Received: 11 September 2022 | Revised: 9 December 2022 | Accepted: 15 December 2022

DOI: 10.1002/asl.1151

RESEARCH ARTICLE

Atmospheric Science Letters RMetS

A deep learning ensemble approach for predicting tropical cyclone rapid intensification

Buo-Fu Chen[1] | **Yu-Te Kuo**[1] | **Treng-Shi Huang**[2]

[1]Center for Weather and Climate Disaster Research, National Taiwan University, Taipei, Taiwan, ROC

[2]Weather Forecast Center, Central Weather Bureau, Taipei, Taiwan, ROC

Correspondence
Buo-Fu Chen, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan, ROC.
Email: bfchen@ntu.edu.tw

Funding information
Central Weather Bureau, Taiwan, Grant/Award Number: 1102056E; Ministry of Science and Technology, Taiwan, Grant/Award Numbers: 109-2625-M-002 -021, 110-2111-M-002-016

Abstract

Predicting rapid intensification (RI) of tropical cyclones (TCs) is critical in operational forecasting. Statistical schemes rely on human-driven feature extraction and predictor correlation to predict TC intensities. Deep learning provides an opportunity to further improve the prediction if data, including satellite images of TC convection and conventional environmental predictors, can be properly integrated by deep neural networks. This study shows that deep learning yields enhanced intensity and RI prediction performance by simultaneously handling the human-defined environmental/TC-related parameters and information extracted from satellite images. From operational and practical perspectives, we use an ensemble of 20 deep-learning models with different neural network designs and input combinations to predict intensity distributions at +24 h. With the intensity distribution based on the ensemble forecast, forecasters can easily predict a deterministic intensity value demanded in operations and be aware of the chance of RI and the prediction uncertainty. Compared with the operational forecasts provided for western Pacific TCs, the results of the deep learning ensemble achieve higher RI detection probabilities and lower false-alarm rates.

KEYWORDS

deep learning, rapid intensification, statistical forecasting, tropical cyclone, tropical cyclone intensity

1 | INTRODUCTION

Tropical cyclone (TC) intensity is one of the most important parameters to predict in TC forecasting and is defined as the 10-m maximum sustained wind speed near the storm center (V_{max}). Intensity forecasting has room for improvement due to the challenge of predicting the occurrence of rapid intensification (RI, DeMaria et al., 2005, 2014; Rappaport et al., 2012), which is defined as an intensity increase surpassing a certain threshold (e.g., 30 kt·day^{-1}) motivated by the 95th percentile of intensity change in climatological data.

Previous studies have shown that complicated and nonlinear physical processes and their interactions across scales can affect TC intensification and trigger extreme RI events, including interactions between a TC and vertical wind shears (Emanuel et al., 2004; Kaplan et al., 2010; Rios-Berrios & Torn, 2017), oceanic fluxes (Bender & Ginis, 2000; Lin et al., 2013), tropospheric thermodynamic profiles (Dunion & Velden, 2004; Hill &

Lackmann, 2009; Tang & Emanuel, 2010), and background wind profiles (Chen, Davis, & Kuo, 2018, 2019, 2021; Finocchio et al., 2016; Onderlinde & Nolan, 2014). Generally, these factors control the convective asymmetries and convective bursts in the TC inner core. Therefore, statistical forecasting schemes use some of these environmental "predictors" and the information of the TC itself to predict TC intensity and RI probabilities. Given the benefit of integrating real-time information and fast calculation, forecast centers widely use statistical schemes, such as the Statistical Hurricane Intensity Prediction Scheme (SHIPS, DeMaria & Kaplan, 1994; DeMaria et al., 2005) and SHIPS RI index (Kaplan et al., 2010).

Deep learning is a candidate that is destined to enhance the power of statistical intensity/RI prediction and has been successfully applied to many geoscience topics (Ebert-Uphoff & Hilburn, 2020; Reichstein et al., 2019); it uses a great amount of data to establish nonlinear relationships for classification and regression or to conduct array transformations based on learning iterations. Convolutional neural networks (CNNs, Chen, Chen et al., 2019; Krizhevsky et al., 2012; Wimmers et al., 2019), recurrent neural networks (RNNs, Bai et al., 2020; Shi et al., 2015), and generative adversarial networks (GANs, Chen, Chen, & Chen, 2021; Chen, Davis, & Kuo, 2021; Goodfellow et al., 2014) are commonly employed. In terms of input data, the advantages of deep learning are its ability to process large data arrays (e.g., photos, sounds, and videos), the flexibility of using a large amount of multivariable data, and the capacity of the model to use any data combination to obtain the best results. Nevertheless, in many cases, humans need to select the initial pool of data and whether to organize the data in a specific way.

Deep learning shows great potential for RI prediction. Li et al. (2017) and Chandra (2017) input TC information into RNN-type models to predict the probability of RI but failed to achieve adequate performance for satisfying operational requirements due to the lack of environmental factors. Cloud et al. (2019) and Xu et al. (2021) deployed multilayer perceptron (MLP) to predict Atlantic and eastern Pacific TC intensity using environmental predictors, which are calculated based on numerical model reforecast data or from the statistical–dynamical SHIPS database. Yang et al. (2020) used long short-term memory (LSTM), an advanced RNN model, to analyze environmental factors similar to those in SHIPS. These deep-learning models exhibited comparable or better RI probability prediction skills than the operational SHIPS RI index (Kaplan et al., 2010).

Although favorable atmospheric and oceanic conditions are generally necessary for RI, perfect environmental conditions still do not guarantee an RI onset (Titley & Elsberry, 2000). An RI event is usually triggered by TC internal dynamics (e.g., Miyamoto & Nolan, 2018; Rogers et al., 2013), typically accompanied by inner-core convective bursts or axisymmetrization of convection. Thus, a successful RI prediction technique must accurately depict both environmental conditions and TC-scale features, such as the distribution of precipitation or convection. As satellite observations may capture these convective features, Bai et al. (2020) used satellite TC images collected in the past 24 h to predict RI probabilities; they proposed a convolutional LSTM model, using convolution layers to extract features from the input images and handling the evolution of spatial–temporal features with LSTM cells to obtain the final predicted RI probability.

Previous studies have demonstrated the feasibility of using either TC-related/environmental predictors (Cloud et al., 2019; Yang et al., 2020) or satellite imagery (Bai et al., 2020) for deep learning-based RI prediction. Our current work and a recent study (Griffin et al., 2022) further explore whether deep learning can effectively integrate different data types to improve prediction. Griffin et al. (2022) used both satellite images and environmental parameters to predict RI for Atlantic and eastern Pacific TCs based on CNNs; their work demonstrates that the two-dimensional features within the satellite imagery yield better 24-h indicators of RI. However, while previous techniques only predict RI probabilities, we further argue that it is practically important in operational forecasting to provide the +24-h V_{max} distribution, from which the chance of RI can be revealed, as forecasters need to predict deterministic intensity values while aware of the prediction uncertainty that is critical to disaster-related decision making. Therefore, this study proposes a cluster of deep-learning models that predict TC V_{max} distributions at +24 h. The proposed model leverages the power of deep learning to simultaneously analyze conventional predictors and satellite observations.

2 | THE DATA-DRIVEN DEEP LEARNING MODEL

2.1 | Data

To integrate conventional predictors and satellite observations for intensity prediction, we use data from two sources: the SHIPS developmental database[1] and the online-released benchmark TC imagery dataset for intensity regression (TCIR, Chen, Chen, & Lin, 2018).[2] This study collects a dataset of 1379 global TCs during 2003–17, including 3-hourly satellite infrared (IR), water vapor (WV), and passive microwave rain rate (PMW) TC images, with a horizontal resolution of 0.07° latitude/longitude.

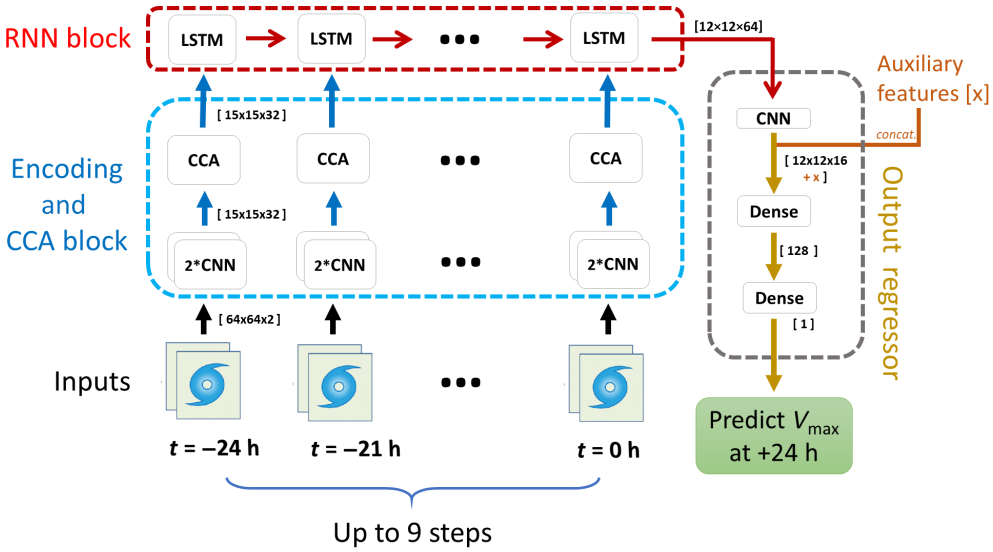

FIGURE 1 Schematic for the deep learning TC intensity prediction model. [H × W × D] indicates the height, width, and depth (channels) of the output feature maps. The auxiliary features (e.g., TC information and SHIPS parameters) are concatenated with the features in the output regressor.

Additionally, postseason-analyzed TC-related information collected/derived from the TCIR dataset is used, including TC locations, translation speeds, ocean basins, distances to the coastline, local times, and most importantly, the best-track V_{max} values, which are used as the labeled data for supervised learning. The TCIR dataset collected best-track data from the Joint Typhoon Warning Center (JTWC) and the revised Atlantic hurricane database (HURDAT2).

Regarding the SHIPS environmental parameters (DeMaria et al., 2005, 2014; DeMaria & Kaplan, 1994), this study collects eight often-used environmental parameters suggested by previous studies (e.g., Kaplan et al., 2015, table 3), including the 200-hPa divergence after vortex removal (D200), TC potential intensity (POT), 850–700-hPa relative humidity (RHLO), shear-related parameters (SHRD, SHR_x, SHR_y, and SHRG), and sea surface temperature (RSST).

The data is split into three parts: training (2003–2014), validation (2015–2016), and testing (2017) datasets. Note that the disjoint testing data allow for evaluating the model's generalizability as they are not involved in model training and fitting model weights. In addition, operational intensity forecasts for the 2017 western North Pacific TCs are collected from the JTWC, the Japan Meteorological Agency (JMA), and the Central Weather Bureau (CWB) to evaluate the model performance.

2.2 | Model design

Convolution layers are first used to autonomously extract the essential features from the satellite images (Figure 1). Subsequently, an RNN block with convolutional LSTM is deployed to handle the feature evolution during the selected period (Figure 1, top part). The final part of the model (Figure 1, output regressor), consisting of a convolution layer and two dense layers, conducts feature-to-intensity regression, predicting V_{max} at +24 h.

As previous studies (e.g., Bai et al., 2020; Chen, Chen, & Lin, 2018; Chen, Chen, & Chen, 2021) used IR and PMW images for TC intensity estimation/prediction, the control model in this study uses IR and mimic PMW (PMW*) images as the model inputs, with dimensions of 64*64*2*t centered at the TC location (Figure 1, input). The PMW* images are generated by the hybrid GAN-CNN model of Chen, Chen, and Chen (2021). Notably, the PMW* images can be retrieved from the IR and WV imagery and thus used for real-time forecasts because the collected PMW rain rate data are not a real-time product. Chen, Chen, and Chen (2021) developed this deep generative model following the concept of Olander and Velden (2009) that a PMW image can be emulated by plotting the difference between IR and WV images with a specifically designed color scale. Moreover, the environmental/TC-related predictors are used as auxiliary features by concatenating them onto the flattened features before passing the features through the dense layers.

The input images are first passed through two convolution layers in the encoding block (Figure 1, blue dashed box). As the convolution operator is executed on neighboring pixels, convolution layers are good at handling data with local spatial dependencies and extracting higher-level features for analyzing intensities. Additionally, batch normalization (Ioffe & Szegedy, 2015) is applied with these layers.

The second part of the encoding block is the cross-channel attention (CCA) module proposed by Bai et al. (2020). The CCA module was inspired by the Dvorak-type schemes for intensity estimation (Olander et al., 2021; Olander & Velden, 2019), in which scene-type analysis is

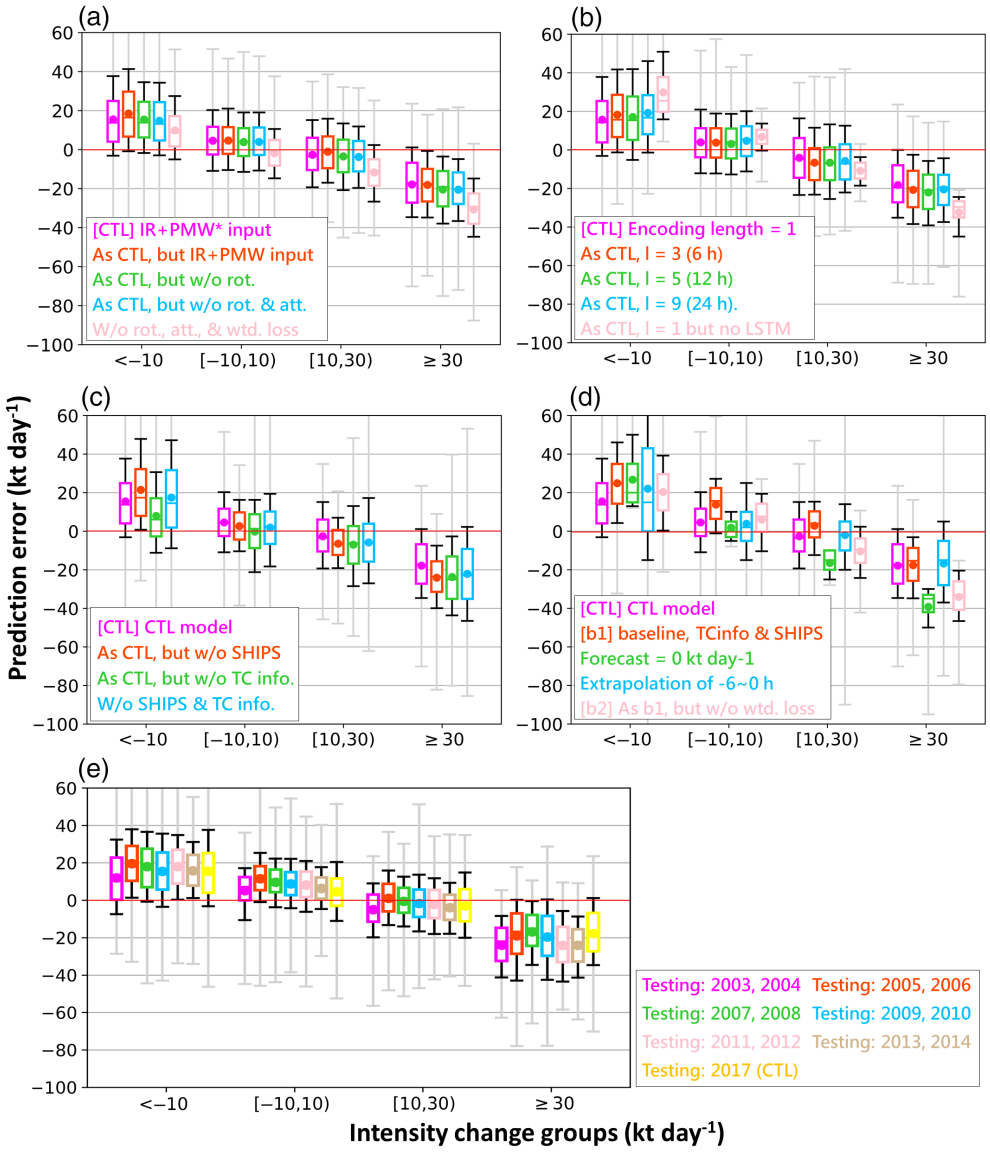

FIGURE 2 Error distributions of the predicted +24-h intensities (Y-axis) for various intensification groups (X-axis) yielded by (a) the control model (CTL, purple) and the models in the ablation tests concerning the weighted loss, CCA, and input rotation components, (b) models with different input durations, (c) the models in the ablation tests concerning the TC information and SHIPS parameters, and (d) the control model (purple) and the baseline schemes not using satellite images. The box and whisker plot indicates the mean (dot), median (middle bar), 75th and 25th percentiles (box), 90th and 10th percentiles (black whiskers), and maxima and minima (gray whiskers). (e) A similar figure displaying the error distributions of the +24-h V_{max} predicted by the seven models in the leave-2-year-out cross-validation experiment

performed to highlight critical cloud features (e.g., eyes, central dense overcast patterns, and shear-induced asymmetry) related to TC intensity. Specifically, the CCA module calculates a two-dimensional importance weighting mask and then applies it via the Hadamard product to multiple feature maps obtained from the previous encoding convolution layer. The importance mask ranging from 0 to 1 is implemented as a two-layer CNN with a sigmoid function for normalization. Therefore, the CCA module facilitates an overall importance evaluation concerning where in the hidden map the model should focus on with more attention.

The convolutional LSTM (ConvLSTM, Shi et al., 2015) cells are applied to analyze the evolved features extracted by the encoding block. A ConvLSTM algorithm is an LSTM module coupled with CNNs, replacing all matrix multiplications with convolution operations to model the spatial and temporal aspects of data. This study tests different input lengths (Figure 2b) from no sequence to a nine-time-spot sequence (24 h).

After the RNN block, features of size [12 × 12 × 64] are passed into the final output regressor (Figure 1, gray dashed box). A convolution layer transforms these features to reduce their dimensions, and environmental/TC-related predictors are concatenated as auxiliary features. Finally, two fully connected dense layers regress the features to the final prediction, V_{max} at +24 h, with dropout adopted to mitigate possible overfitting.

3 | INTEGRATING HUMAN-DEFINED PREDICTORS AND DEEP LEARNING-EXTRACTED SATELLITE FEATURES

This section first describes the control model of this study and presents four experimental comparisons (Section 3.2) to explore the effects on the prediction performance of

deploying various model components and using different data combinations.

3.1 | The control model

The control model uses IR and PMW* as the satellite imagery inputs; it also uses TC-related parameters and SHIPS environmental parameters at $t = 0$ and $t = +24$ h as axillary features. The six TC-related parameters are the TC location, translation speed, ocean basin code, distance to the coastline, local time, and −6–0-h V_{max} change (not used at $t = +24$ h), and the SHIPS parameters are the D200, POT, RHLO, SHRD, SHR_x, SHR_y, SHRG, and RSST. The environmental and TC-related parameters of +24 h are used because current numerical weather models can well grasp the synoptic-scale environment of the next day.

The control model deploys the two CNN encoding layers, the CCA block, and the LSTM layer with input lengths of 1. Although we follow previous studies (Bai et al., 2020; Yang et al., 2020) using LSTM to learn the feature evolution, it is worth noting that, based on additional experiments (Figure 2b, discussed later), the control model ends up using only the current satellite image to achieve the best result.

As the difficulty of predicting RI is partly due to its rarity, a weighted mean squared error (MSE) loss is applied to force the model to give higher weights to intensification events:

$$w = tanh\left(\frac{\delta V_{max} - 20}{10}\right) \times 1000 + 1000.1 \quad (1)$$

$$\text{Weighted MSE} = \frac{1}{N}\sum_{i=1}^{N} w \times (X_i - Y_i)^2 \quad (2)$$

where δV_{max} is the intensification rate, N is the sample number of the training data (or the batch size), X is the model-predicted V_{max} at +24 h, and Y is the target labeled data.

Last, favorable convective features for intensification can be easily revealed in shear-relative coordinates (Chen, Chen, & Chen, 2021; Chen, Davis, & Kuo, 2021; Rogers et al., 2013; Stevenson et al., 2018). Thus, we rotate the input images to align the shear vectors, expecting that the rotation aids the learning result due to better spatial feature extraction.

3.2 | Testing the model components

This subsection explores the effects on the prediction performance of various model configurations and suggests that the control setting generally performs better. Figure 2a–d present four experimental comparisons to evaluate the performance of deep-learning models, including (i) tests of image preprocessing/analyzing components and loss function design, (ii) exploring the input duration and the benefit of LSTM, (iii) incorporating conventional human-selected predictors, and (iv) comparison with baseline schemes.

To compare the models, the error distributions of the predicted +24-h intensities for four intensity change (ΔV_{max}) categories {<−10, [−10,10), [10,30), ≥30} are examined based on the 2017 global TCs from the testing dataset. As the sample ratios for the four categories are 18.8%, 48.0%, 25.3%, and 7.9%, we discuss the models' performance separately for various categories to highlight the RI and weakening TC prediction performance.

First, ablation tests (Figure 2a) are performed for (i) replacing the PMW* input with PMW, (ii) excluding the image rotation regarding the shear, (iii) excluding the CCA module, and (iv) replacing the weighted MSE loss with the simple MSE. The results show that the control model (Figure 2a, purple) using IR + PMW* as input has comparable performance to that of the model using IR + PMW as input (Figure 2a, orange), except for a smaller bias for weakening TCs ($\Delta V_{max} < -10$ kt). Furthermore, if the image rotation operation is not applied (Figure 2a, green) or the CCA block is further excluded (Figure 2a, aqua), the model performance decreases for the [10,30) and ≥30 ΔV_{max} categories, suggesting that these designs help the model better handle the given satellite images.

Comparing the control model with the model using the simple MSE loss (Figure 2a, pink), the weighted loss leads to better prediction performance for intensifying TCs due to the reduced negative biases. A 30% root-mean-square error (RMSE) improvement is achieved for the ≥30 ΔV_{max} category (from 33.7 to 23.5 kt·day^{-1}), and the error distribution of the [10,30) category becomes bias-free. However, applying the weighted MSE and giving special weights to extreme events is, in essence, building a statistical model that overgeneralizes to the extremes of a random distribution. Thus, the control model has a higher positive bias for weakening TCs. Nevertheless, as predicting TC intensification is more important in disaster prevention, we promote the model using the weighted loss here because of its substantial improvement in predicting intensification and its acceptable TC weakening prediction performance reduction (Figure 2a, purple vs. pink).

Models with various input durations are tested (Figure 2b). Comparing the model with nine input frames (Figure 2b, aqua) and the control model that only analyzes the current images (Figure 2b, purple), the control model has better RMSEs for all ΔV_{max} categories and

the smallest negative bias for TC intensification. Furthermore, an additional experiment that omits the ConvLSTM cell from the control model (Figure 2b, pink) yields large negative (positive) biases for intensifying (weakening) TCs, suggesting that even though the control model only handles the images at $t = 0$, the convolution operations within the ConvLSTM cell help transform the feature maps to relatively effective predictors for intensity forecasting.

Figure 2c demonstrates that incorporating conventional human-selected predictors (SHIPS and TC parameters) with satellite images improves the prediction results. Compared with the model excluding auxiliary features (Figure 2c, blue), the control model (Figure 2c, purple) has smaller biases for both the weakening and RI categories.

Compared with some baseline schemes not using satellite images, the control model (Figure 2d, purple) has a much better performance predicting TC weakening than the persistence method of the −6–0-h V_{max} change (Figure 2d, blue). Moreover, we also evaluate the performance of two baseline models[3] using only the TC and SHIPS parameters. Note that the first baseline model (Figure 2d, orange, the b1 model) uses the weighted MSE loss, and the second model uses the simple MSE loss (Figure 2d, pink, the b2 model). The control model performs better than baseline model b1 for weakening TCs but has comparable performance for intensifying TCs, whereas it has better performance than baseline model b2 for all ΔV_{max} categories. Therefore, it is suggested that the information from satellite images helps the control model fight the side effect caused by the weighted loss, leading to a better overall performance than that of baseline models.

3.3 | Leave-2-year-out cross-validation

Although this study primarily uses the data from 2017 for evaluating the model performance, an additional leave-2-year-out cross-validation is performed to ensure the model generality (Figure 2e). Six models are trained using the same validation data (2015 and 2016) but evaluated on different 2-year testing data. The data from the rest of the years are used as the training data.

TABLE 1 Settings of the 20 models used for constructing the ensemble forecasts

	TC and SHIPS parameters; weighted MSE loss; encoding length = 1						
	Attention	**Rotation**	**IR**	**WV**	**PMW***	**VIS***	**Deeper CNN**
ENS_01	X	X	X	X	O	X	X
ENS_02	X	X	X	X	O	O	X
ENS_03	X	X	O	X	O	X	X
ENS_04	O	X	X	X	O	X	X
ENS_05	O	X	X	X	O	O	X
ENS_06	O	X	O	X	O	X	X
ENS_07	X	O	X	X	O	X	X
ENS_08	X	O	X	X	O	O	X
ENS_09	X	O	O	X	O	X	X
ENS_10	O	O	X	X	O	X	X
ENS_11	O	O	X	X	O	O	X
ENS_12	O	O	O	X	O	X	X
ENS_13	X	X	X	X	O	X	O
ENS_14	X	X	X	X	O	O	O
ENS_15	X	X	O	X	O	X	O
ENS_16	X	O	X	X	O	X	O
ENS_17	X	O	X	X	O	O	O
ENS_18	X	O	O	X	O	X	O
ENS_19	X	X	X	O	O	X	X
ENS_20	X	X	O	O	O	O	X

Note: "Attention" indicates that the CCA module is deployed; "rotation" indicates that input image rotation is used; "IR, WV, PMW*, and VIS*" are the employed input satellite channels. Members 13–18 use a deeper CNN for feature extraction. ENS_12 is the control model described in Section 3.1.

These models have biases for the <−10 ΔV_{max} category ranging from ∼11.5 to ∼19.5 kt·day^{-1}, which are smaller than those of the two baseline models (∼26 and ∼ 20 kt·day^{-1}). Additionally, the biases of these models for the ≥30 ΔV_{max} category range from approximately −18 to −22 kt·day^{-1}, these results are better than that of model b2 (∼34 kt·day^{-1}) and comparable to that of model b1 (∼18 kt·day^{-1}). Moreover, the averaged RMSE of these seven models is 15.81 kt·day^{-1}, with a standard deviation of 0.91 kt·day^{-1}. The models in the cross-validation experiments generally outperform the baseline models, b1 and b2, which have overall RMSEs of 19.0 and 17.3 kt·day^{-1}, respectively. It is thus suggested that the proposed control model setting does not overfit the data, has consistent performance across different years, and could be stable in future applications.

4 | ENSEMBLE APPROACH AND VERIFICATION

4.1 | The ensemble approach

We propose using a cluster of 20 deep-learning models to predict the V_{max} distribution at +24 h. Table 1 shows

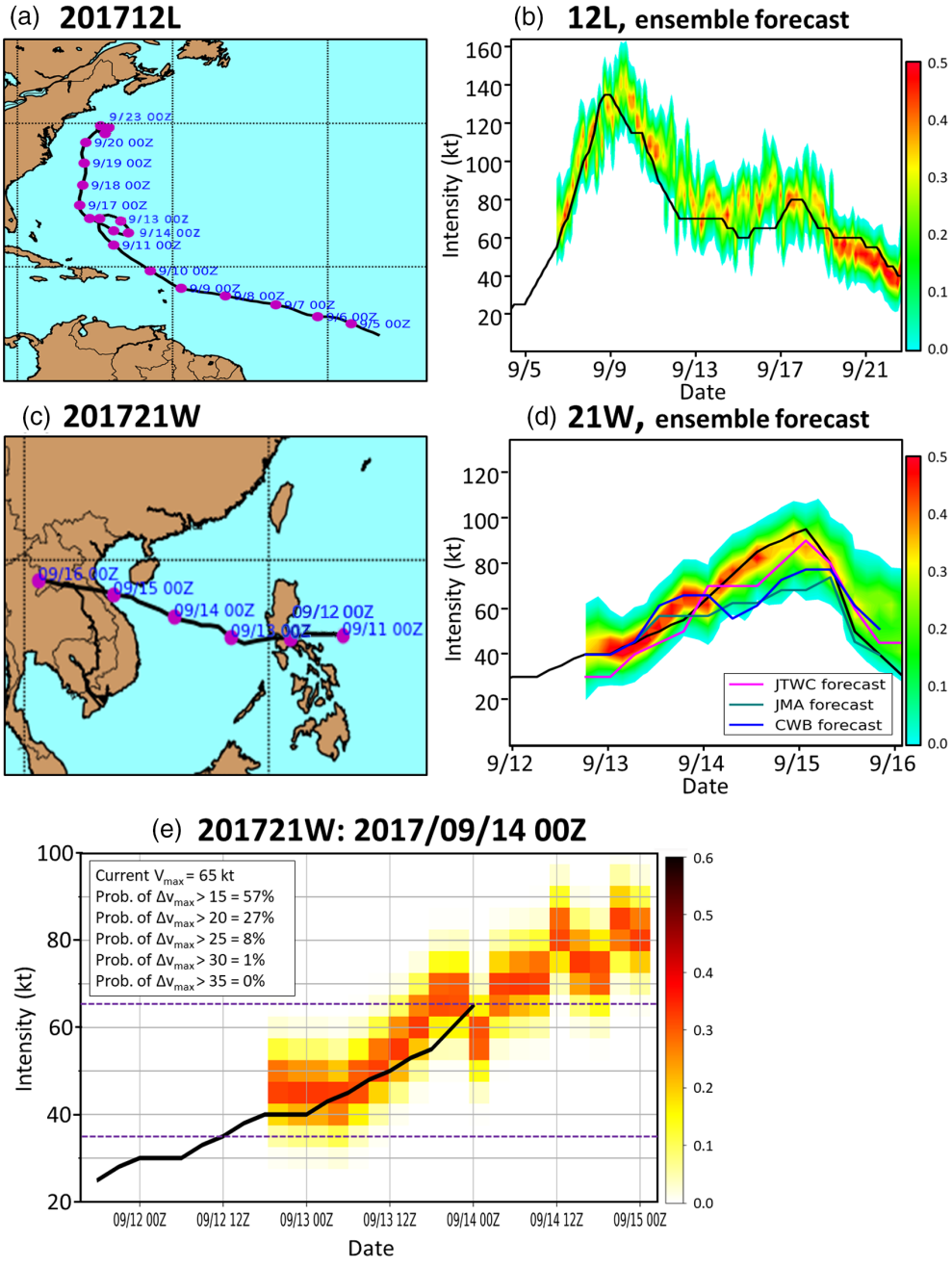

FIGURE 3 (a) The track of TC 201712 L. (b) The predicted V_{max} distribution (shading, {summation = 1}) at +24 h made by the deep learning ensemble. The black line indicates the best-track intensity. (d), (e) Same as (a) and (b) except for TC 201721 W with the operational forecasts from the JTWC, JMA, and CWB. (e) A demo product of the ensemble intensity prediction produced for TC 201721 W at 00:00 UTC 14 September 2017; it shows the predicted +24-h V_{max} probabilities (color bar) since 1800 UTC 12 to 0000 UTC 15 September. Note that the plot also shows estimated intensities (black line), which are the initial intensities of the predictions, from 1800 UTC 11 to 0000 UTC 14 September. The RI probabilities for various thresholds are also shown in the upper-left inset.

these 20 models, which have different neural network designs and handle various input combinations.

While the control model uses IR and PMW* images, the ensemble models use various input combinations of IR, WV, PMW*, and mimic visible (VIS*) images. Similar to producing PMW* images, the VIS* images are generated by the hybrid GAN-CNN model of Chen, Davis, and Kuo (2021). Furthermore, a deeper input encoding block is used by members 13–18 for feature extraction. This block has a similar layer structure to that in Wimmers et al. (2019), in which a CNN was used to analyze TC microwave imagery and estimate the intensity. Although these ensemble members do not significantly outperform the models with the two-layer encoder shown in Figure 1, it is expected that the deeper encoder extracts some different features to make its predictions and thus enhance the overall prediction stability of the ensemble.

The continuous ranked probability score (CRPS, Zamo & Naveau, 2018) is the quadratic difference between the forecasted cumulative distribution function [CDF, $F(y)$] and the empirical CDF of the scalar observation [$\mathbb{1}(x \geq y)$]:

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}(x \geq y))^2 dy.$$

We calculate the averaged CRPSs of the V_{max} distribution of all samples for the ensemble prediction, that is, 9.25 kt^2, which is smaller than those for both the control model (12.59 kt^2) and baseline model b1 (16.14 kt^2), suggesting that the ensemble prediction has more predictive power than models only providing deterministic forecasts.

Examples of this new approach are shown. For hurricane 201,712 L, the V_{max} distribution predicted by the ensemble model (Figure 3b, shading) is fairly aligned with the best-track during the intensification period (7–9 September), but an overpredicted V_{max} appears during the steady-state period (13–17 September). Additionally, for Typhoon 201,721 W (Figure 3d), the deep learning ensemble successfully predicts the intensity evolution, while all forecast centers missed the near-shore RI on 14 September. Notably, the ensemble captures the RI's magnitude, maxima, and timing.

On the other hand, some possible reasons for bad performances are discussed. We notice that the model sometimes misses predicting intensification for some smaller western North Pacific TCs, of which the intensification may be affected more by TC internal dynamics than by TC-environment interaction. In the worst cases, the proposed model predicts the same intensity as the current V_{max} (i.e., $\Delta V_{max} = 0$). Second, for some extreme RI and weakening cases ($\Delta V_{max} > 50$ or < 50 kt·day^{-1}), the model lacks the ability to predict these extremely rare events. Lastly, TC center positioning is critical to the model performance; the displacement of the TC center usually leads to erroneously lower (or negative) predicted intensification rates.

Figure 3e is a demo product of the ensemble intensity prediction for Typhoon 201,721 W, and the shading color shows the V_{max} probabilities, which come from the group of 20 deterministic predictions. Although the current approach can reveal the prediction uncertainty, future studies should further test various member combinations or member generation methods to explore and improve the relationship between ensemble forecast spread and skill.

4.2 | Evaluating the model performance

This subsection presents a verification[4] against forecasts from the JTWC, JMA, and CWB for the 2017 western

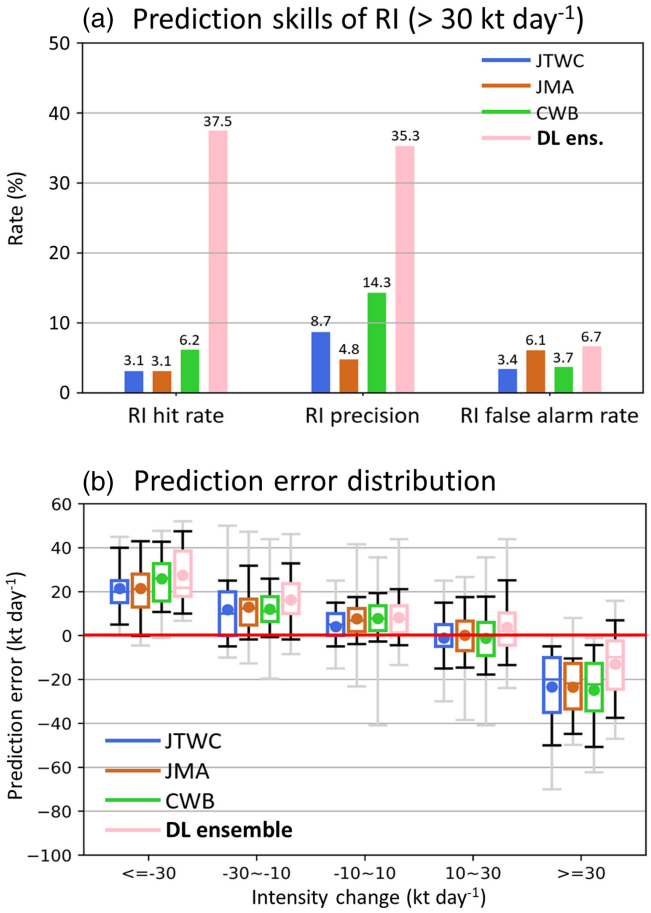

FIGURE 4 Verification of the deep learning ensemble mean against the operational forecasts from the JTWC, JMA, and CWB based on the testing dataset: (a) hit rates, precision, and false-alarm rates for RI cases over 30 kt·day^{-1}; (b) prediction errors of the 24-h intensity differences for various intensification groups (similar to Figure 2)

North Pacific TCs. For RI prediction (Figure 4a), the deep learning ensemble mean has a much higher hit rate ($=TP/[TP+FN]$, 37.5%) and precision ($=TP/[TP+FP]$, 35.3%) for the >30-kt·day^{-1} RI threshold, demonstrating an enhanced capability to detect RI. Moreover, the proposed ensemble approach keeps the false-alarm rate ($=FP/[FP+TN]$) under 10%, which is comparable to that of the operational forecasts. Figure 4b shows the prediction error distributions for various intensification groups. The mean error of the ensemble (Figure 4b, pink dot) is approximately 50% of that of the other forecasts for RI cases, but its ability to predict TC weakening is slightly decreased with slower weakening rates. Of note, the seven deterministic deep-learning models in the leave-2-year-out cross-validation (Section 3c) have averaged precision and false-alarm rates of 30.9% and 3.4%, respectively, with standard deviations of 8.4% and 1.7%. They are better than the subjective operational forecasting from weather centers. Along with this paper, Table S1 discusses other scores measuring the prediction performance, such as the Brier skill score and Heidke skill score, against operational forecasts.

5 | CONCLUSIONS

Conventional RI forecasting utilizes high-level statistical predictors that summarize atmospheric data. The high-level physical meaning carried by predictors allows humans to comprehend and design correlations accordingly. However, the loss of detail in the predictors may be the bottleneck for improving prediction.

This study explores the feasibility of deep learning for intensity and RI prediction by simultaneously utilizing conventional predictors and autoextracted features from satellite observations. Domain knowledge-inspired treatments (i.e., the weighted loss, the CCA module, and rotation for aligning the shear vectors) help the model better learn from the data. Moreover, a novel deep learning ensemble approach is proposed for TC intensity prediction, as the best single model is perhaps not the best solution for operational TC intensity forecasting, given that the physical processes that trigger RI are too nonlinear. This ensemble approach involves different neural network designs and input data combinations. The 20 ensemble members separately predict V_{max} at +24 h and format the V_{max} distribution. It is suggested that various ensemble members presumably learn various RI omens; thus, the V_{max} distribution is more useful than a single deep learning model.

Compared with the intensity forecasts provided for the 2017 western Pacific TCs, the V_{max} forecasts of the proposed deep learning ensemble achieve a higher RI detection probability and keep the false-alarm rate acceptable. Moreover, the ensemble approach helps forecasters easily predict the deterministic intensities demanded in operation, while aware of the chance of RI and prediction uncertainty. Our ongoing work involves making this new technique operational in the CWB. With forecasters' feedback and real-time operational verifications/comparisons, the forecasting guidance for the deep learning techniques will be completed, leading to practical improvements in operational forecasting.

AUTHOR CONTRIBUTIONS

Buo-Fu Chen: Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; supervision; visualization; writing – original draft; writing – review and editing. **Yu-Te Kuo:** Conceptualization; data curation; formal analysis; investigation; methodology; validation; visualization. **Treng-Shi Huang:** Conceptualization; funding acquisition; resources; supervision.

ACKNOWLEDGEMENTS

The authors appreciate the feedback from CWB forecasters. Computing resources for this study were mainly provided by the Center for Weather Climate and Disaster Research, National Taiwan University. This project was funded by Grant MOST 109-2625-M-002-021 and MOST 111-2111-M-002-016 of the Ministry of Science and Technology, Taiwan, and Project 1102056 E of the Central Weather Bureau, Taiwan.

DATA AVAILABILITY STATEMENT

The TC satellite images and TC records used in this study can be downloaded from Tropical Cyclone for Image-to-intensity Regression Dataset website: https://www.csie.ntu.edu.tw/~htlin/program/TCIR/ and https://www.csie.ntu.edu.tw/~htlin/program/TCRISI/. The SHIPS developmental database is available at https://rammb2.cira.colostate.edu/research/tropical-cyclones/ships/. Samples of the compiled dataset for model training and validation and the code can be found at https://github.com/kuoyute/TCRI?fbclid=IwAR0P6UcnZcVr3y3KSUIaOUA8s7dpgFq3VZgK19GDkdquXDrlJuo6FDBaT-A.

ORCID

Buo-Fu Chen https://orcid.org/0000-0002-6722-7731

ENDNOTES

[1] SHIPS developmental dataset: https://rammb2.cira.colostate.edu/research/tropical-cyclones/ships/

[2] Tropical Cyclone for Image-to-intensity Regression dataset: https://www.csie.ntu.edu.tw/~htlin/program/TCIR/ (Chen, Chen, &

Lin, 2018) and https://www.csie.ntu.edu.tw/~htlin/program/TCRISI/ (Bai et al., 2020)

[3] This model is identical to the output regressor, as shown in Figure 1, except for that the output of the convolution layer is replaced with 0.

[4] The JTWC best-track V_{max} is used as the ground truth, and the 10-min-averaged V_{max} values from the JMA and CWB are converted to 1-min-averaged V_{max} values with multiples of 0.88^{-1} (Harper et al., 2008).

REFERENCES

Bai, C.-Y., Chen, B.-F. & Lin, H.-T. (2020) Benchmarking tropical cyclone rapid intensification with satellite images and attention-based deep models. In: *Joint European conference on machine learning and knowledge discovery in databases*. Cham: Springer, pp. 497–512.

Bender, M.A. & Ginis, I. (2000) Real-case simulations of hurricane-ocean interaction using a high-resolution coupled model: effects on hurricane intensity. *Monthly Weather Review*, 128, 917–946.

Chandra, R. (2017) Towards prediction of rapid intensification in tropical cyclones with recurrent neural networks. *arXiv*, arXiv: 1701.04518.

Chen, B., Chen, B.-F. & Chen, Y.-N. (2021) Real-time tropical cyclone intensity estimation by handling temporally heterogeneous satellite data. In *Proceedings of the AAAI conference on artificial intelligence*, 35(17), 14721–14728.

Chen, B., Chen, B.-F. & Lin, H.-T. (2018) Rotation-blended CNNs on a new open dataset for tropical cyclone image-to-intensity regression. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (KDD'18)*. London: Association for Computing Machinery, p. 10. Available from: https://doi.org/10.1145/3219819.3219926

Chen, B.-F., Chen, B., Lin, H. & Elsberry, R.L. (2019) Estimating tropical cyclone intensity by satellite imagery utilizing convolutional neural networks. *Weather and Forecasting*, 34, 447–465. Available from: https://doi.org/10.1175/WAF-D-18-0136.1

Chen, B.-F., Davis, C.A. & Kuo, Y.-H. (2018) Effects of low-level flow orientation and vertical shear on the structure and intensity of tropical cyclones. *Monthly Weather Review*, 146, 2447–2467. Available from: https://doi.org/10.1175/MWR-D-17-0379.1

Chen, B.-F., Davis, C.A. & Kuo, Y.-H. (2019) An idealized numerical study of shear-relative low-level mean flow on tropical cyclone intensity and size. *Journal of the Atmospheric Sciences*, 76, 2309–2334. Available from: https://doi.org/10.1175/JAS-D-18-0315.1

Chen, B.-F., Davis, C.A. & Kuo, Y.-H. (2021) Examination of the combined effect of deep-layer vertical shear direction and lower-tropospheric mean flow on tropical cyclone intensity and size based on the ERA5 reanalysis. *Monthly Weather Review*, 149(12), 4057–4076.

Cloud, K.A., Reich, B.J., Rozoff, C.M., Alessandrini, S., Lewis, W. E. & Delle Monache, L. (2019) A feed forward neural network based on model output statistics for short-term hurricane intensity prediction. *Weather and Forecasting*, 34(4), 985–997.

DeMaria, M. & Kaplan, J. (1994) A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Weather and Forecasting*, 9, 209–220.

DeMaria, M., Mainelli, M., Shay, L.K., Knaff, J.A. & Kaplan, J. (2005) Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). *Weather and Forecasting*, 20, 531–543.

DeMaria, M., Sampson, C.R., Knaff, J.A. & Musgrave, K.D. (2014) Is tropical cyclone intensity guidance improving? *Bulletin of the American Meteorological Society*, 95, 387–398.

Dunion, J.P. & Velden, C.S. (2004) The impact of the Saharan air layer on Atlantic tropical cyclone activity. *Bulletin of the American Meteorological Society*, 85, 353–365.

Ebert-Uphoff, I. & Hilburn, K. (2020) Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, 101(12), E2149–E2170.

Emanuel, K.A., DesAutels, C., Holloway, C. & Korty, R. (2004) Environmental control of tropical cyclone intensity. *Journal of the Atmospheric Sciences*, 61, 843–858.

Finocchio, P.M., Majumdar, S.J., Nolan, D.S. & Iskandarani, M. (2016) Idealized tropical cyclone responses to the height and depth of environmental vertical wind shear. *Monthly Weather Review*, 144, 2155–2175.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. et al. (2014) Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.

Griffin, S.M., Wimmers, A. & Velden, C.S. (2022) Predicting rapid intensification in North Atlantic and eastern North Pacific tropical cyclones using a convolutional neural network. *Weather and Forecasting*, 37(8), 1333–1355.

Harper, B.A., Stroud, S.A., McCormack, M. & West, S. (2008) A review of historical tropical cyclone intensity in North-Western Australia and implications for climate change trend analysis. *Australian Meteorological Magazine*, 57, 121–141.

Hill, K.A. & Lackmann, G.M. (2009) Influence of environmental humidity on tropical cyclone size. *Monthly Weather Review*, 137, 3294–3315.

Ioffe, S. & Szegedy, C. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. *In International conference on machine learning*, 448–456.

Kaplan, J., DeMaria, M. & Knaff, J.A. (2010) A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Weather and Forecasting*, 25, 220–241. Available from: https://doi.org/10.1175/2009WAF2222280.1

Kaplan, J., Rozoff, C.M., DeMaria, M., Sampson, C.R., Kossin, J.P., Velden, C.S. et al. (2015) Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Weather and Forecasting*, 30(5), 1374–1396.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012) Imagenet classification with deep convolutional neural networks. *Communications of the ACM 60*, 6, 84–90.

Li, Y., Yang, R., Yang, C., Yu, M., Hu, F. & Jiang, Y. (2017) Leveraging LSTM for rapid intensifications prediction of tropical cyclones. *ISPRS annals of photogrammetry, remote sensing & spatial information sciences*, 4.

Lin, I.-I., Goni, G.J., Knaff, J.A., Forbes, C. & Ali, M.M. (2013) Ocean heat content for tropical cyclone intensity forecasting and its impact on storm surge. *Natural Hazards*, 66, 1481–1500.

Miyamoto, Y. & Nolan, D.S. (2018) Structural changes preceding rapid intensification in tropical cyclones as shown in a large ensemble of idealized simulations. *Journal of the Atmospheric Sciences*, 75(2), 555–569.

Olander, T.L. & Velden, C.S. (2009) Tropical cyclone convection and intensity analysis using differenced infrared and water vapor imagery. *Weather and Forecasting*, 24(6), 1558–1572.

Olander, T.L. & Velden, C.S. (2019) The advanced Dvorak technique (ADT) for estimating tropical cyclone intensity: update and new capabilities. *Weather and Forecasting*, 34(4), 905–922.

Olander, T., Wimmers, A., Velden, C. & Kossin, J.P. (2021) Investigation of machine learning using satellite-based advanced Dvorak technique analysis parameters to estimate tropical cyclone intensity. *Weather and Forecasting*, 36(6), 2161–2186.

Onderlinde, M.J. & Nolan, D.S. (2014) Environmental helicity and its effects on development and intensification of tropical cyclones. *Journal of the Atmospheric Sciences*, 71, 4308–4320.

Rappaport, E.N., Jiing, J.-G., Landsea, C.W., Murillo, S.T. & Franklin, J.L. (2012) The joint hurricane test bed: its first decade of tropical cyclone research-to-operations activities reviewed. *Bulletin of the American Meteorological Society*, 93(3), 371–380.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. & Carvalhais, N. (2019) Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743), 195–204.

Rios-Berrios, R. & Torn, R. (2017) Climatological analysis of tropical cyclone intensity changes under moderate vertical wind shear. *Monthly Weather Review*, 145, 1717–1738. Available from: https://doi.org/10.1175/MWR-D-16-0350.1

Rogers, R., Reasor, P. & Lorsolo, S. (2013) Airborne Doppler observations of the inner-core structural differences between intensifying and steady-state tropical cyclones. *Monthly Weather Review*, 141(9), 2970–2991.

Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K. & Woo, W.C. (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, Vol. 1, pp. 802–810.

Stevenson, S.N., Corbosiero, K.L., DeMaria, M. & Vigh, J.L. (2018) A 10-year survey of tropical cyclone inner-core lightning bursts and their relationship to intensity change. *Weather and Forecasting*, 33(1), 23–36.

Tang, B. & Emanuel, K. (2010) Midlevel ventilation's constraint on tropical cyclone intensity. *Journal of the Atmospheric Sciences*, 67, 1817–1830.

Titley, D.W. & Elsberry, R.L. (2000) Large intensity changes in tropical cyclones: a case study of Supertyphoon Flo during TCM-90. *Monthly Weather Review*, 128(10), 3556–3573.

Wimmers, A., Velden, C. & Cossuth, J.H. (2019) Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review*, 147(6), 2261–2282.

Xu, W., Balaguru, K., August, A., Lalo, N., Hodas, N., DeMaria, M. et al. (2021) Deep learning experiments for tropical cyclone intensity forecasts. *Weather and Forecasting*, 36(4), 1453–1470.

Yang, Q., Lee, C.Y. & Tippett, M.K. (2020) A long short-term memory model for global rapid intensification prediction. *Weather and Forecasting*, 35(4), 1203–1220.

Zamo, M. & Naveau, P. (2018) Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2), 209–234.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chen, B.-F., Kuo, Y.-T., & Huang, T.-S. (2023). A deep learning ensemble approach for predicting tropical cyclone rapid intensification. *Atmospheric Science Letters*, *24*(5), e1151. https://doi.org/10.1002/asl.1151