A Long Short-Term Memory Model for Global Rapid Intensification Prediction

QIDONG YANG

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

CHIA-YING LEE

Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York

MICHAEL K. TIPPETT

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

(Manuscript received 1 October 2019, in final form 21 April 2020)

ABSTRACT

Rapid intensification (RI) is an outstanding source of error in tropical cyclone (TC) intensity predictions. RI is generally defined as a 24-h increase in TC maximum sustained surface wind speed greater than some threshold, typically 25, 30, or 35 kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$). Here, a long short-term memory (LSTM) model for probabilistic RI predictions is developed and evaluated. The variables (features) of the model include storm characteristics (e.g., storm intensity) and environmental variables (e.g., vertical shear) over the previous 48 h. A basin-aware RI prediction model is trained (1981–2009), validated (2010–13), and tested (2014–17) on global data. Models are trained on overlapping 48-h data, which allows multiple training examples for each storm. A challenge is that the data are highly unbalanced in the sense that there are many more non-RI cases than RI cases. To cope with this data imbalance, the synthetic minority-oversampling technique (SMOTE) is used to balance the training data by generating artificial RI cases. Model ensembling is also applied to improve prediction skill further. The model's Brier skill scores in the Atlantic and eastern North Pacific are higher than those of operational predictions for RI thresholds of 25 and 30 kt and comparable for 35 kt on the independent test data. Composites of the features associated with RI and non-RI situations provide physical insights for how the model discriminates between RI and non-RI cases. Prediction case studies are presented for some recent storms.

1. Introduction

Tropical cyclones (TCs) cause death and bring severe economic damage to society (Geiger et al. 2016; Peduzzi et al. 2012). TC forecasts are a key element in managing TC risk, and considerable effort has been made to improve their accuracy and timeliness. Two important parts of a TC forecast are where a storm will go (track) and how strong its winds will be (intensity). In recent years, improvements in track forecasts have been more striking than improvements in intensity forecasts (DeMaria et al. 2014, and references therein). One notable challenge for intensity forecasts is the inability to accurately predict rapid intensification (RI), which is defined as a 24-h increase in maximum sustained surface wind speed greater than some threshold, typically 25, 30, or 35 kt

($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$; Rozoff et al. 2015). The difficulty in predicting RI has been documented in numerous studies (e.g., DeMaria 1996; Rappaport et al. 2009; Yang 2016). The inability to accurately predict RI reflects an incomplete understanding of the physical mechanisms that are directly responsible for its occurrence. RI is also important in the context of climate and climatological risk because almost all of the strongest storms undergo RI during their lifetime (Lee et al. 2016).

In previous studies, RI has been analyzed from at least three perspectives: inner-core, oceanic, and large-scale processes. The potential relationship between RI and storm inner-core processes has been studied in several works. Willoughby et al. (1982) demonstrated that storms experience substantial intensity increases while outer eyewalls are contracting during a concentric eyewall cycle. Sitkowski and Barnes (2009) identified a spiraling in the eyewall of storm Guillermo (1997) as

Corresponding author: Qidong Yang, qy2216@columbia.edu

DOI: 10.1175/WAF-D-19-0199.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

being responsible for the generation of a small and complete eye that was speculated to trigger a RI process. The upper-ocean structure is another crucial factor. A warm ocean eddy can prevent strong storm-induced sea surface temperature cooling (Wu et al. 2007), and thus promote RI processes (Hong et al. 2000; Cione and Uhlhorn 2003). The importance of large-scale environment forcing on RI is well studied as well. Frank and Ritchie (2001) showed that storms in model simulations are more likely to undergo RI when embedded in regions with low vertical wind shear, which is consistent with DeMaria (1996) and Tang and Emanuel (2012a). Hanley et al. (2001) showed that RI processes occur more often when there is no interaction between the TC and an upper-level trough or cold-core low.

Identification of factors that are important for RI has led to the development of a series of statistical prediction models. First, Kaplan and DeMaria (2003) examined statistical differences in the large-scale environment and storm characteristics for RI and non-RI cases in the Atlantic to determine which factors are informative for prediction, then used those predictors to construct a regression model to predict the probability of RI. Second, Kaplan et al. (2010) leveraged a more complex statistical method to build separate probabilistic RI models for the Atlantic and eastern North Pacific basins. The predictors used included ones proven effective in their previous work (Kaplan and DeMaria 2003), as well as predictors derived from satellite data. Third, Rozoff et al. (2015) used predictors from earlier work (Kaplan et al. 2010) and ones calculated from passive microwave observations in logistic regression models for predicting RI probabilities in the Atlantic and eastern North Pacific basins. Fourth, Kaplan et al. (2015) developed a consensus RI model that combined various models. The consensus model takes strengths from each individual model and outperforms them all. Although RI prediction skill has increased with the introduction of new predictors and methods, the extent to which further improvement is possible is unclear since there may be fundamental limits on the predictability of TC intensity (Brown and Hakim 2013; Judt and Chen 2016; Emanuel and Zhang 2016).

In recent years, machine learning techniques have made great progress in the fields of natural language processing and computer vision, as well as in the physical sciences. Successes in the physical sciences include, but are not limited to, machine learning based methodologies for quantifying uncertainty (Yang and Perdikaris 2019), ocean data inference and subgrid parameterization (Bolton and Zanna 2019), and climate subgrid process representation (Rasp et al. 2018). This success results in part from the flexibility of using a large number

of parameters, which allows the models to approximate any continuous function.

Some researchers have tried to take advantage of the flexibility of machine learning models to make RI predictions. Li et al. (2017) utilized long short-term memory (LSTM; Hochreiter and Schmidhuber 1997) to train a RI model on the Statistical Hurricane Intensity Prediction Scheme (SHIPS) dataset for Atlantic hurricanes (DeMaria and Kaplan 1994). The SHIPS dataset contains predictors from reanalysis fields as well as satellite-derived variables. Yang (2016) conducted a systematic RI classification investigation on the SHIPS dataset to find the optimal combination of machine learning models and RI predictors. The investigated models included Naïve Bayes, decision trees, and their ensembles. Mercer and Grimes (2015) used unsupervised methods for RI predictor selection and incorporated winds, temperature, moisture, and geopotential height as predictors in a support vector machine (SVM; Cortes and Vapnik 1995) to predict RI probabilities. Despite the good function approximation capability of machine learning models, deterministic performance indicators showed limited progress in the aforementioned works. Compared to the work by Kaplan et al. (2010), the methods proposed by Li et al. (2017) and Yang (2016) modestly improved the probability of detection of RI in Atlantic given the same false alarm ratio in the work (Kaplan et al. 2010).

There are many possible reasons for the limited success of previous machine learning efforts, such as fundamental predictability limits (chaos), and data/methodological limitations. Focusing on the machine learning perspective, we speculate that the following issues might be relevant. First, the machine learning models might be overfitting to the training sets, which would lead to poor performance on independent data. A large amount of data is typically required to fit the model's large number of parameters. Models in Li et al. (2017) and Yang (2016) were trained only on the Atlantic dataset, which limits the amount of storm data. Second, the number of RI cases is significantly smaller than the number of non-RI cases because of the rarity of RI, which means the training dataset is highly unbalanced. Due to the small number of RI cases, the RI cases contribute little to the overall penalty. Thus, nonlinear optimization methods may settle on models whose probability predictions show little sharpness (deviation from the base rate) or that, in extreme cases, never predict RI if predictive accuracy is emphasized (Chawla 2010). While the problem of unbalanced data was not discussed in Yang (2016) or Mercer and Grimes (2015), Li et al. (2017) attempted to address it via a weighted loss function approach, with limited success. Last, although

the SHIPS dataset (Yang 2016) includes some predictors with explicit time sequence information such as the intensity change over the previous 12 h, as well as forecast information, longer prior sequences of predictor information might provide additional utility for RI prediction.

Here we propose a machine learning method for RI prediction that is tailored to address these challenges. Global storm data and rolling window data preprocessing are used to provide as many samples as possible for model training; the synthetic minority-oversampling technique (SMOTE; Chawla et al. 2002) is used to cope with data imbalance; and longer predictor sequences are fed to LSTM to help extract additional temporal relation information. The paper is organized as follows. The datasets and the RI predictor selection process are discussed in section 2. This section also describes the rolling window data preprocessing technique and SMOTE. Section 3 gives the details of our RI prediction model. Model performance is reported in section 4, along with case studies. A model interpretation method is introduced in section 5 to understand our model's prediction behavior. Finally, we conclude with a summary.

2. Data preparation

a. Storm, environmental, and comparison data

Two datasets, similar to the ones in Lee et al. (2015, 2016), are used in this study to develop and test the model. The first dataset is used for RI labeling and storm feature calculations and is taken from the National Hurricane Center (NHC; Landsea and Franklin 2013) and the Joint Typhoon Warning Center best track data (JTWC; Chu et al. 2002). The second dataset contains the environmental features, which are calculated using the monthly European Centre for Medium-Range Weather Forecasts interim reanalysis (ERA-Interim; Dee et al. 2011). Although the monthly averaged environmental features contain less detailed information than the higher temporal resolution data that are used in operational predictions, Lee et al. (2015) showed that in a statistical model for intensity prediction there was only a modest reduction in error when using daily environmental data instead of monthly averages, and our experiments support this statement, showing no better performance achieved using daily data. Therefore, we present results here using monthly environmental data and discuss the impact of using higher-frequency data in the Conclusions. The dataset includes storms from five basins: Atlantic, eastern North Pacific (including central North Pacific), northern Indian Ocean, western North Pacific, and Southern Hemisphere. Each storm was

TABLE 1. The number of RI case in the training, validation, and test datasets for thresholds of 25, 30, and 35 kt.

Dataset	25-kt threshold	30-kt threshold	35-kt threshold
Training set	4231	2701	1759
Validation set	319	205	139
Test set	671	496	358

recorded in a time-sequential format with descriptions of the storm's status and its surrounding environment conditions every 6 h. Because we use monthly average environmental fields (centered at 15th of the month) interpolated to daily, changes in the storm environment from one time step to the next are primary due only to the change in storm location. We use storms whose life-time maximum intensity reached at least tropical storm strength (34 kt). Storm intensity and environmental features are calculated from their genesis (the first record) to landfall. We do not include information after landfall. Extratropical stages are also included. More details of the dataset are discussed in Lee et al. (2015).

The model is trained using data from the period 1981–2009 (training dataset). Hyperparameters (e.g., model configuration and learning rate) are tuned on data from the period 2010–13 (validation dataset). Data from the period 2014–17 (test dataset) are used for model evaluation and are completely independent of the training and validation datasets. There are 51 629, 4007, and 6857 samples in the training, validation, and testing datasets, respectively. The information regarding the number of RI cases is summarized in Table 1. Model predictions during the period 2014–17 for the Atlantic and eastern North Pacific are compared with the operational SHIPS Rapid Intensification Index (SHIPS RII).

b. Feature engineering

A total of 35 features (8 storm features, 22 environmental features, and 5 basins) are included in the model. Storm features represent the storm status at the time steps prior to the initialization time, including storm maximum wind, minimum sea level pressure, wind speed changes, basin, and so on (Table 2). Environment features describe the large-scale conditions near the storm and include ocean temperature, wind shear, and so on (Tables 3 and 4). In these two tables, we can see that some variables are averaged over a small disk and others over a large annulus. The former ones contain storm structure information while variables averaged over a larger annulus describe environment features in which the storms are embedded. Now we briefly discuss some of the features.

One of the most important environment controls on TC behavior is maximum potential intensity (MPI;

TABLE 2. Storm features.

Notation	Feature
$\text{Speed}_{\text{max}}$	Storm center max wind speed
$\text{Cos}_{\text{trans}}$	Cosine of storm translation direction
$\text{Sin}_{\text{trans}}$	Sine of storm translation direction
$\text{Speed}_{\text{trans}}$	Storm translation speed
dS/dt_{24}	24-h change of storm max wind speed
dS/dt_{18}	18-h change of storm max wind speed
dS/dt_{12}	12-h change of storm max wind speed
dS/dt_6	6-h change of storm max wind speed
Location	Indicator of the basin

Camargo et al. 2009, 2007). Here, we used the MPI definition developed in Emanuel (1988) and modified in Emanuel (1995) and Bister and Emanuel (2002). MPI calculated using sea surface temperature is a function of outflow temperature which is defined as the 200 hPa temperature away from the storm center. The temperature at 200 hPa ($T_{\text{air}200\text{mb}}$) is also included as a separate feature (Kaplan and DeMaria 2003). MPI also depends on sea surface temperature. Storm-centered and area-averaged values of ocean temperature averaged over the top 100 m are included as predictors (Price 2009). Vertical wind shear variables (meridional, zonal, and annulus averaged) are included since strong vertical wind shear has a negative impact on the TC intensification process (DeMaria 1996; Tang and Emanuel 2012b). Annulus averages are computed over the annulus region extending 200–800 km from the storm center to capture surrounding conditions. Moisture is another useful predictor for the RI process. Kaplan and DeMaria (2003) found that storms were more likely to

undergo RI when they were embedded in moister environments. For our model, storm-center and annulus-averaged high-level relative humidity ($\text{RH}_{500-300\text{mb}}$) and low-level relative humidity ($\text{RH}_{850-700\text{mb}}$) are considered. Instead of convective instability used in DeMaria and Kaplan (1994), we considered storm-centered and annulus averages of conditional instability defined as the vertical gradient of saturated equivalent potential temperature ($d\theta_e/dz$). Storms that undergo RI often show indications of intensity changes at previous time steps, and previous changes in intensity (dS/dt_6 , dS/dt_{12} , dS/dt_{18} , and dS/dt_{24}) are also included. Finally, the model is global, and every storm has a location feature vector with five Boolean elements to indicate the basin.

To interpret our model more comprehensively in section 5, we introduce two passive features: the norm of vertical wind shear (Shear) and the difference between potential max intensity and current storm intensity ($P - S$). Experiments demonstrate that the inclusion of these two features does not add skill to our model because these two variables are dependent on other features that we have already included in the model. Therefore, they are not used for model training but only for model interpretation.

c. Data preprocessing

To address the data insufficiency problem, we first split each storm into overlapping 48-h time chunks. This allows multiple training examples to be yielded from each storm. This method is called “rolling window with overlapping,” and is illustrated in Fig. 1. A storm’s lifetime consists of four components in this figure:

TABLE 3. Environment features. Variables presented here are either derived at storm center or averaged over a disk or over an annulus. Disk-averaged variables are averaged over the area within 500 km of the storm center while annulus-averaged ones are averaged over a ring-shape area with inner radius of 200 km and outer radius 800 km from storm center.

Notation	Feature
$\text{PI}_{\text{pressure}}$	Potential intensity of minimum sea level pressure at the storm center
$\text{PI}_{\text{pressure}_d}$	Potential intensity of minimum sea level pressure, disk averaged
PI_{speed}	Potential intensity (kt) at storm center
$\text{PI}_{\text{speed}_d}$	Potential intensity (kt), disk averaged
$T_{\text{ocean}0-100\text{m}}$	Ocean temperature averaged over top 100 m, at storm center
$T_{\text{ocean}0-100\text{m}_d}$	$T_{\text{ocean}0-100\text{m}}$ but averaged over a 100 km storm-centered, disk averaged
$T_{\text{air}200\text{mb}}$	Temperature at 200 hPa at storm center
$T_{\text{air}200\text{mb}_a}$	$T_{\text{air}200\text{mb}}$, annulus averaged
Shear_z	Vertical wind shear (zonal) between 850 and 200 hPa at storm center
Shear_{z_a}	Shear_z , annulus averaged
Shear_m	Vertical wind shear (meridional) between 850 and 200 hPa at storm center
Shear_{m_a}	Shear_m , annulus averaged
$\text{Div}_{200\text{mb}}$	Divergence field at 200 hPa
$\text{Div}_{200\text{mb}_a}$	$\text{Div}_{200\text{mb}}$, annulus averaged
$d\text{PI}_{\text{pressure}}/dt_{12}$	Change in potential intensity of minimum sea level pressure over 12 h
$d\theta_e/dz$	Vertical gradient of saturated equivalent potential temperature averaged from boundary layer to 300 hPa at storm center

TABLE 4. Environment features (continued).

Notation	Feature
$d\theta_e/dz_a$	$d\theta_e/dz$, annulus averaged
$RH_{850-700mb}$	Low-level relative humidity (850–700 hPa) at storm center
$RH_{850-700mb_a}$	$RH_{850-700mb}$, annulus averaged
$RH_{500-300mb}$	High-level relative humidity (500–300 hPa)
$RH_{500-300mb_a}$	$RH_{500-300mb}$, annulus averaged
Landmask	Storm location: land or ocean
$P - S$	$PI_{speed_d} - Speed_{max}$
Shear	Norm of shear vector

time, environment features, storm features, and intensity. Time defines the time steps in a storm's lifetime, and intensity is used to construct RI/no-RI labels. The rolling windows are overlapping 48-h windows. The data preprocessing starts with a black window covering the period from time step 0 to time step 7. The features at the eight time steps become the first sample of the storm denoted as s_1 , and its label y_1 indicating whether RI occurs in the next 24 h is defined as $y_1 = 1[(Speed_{max11} - Speed_{max7}) \geq \text{threshold}]$; the function $1[\cdot]$ is one when its argument is true and zero otherwise. The RI/no-RI label y_1 is the quantity to be predicted. Thus, the first sample for a given storm will be the features of the first 48 h (8 time steps) of the storm. This sample is then used to predict whether RI occurs in the next 24 h (hour 72 of the storm). In the training dataset, the RI threshold used is 25 kt due to the limited number of RI cases for higher thresholds. When testing the model's performance for different thresholds, thresholds of 25, 30, or 35 kt can be used. Pairs of consecutive windows are then shifted by one time step of 6 h. Thus, the window following the black window is the red one. The red window contains the features from time step 1 to time step 8, which become the second sample s_2 with its associated label $y_2 = 1[(Speed_{max12} - Speed_{max8}) \geq \text{threshold}]$. Similarly, the green window covers the third sample s_3 with label $y_3 = 1[(Speed_{max13} - Speed_{max9}) \geq \text{threshold}]$. Following this procedure, the windows roll to the end of the storm's lifetime, and those samples without 4 future time steps are not included in our dataset since we do not have access to their RI labels. In the end, each storm record is divided into multiple overlapping samples. Every sample s_i consists of the features during eight time steps: $s_i = \{\mathbf{x}_{j_i}, \dots, \mathbf{x}_i, \dots, \mathbf{x}_{j_i+7}\}$, where we call the vector $\mathbf{x}_t$ of length d a "description" that denotes the features at time step t , and d is the number of features included in the single time step's description. The term j is a list whose i th element j_i indicates the starting time point of sample s_i in a storm. In what follows, $d = 36$, and it will sometimes be convenient to consider $\mathbf{x}_t$ as a row vector of length 36, which is the

*one time step means six hours

Time*	Environment features	Storm features	Intensity
0			V_0
1			V_1
2	...	...	V_2
3			V_3
4			V_4
5			V_5
6	...	...	V_6
7			V_7
8			V_8
9			V_9
10	...	...	V_{10}
11			V_{11}
12			V_{12}
...			...

FIG. 1. Rolling window data preprocessing.

number of features (30 storm/environment features and the location feature vector of length 5; one of the environment features, Landmask, occupies two Boolean elements to indicate two types of storm location: land or ocean).

d. Data balance

To cope with the data imbalance issue (RI is relatively rare), we applied the synthetic minority-oversampling technique (SMOTE; Chawla et al. 2002) on the pre-processed dataset. For the convenience of visualization, Fig. 2 illustrates the underlying idea of SMOTE in a two-dimensional space instead of in the original high-dimensional feature space.

Figure 2a shows two sample distributions, with yellow points representing RI cases and blue points representing non-RI cases. The rarity of RI cases is reflected by the smaller number of yellow points compared to the number of blue points. To apply SMOTE, we randomly choose an RI case denoted as the black point in Fig. 2b. Second, the chosen RI case is connected with its nearest RI neighbors denoted as green points, which are determined by Euclidean distance in the feature space. The number of nearest neighbors used is a hyper-parameter of this technique, which can be adjusted to achieve the

FIG. 2. For the convenience of visualization, high-dimensional feature space is represented in a two-dimensional space. (a) Two sample distributions, where blue points represent non-RI cases and yellow points represent RI cases. (b) A RI case is chosen denoted as a black point, and it is connected to its two nearest neighbors denoted as green points by dashed lines. (c) Two RI cases denoted as red points are oversampled on those dashed lines.

best performance. In our task, experiments (validation data) demonstrate that five is the best choice. For simplicity, only two green points are drawn in Fig. 2b. Last, a point is linearly interpolated randomly on each connected line, and they are regarded as oversampled (artificial) RI cases (red points in Fig. 2c). Following the above procedures, the original unbalanced dataset consisting of 4231 RI cases and 47 398 no-RI cases was balanced to form a dataset with 47 398 cases for both RI and no-RI.

3. Rapid intensification prediction

a. Model structure

The model used for RI prediction has two parts: long short-term memory (LSTM) and classifier as shown in Fig. 3. LSTM is used as a feature extractor. A sample s_i with its eight descriptions $\{\mathbf{x}_{j_i}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{j_i+7}\}$ is fed to the LSTM one by one in time sequence. After inputting a description $\mathbf{x}_t$, the LSTM's current state $\mathbf{h}_t$ is generated. The state vector $\mathbf{h}_t$ has length k where k is the size of the hidden state and takes information not only from current time step's description $\mathbf{x}_t$ but also from the previous time step's state $\mathbf{h}_{t-1}$. As descriptions are input time sequentially, the state of the LSTM updates, which allows LSTM to capture the temporal characteristics of the input samples. After feeding in all the descriptions in sample s_i , LSTM outputs the final state $\mathbf{h}_{j_i+7}$. This state is a condensed representation of s_i and should include the most important information for predicting RI. In other words, $\mathbf{h}_{j_i+7}$ is the feature vector that LSTM extracts from s_i , summarized as $\mathbf{h}_{j_i+7} = \text{LSTM}(s_i)$.

Then the extracted feature vector is input to the model's second part, which is the classifier that will output the probability of RI. Here, the classifier is a one-layer fully connected neural network (Jain et al. 1996), which is comprised of a neuron. It computes a linear

combination of the extracted feature from the LSTM. The weights in the linear combination are learned as in a linear regression model; then the neuron applies a nonlinear activation function—sigmoid activation function. The function forces the output to be in the range from 0 to 1, which is the probability $P(\text{RI}_i)$ of RI (25-kt threshold) occurring in the following 24 h based on the sample s_i : $P(\text{RI}_i) = \text{Classifier}[\text{LSTM}(s_i)]$. RI probabilities for thresholds of 30 and 35 kt are computed in an offline calibration procedure by logistic regression which will be discussed in section 3e.

b. Long short-term memory

The ability of LSTM to encode temporal relationships comes from its unique gating mechanism, which makes it suitable for temporal extrapolation and temporal feature extraction. In short, LSTM extracts the feature $\mathbf{h}_{j_i+7}$ from the input sample s_i , and thus LSTM can be viewed as a complicated function with $\mathbf{h}_{j_i+7} = \text{LSTM}(s_i)$. In this

FIG. 3. The model consists of two parts: LSTM and classifier. A sample s_i with its eight descriptions $\{\mathbf{x}_{j_i}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{j_i+7}\}$ is fed to the LSTM. As descriptions are input to LSTM one by one, LSTM's current state updates from $\mathbf{h}_{j_i}$ until $\mathbf{h}_{j_i+7}$. Then $\mathbf{h}_{j_i+7}$ as an extracted feature vector is fed to the classifier and RI probability is finally obtained.

FIG. 4. Long short-term memory structure. The terms I , F , and O indicate the input gate, forget gate, and output gate, respectively. The terms $\mathbf{x}_{t-1}$, $\mathbf{x}_t$, and $\mathbf{x}_{t+1}$ represent consecutive three time steps' descriptions. The terms $\mathbf{h}$ and $\mathbf{c}$ are a time step's state and memory cell. The $\tanh$ is the hyperbolic-tangent activation function, and σ is the sigmoid activation function. The $\times$ and $+$ marks mean element-wise multiplication and addition, respectively.

section, we describe the basic structure and mathematics of the LSTM and refer readers to Goodfellow et al. (2016) for further details. Readers who are not interested in LSTM details can skip this section.

To understand its feature extraction details, LSTM's structure is shown in Fig. 4. In addition to the state variable $\mathbf{h}_t$, LSTM has a memory cell variable $\mathbf{c}_t \in \mathbb{R}^{1 \times k}$ to store information from the storm descriptions. LSTM has three gates that control the stream of state information: a forget gate $F \in \mathbb{R}^{1 \times k}$, an input gate $I \in \mathbb{R}^{1 \times k}$, and an output gate $O \in \mathbb{R}^{1 \times k}$. Those gates at time step t are updated by the following equations:

$$\begin{aligned} I_t &= \text{sigmoid}(\mathbf{x}_t \cdot \mathbf{W}_{xi} + \mathbf{h}_{t-1} \cdot \mathbf{U}_{hi} + \mathbf{b}_i) \\ F_t &= \text{sigmoid}(\mathbf{x}_t \cdot \mathbf{W}_{xf} + \mathbf{h}_{t-1} \cdot \mathbf{U}_{hf} + \mathbf{b}_f) \\ O_t &= \text{sigmoid}(\mathbf{x}_t \cdot \mathbf{W}_{xo} + \mathbf{h}_{t-1} \cdot \mathbf{U}_{ho} + \mathbf{b}_o). \end{aligned} \quad (1)$$

The sigmoid is an activation function mapping from $\mathbb{R}$ to $(0, 1)$; the $\mathbf{W} \in \mathbb{R}^{d \times k}$ and $\mathbf{U} \in \mathbb{R}^{k \times k}$ are weight matrices, and the $\mathbf{b} \in \mathbb{R}^{1 \times k}$ are bias terms; “ $\cdot$ ” is the usual matrix multiplication. In the LSTM unit, the forget gate and the input gate use the current storm description $\mathbf{x}_t$ and the previous state $\mathbf{h}_{t-1}$ to control the information kept in the memory cell $\mathbf{c}_t$ as described by the equation:

$$\mathbf{c}_t = F_t \odot \mathbf{c}_{t-1} + I_t \odot \tilde{\mathbf{c}}_t, \quad (2)$$

where $\odot$ is the Schur product (element-wise multiplication), and $\tilde{\mathbf{c}}_t$ is a memory cell candidate. The forget gate F controls how much information from the last memory cell to forget, and the input gate I decides how much information from the candidate memory cell to keep. The memory candidate is computed from

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{x}_t \cdot \mathbf{W}_{xc} + \mathbf{h}_{t-1} \cdot \mathbf{U}_{hc} + \mathbf{b}_c), \quad (3)$$

where again the $\mathbf{W}_{xc} \in \mathbb{R}^{d \times k}$ and $\mathbf{U}_{hc} \in \mathbb{R}^{k \times k}$ are weights, $\mathbf{b}_c \in \mathbb{R}^{1 \times k}$ is a bias term, and $\tanh$ is the activation function. The output gate O_t is the last gate and controls the state update by determining how much information from the memory cell goes into the new state: $\mathbf{h}_t = O_t \odot \tanh(\mathbf{c}_t)$.

c. Model training

Before the model can be used for RI prediction, its parameters must be estimated using the training dataset, which are the data from the period 1981–2009. The parameters consist of two sets: the parameters in LSTM and the parameters of the classifier. As mentioned before, we estimate the model parameters for the case where RI is defined as intensification greater than 25 kt in 24 h since there are substantially fewer RI cases at higher thresholds. In the model training, the first step is to initialize the parameters of the LSTM and the classifier with random values. Second, the model takes in training data samples s_i and outputs predicted RI probability $P(\text{RI}_i)$ for all i . Third, the distance between the predicted RI probabilities and the RI labels y_i is measured by a loss function, which is logarithm loss in our case. Optimization of the loss function yields the model parameters. The training steps can be expressed by the following formula:

$$\argmin_{\theta} \frac{1}{N} \sum_{i=1}^N L\{\text{Classifier}[\text{LSTM}(s_i)], y_i\}, \quad (4)$$

where θ refers to the trainable parameters from both the LSTM and the classifier. N is the number of samples, and L is the loss function used to measure the difference between predictions and observations. Because our task is a standard binary classification, the loss function used

TABLE 5. Global and basin values of Brier skill score (BSS) and RI cases for RI threshold of 25 kt.

Basins	Cases	Ratio	BSS
Global	671	0.036	0.241
Atlantic	78	0.087	0.197
Eastern North Pacific	165	0.024	0.302
Northern Indian Ocean	24	0.133	0.337
Southern Hemisphere	138	0.000	0.130
Western North Pacific	266	0.040	0.258

is logarithm loss by convention, which is often called cross-entropy loss (De Boer et al. 2005) in machine learning and the negative of the log or ignorance score in weather applications (Roulston and Smith 2002). Explicitly, for a forecast probability p and occurrence y , the cross-entropy loss is

$$L(p, y) = -y \log p - (1 - y) \log(1 - p). \quad (5)$$

When RI occurs ($y = 1$), the loss function becomes $-\log p \in [0, \infty)$ given probability $p \in [0, 1]$, which is made small to zero by making p as close as possible to one. When RI does not occur ($y = 0$), the loss function becomes $-\log(1 - p) \in [0, \infty)$ given probability $p \in [0, 1]$, which is made small by making p as close as possible to zero. As Eq. (4) shows, the parameters θ are adjusted to reduce the average (over all training cases) of the loss function L . There are many optimizers available to solve this optimization problem such as Adam, RMSprop, and Adagrad. Experiments found that RMSprop by Tieleman and Hinton (2012) works well here making training process more stable.

d. Model ensemble

To improve predictive performance further, we trained two models and combined their predictions. The following formula shows the prediction mixture formula:

$$P(\text{RI}) = \text{ratio } M_+(s_i) + (1 - \text{ratio}) M_-(s_i), \quad (6)$$

where M_+ is a positive model trained on a SMOTE-balanced dataset, and M_- is a negative model trained on the originally unbalanced dataset. Both positive model and negative model have the same structure as shown in Fig. 3 and are trained on 25-kt threshold RI dataset. The best ratio for combining the models for each threshold is obtained from the training dataset of the corresponding RI threshold and summarized in Tables 5–7. Using the ratio associated with each threshold, the ensemble model for each threshold is created. The values of the ratio are quite small for the 25-kt threshold (Table 5), which is an indication of the

TABLE 6. As in Table 5, but for a RI threshold of 30 kt. BSSc is the Brier skill score for the calibrated (logistic regression) predictions.

Basins	Cases	Ratio	BSS	BSSc
Global	496	0.000	0.194	0.198
Atlantic	60	0.000	0.172	0.163
Eastern North Pacific	122	0.000	0.264	0.298
Northern Indian Ocean	18	0.024	0.257	0.231
Southern Hemisphere	99	0.000	0.053	0.055
Western North Pacific	197	0.000	0.219	0.211

modest improvement due to using the SMOTE-balanced dataset. The ratios are almost zero for the higher thresholds (Tables 6 and 7), indicating that SMOTE contributes little to ensemble models' predictions for higher thresholds. Recalling that the two components of the ensemble models are fit using the 25-kt threshold, and that the ratio for each threshold is estimated from the corresponding threshold's training data, this differing behavior of the ratio depending on threshold indicates that the SMOTE-balanced low-threshold dataset does not capture well the distribution of the higher threshold datasets.

e. Offline calibration for thresholds of 30 and 35 kt

We explored three approaches for generating predictions of RI for the thresholds of 30 and 35 kt. The first one is to directly use the ensemble model associated with each threshold (30 and 35 kt). However, it is not satisfying since the probability of intensifying 35 kt or more should be less than that of intensifying 25 kt or more, and that is not the case for the two components of the ensemble models, which are trained on 25-kt threshold RI dataset. The second approach is to directly train a new ensemble model on 30/35-kt threshold RI dataset, where two components of each ensemble model are also trained on associated threshold's dataset. The third approach is to view the probability of intensification greater than 25 kt as an extracted feature, which then is used to predict the probability of exceedance of the higher thresholds via probit transformation and logistic regression. Since the LSTM performs effectively on 25-kt threshold RI dataset, we see it as having

TABLE 7. As in Table 5, but for a RI threshold of 35 kt.

Basins	Cases	Ratio	BSS	BSSc
Global	358	0.000	0.091	0.165
Atlantic	40	0.000	0.125	0.149
Eastern North Pacific	89	0.000	0.126	0.255
Northern Indian Ocean	13	0.000	0.137	0.140
Southern Hemisphere	67	0.000	−0.103	0.016
Western North Pacific	149	0.000	0.141	0.181

extracted all predictive information from input storm features and environment features and transformed it into the predicted probability. That is, the predicted probability includes all the predictive information from inputs. This is the basis for using the prediction probability to fit logistic regressions. Although the achieved prediction performance by the third approach is comparable with that of the second approach, the third approach is preferred for two reasons. The first reason is that the interpretation of LSTM model is complicated. Using the third approach means that only a single pair of LSTM models needs to be interpreted. The second reason is that LSTM models are more computationally expensive to train than logistic regression models.

In the third approach, we fit a logistic regression for each basin whose input is the probit-transformed 25-kt exceedance probabilities and whose outputs are the probability of exceeding 30- and 35-kt thresholds on unbalanced training datasets. Probit transformation is defined as $\sqrt{2} \operatorname{erf}^{-1}[2P(\text{RI}) - 1]$. The primary reason for the probit transformation is to convert the probability input to have larger range so that logistic regression output has the full range 0–1 as it should. Although the number of cases is small, the number of parameters to be fit in logistic regressions is substantially smaller—two for each basin and threshold.

4. Model evaluation

a. Skill scores

The ensemble models are evaluated on independent (testing) storm data during the period 2014–17. The testing data are not used in any part of the development of the ensemble models. To have a better sense of our models' performance, the quality of the probabilistic predictions is measured by the Brier skill score (BSS; [Brier 1950](#)). Positive values of BSS indicate that the RI predictions are better than reference predictions, which means that mean squared error between RI predictions and observations is smaller than the error between reference predictions and observations, while negative values would indicate the opposite. The reference prediction is the basin-dependent RI base rate, computed from storm data during the period 1981–2009. BSS values were computed globally and by basins for each RI threshold. BSS values are positive globally, in all basins, and for all three RI thresholds ([Tables 5–7](#)). BSS values decrease in all basins as the RI threshold increases, likely in part due to the two components of ensemble models being trained on the 25-kt threshold RI dataset, though the Brier score has been noted to depend on event frequency ([Stephenson et al. 2008](#)). Comparing

BSS values across basins at the same threshold, our model performs best in eastern North Pacific and northern Indian Ocean basins, and worst in the Southern Hemisphere. The performance in Atlantic is the most stable one as the RI threshold increases because most RI cases in Atlantic are above high thresholds. Over half of the 25-kt RI cases there are 35-kt RI cases.

A natural question is how the performance of the ensemble models compares with that of previously proposed methods and real-time predictions. First, we compare the BSS of the ensemble models in Atlantic and eastern North Pacific with that of previously published methods, noting that this is an imperfect comparison since different independent (test) periods are used. Then we compare the BSS of the ensemble models with that of real-time predictions on the same period.

[Kaplan et al. \(2010\)](#) derived two RI prediction models and computed their BSS on independent data during 2006 and 2007. For Atlantic basin, the best model's BSS values were 0.068, 0.165, and 0.145 for thresholds of 25, 30, and 35 kt, respectively. These BSS values in independent data were substantially lower than the ones in the training data, which can be an indication that training data and test data have dissimilar distributions or that there is overfitting. For the eastern North Pacific, the best model's BSS values in independent data (2006–07) were 0.167, 0.14, and 0.036. For both basins, the BSS values of the ensemble models here are similar or higher, but are computed on the different independent period (2014–17). [Rozoff et al. \(2015\)](#) developed four models for RI prediction, with some including passive microwave predictors, and assessed their performance on independent data from 2004 to 2013. For the Atlantic basin, the best model's BSS values were 0.205, 0.175, and 0.097. For the eastern North Pacific, the best model's BSS values on independent data (2006–07) were 0.227, 0.168, and 0.13. Again, for both basins, the BSS values of the ensemble models here are similar or higher, but are computed on the different independent period (2014–17). The best models in [Rozoff et al. \(2015\)](#) were those that used passive microwave predictors, which agrees with [Kieper and Jiang \(2012\)](#) and suggests additional room for improvement in the model here. Overall, the above comparison demonstrates that the ensemble models have a level of prediction skill that is comparable with that of previously proposed methods.

To exclude the effect of different testing periods on model evaluation, we compare the ensemble model predictions with SHIPS real-time predictions for the same period 2014–17. One advantage we have over SHIPS is that we are using the best track and reanalysis data while SHIPS is using the working best track and the analysis/forecast fields that are available in real time.

FIG. 5. BSS comparison between our proposed method and SHIPS real-time prediction model in Atlantic and eastern North Pacific. BSS values are computed from storm data of the same time period (2014–17). The blue column and yellow column are the BSS values of the SHIPS model in the Atlantic and eastern North Pacific, respectively. The orange column and purple column represent our ensemble models' performance in the Atlantic and eastern North Pacific without calibration (logistic regression), respectively. The gray column and green column show our ensemble models' BSS values after calibration in the Atlantic and eastern North Pacific, respectively.

On the other hand, SHIPS uses more predictors and high-frequency data instead of monthly averages. More predictors and high-frequency data provide the SHIPS model with more comprehensive information regarding the storm and its surrounding environment, which gives it better a chance to make correct predictions. Overall, it is not immediately clear whether the model here is at an advantage or disadvantage. The details of the comparison between our model predictions with SHIPS real-time predictions are shown in Fig. 5 for the Atlantic and eastern North Pacific. For the 25-kt threshold, our ensemble model's performance is better than the SHIPS real-time predictions in both the Atlantic and eastern North Pacific. To evaluate the significance of this relationship, we run a sign test on the 25-kt threshold Atlantic and eastern North Pacific test datasets (DelSole and Tippett 2014). The sign test uses the Brier score as the evaluation criterion, and the null hypothesis is that our ensemble model has the same performance as SHIPS for the 25-kt threshold. The resultant p values are 1.463×10^{-21} and 1.242×10^{-77} for the Atlantic and eastern North Pacific, respectively, which is a strong indication of statistical significance. As the RI threshold increases, there is some obvious degradation of our models' performance. For the RI threshold of 35 kt, the performance of the ensemble model is slightly worse than the SHIPS real-time predictions. A similar sign test is performed on the 35-kt threshold Atlantic and eastern North Pacific test datasets. The resultant p values indicate the relationship is not as statistically significant as the relationship on the 25-kt threshold Atlantic and

eastern North Pacific test datasets. Figure 5 also shows that the two calibrations for thresholds of 30 and 35 kt have overall good performance. In eastern North Pacific, the BSS increases from 0.264 (0.126) to 0.298 (0.255) for thresholds of 30 (35) kt. The BSS for the 35-kt threshold more than doubles. The calibrated models outperform SHIPS for the 30- and 35-kt thresholds in the eastern North Pacific significantly. Similar sign tests are also performed to evaluate the relationship for the 30 and 35-kt thresholds in the eastern North Pacific. The resultant p values are 2.222×10^{-102} and 7.085×10^{-117} for the 30 and 35-kt thresholds, again strong indications of statistical significance. In the Atlantic, BSS goes from 0.172 (0.125) to 0.163 (0.149) with calibration for the 30- and 35-kt thresholds, respectively. The BSS in the Atlantic for 30 kt suffered from slight degradation. One possible reason for the large increase of BSS in eastern North Pacific but the slight decrease of BSS in Atlantic is that the environment climatology in the eastern North Pacific during 2014–17 is similar to that during 1981–2009, while the environment climatology in Atlantic during 2014–17 is different.

Besides BSS, we also evaluate our models via the reliability diagram and performance diagram. In Fig. 6, most of the predicted probabilities are in the range from 0 to 0.4. This is reasonable because RI is such a rare event. As the threshold goes up, the number of mean predicted values at the range from 0 to 0.1 increases, which means ensemble models are correctly calibrated to lower prediction probabilities for 30- and 35-kt thresholds. In the upper panel of Fig. 6, the blue line

FIG. 6. (top) Reliability diagram for five models in global scale: ensemble models for 25-, 30-, and 35-kt thresholds and calibrated ensemble models for 30- and 35-kt thresholds denoted as blue, orange, green, red, and purple. Only squares associated with bins to which over 50 predictions belong are plotted. The dashed line represents the perfect reliability as standard. (bottom) The prediction count diagram. The five models are denoted using the same colors as the top panel.

(for the 25-kt model) is right on top of the dashed line, while the orange line and green line (for the 30- and 35-kt models) are always under the dashed line, suggesting that the ensemble models' performance degrades as threshold increases. This is reasonable as two of the models' components are trained on 25-kt threshold. Conversely, the red and purple lines (the calibrated 30- and 35-kt models) are above the dashed line indicating underconfidence. Overall, the reliability of the forecast probabilities is fairly good, and the logistic regression procedure has a positive effect on improving ensemble models' performance for high thresholds.

Figure 7 summarizes the probability of detection (POD), success ratio (SR), and critical success index (CSI), and bias for our models. For uncalibrated ensemble models, the ensemble model for the 25-kt threshold (the red dot) basically has the best performance. It

gives better scores for SR, CSI, and bias than other uncalibrated ensemble models. As the threshold increases, uncalibrated ensemble model's performance degrades, which is indicated by yellow dot and green dot being located farther to the upper left than the red dot. Comparing models for the same threshold before and after calibration, the calibrated model achieves better performance. For example, the calibrated model for the 30-kt threshold (the orange dot) gives better CSI, bias, and SR than the corresponding uncalibrated model (the green dot). This illustrates that the calibration can help to improve model's discriminant performance.

b. Case studies

To give an indication of the character of the ensemble model predictions, we show four case studies from the

FIG. 7. The performance diagram summarizes critical success index (CSI), success ratio (SR), probability of detection (POD), and bias. Dashed lines represent bias scores, while colored contour fields are CSI. The performance of five models in global scale: ensemble models for 25-, 30-, and 35-kt thresholds and calibrated ensemble models for 30- and 35-kt thresholds, which are represented by red, green, yellow, orange, and pink dots, is demonstrated on this diagram. Each dot is chosen by maximizing the difference between POD and false alarm ratio (FAR) for each model.

test data (Fig. 8). They are three hurricanes (Maria, Harvey, Patricia) and one typhoon (Meranti). For the Atlantic and eastern North Pacific cases we also show the SHIPS real-time predictions. Overall, our model predicts higher probabilities when RI is observed than when RI is not observed. When comparing performance among the four cases, we notice that our model has better performance on Maria, Patricia, and Meranti than on Harvey. One possible reason is that Maria, Patricia, and Meranti all have relatively steady intensification process which provides useful time sequence information to predict intensity changes. Since intensity change (derivative of intensity with respect to time) plays an important role in RI prediction, it is plausible that RI of storms with steady intensification processes can be better predicted. Like our model, the SHIPS model produces higher probabilities when RI is observed. However, our model predicts lower RI probabilities than SHIPS when RI is not observed, especially for Maria and Harvey. Most of the green patches, representing the predictions of our model, are under the purple lines which indicate the SHIPS predictions. The Patricia figure demonstrates there is a

delay between the SHIPS RI predictions and RI occurrences, while our ensemble model predicts RI promptly. These cases are consistent with our model being more skillful (higher BSS) than the SHIPS model in these basins.

5. Model interpretation

Composite feature images

To interpret the basic behaviors of the model, we examine composite feature images associated with RI and non-RI cases for the 25-kt threshold. The positive composite feature image (Image_p) and the negative composite feature image (Image_n) are averages of the samples when the model predicts the probability of RI to be greater than the cutoff and when the model predicts the probability of RI to be less than the cutoff, respectively. The cutoff is chosen from the receiver operating characteristic (ROC) curve in the training dataset by maximizing Peirce score, which is the difference between true positive rate and false positive rate. Composite images are defined by the following two equations:

FIG. 8. Probabilistic RI predictions from the ensemble model (green) for Hurricanes Maria, Harvey, Patricia and Typhoon Meranti. Black lines show the observed storm intensity and gray patches label when RI greater than 30 kt occurs. Purple lines are SHIPS real-time RI probabilistic predictions for Maria, Harvey, and Patricia for the 30-kt threshold. There is no SHIPS real-time prediction for Typhoon Meranti. The x axis represents the lifetime of the hurricanes and typhoon. The inset figures show storm tracks.

$$\text{Image}_p = \frac{1}{N_1} \sum s_i 1[P(\text{RI}_i) > \text{cutoff}]$$

$$\text{Image}_n = \frac{1}{N_2} \sum s_i 1[P(\text{RI}_i) \leq \text{cutoff}], \quad (7)$$

where N_1 and N_2 are the numbers of samples predicted as RI cases and non-RI cases, respectively. Image_p describes how storms behave on average in the 48 h before RI is predicted. Image_n describes how storms behave on average in the 48 h before RI is not predicted. These composite images can be computed globally and by basin. The difference image $\text{Image}_d = \text{Image}_p - \text{Image}_n$ provides information as to how, on average, the ensemble model discriminates RI cases from non-RI cases. Features are standardized to have zero mean and unit variance. Figure 9 shows the difference image for each basin and for the globe.

The Atlantic difference image shows that storms in Atlantic that are predicted to undergo RI move into environments with higher ocean temperatures and are intensifying faster than those that are not predicted to undergo RI, as indicated by ocean temperature stripes and intensification rate stripes that become redder over the 48-h period prior to prediction. This conclusion is

consistent with Rozoff et al. (2015). The predictor $P - S$ is positive, which agrees with the work by Kaplan and DeMaria (2003) showing the intensity of storms undergoing RI are farther away from their potential maximum than those not going through RI. The $P - S$ stripe becomes a lighter red near the end of the 48-h period, which is physically reasonable as well. Storms often undergo small intensification before RI. The Shear stripe is constantly blue at the bottom of the image and demonstrates that the storms tend to be in low vertical wind shear situations before RI, in agreement with Tang and Emanuel (2012b). Rows corresponding to moisture variables are red and near the top of the image, reflecting the known preference of moister environments for RI (Kaplan and DeMaria 2003). The conclusions from the Atlantic are nearly the same as for the eastern North Pacific, with ocean temperatures and upper-level outflow temperatures being exceptions. There are no pronounced differences in these temperatures between RI cases and non-RI cases in eastern North Pacific, which may reflect the differing environments of the basins. Another notable characteristic in the eastern North Pacific image is that divergence stripes are redder than those in Atlantic image. It is not surprising for the $\text{Div}_{200\text{mb}}$ stripe to be red because higher divergence is favorable for RI (Rozoff et al. 2015).

In the northern Indian Ocean, storms are more likely to undergo RI in environments that are relatively more humid and have larger 200-hPa divergence. The red strips in Fig. 9 for these two variables become less red with increasing time step, meaning that the values of divergence and humidity tend to decrease (but are still anomalously positive) prior to RI. The reddish strip for $T_{\text{ocean0-100m}}$ and the almost white Shear stripe do match our understanding that RI is more likely to occur over warm ocean and under low shear, but the closeness to

Since each basin has its own climatological distribution of environments, the predictor pattern favorable for RI is also different. Colors of some variables are mildly inconsistent across basins, such as $d\theta_e/dz$, the conditional instability variable. To compare the difference images, the pattern correlations between each basin's difference images are computed via cosine similarity (Fig. 10). The lowest similarity is between the difference images for the Atlantic and northern Indian Ocean. To

FIG. 10. Pattern correlation between the global and basin difference images via cosine similarity.

compare the image differences for these two basins in more detail we plot them with the same feature ordering (Fig. 11). There are a few notable points. First, the stripes for the shear variables are bluer in the Atlantic difference image than in the northern Indian Ocean difference image. When computing the average of Shear over eight time steps, the value of Atlantic image is less than that of northern Indian Ocean image by 0.55. Shear is a less informative feature for predicting RI in the northern Indian Ocean than in the Atlantic. Second,

ocean temperature stripes in Atlantic difference image are redder than in the northern Indian Ocean difference image. For example, the average of $T_{\text{ocean}_{0-100\text{m}}}$ over eight time steps for Atlantic is 0.52 greater than that for northern Indian Ocean, indicating that in the Atlantic relatively higher ocean temperature is more favorable for RI. This may be because the gradient of ocean temperature is weaker in the Indian Ocean due to both the limited latitude expansion and the deep ocean circulations. Third, humidity variable stripes are redder in northern Indian Ocean image. The northern Indian Ocean averages of $\overline{RH_{850-700\text{mb}}}$, $\overline{RH_{850-700\text{mb}_a}}$, $\overline{RH_{500-300\text{mb}}}$, and $\overline{RH_{500-300\text{mb}_a}}$ are greater than Atlantic ones by 0.06, 0.04, 0.16, and 0.26, respectively. This shows that a moister environment is informative for predicting RI in the northern Indian Ocean. Interestingly, the difference between potential intensity and storm intensity, which is a good discriminator for RI in the Atlantic, is much less so on average in the northern Indian Ocean.

6. Summary and conclusions

A probabilistic rapid intensification (RI) prediction model using long short-term memory (LSTM) has been developed and evaluated. A single model predicts the probability of RI in five basins: Atlantic, eastern North Pacific, northern Indian Ocean, western North Pacific,

FIG. 11. Image difference ($\text{Image}_d = \text{Image}_p - \text{Image}_n$) for the Atlantic and northern Indian basins. Features (rows) are in the same order to highlight differences.

and Southern Hemisphere. The model produces predictions for intensification exceeding 25, 30, and 35 kt over the next 24 h. To deal with the relatively small sample size and the imbalance between the number of RI and non-RI cases, rolling window with overlapping and SMOTE are applied. Prediction skill is further improved by model ensembling.

The RI prediction model is trained on data from 1981 to 2009 and validated on data from 2010 to 2013. Independent storm data from 2014 to 2017 are used to evaluate model performance. The Brier skill score (BSS) is calculated globally and at the basin level for three RI thresholds. Model performance is also diagnosed via performance and reliability diagrams. Despite using monthly averaged environmental data, our model demonstrates prediction skill that is comparable to methods using higher temporal resolution data. For example, our model achieves BSS in Atlantic and eastern North Pacific for thresholds of 25 and 30 kt that are higher than those of the Statistical Hurricane Intensity Prediction Scheme (SHIPS) real-time prediction model where both models are evaluated on the same time period. For the 35-kt threshold, our model achieves almost equivalent performance to SHIPS model.

To demonstrate our model's good performance in practical applications, we examined in detail the RI predictions along the lifetime of three hurricanes (Maria, Harvey, and Patricia) and one typhoon (Meranti). Overall, our model predicted higher probabilities when RI is observed. When comparing prediction performance among those four cases, we notice that our model did a better job on Patricia, Meranti, and Maria than Harvey. A possible reason for this difference is that Harvey had a relatively unsteady intensification process, which reduces the utility of intensification rate predictors.

Finally, we examined composite feature difference images for each basin and at global scale to understand the important features in the model. The composite images show patterns that are consistent with the current understanding of RI. For example, RI is likely when storms are in low vertical wind shear environments and when their intensity is below potential maximum. The composite images differ most between the Atlantic and northern Indian Ocean basins. The difference can be summarized as: RI cases and non-RI cases in the northern Indian Ocean almost have the same shear pattern; however, in the Atlantic non-RI cases have larger shear than RI cases; RI cases in the Atlantic occur with relatively higher ocean water temperature than the northern Indian Ocean; RI cases come along with moister ocean environment in the northern Indian Ocean than in the Atlantic.

While results achieved by our model are encouraging and composite image interpretation shows our model has captured some RI patterns which are consistent with current understandings, there are still multiple aspects worth exploring. First, the data used for training are monthly averages. To know exactly if using higher resolution data can improve our model's performance, further experiments are needed. Second, SMOTE is a simple oversampling method and more complex techniques such as adversarial generative networks should be experimented with in the future. In the end, the quality of the data we can obtain in real time may differ from that of the training data. Thus, the performance of our model in real-time situations remains to be fully determined.

Acknowledgments. We thank the three anonymous reviewers for their valuable comments. This research was supported by the Columbia Initiative on Extreme Weather and Climate. Lee is supported by a Columbia Center for Climate and Life Fellowship.

Data availability statement: ERA-Interim data are available at <http://apps.ecmwf.int/datasets/data/interimfull-moda/levtype5pl/>. Best track data are available at <http://www.nhc.noaa.gov/data/#hurdat> and http://www.usno.navy.mil/NOOC/nmfc-ph/RSS/jtwc/best_tracks for HURDAT2 and JTWC dataset. SHIPS data are available at http://rammb.cira.colostate.edu/research/tropical_cyclones/ships/. SHIPS Rapid Intensification Index (SHIPS RII) is available at <http://hurricanes.ral.ucar.edu/realtime/plots/>.

REFERENCES

- Bister, M., and K. A. Emanuel, 2002: Low frequency variability of tropical cyclone potential intensity. 1. Interannual to interdecadal variability. *J. Geophys. Res.*, **107**, 4801, <https://doi.org/10.1029/2001JD000776>.
- Bolton, T., and L. Zanna, 2019: Applications of deep learning to ocean data inference and subgrid parameterization. *J. Adv. Model. Earth Syst.*, **11**, 376–399, <https://doi.org/10.1029/2018MS001472>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Brown, B. R., and G. J. Hakim, 2013: Variability and predictability of a three-dimensional hurricane in statistical equilibrium. *J. Atmos. Sci.*, **70**, 1806–1820, <https://doi.org/10.1175/JAS-D-12-0112.1>.
- Camargo, S. J., K. A. 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>.
- , M. C. Wheeler, and A. H. Sobel, 2009: Diagnosis of the MJO modulation of tropical cyclogenesis using an empirical index. *J. Atmos. Sci.*, **66**, 3061–3074, <https://doi.org/10.1175/2009JAS3101.1>.
- Chawla, N. V., 2010: Data mining for imbalanced datasets: An overview. *Data Mining and Knowledge Discovery Handbook*,

- O. Maimon and L. Rokach, Eds., Springer, 875–886, https://doi.org/10.1007/978-0-387-09823-4_45.
- , K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002: SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357, <https://doi.org/10.1613/jair.953>.
- Chu, J.-H., C. R. Sampson, A. S. Levine, and E. Fukada, 2002: The joint typhoon warning center tropical cyclone best-tracks, 1945–2000. Tech. Rep. NRL/MR/7540-02, 16 pp., <https://www.metoc.navy.mil/jtwc/products/best-tracks/tc-bt-report.html>.
- Cione, J. J., and E. W. Uhlhorn, 2003: Sea surface temperature variability in hurricanes: Implications with respect to intensity change. *Mon. Wea. Rev.*, **131**, 1783–1796, <https://doi.org/10.1175/2562.1>.
- Cortes, C., and V. Vapnik, 1995: Support-vector networks. *Mach. Learn.*, **20**, 273–297, <https://doi.org/10.1007/BF00994018>.
- De Boer, P.-T., D. P. Kroese, S. Mannor, and R. Y. Rubinstein, 2005: A tutorial on the cross-entropy method. *Ann. Oper. Res.*, **134**, 19–67, <https://doi.org/10.1007/s10479-005-5724-z>.
- 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>.
- DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, <https://doi.org/10.1175/MWR-D-14-00045.1>.
- DeMaria, M., 1996: The effect of vertical shear on tropical cyclone intensity change. *J. Atmos. Sci.*, **53**, 2076–2088, [https://doi.org/10.1175/1520-0469\(1996\)053<2076:TEOVSO>2.0.CO;2](https://doi.org/10.1175/1520-0469(1996)053<2076:TEOVSO>2.0.CO;2).
- , and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, [https://doi.org/10.1175/1520-0434\(1994\)009<0209:ASHIPS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2).
- , C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, <https://doi.org/10.1175/BAMS-D-12-00240.1>.
- Emanuel, K. A., 1988: The maximum intensity of hurricanes. *J. Atmos. Sci.*, **45**, 1143–1155, [https://doi.org/10.1175/1520-0469\(1988\)045<1143:TMIOH>2.0.CO;2](https://doi.org/10.1175/1520-0469(1988)045<1143:TMIOH>2.0.CO;2).
- , 1995: Sensitivity of tropical cyclones to surface exchange coefficients and a revised steady-state model incorporating eye dynamics. *J. Atmos. Sci.*, **52**, 3969–3976, [https://doi.org/10.1175/1520-0469\(1995\)052<3969:SOTCTS>2.0.CO;2](https://doi.org/10.1175/1520-0469(1995)052<3969:SOTCTS>2.0.CO;2).
- , and F. Zhang, 2016: On the predictability and error sources of tropical cyclone intensity forecasts. *J. Atmos. Sci.*, **73**, 3739–3747, <https://doi.org/10.1175/JAS-D-16-0100.1>.
- Frank, W. F., and L. Ritchie, 2001: Effects of vertical wind shear on the intensity and structure of numerically simulated hurricanes. *Mon. Wea. Rev.*, **129**, 2249–2269, [https://doi.org/10.1175/1520-0493\(2001\)129<2249:EOVWSO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2249:EOVWSO>2.0.CO;2).
- Geiger, T., K. Frieler, and A. Levermann, 2016: High-income does not protect against hurricane losses. *Environ. Res. Lett.*, **11**, 084012, <https://doi.org/10.1088/1748-9326/11/8/084012>.
- Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press, 800 pp.
- Hanley, D., J. Molinari, and D. Keyser, 2001: A composite study of the interactions between tropical cyclones and upper-tropospheric troughs. *Mon. Wea. Rev.*, **129**, 2570–2584, [https://doi.org/10.1175/1520-0493\(2001\)129<2570:ACSOTI>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2570:ACSOTI>2.0.CO;2).
- Hochreiter, S., and J. Schmidhuber, 1997: Long short-term memory. *Neural Comput.*, **9**, 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Hong, X., S. W. Chang, S. Raman, L. Shay, and R. Hodur, 2000: The interaction between Hurricane Opal (1995) and a warm core ring in the Gulf of Mexico. *Mon. Wea. Rev.*, **128**, 1347–1365, [https://doi.org/10.1175/1520-0493\(2000\)128<1347:TIBHOA>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<1347:TIBHOA>2.0.CO;2).
- Jain, A. K., J. Mao, and K. Mohiuddin, 1996: Artificial neural networks: A tutorial. *Computer*, **29**, 31–44, <https://doi.org/10.1109/2.485891>.
- Judt, F., and S. S. Chen, 2016: Predictability and dynamics of tropical cyclone rapid intensification deduced from high-resolution stochastic ensembles. *Mon. Wea. Rev.*, **144**, 4395–4420, <https://doi.org/10.1175/MWR-D-15-0413.1>.
- Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108, [https://doi.org/10.1175/1520-0434\(2003\)018<1093:LCORIT>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2).
- , —, and J. A. Knaff, 2010: A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **25**, 220–241, <https://doi.org/10.1175/2009WAF2222280.1>.
- , and Coauthors, 2015: Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Wea. Forecasting*, **30**, 1374–1396, <https://doi.org/10.1175/WAF-D-15-0032.1>.
- Kieper, M. E., and H. Jiang, 2012: Predicting tropical cyclone rapid intensification using the 37 GHz ring pattern identified from passive microwave measurements. *Geophys. Res. Lett.*, **39**, L13804, <https://doi.org/10.1029/2012GL052115>.
- Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, <https://doi.org/10.1175/MWR-D-12-00254.1>.
- Lee, C.-Y., M. K. Tippett, S. J. Camargo, and A. H. Sobel, 2015: Probabilistic multiple linear regression modeling for tropical cyclone intensity. *Mon. Wea. Rev.*, **143**, 933–954, <https://doi.org/10.1175/MWR-D-14-00171.1>.
- , —, A. H. Sobel, and S. J. Camargo, 2016: Rapid intensification and the bimodal distribution of tropical cyclone intensity. *Nat. Commun.*, **7**, L10625, <https://doi.org/10.1038/ncomms10625>.
- Li, Y., R. Yang, C. Yang, M. Yu, F. Hu, and Y. Jiang, 2017: Leveraging LSTM for rapid intensifications prediction of tropical cyclones. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, **4**, 101–105, <https://doi.org/10.5194/isprs-annals-IV-4-W2-101-2017>.
- Mercer, A., and A. Grimes, 2015: Diagnosing tropical cyclone rapid intensification using kernel methods and reanalysis datasets. *Procedia Comput. Sci.*, **61**, 422–427, <https://doi.org/10.1016/j.procs.2015.09.179>.
- Peduzzi, P., B. Chatenoux, H. Dao, A. D. Bono, C. Herold, J. Kossin, F. Mouton, and O. Nordbeck, 2012: Global trends in tropical cyclone risk. *Nat. Climate Change*, **2**, 289–294, <https://doi.org/10.1038/nclimate1410>.
- Price, J. F., 2009: Metrics of hurricane-ocean interaction: Vertically-integrated or vertically-averaged ocean temperature? *Ocean Sci.*, **5**, 351–368, <https://doi.org/10.5194/os-5-351-2009>.
- Rappaport, E. N., and Coauthors, 2009: Advances and challenges at the National Hurricane Center. *Wea. Forecasting*, **24**, 395–419, <https://doi.org/10.1175/2008WAF2222128.1>.
- Rasp, S., M. S. Pritchard, and P. Gentile, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, <https://doi.org/10.1073/pnas.1810286115>.
- Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, [https://doi.org/10.1175/1520-0493\(2002\)130<1653:EPFUIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2).

- Rozoff, C. M., C. S. Velden, J. Kaplan, J. P. Kossin, and A. J. Wimmers, 2015: Improvements in the probabilistic prediction of tropical cyclone rapid intensification with passive microwave observations. *Wea. Forecasting*, **30**, 1016–1038, <https://doi.org/10.1175/WAF-D-14-00109.1>.
- Sitkowski, M., and G. M. Barnes, 2009: Low-level thermodynamic, kinematic, and reflectivity fields of Hurricane Guillermo (1997) during rapid intensification. *Mon. Wea. Rev.*, **137**, 645–663, <https://doi.org/10.1175/2008MWR2531.1>.
- Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteor. Appl.*, **15**, 41–50, <https://doi.org/10.1002/met.53>.
- Tang, B., and K. Emanuel, 2012a: Sensitivity of tropical cyclone intensity to ventilation in an axisymmetric model. *J. Atmos. Sci.*, **69**, 2394–2413, <https://doi.org/10.1175/JAS-D-11-0232.1>.
- , and —, 2012b: A ventilation index for tropical cyclones. *Bull. Amer. Meteor. Soc.*, **93**, 1901–1912, <https://doi.org/10.1175/BAMS-D-11-00165.1>.
- Tieleman, T., and G. Hinton, 2012: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 31 pp., https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
- Willoughby, H. E., J. A. Clos, and M. G. Shoreibah, 1982: Concentric eye walls, secondary wind maxima, and the evolution of the hurricane vortex. *J. Atmos. Sci.*, **39**, 395–411, [https://doi.org/10.1175/1520-0469\(1982\)039<0395:CEWSWM>2.0.CO;2](https://doi.org/10.1175/1520-0469(1982)039<0395:CEWSWM>2.0.CO;2).
- Wu, C.-C., C.-Y. Lee, and I.-I. Lin, 2007: The effect of the ocean eddy on tropical cyclone intensity. *J. Atmos. Sci.*, **64**, 3562–3578, <https://doi.org/10.1175/JAS4051.1>.
- Yang, R., 2016: A systematic classification investigation of rapid intensification of Atlantic tropical cyclones with the SHIPS database. *Wea. Forecasting*, **31**, 495–513, <https://doi.org/10.1175/WAF-D-15-0029.1>.
- Yang, Y., and P. Perdikaris, 2019: Adversarial uncertainty quantification in physics-informed neural networks. *J. Comput. Phys.*, **394**, 136–152, <https://doi.org/10.1016/j.jcp.2019.05.027>.