZHANG AND WUYIN LIN

Brookhaven National Laboratory, Brookhaven, New York

YANLUAN LIN

Ministry of Education Key Laboratory for Earth System Modeling, and Department for Earth System Science, Tsinghua University, Beijing, China

MINGHUA ZHANG AND HAIYANG YU

School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York

KATHY CAO

Commack High School, Commack, New York

WEI XUE

Ministry of Education Key Laboratory for Earth System Modeling, and Department for Earth System Science, Tsinghua University, and Department of Computer Science and Technology, Tsinghua University, Beijing, China

(Manuscript received 6 December 2018, in final form 4 June 2019)

ABSTRACT

Tropical cyclone (TC) genesis is a problem of great significance in climate and weather research. Although various environmental conditions necessary for TC genesis have been recognized for a long time, prediction of TC genesis remains a challenge due to complex and stochastic processes involved during TC genesis. Different from traditional statistical and dynamical modeling of TC genesis, in this study, a machine learning framework is developed to determine whether a mesoscale convective system (MCS) would evolve into a tropical cyclone. The machine learning models 1) are built upon a number of essential environmental predictors associated with MCSs/TCs, 2) predict whether MCSs can become TCs at different lead times, and 3) provide information about the relative importance of each predictor, which can be conducive to discovering new aspects of TC genesis. The results indicate that the machine learning classifier, AdaBoost, is able to achieve a 97.2% F1-score accuracy in predicting TC genesis over the entire tropics at a 6-h lead time using a comprehensive set of environmental predictors. A robust performance can still be attained when the lead time is extended to 12, 24, and 48 h, and when this machine learning classifier is separately applied to the North Atlantic Ocean and the western North Pacific Ocean. In contrast, the conventional approach based on the genesis potential index can have no more than an 80% F1-score accuracy. Furthermore, the machine learning classifier suggests that the low-level vorticity and genesis potential index are the most important predictors to TC genesis, which is consistent with previous discoveries.

1. Introduction

Tropical cyclones (TCs) are one of the most disastrous extreme weather events, which can induce huge damages with heavy precipitation and strong winds (Chan 2005). Understanding TC genesis has long been a research interest of the broader scientific community (Gray 1968; Emanuel 1989; Gray 1998; Peng et al. 2012). In the absence of consistency among theories, predicting TC genesis remains a challenging subject.

TC genesis generally has two stages in nature (Holland 1995; Briegel and Frank 1997; Ritchie and Holland 1999). The first stage is the transition from a tropical disturbance to a tropical depression. The second

Corresponding author: Wuyin Lin, wlin@bnl.gov

DOI: 10.1175/WAF-D-18-0201.1

stage is the development from a tropical depression to a tropical storm. Most previous studies focused on the second stage, during which the warm-core structure of TCs has been established. The mechanisms governing the second-stage transition include the convective instability of the second kind (CISK; Charney and Eliassen 1964; Emanuel 1991) and wind-induced surface heat exchange (WISHE; Emanuel 1986). The first stage, where the genesis processes begin with weak or unorganized disturbances, is more stochastic and not well understood. In fact, even the CISK and WISHE mechanisms for the second stage are questionable (Craig and Gray 1996). A complete understanding of the processes that cause TC genesis remains elusive despite decades of research (Halperin et al. 2013).

Nevertheless, over the decades, many environmental factors that influence TC genesis have been recognized. Palmen (1948) finds that the SST should be greater than 26°C for TC genesis. McBride and Zehr (1981) indicate that low vertical wind shear benefits tropical cyclone genesis. Gray (1998) summarizes several environmental factors crucial for TC genesis: an adequately deep layer of warm ocean; a conditionally unstable atmosphere, which is almost always presented; higher midtroposphere moisture; and stronger low-level vorticity. Motivated by the work of Gray (1979), Emanuel and Nolan (2004) developed the genesis potential index (GPI) for quantitative assessment of TC genesis potential given environmental conditions. The GPI takes most of the commonly recognized environmental factors into account, including low-level vorticity, vertical wind shear, midtroposphere humidity, and conditional instability of the atmosphere. Following Emanuel (1986, 1989) and Emanuel and Nolan (2004), it is calculated as

$$\mathrm{GPI} = \|10^5\eta\|^{3/2}(1+0.1V_{\mathrm{shear}})^{-2}\left(\frac{H}{50}\right)^3\left(\frac{V_{\mathrm{pot}}}{70}\right)^3, \qquad (1)$$

where η is the absolute vorticity (s^{-1}) at 850 hPa, V_{shear} ($m\,s^{-1}$) is the magnitude of the vertical wind shear between 850 and 200 hPa, H is the relative humidity at 600 hPa, and V_{pot} ($m\,s^{-1}$) represents the TC maximum potential intensity. The GPI has been widely used to investigate the spatial distribution and annual cycle of TC genesis at the hemisphere-scale (Camargo et al. 2007) and as a proxy to evaluate the performance of climate models in simulating TCs and their temporal variabilities (Yokoi et al. 2009). However, it is rarely used to investigate individual TC genesis. Although TC genesis events usually occur over regions having a high GPI value, considering all high GPI value as TC genesis would lead to a large number of overpredictions (Camargo et al. 2007).

In recent years, an increasing body of work has focused on predicting developing or nondeveloping TCs using statistical and machine learning models. Hennon and Hobgood (2003) use a linear discriminant analysis to predict whether the cloud clusters will evolve into tropical cyclone over the Atlantic basin. They find that the daily genesis potential and latitude are the most significant predictors. However, the linear method only gets a relatively low classification performance when the lead time is 6 h. Hennon et al. (2005) improve the classification performance using the neural network method, which achieves higher reliability and robustness than the linear discriminant analysis does. But the best performance remains similar to the linear method. Zhang et al. (2015) investigate the tropical cyclone genesis from tropical disturbances in the western North Pacific using the decision tree model. The optimal classification accuracy achieves 84.6% with 24-h lead time. Through the classification method, they find that the maximum 800-hPa relative vorticity, SST, precipitation rate, divergence averaged between 1000- and 500-hPa levels, and 300-hPa air temperature anomaly have the largest influences on the TC genesis.

In addition to the environmental conditions critically influencing the TC genesis process, Gray (1998) emphasizes the important roles of mesoscale convective systems (MCSs) as contributors to TC genesis. MCSs are organized thunderstorms with contiguous rain regions. Their scales are larger than the individual thunderstorms but smaller than tropical or extratropical cyclones (Rickenbach and Rutledge 1998). MCSs have been found to have a strong connection with TC genesis in the monsoon trough and tropical waves (Lander 1994; McBride 1995; Chen et al. 2004; Lu et al. 2012; Lee et al. 2008). Two major TC genesis mechanisms, the top-down (Chen and Frank 1993) and the bottom-up pathways (Zehr 1992), both suggest that MCSs are of great importance in the TC genesis processes.

Modern weather predictions strongly depend on and enormously benefit from near-full-time satellite surveillance of Earth's atmosphere and weather systems. MCSs can be identified and tracked based on satellite measurements of cloud properties (e.g., Huang et al. 2018). A TC genesis prediction algorithm relating MCSs to TC genesis has a strong potential to provide another means to enhance TC prediction skills around the globe. The purpose of this study is to determine whether an MCS will evolve into a TC using machine learning classifiers. The machine learning algorithms extract essential properties associated with MCSs/TCs, such as spatial and temporal information, low-level vorticity, relative humidity, wind shear, TC potential intensity, MCS brightness temperature, propagation

speed and direction of MCS, and the GPI of the environment. The possibility of an MCS becoming a TC is predicted using the machine learning methods at different lead times up to 48 h prior to the TC genesis. Several classifiers are evaluated, including the linear methods: linear logistics (Kutner et al. 2004) and naive Bayes (Friedman et al. 1997); the nonlinear methods: *k*-nearest neighbors (Keller et al. 1985), support vector machines (Cortes and Vapnik 1995), decision trees (Quinlan 1987), multilayer perceptron (Rumelhart et al. 1985), and quadratic discriminant analysis (McLachlan 2004); and nonlinear ensemble methods: random forest (Liaw and Wiener 2002) and AdaBoost (Freund and Schapire 1997).

Our purpose is to use machine learning classifiers to predict the TC genesis from MCS anywhere in the tropics. In this work, in addition to predicting TC genesis in the tropics as a whole, we separately evaluate the performance over the two most active basins: the western North Pacific (WNP) and the North Atlantic (NA). This study also quantifies the relative importance of those environmental predictors to TC genesis to gain a deeper understanding of the TC genesis mechanism.

The remainder of this paper is organized as follows. Section 2 describes the training data and predictors. Section 3 is used to present the machine learning framework and the performance metrics. Section 4 presents the performance of machine learning methods and comparison with previous studies. A summary is given in section 5.

2. Data and predictors

The data used for the machine learning models include a tropics-wide MCS database, which contains the track, intensity, size (area covered), eccentricity, and lifetime for each identified MCS. The MCS dataset was obtained based on long-term (1985–2008) tropical 3-hourly brightness temperature from the Cloud Archive User Service (CLAUS) project to identify and track MCS by a novel method, which can capture small and fast-moving MCSs. As a consequence, this MCS dataset contains more MCS samples than previous methods (Huang et al. 2018). The brightness temperature and area-coverage threshold used for this dataset are 233 K and 5000 km^2, respectively. We use the 6-hourly European Centre for Medium-Range Weather Forecasts (ECMWF) interim reanalysis (ERA-Interim; Dee et al. 2011) at 0.25° × 0.25° resolution for various environmental predictors. Besides global tropical region, two basins during their TC seasons (from May to November) are also investigated. They are the North Atlantic Ocean (NA; 0°–30°N, 100°–20°W) and the western North Pacific Ocean (WNP; 0°–30°N, 100°E–180°), respectively.

TABLE 1. List of environmental predictors for TC genesis.

Predictors	Abbreviation	Dataset source
Longitude	lon	MCS dataset
Latitude	lat	MCS dataset
Month	mon	MCS dataset
Genesis potential index	GPI	ERA-Interim
850-hPa vorticity	Vort850	ERA-Interim
600-hPa relative humidity	Q600	ERA-Interim
Vertical wind shear from 850 to 200 hPa	wind_shear	ERA-Interim
TC potential intensity	TC_PI	ERA-Interim
MCS area coverage	MCS_size	MCS dataset
Average BT of all pixels with an MCS	MCS_avgBT	MCS dataset
Lowest BT of all pixels with an MCS	MCS_minBT	MCS dataset
Propagation speed of an MCS	MCS_speed	MCS dataset
Movement direction of an MCS	MCS_direct	MCS dataset

The machine learning classification models are trained using these datasets. The training data contains a number of MCS/TC environmental predictors summarized in Table 1. Spatial and temporal information are included to reflect geophysical and seasonal relationships of MCS and TCs to the environment, as well as the Coriolis parameter, which is also a necessary factor for TC genesis. Large-scale thermodynamic and dynamical predictors consist of relative humidity in midtroposphere, low-level vorticity, wind shear, and the TC potential intensity, as well as GPI (Gray 1998; McBride and Zehr 1981; Holland 1995; DeMaria et al. 2001; Emanuel and Nolan 2004). The proprieties of MCS are also considered and used for the machine learning models. They are selected based on Hennon and Hobgood (2003), Hennon et al. (2005), Zhang et al. (2015), and Huang et al. (2018).

Because the definition of TC genesis is ambiguous (Horn et al. 2014), it is hard to exactly determine the time of TC genesis. In this study, we assume that TC genesis occurs at the initial appearance of a TC record in the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2010), and we define it as a TC genesis point. If a genesis point is near the trajectory of an MCS within a 2° distance in both longitude and latitude, the MCS is considered to have evolved into a TC and labeled as a positive case. Otherwise, the MCS is considered to fail to become a TC and labeled as a negative case. These two class of cases are used to train and test the machine learning classifiers.

For the positive cases, all environmental predictors in Table 1 are obtained at various lead times before the TC

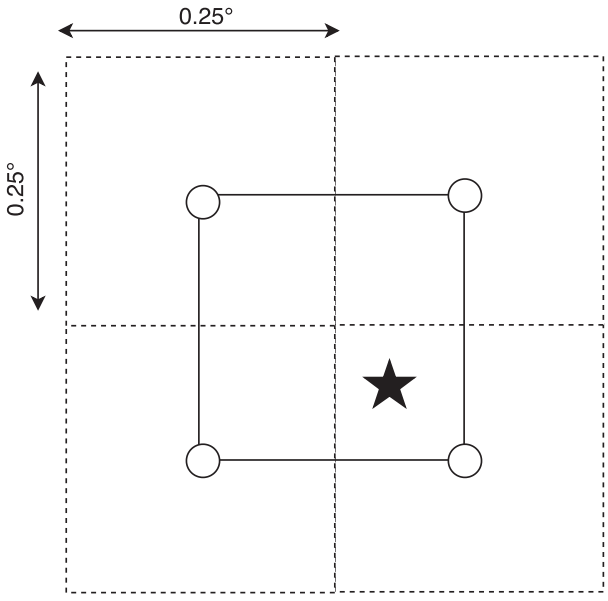

FIG. 1. Acquiring the large-scale predictors by averaging the ERA-Interim grid points. The star is the location of a point in an MCS trajectory. The four open circles are the ERA-Interim grid points at 0.25° resolution. The large-scale predictors are computed by averaging these four points.

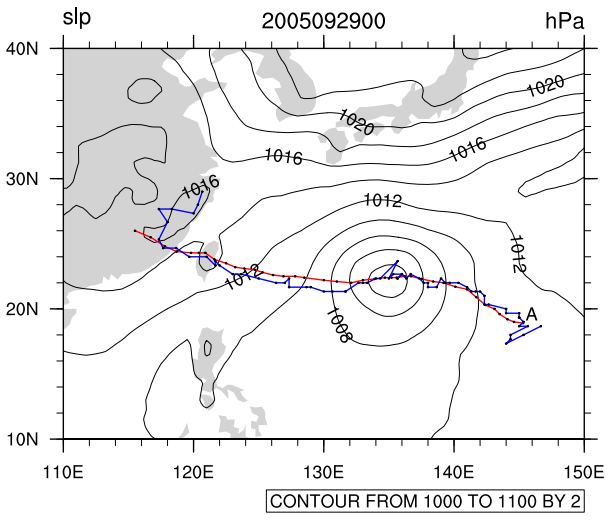

FIG. 2. The low of a snapshot sea level pressure from ERA-Interim. Point A is the TC genesis point. Red line is the trajectory of TC from IBTrACS. Blue line is the trajectory of MCS from MCS dataset.

genesis (6, 12, 24, and 48 h). The negative cases represent those MCSs that cannot evolve into TCs at any lead time. They are randomly selected from the MCS datasets, excluding those that have been identified as positive cases. In this study, the machine learning models at different lead times are trained and tested separately. For predicting TC genesis at a specific lead time, it only requires environmental predictors at a single time, correspondingly. The spatial–temporal predictors and the properties of MCS at specified lead times are obtained from the reanalysis and the MCS dataset by backtracking the corresponding lead times. The large-scale environment predictors are computed by averaging the neighboring area of the TC genesis point. The area is comprised of four ERA-Interim grid points surrounding the TC location, as illustrated in Fig. 1.

The data required by this study are from the combination of three sources (i.e., MCS dataset, IBTrACS, and ERA-Interim). Figure 2 is used to illustrate the consistency among them. Point A, the start point of TC track data, is the TC genesis point. After that, the trajectories of MCS and TC are essentially overlapped with each other. As expected, a snapshot of the low sea level pressure centers at the trajectories and moves along them. This indicates that the three observation datasets capture the same TC and the data sampled from different sources are consistent.

3. Methodology

a. *Machine learning framework*

This study aims to classify developing or nondeveloping TCs from MCSs. The machine learning models (i.e., classifiers) map the TC/MCS environmental predictors into a two-category discrimination. The classifiers are trained from the labeled historical observational data. Well-trained machine learning models can be used to predict future TC genesis.

Figure 3 illustrates the workflow of the machine learning classification in terms of identifying TC genesis. The historical data comprises of environment predictors and predictands (i.e., positive and negative labels) from February 1985 to November 2008 described in section 2. It is divided into two parts: the training set and the testing set. The training dataset is used to build the machine learning classifiers and the testing dataset is used to evaluate the performance. In this study, various linear, nonlinear, and nonlinear ensemble classifiers (Table 2) are tested and evaluated.

A classifier is linear if its decision boundary is a line, a plane, or a hyperplane. Otherwise, it belongs to the nonlinear category. Logistic regression and naive Bayes algorithms are the linear classifiers (Kutner et al. 2004; Friedman et al. 1997). They map the predictors into a probability value and make the prediction by a threshold. The probability information is able to help interpret and understand the physics behind the machine learning models. But, due to the linearity, they cannot work well for nonlinear classification problems. The *k*-nearest neighbor (KNN) is one of the simplest machine learning

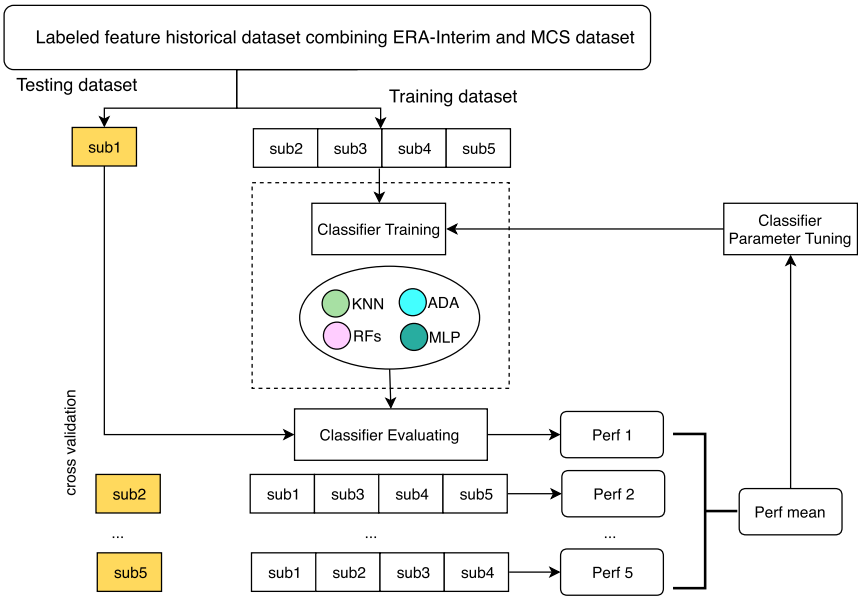

FIG. 3. Flow diagram of TC genesis prediction using machine learning. The predictors dataset is divided into two parts: the training set and the testing set. The training dataset is used to build the machine learning classifiers and the testing dataset is used to evaluate the performance. The fivefold cross-validation method is used to test the capability of each classifier. The parameters of all classifiers are well tuned based on the cross-validation performance.

classifiers (Keller et al. 1985). It stores all training data samples and makes predictions for the testing data by counting the classes among the k closest neighbors. KNN works well with a small number of training data but would be slow when they become large. It cannot learn the generalized knowledge from the training data so it is not expected to be robust for noisy data. Decision tree mimics how humans think and make decisions based on a series of rules that are organized in a tree shape (Quinlan 1987). These rules make it possible to see the logical relationship of input predictors and acquire knowledge due to the good interpretability. But it is prone to overfitting because it would generate too many branches for noisy data. Support vector machine (SVM) works by finding the maximum-margin hyperplane that separates the predictors into various classes (Cortes and Vapnik 1995). Kernel SVM uses a kernel function to transform the original input into a higher-dimensional space so as to gain nonlinear classification capability. But the performance of SVM is heavily sensitive to the choice of kernel functions and the kernel parameters. Overtuning the parameters can lead to overfitting. Multilayer perceptron (MLP) is a kind of artificial neural networks, which is consist of an input layer, some hidden layers, and an output layer (Rumelhart et al. 1985). Each layer is made up of multiple neural nodes. They are independent in the same layer and connect by links in different layers. The MLP, like other neural network methods, such as the convolutional neural network and the recursive neural network, provides the best solution for natural language processing, speech recognition, and image recognition. The disadvantage is its weak interpretation that makes it hard to understand why neural networks achieve the output. This algorithm usually requires much more training data than other machine learning algorithms. Quadratic discriminant analysis (QDA) are also based on the Bayes's theorem and provide probability information (McLachlan 2004). It makes classification through a quadratic decision boundary so that it has the nonlinear classification capability. But QDA requires

TABLE 2. Classifiers used in this study.

Category	Classifier	Abbreviation
Linear	Logistic regression	Logist
	Naïve Bayes	NB
Nonlinear	Decision tree	DTree
	k-nearest neighbors	KNN
	Multilayer perceptron	MLP
	Quadratic discriminant analysis	QDA
	Support vector machine	SVM
Nonlinear ensemble	AdaBoost	ADA
	Random forests	RFs

the dataset to follow a multivariate normal distribution, which is hard for most applications.

The nonlinear ensemble classifiers construct a set of nonlinear classifiers and then make a decision by calculating the weighted average of individual predictions. The generalization capability of an ensemble classifier is usually stronger than that of a nonensemble model. Random forest and AdaBoost are the two most common ensemble methods. Random forest is the bagging model, which trains the base models separately (Liaw and Wiener 2002). Each base model is trained by a random subset of the data. AdaBoost is the boosting model, which trains the base models iteratively (Freund and Schapire 1997). Each individual model learns from mistakes made by the previous step. However, the base models in AdaBoost are trained sequentially so that it is difficult to parallelize this algorithm.

As a comprehensive assessment of the predictive skill of machine learning models for TC genesis, we intend to apply all the abovementioned classifiers in this study.

The cross-validation technique (Kohavi 1995) is used to objectively evaluate the performance of a machine learning model. It ensures that each sample of the dataset can be used for training and testing, leading to high confidence in the general capability of the machine learning classification algorithms. The *k*-fold and leave-one-out (LOO) are the two common cross-validation methods. The *k*-fold cross validation divides the total dataset into k equal subsets and then the training and testing are performed for k iterations. In each iteration, one subset is utilized for validation while the remaining subsets (i.e., $k - 1$ subsets) are used for the model training. The prediction metric is calculated by the mean of performance criteria of each iteration. LOO conducts validation by using just one point as the testing data and the rest as the training dataset. Thus, for the dataset with n samples, the classifiers are trained and tested n times. If n is very large, LOO is much more computational expensive than *k*-fold. In LOO, the training dataset resemble each other, which increases variance for testing process. Therefore, we use a fivefold cross-validation method to test the capability of each classifier in this study.

The free parameters involved in the above algorithms are set as follows. In logistic regression, the parameter of inverse of regularization strength C is set to 1.0×10^5. In SVM, the kernel used is radial basis function (RBF). The penalty value C is set to be 1800. The gamma parameter is set to be 1/13. In KNN, the parameter K is set to be 5. In MLP, the hidden layer size is 50. The L2 penalty parameter alpha is 1.0×10^{-4}. The batch size is 200. The learning rate is 0.1. In AdaBoost and random forest, the number of trees is 600. These parameter values are tuned and determined using grid search with cross validation to find the optimization parameters. Grid search is a method of performing parameter optimization to determine the optimal values for a given machine learning model. It exhaustively searches the parameter values through a manually specified subset of the parameter space of a learning method.

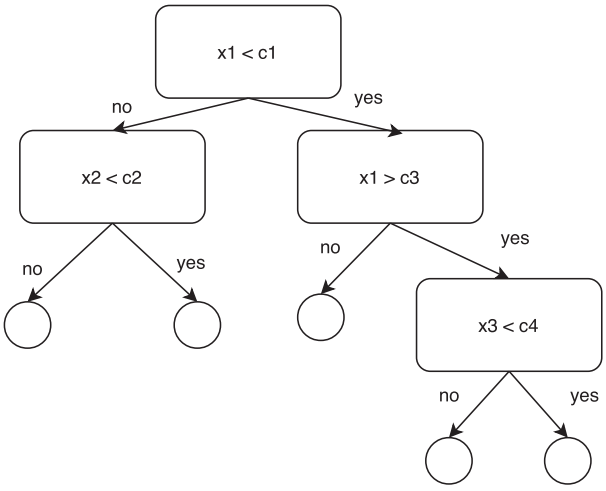

FIG. 4. A decision tree example to illustrate the structure; x1, . . . , n is the predictor and c1, . . . , n is the classification criterion.

b. *Importance of predictors*

Tree-based classifiers, such as decision tree, random forests, and AdaBoost used in this study, not only have good performance but also provide the capability to quantify the contribution of each critical predictor to the performance of these classifiers, also called the importance of predictors (Breiman 2001; Breiman et al. 2017). The decision tree method is the foundation of the random forests and AdaBoost methods. An example of decision tree is presented in Fig. 4. It is built through a recursive procedure that partition the training dataset into smaller subsets. In the tree structure, the leaf nodes represent the final classification results. The nonleaf (branch) nodes break down the dataset into two subtrees based on a threshold value of one of the predictors.

Choosing an appropriate variable predictor at each branch node is crucial for constructing the tree. The selection principle is to make the subset belong to the same category (i.e., the chosen predictor makes the subset have a high purity). In the decision tree method, the best splitting predictor s in the current branch node t maximizes the decrease of impurity i, as measured by $\Delta i(s, t)$

$$s^* = \text{argmax}_s \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R), \qquad (2)$$

where t_L and t_R are the divided subsets; p_L and P_R are the subset proportions of the total dataset; and the

impurity measures i can be the variance, the Shannon entropy, or the Gini index (Breiman 2001; Breiman et al. 2017). Breiman (2001) proposes the mean decrease impurity importance (MDI) to measure the importance of predictor x_m by averaging the weighted impurity decrease with regard to this predictor over all nodes in the tree and then averaged over all trees in the forest by using an ensemble-based algorithm:

$$\mathrm{MDI}(x_m) = \frac{1}{N_T}\sum_{T}\sum_{t\in T:v(s_t)=x_m} p(t)\Delta i(s,t), \tag{3}$$

where N_T is the number of trees in the forest T, with $N_T = 1$ for a decision tree; $p(t)$ is the fraction of subset at node t in a tree, and $v(s_t)$ is the predictor value used to determine the partition.

c. GPI classifier

GPI has been widely accepted to have a close association with TC genesis. As a contrast to machine learning methods, we also evaluate the classifier based on GPI threshold. GPI takes into account various environmental factors influencing TC genesis.

Figure 5 presents the kernel density estimation (KDE) of GPI of negative cases and positive cases described in section 2 under different lead times. KDE is a nonparameter widespread method to estimate probability density functions (PDFs) based on the sampled dataset (Rosenblatt 1956; Parzen 1962). The PDF is approximated through the superposition of kernel functions. Here Gaussian kernel is selected. Most of the negative cases lie on the lowest interval, from 0 to 0.025 in the global tropical region and the other two basins. In contrast, the density of positive cases is consistently greater than that of negative cases at higher values of GPI. In other word, the GPI values of positive cases are highly likely to be greater than that of negative cases. Therefore, while the index is often used to statistically count the number of TCs in space and time, it can also be used to distinguish the developing and nondeveloping TC from an MCS.

In Fig. 5, the point of intersection between negative and positive cases is about 0.025. Consequently, a GPI threshold in the neighborhood of this value can be set as a classification boundary.

d. Performance metrics

In the field of machine learning classification, a confusion matrix (or contingency table as used in verification of weather forecast) is a 2×2 table that summarizes the four possible outcomes of a two-category classification and is used to describe the performance of a classifier. Table 3 presents the correspondence between the

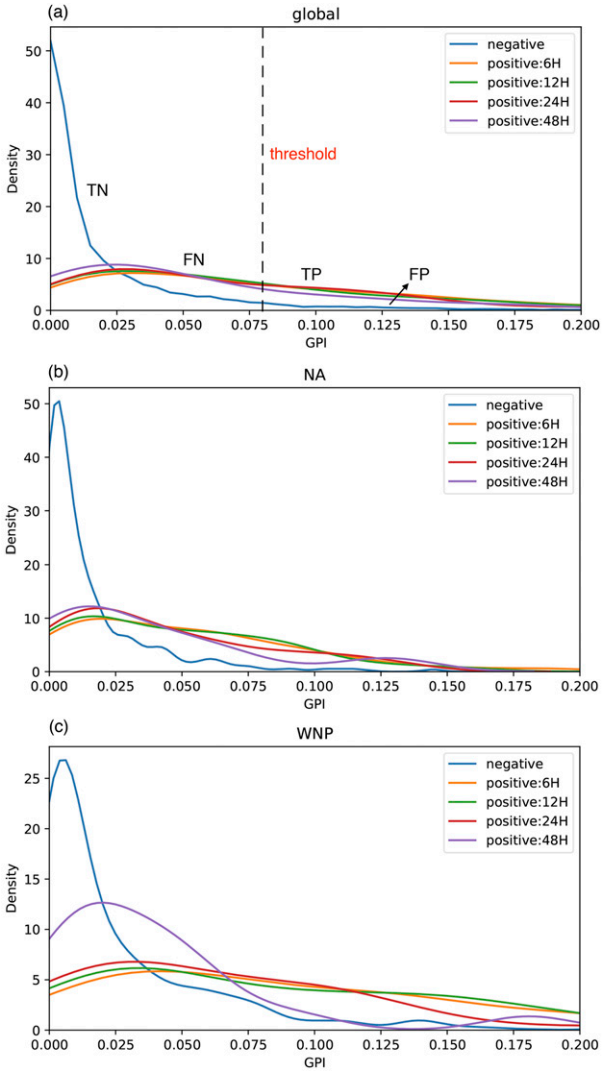

FIG. 5. The GPI KDE of negative cases and positive cases under various lead times at (a) the global tropics, (b) NA, and (c) WNP. The four possible predictions in the contingency table are marked in (a). TN and FN are labeled when the GPI values of the negative cases and the positive cases are lower than the threshold, respectively. Otherwise, FP and TP are labeled.

confusion matrix and the contingency table. Predictions from the testing set are labeled as true positive (TP)/hit or true negative (TN)/correct negative if the classifiers correctly predict an MCS will develop or not develop into a TC at a lead time. Those that incorrectly predict developing or nondeveloping are treated as false positive (FP)/false alarm or false negative (FN)/miss.

A number of performance criteria can be calculated based on the confusion matrix. Precision P and recall R are the two of them and defined as:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{4}$$

TABLE 3. The confusion matrix/contingency table showing the four possible outcomes for a two-class developing and non-developing problem.

	Predicted	
Actual	Yes	No
Yes	TP (Hit)	FN (Miss)
No	FP (False alarm)	TN (Correct negative)

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

Precision is the ratio between the number of true positives and the total number of predicted positive cases. It measures the quality of the classifiers in predicting true positive cases. A low value of precision represents a large number of false alarm cases. On the other hand, recall is the ratio between the true positives and the total number of actual positive classes. It measures the completeness of classifier prediction. A low value of recall represents large number of missing cases. The ideal situation is when precision and recall are 1.0. But, in practice, this is impossible to achieve. It is possible to have a perfect recall (simply predict that all are positive), but a horrible precision value. Similarly, it is easy to increase precision alone (just predicting to be positive only for those samples that are probably actual positive cases.), but this will most likely come with a poor recall. A more robust performance criterion is to take both P and R into account, as in the F1-score defined as the harmonic mean of precision and recall [Eq. (6)]. Only when both precision and recall are high, will the F1-score achieve a high value:

$$\text{F1} = \frac{2PR}{P + R}. \tag{6}$$

4. Results

a. *Performance of GPI threshold classifier*

We first show the performance of GPI threshold classifier over the global tropical region and the basins of WNP and NA at different lead times (Figs. 6–8). A number of threshold values are tested based on the KDE distribution in Fig. 5. All those with GPI values larger than this threshold would be regarded as positive cases, and those with GPI values lower than the threshold are negative cases. The four possible predictions in the contingency table (shown in Table 3) are marked in Fig. 5a. The F1 score of the 6-h lead time achieves about 70%, 80%, and 75% for global, NA, and WNP regions, respectively. When the lead time is extended to 12, 24, and 48 h, the performance deteriorates significantly.

The best performance of the GPI-based classifier is obtained using the threshold of 0.05, 0.02, and 0.075 for the global tropics, NA and WNP, respectively. They are near or to the right of the intersection point between the distributions for the positive and negative cases shown in Fig. 5. With increased threshold values, the number of TP and FP decrease, while the number of FN and TN increase. According to Eq. (5), the recall performance becomes worse because of increased FN and decreased

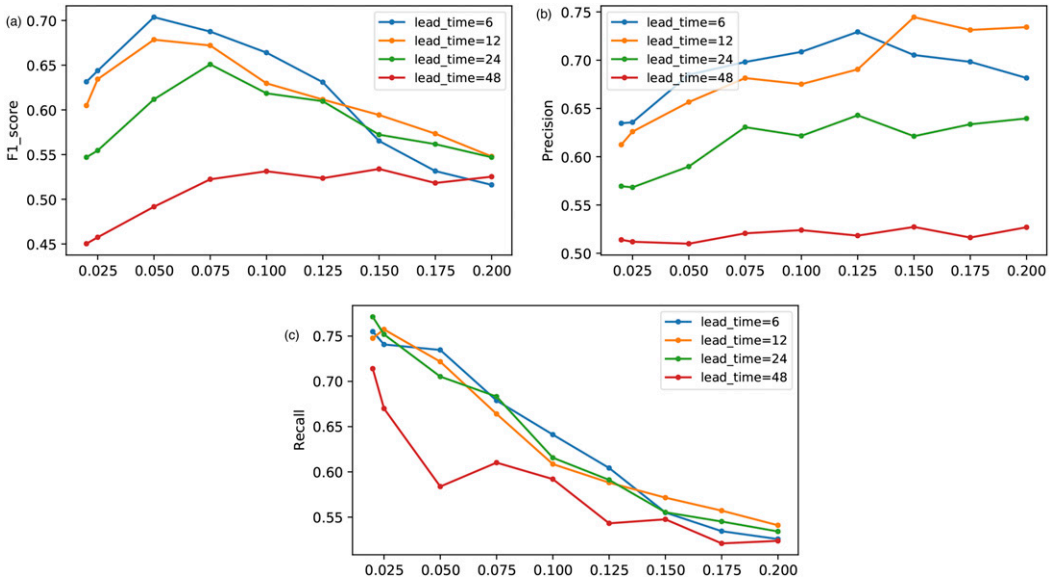

FIG. 6. Performance of GPI threshold classifier in the global tropics at various lead times, including (a) F1 score, (b) precision, and (c) recall.

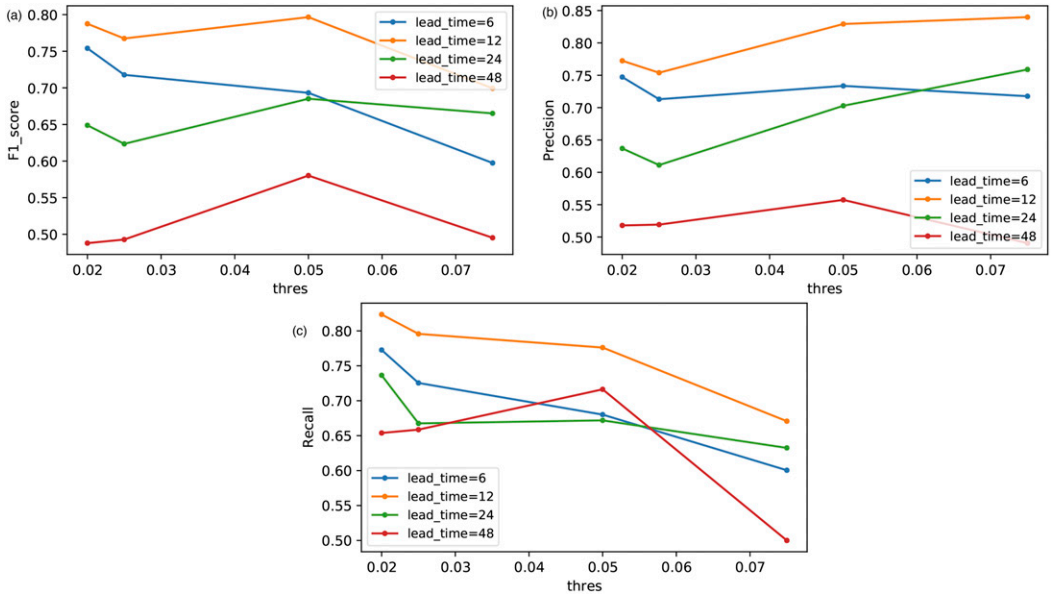

FIG. 7. Performance of GPI threshold classifier in the North Atlantic Ocean at various lead times, including (a) F1 score, (b) precision, and (c) recall.

TP. Although both TP and FP decrease, the decrease in the number of FP is more substantial, resulting in a better precision.

b. Machine learning classifiers

The results of machine learning classifiers (Table 2) are derived from the historical observation dataset described in section 2 using the predictor variables in Table 1. These machine learning classifiers are independently trained and tested in the global tropical region, NA and WNP areas at different lead times. Fivefold cross-validation technology described in section 3 is used to evaluate the performance of each classifiers. Namely, 80% of historical data are used to train the classifiers, leaving 20% to validate the performance. Table 4 lists the number of negative and positive samples. The numbers in the ''negative'' row represent the number of negative samples in global tropical, NA and WNP areas.

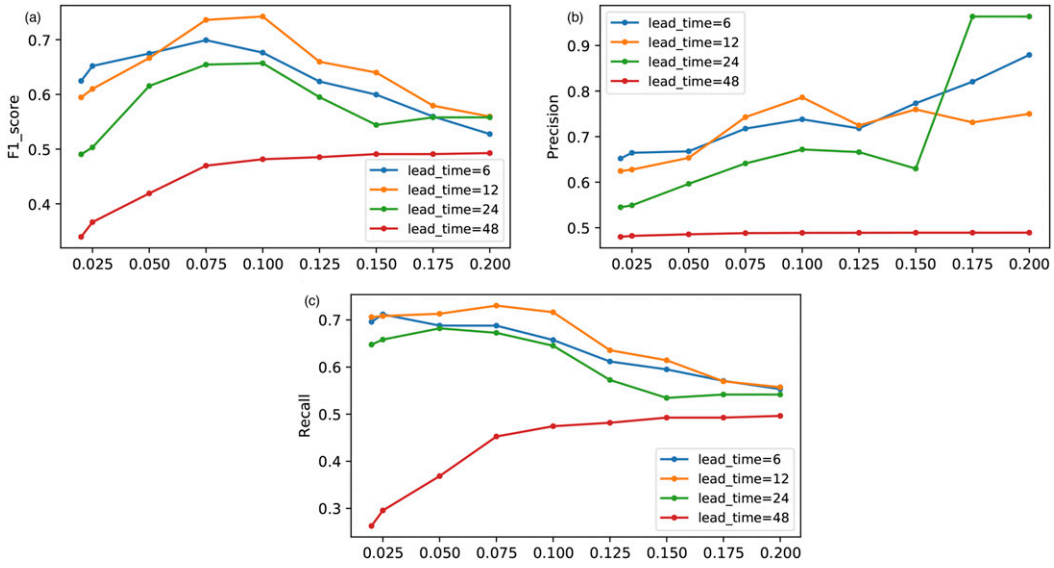

FIG. 8. Performance of GPI threshold classifier in the western North Pacific Ocean at various lead times, including (a) F1 score, (b) precision, and (c) recall.

TABLE 4. Number of negative and positive samples at 6-, 12-, 24-, and 48-h lead times in the global tropical, NA, and WNP areas.

	Global	NA	WNP
Negative	6238	510	683
Positive, 6 h	1084	240	209
Positive, 12 h	806	178	157
Positive, 24 h	421	90	73
Positive, 48 h	89	17	17

The second to fifth rows list the number of positive samples at different lead times.

Figure 9 presents the F1 score accuracy of the machine learning classifiers at 6-, 12-, 24-, and 48-h lead times in the three regions. Among the selected machine learning classifiers, ADA consistently achieves the best performance at different lead times and in different regions. Note that the uncertain parameters in all machine learning classifiers used here are well tuned. The ADA algorithm stands out likely because it is a boosting technique that combines a number of weak classifiers (in this study, we use 600 decision trees for ensemble) into a single strong classifier (Freund and Schapire 1997). The results in this study confirm the conclusion of Breiman (1998) that ADA is better than decision tree, neural networks, and logistic regression in various applications. When the lead times are closer to the TC genesis point (i.e., from 48 to 6 h), most of classifiers have increasing capability to discriminate whether or not an MCS develops into a TC. For the global tropical area, ADA achieves the peak performance of 97.2% F1-score accuracy, at 6-h lead time. In comparison, the best performance of the GPI threshold method is only 75.5%. The results of other lead times confirm this conclusion that ADA outperforms the GPI threshold method, which is also true for NA and WNP regions. For all the classifiers, precision is greater than recall, shown in Fig. 10–11. According to Eqs. (4) and (5), this means the number of false alarm cases (FP) is lower than the number of misses (FN). In other words, the machine learning classifiers have a greater tendency to misclassify the developing ones than the nondeveloping ones, resulting in fewer classified TCs. Consistent performance is found in the WNP and NP areas at all lead times.

c. *Importance of predictors*

Using ensemble tree-based machine learning classifiers, we calculate the relative importance index of individual predictors, which adds up to 1.0 for all the predictors. The index is able to quantify the contribution of each predictor to the prediction of TC genesis. This is done for the ADA and RFs algorithms and the results

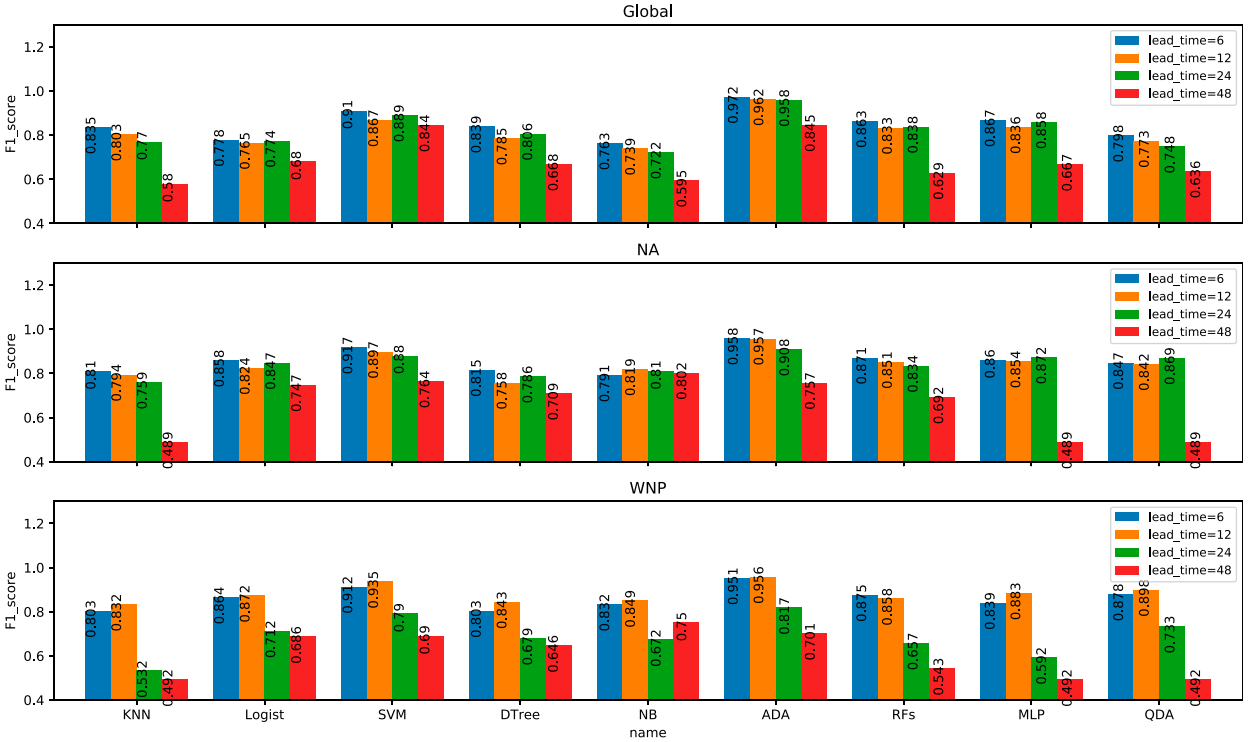

FIG. 9. The F1 score of machine learning classifiers shown in Table 2 in the global tropical, North Atlantic Ocean, and western North Pacific Ocean at various lead times.

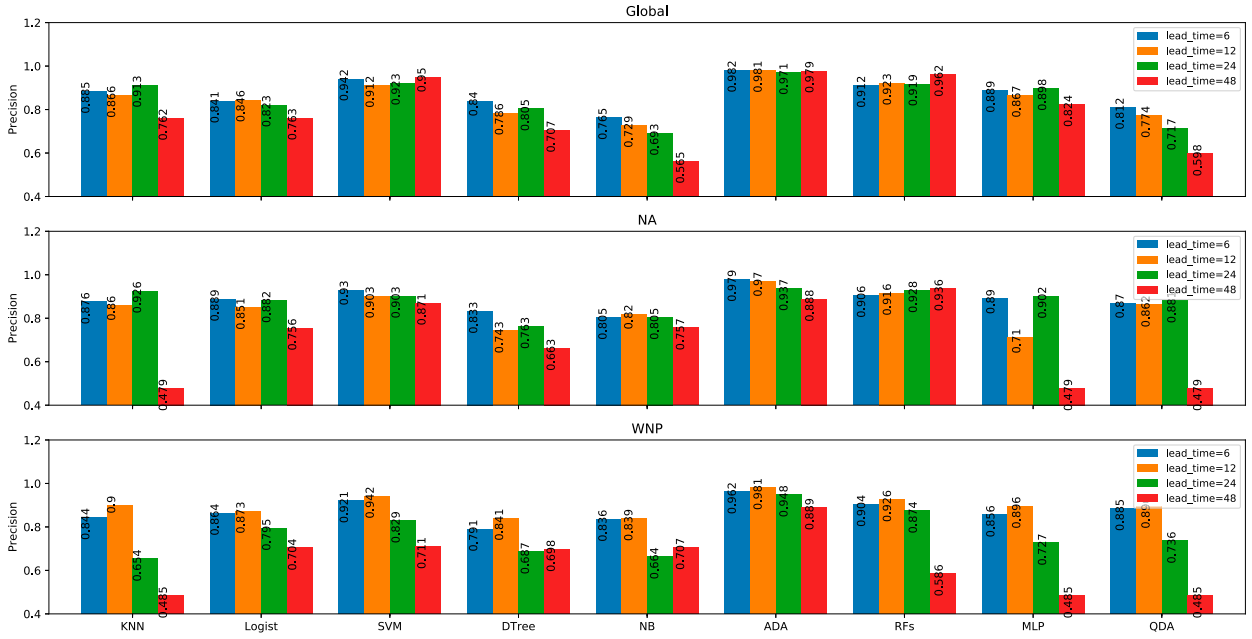

FIG. 10. The precision of machine learning classifiers shown in Table 2 in the global tropical, North Atlantic Ocean, and western North Pacific Ocean at various lead times.

are shown in Figs. 12–14. Overall, it is found that the GPI and vorticity at 850 hPa play the most important roles during the TC genesis processing in the global tropics and the two regional basins. This is not unexpected but does confirm the strength of using machine learning framework. GPI has been well accepted as an important index to evaluate TC genesis. Although it is comprised of other predictors (absolute vorticity, vertical shear, relative humidity, and potential intensity), in a nonlinear way, it is seen as just another predictor by the machine learning models. That the GPI is determined as a more important predictor demonstrates the

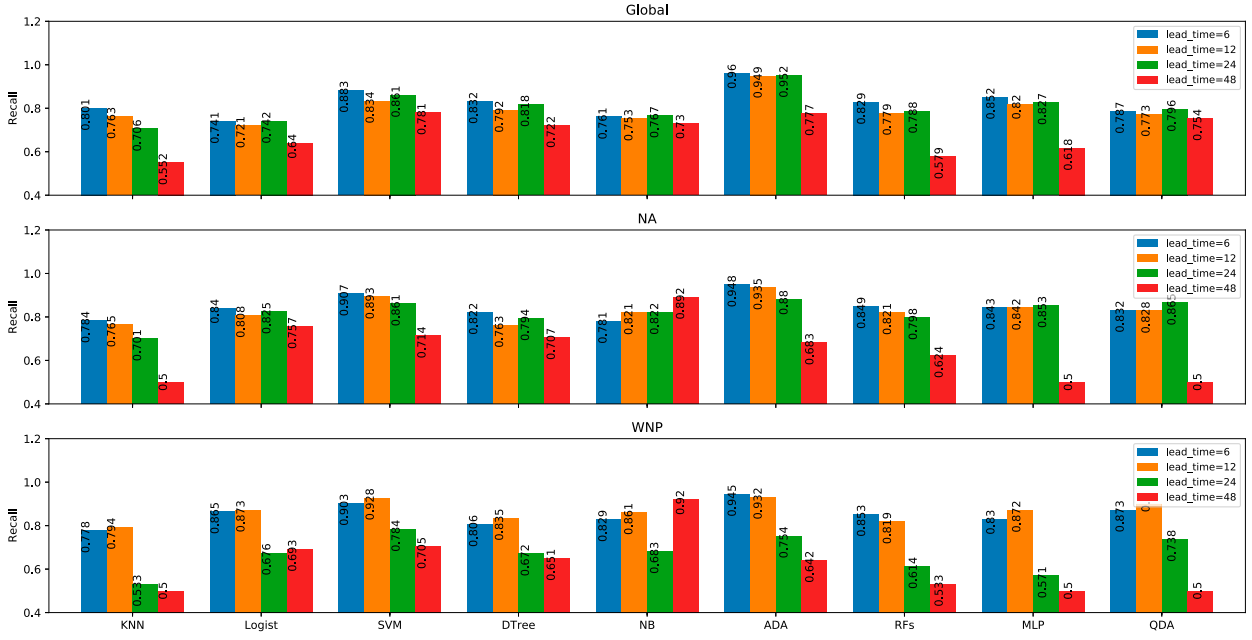

FIG. 11. The precision of machine learning classifiers shown in Table 2 in the global tropical, North Atlantic Ocean, and western North Pacific Ocean at various lead times.

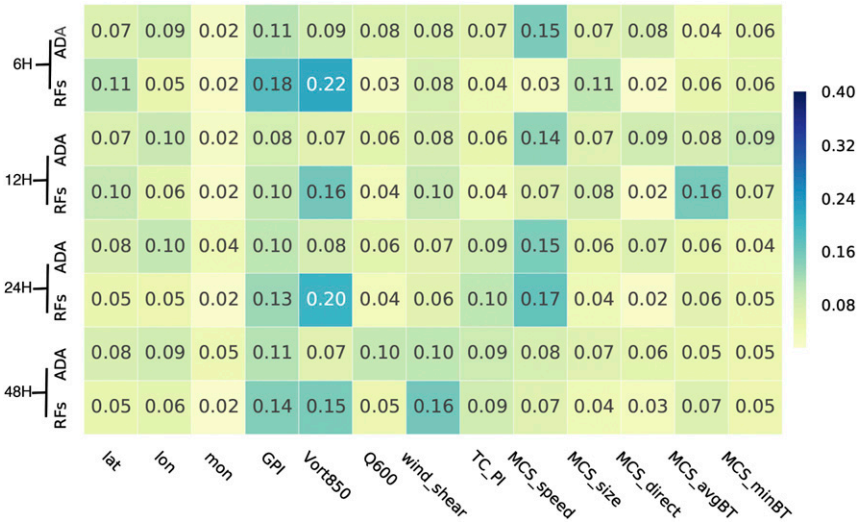

FIG. 12. The relative importance index (MDI) of individual predictors based on the methods of AdaBoost and Random forests in the global tropics at various lead times.

usefulness of GPI in the TC genesis prediction models, and further suggests that a well-vested aggregated index developed by human experts can still play a very effective role even in a modern machine learning model.

Meanwhile, both top-down and bottom-up mechanisms emphasize the low-level vorticity as a potential precursor for TCs. Usually, TCs are generated when the vortices are enhanced and become self-sustainable, which are often in companion with bursts of deep convection with a curved pattern of organized clouds (Zehr 1992). Furthermore, wind shear and TC potential intensity are found to be also important to the TC genesis in NA and WNP basins, respectively. The physical verification and interpretation will be studied in a future work.

5. Summary and discussion

This study presents a machine learning framework to classify whether an MCS will evolve into a TC. The machine learning classifiers are built on environmental predictors and MCS properties that are known to influence TC genesis. The AdaBoost classifier achieves a 97.2% F1-score accuracy at the 6-h lead time and has a robust performance as the lead time is extended to as much as 48 h. The AdaBoost outperforms the traditional GPI approach and other machine learning classifiers.

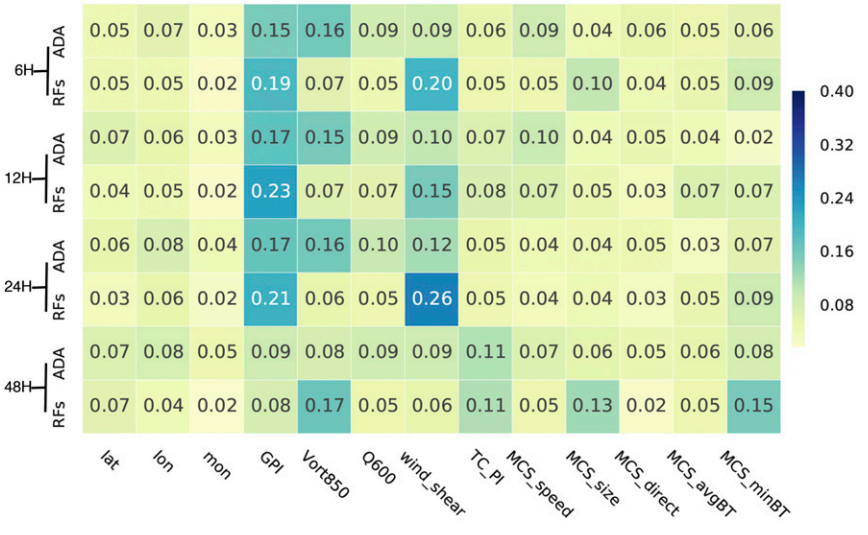

FIG. 13. As in Fig. 10, but for the North Atlantic Ocean.

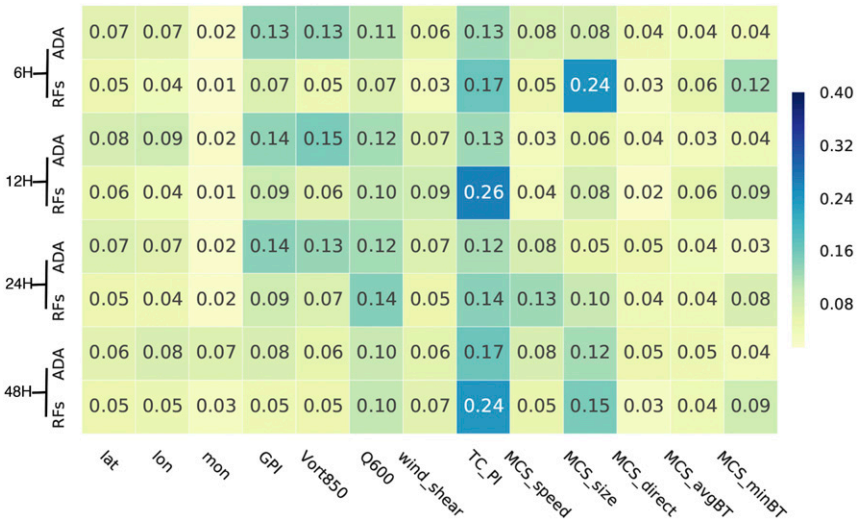

FIG. 14. As in Fig. 10, but for the western North Pacific Ocean.

Although every machine learning classifier tends to suit particular problem types better than others, AdaBoost often stands out with robust performance in various applications. It is a boosting iterative ensemble algorithm, which is a way of combing a couple of weak performance classifiers into a single strong predictive one. At each step, a new weak classifier is trained and the distribution of training set will adjust so that the wrong-predicted samples receive attention in the next iteration. This process is able to effectively reduce the predictive bias since it enables to build a strong ensemble classifier even if the base weak classifiers do not have good generalization ability.

GPI appears to have little predictive skill with a long lead time. Two tree-based classifiers used are also used to find the relative importance of the many predictors relevant to TC genesis. It is found that the low-level vorticity and GPI play the most important roles for TC genesis. These findings are consistent with previous studies. Results suggest that the machine learning framework not only attains a very high predictive skill but also has the potential to reveal new physical mechanisms of TC genesis.

The machine learning framework established in this study has many implications and potential usage in addition to the traditional TC forecast. It can be applied to other weather and climate phenomena, such as classifying prestage conditions leading to MCS genesis and different categories of TCs/hurricanes. It can also be applied to the investigation of physical parameterization schemes, such as convection and cloud schemes. Meanwhile, new discoveries from the machine learning framework can be further verified through physics-based interpretation, having the potential to lead to a deeper understanding of the genesis mechanism of TCs. We plan to extend such applications in separate works.

Acknowledgments. This work is supported by the CMDV Project to Brookhaven National Laboratory under Contract DE-SC0012704 and Brookhaven National Laboratory's Laboratory Directed Research and Development (LDRD) Project 21090.

REFERENCES

Breiman, L., 1998: Arcing classifier (with discussion and a rejoinder by the author). *Ann. Stat.*, **26**, 801–849, https://doi.org/10.1214/aos/1024691079.

——, 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

——, J. H. Friedman, R. A. Olshen, and C. J. Stone, 2017: *Classification and Regression Trees*. Routledge, 368 pp.

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, https://doi.org/10.1175/1520-0493(1997)125<1397:LSIOTC>2.0.CO;2.

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

Chan, J. C., 2005: Interannual and interdecadal variations of tropical cyclone activity over the western North Pacific. *Meteor. Atmos. Phys.*, **89**, 143–152, https://doi.org/10.1007/s00703-005-0126-y.

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

Chen, S. S., and W. M. Frank, 1993: A numerical study of the genesis of extratropical convective mesovortices. Part I: Evolution and dynamics. *J. Atmos. Sci.*, **50**, 2401–2426, https://doi.org/10.1175/1520-0469(1993)050<2401:ANSOTG>2.0.CO;2.

Chen, T.-C., S.-Y. Wang, M.-C. Yen, and W. A. Gallus Jr., 2004: Role of the monsoon gyre in the interannual variation of tropical cyclone formation over the western North Pacific.

Wea. Forecasting, **19**, 776–785, https://doi.org/10.1175/1520-0434(2004)019<0776:ROTMGI>2.0.CO;2.

Cortes, C., and V. Vapnik, 1995: Support-vector networks. *Mach. Learn.*, **20** (3), 273–297.

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

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

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

Emanuel, K., 1986: An air–sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. *J. Atmos. Sci.*, **43**, 585–605, https://doi.org/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, https://doi.org/10.1175/1520-0469(1989)046<3431:TFANOT>2.0.CO;2.

——, 1991: The theory of hurricanes. *Annu. Rev. Fluid Mech.*, **23**, 179–196, https://doi.org/10.1146/annurev.fl.23.010191.001143.

——, and D. S. Nolan, 2004: Tropical cyclone activity and the global climate system. *26th Conf. on Hurricanes and Tropical Meteorolgy*, Miami, FL, Amer. Meteor. Soc., 10A.2, https://ams.confex.com/ams/26HURR/techprogram/paper_75463.htm.

Freund, Y., and R. E. Schapire, 1997: A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139, https://doi.org/10.1006/jcss.1997.1504.

Friedman, N., D. Geiger, and M. Goldszmidt, 1997: Bayesian network classifiers. *Mach. Learn.*, **29**, 131–163, https://doi.org/10.1023/A:1007465528199.

Gray, W. M., 1968: Global view of the origin of tropical disturbances and storms. *Mon. Wea. Rev.*, **96**, 669–700, https://doi.org/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, https://doi.org/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, https://doi.org/10.1175/WAF-D-13-00008.1.

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, https://doi.org/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, https://doi.org/10.1175/WAF890.1.

Holland, G., 1995: Scale interaction in the western Pacific monsoon. *Meteor. Atmos. Phys.*, **56**, 57–79, https://doi.org/10.1007/BF01022521.

Horn, M., and Coauthors, 2014: Tracking scheme dependence of simulated tropical cyclone response to idealized climate simulations. *J. Climate*, **27**, 9197–9213, https://doi.org/10.1175/JCLI-D-14-00200.1.

Huang, X., C. Hu, X. Huang, Y. Chu, Y.-H. Tseng, G. J. Zhang, and Y. Lin, 2018: A long-term tropical mesoscale convective systems dataset based on a novel objective automatic tracking algorithm. *Climate Dyn.*, **51**, 3145–3159, https://doi.org/10.1007/s00382-018-4071-0.

Keller, J. M., M. R. Gray, and J. A. Givens, 1985: A fuzzy K-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybern.*, **SMC-15**, 580–585, https://doi.org/10.1109/TSMC.1985.6313426.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS) unifying tropical cyclone data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

Kohavi, R., 1995: A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95 Proc. 14th Int. Joint Conf. on Artificial Intelligence*, Vol. 2, Montreal, Quebec, Canada, AAAI, 1137–1145.

Kutner, M. H., C. Nachtsheim, and J. Neter, 2004: *Applied Linear Regression Models*. McGraw-Hill/Irwin, 701 pp.

Lander, M. A., 1994: Description of a monsoon gyre and its effects on the tropical cyclones in the western North Pacific during August 1991. *Wea. Forecasting*, **9**, 640–654, https://doi.org/10.1175/1520-0434(1994)009<0640:DOAMGA>2.0.CO;2.

Lee, C.-S., K. K. Cheung, J. S. Hui, and R. L. Elsberry, 2008: Mesoscale features associated with tropical cyclone formations in the western North Pacific. *Mon. Wea. Rev.*, **136**, 2006–2022, https://doi.org/10.1175/2007MWR2267.1.

Liaw, A., and M. Wiener, 2002: Classification and regression by randomforest. *R News*, **2** (3), 18–22.

Lu, X., K. K. Cheung, and Y. Duan, 2012: Numerical study on the formation of Typhoon Ketsana (2003). Part I: Roles of the mesoscale convective systems. *Mon. Wea. Rev.*, **140**, 100–120, https://doi.org/10.1175/2011MWR3649.1.

McBride, J. L., 1995: Tropical cyclone formation. *Global Perspectives on Tropical Cyclones*, R. Elsberry, Ed., WMO, 63–105.

——, and R. Zehr, 1981: Observational analysis of tropical cyclone formation. Part II: Comparison of nondeveloping versus developing systems. *J. Atmos. Sci.*, **38**, 1132–1151, https://doi.org/10.1175/1520-0469(1981)038<1132:OAOTCF>2.0.CO;2.

McLachlan, G., 2004: *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 526 pp.

Palmen, E., 1948: On the formation and structure of tropical hurricanes. *Geophysica*, **3** (1), 26–38.

Parzen, E., 1962: On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065–1076, https://doi.org/10.1214/aoms/1177704472.

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, https://doi.org/10.1175/2011MWR3617.1.

Quinlan, J. R., 1987: Simplifying decision trees. *Int. J. Man Mach. Stud.*, **27**, 221–234, https://doi.org/10.1016/S0020-7373(87)80053-6.

Rickenbach, T. M., and S. A. Rutledge, 1998: Convection in TOGA COARE: Horizontal scale, morphology, and rainfall production. *J. Atmos. Sci.*, **55**, 2715–2729, https://doi.org/10.1175/1520-0469(1998)055<2715:CITCHS>2.0.CO;2.

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, https://doi.org/10.1175/1520-0493(1999)127<2027:LSPAWT>2.0.CO;2.

Rosenblatt, M., 1956: Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, **27**, 832–837, https://doi.org/10.1214/aoms/1177728190.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1985: Learning internal representations by error propagation. Tech. Rep., Institute for Cognitive Science, University of California, San Diego, La Jolla, CA, 49 pp.

Yokoi, S., Y. N. Takayabu, and J. C. Chan, 2009: Tropical cyclone genesis frequency over the western North Pacific simulated in medium-resolution coupled general circulation models. *Climate Dyn.*, **33**, 665–683, https://doi.org/10.1007/s00382-009-0593-9.

Zehr, R. M., 1992: Tropical cyclogenesis in the western North Pacific. Ph.D. thesis, Colorado State University, 189 pp.

Zhang, W., B. Fu, M. S. Peng, and T. Li, 2015: Discriminating developing versus nondeveloping tropical disturbances in the western North Pacific through decision tree analysis. *Wea. Forecasting*, **30**, 446–454, https://doi.org/10.1175/WAF-D-14-00023.1.