Estimating Tropical Cyclone Intensity by Satellite Imagery Utilizing Convolutional Neural Networks

BUO-FU CHEN

National Center for Atmospheric Research, Boulder, Colorado

BOYO CHEN AND HSUAN-TIEN LIN

Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

RUSSELL L. ELSBERRY

Department of Meteorology, Naval Postgraduate School, Monterey, California, and Trauma, Health, and Hazards Center, University of Colorado–Colorado Springs, Colorado Springs, Colorado

(Manuscript received 14 August 2018, in final form 17 February 2019)

ABSTRACT

Accurately estimating tropical cyclone (TC) intensity is one of the most critical steps in TC forecasting and disaster warning/management. For over 40 years, the Dvorak technique (and several improved versions) has been applied for estimating TC intensity by forecasters worldwide. However, the operational Dvorak techniques primarily used in various agencies have several deficiencies, such as inherent subjectivity leading to inconsistent intensity estimates within various basins. This collaborative study between meteorologists and data scientists has developed a deep-learning model using satellite imagery to estimate TC intensity. The conventional convolutional neural network (CNN), which is a mature technology for object classification, requires several modifications when being used for directly estimating TC intensity (a regression task). Compared to the Dvorak technique, the CNN model proposed here is objective and consistent among various basins; it has been trained with satellite infrared brightness temperature and microwave rain-rate data from 1097 global TCs during 2003–14 and optimized with data from 188 TCs during 2015–16. This paper also introduces an upgraded version that further improves the accuracy by using additional TC information (i.e., basin, day of year, local time, longitude, and latitude) and applying a postsmoothing procedure. An independent testing dataset of 94 global TCs during 2017 has been used to evaluate the model performance. A root-mean-square intensity difference of 8.39 kt ($1\,\text{kt} \approx 0.51\,\text{m s}^{-1}$) is achieved relative to the best track intensities. For a subset of 482 samples analyzed with reconnaissance observations, a root-mean-square intensity difference of 8.79 kt is achieved.

1. Introduction

Tropical cyclone (TC) intensity, which is defined as the maximum sustained surface wind near the TC center, is an important parameter that needs to be accurately estimated in TC forecasting and disaster warning/management. For instance, Zhai and Jiang (2014) proposed that normalized hurricane economic losses approximately follow a power-law relation with TC intensity, in which the exponent generally ranges between 4 and 12. In addition, an accurate estimation of the intensity could lead to a better initialization of the numerical models, and thus lead

to better forecasts. More accurate intensity estimation may also contribute to a forecast of rapid intensification, which is one of the most challenging forecast issues and remains the highest operational forecasting priority at the National Hurricane Center (Rappaport et al. 2012; DeMaria et al. 2014).

A major problem for intensity estimation and forecasts has been the lack of in situ observations, which are difficult because TCs spend most of their lifetime over the open ocean, where only a few surface observations are available on small islands and from buoys. Although reconnaissance, research aircraft with radar, dropsondes, and other instruments provide high-quality observations, such aircraft missions are expensive and are

Corresponding author: Buo-Fu Chen, bfchen777@gmail.com

only available in the Atlantic and eastern North Pacific when a hurricane is near the U.S. mainland. Therefore, satellite remote sensing observations are the primary source of TC information, due to their global coverage and high temporal frequency. Although satellite remote sensing does not provide a direct measurement of the wind near the surface, satellite images of cloud, water vapor, and precipitation serve as proxies for estimating TC intensity (e.g., Cecil and Zipser 1999; Velden et al. 2006; Olander and Velden 2009; Ritchie et al. 2012).

For over 40 years, the Dvorak (1975, 1984) technique has been the primary source of TC intensity estimates worldwide, especially when aircraft reconnaissance data are not available (Velden et al. 2006). The Dvorak technique involves subjectively identifying central and banded cloud features in color-enhanced infrared images (i.e., scene-type determination) and linking these cloud patterns to TC intensity by lookup tables. During the last two decades, several revised Dvorak techniques have been developed (Velden et al. 1998). For example, the Olander and Velden (2007) advanced Dvorak technique (ADT) is currently used for operational TC intensity estimation. The ADT reduces the subjectivity by using an objective storm center determination scheme and computer-based algorithms for recognizing cloud features, to which linear regression is applied to estimate TC intensity.

As the basic idea of using satellite imagery to estimate TC intensity is that intensity is strongly related to cloud patterns in the images, other parameters calculated from satellite infrared images have been proposed to correlate with TC intensity: (i) deviation angle variance (DAV), which determines the degree of symmetry of the TC by evaluating the gradients of cloud-top temperatures (Ritchie et al. 2012, 2014); (ii) mean and standard deviation of cloud-top temperatures in 14 radial rings around a TC (Fetanat et al. 2013); and (iii) slope of TC inner-core cloud tops (Sanabia et al. 2014). Furthermore, various regression methods have been applied to TC intensity estimation: (i) Ritchie et al. (2012) used a nonlinear sigmoid equation to describe the relationship between DAV and TC intensity; (ii) Fetanat et al. (2013) used the k-nearest-neighbors algorithm to identify analog TCs that exhibit similar cloud features; (iii) Zhao et al. (2016) introduced a multiple linear regression model that utilizes seven different parameters of TC cloud characteristics; and (iv) Zhang et al. (2016) introduced a machine-learning method, called the relevance vector machine, to estimate TC intensity.

In addition to geostationary satellite infrared imagery, temperature anomaly profiles associated with the TC warm core, observed by the Advanced Microwave Sounding Unit (AMSU), have been applied for intensity estimation (Spencer and Braswell 2001; Demuth et al. 2004). The

AMSU, which is in low Earth orbit (~810 km above the surface vs ~36 000 km for geostationary satellites), detects earth/atmosphere emitted radiation in the microwave portion of the electromagnetic spectrum. The brightness temperatures of various AMSU channels can be used to determine the temperature anomaly associated with the TC warm core, which has a strong relationship to the TC intensity. As each of these techniques has pros and cons, the state-of-the-art Satellite Consensus (SATCON) blends the ADT estimate and other estimates based on polar-orbiting satellite overpass, including AMSU, to produce an ensemble estimate of TC intensity worldwide (Herndon et al. 2010; Velden and Herndon 2014; Herndon and Velden 2018). Specifically, SATCON utilizes a statistically derived weighting scheme that maximizes (minimizes) the strength (weaknesses) of each technique to produce a consensus intensity estimate for a variety of TC structures. Although SATCON will only provide an updated estimate when a polar-orbiting member overpass of a target TC becomes available, it is one of the best techniques with a root-mean-square intensity error under 10 kt (1 kt $\approx$ 0.51 m s^{-1}).

Each of these techniques has issues, such as inconsistency across basins and uncertainties arising from inherent subjectivity or complicated finetuning procedures. Several studies have shown that the Dvorak technique is negatively affected by the inherent subjectivity of storm center selection and scene-type determinations. Nakazawa and Hoshino (2009) documented differences between the Dvorak techniques as applied by the Japan Meteorological Agency and by the Joint Typhoon Warning Center (JTWC) for western North Pacific TCs from 1987 to 2006. Nakazawa and Hoshino (2009) also showed that the JTWC tends to estimate a faster intensification rate before the mature stage and a slower or delayed start of the weakening stage. Maskey et al. (2018) provided an example of widely different estimates among the U.S. agencies: "the 15 UTC 10 October 2017 National Hurricane Center discussion for Tropical Storm Ophelia noted that the Dvorak intensity estimates ranged from T2.3/33 kt (by UW-CIMSS[1]) to T3.0/45 kt (by TAFB[2]) to T4.0/65 kt (by SAB[3]). In this particular case, human experts at TAFB and SAB differed by 20 kts in their Dvorak analyses, and the automated version at the University of Wisconsin was 12 kt lower than either of them."

---

[1] Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin–Madison.

[2] Tropical Analysis and Forecast Branch, National Hurricane Center.

[3] Satellite Analysis Branch; National Environmental Satellite, Data, and Information Service.

In summary, most of the current techniques for TC intensity estimation rely upon feature-engineering to transform low-level satellite imagery into high-level human-constructed features. Even for the most experienced meteorologists and forecasters, it is still hard to identify if a feature is suitable for intensity regression for all TCs in various life stages, environments, and basins. In addition, only a few features (usually less than 10) may be finally used in the regression models.

This collaborative study between meteorologists and data scientists proposes a deep-learning model to address the need for an automated, objective, and end-to-end intensity estimation technique. Since AlexNet, which established the baseline architecture of convolutional neural networks for image recognition used today, was proposed in 2012 (Krizhevsky and Hinton 2009; Krizhevsky et al. 2012), deep-learning algorithms have flourished. These algorithms include convolution neural networks (CNN), recurrent neural networks (RNN), and generative adversarial networks (GAN). A CNN consists of convolutional layers that extract spatial features from the input image and fully connected layers with simple computational units that learns discriminative features to improve prediction of the target phenomenon, without relying on human intelligence to identify which features are most important (e.g., He et al. 2016). Therefore, a CNN could be applied for extracting features form satellite infrared images just as meteorologists determine cloud patterns (e.g., the eye of a hurricane) that are related to TCs within certain intensity ranges. Then, the CNN will use these features as predictors to estimate TC intensity.

Recent studies (Pradhan et al. 2018; B. Chen et al. 2018; Velden and Cossuth 2019) have applied CNN to estimate TC intensity. Pradhan et al. (2018) used CNN to classify TC images into eight categories (i.e., not a TC, tropical depression, tropical storm, and five Saffir–Simpson scales for hurricane intensity). The intensity intervals between these eight categories range from 12 to 29 kt. Although the Pradhan study demonstrated the potential of applying CNN for intensity estimation, they studied only 98 hurricanes in the eastern North Pacific and Atlantic. Moreover, their training data and validation data were from the same hurricanes, and thus these datasets were correlated. Therefore, the capability of CNN for intensity estimation has not yet been demonstrated.

Instead of solving this task through classification, our previous study (B. Chen et al. 2018) proposed a preliminary CNN model that estimates TC intensity as a regression task. In this study, the previous model is extended and optimized for global TC intensity estimation with comprehensive verification. A total of 1097 TCs during 2003–14 were used for training the model, and the model was further optimized with data from 188 TCs during

2015–16. Subsequently, an independent testing dataset of 94 global TCs during 2017 was used to evaluate the model performance and compare the model performance with the operational ADT, AMSU, and SATCON techniques. The datasets used in this study and the CNN model design are described in section 2. Some optimizations of the preliminary model (B. Chen et al. 2018) to make it suitable for global usage are described in section 3. The performance of the CNN model is evaluated in section 4, and conclusions are given in section 5.

2. Data and the CNN design

a. Data

As satellite observations have been used for estimating TC intensity (e.g., Cecil and Zipser 1999; Velden et al. 2006; Olander and Velden 2009), this study utilizes the Gridded Satellite dataset (GridSat; Knapp et al. 2011; Inamdar and Knapp 2015) and the passive-microwave rain rate derived from the Climate Prediction Center morphing technique (CMORPH; Joyce et al. 2004). GridSat is a long-term dataset of geostationary meteorological satellite observations, including infrared (IR1), water vapor (WV), and visible channel (VIS) brightness temperatures. This global dataset is available every 3 h since 1981 and with a resolution of 0.07° latitude/longitude. Since the resolution and quality from different satellites have been calibrated, the GridSat data are suitable for training a CNN model for estimating TC intensity in different basins. Because the VIS observations are only available during daylight hours, only the GridSat IR1 and WV data were used in the current study. The CMORPH technique provides a passive microwave (PMW) rain-rate dataset that is derived from low-Earth-orbit microwave satellite observations, and these observations are translated via spatial propagation information obtained from geostationary IR1 data. CMORPH PMW has global coverage, with a resolution of 0.25° latitude/longitude, and is available every 3 h since 2003. For this study, the CMORPH data were regridded to 0.07° by linear interpolation to unify the resolution of the input arrays for the CNN model. Although the CMORPH rain rate improves model performance (as described later in section 3b), the CMORPH version used in this study is produced at a 24-h latency, because it is a blend of microwave imager data before and after the valid time. For real-time operational use, the model presented in this paper would have a 24-h latency. A future version will test a modification: adopting the nearest PMW rain rate observation within the past 1.5 h.

Examples of IR1, WV, and PMW observations of Typhoon Champi (2015) are shown in Fig. 1. At 0600 UTC 17 October, the GridSat IR1 image (Fig. 1a) and WV
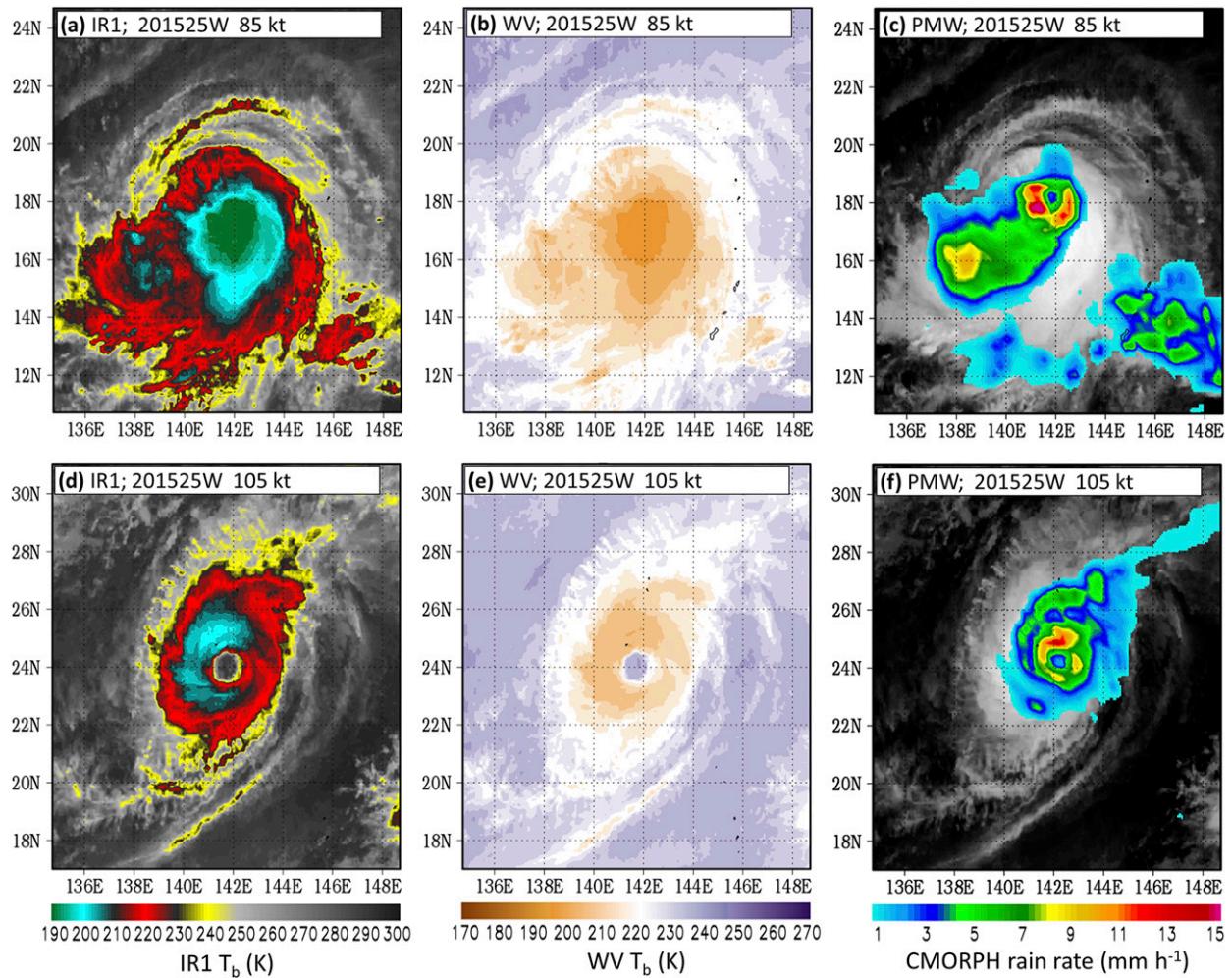
FIG. 1. GridSat (a) IR1 and (b) WV brightness temperature images and (c) CMORPH PMW rain rate (shaded color over grayscale IR1 brightness temperature) of Typhoon Champi (201525W) at 0600 UTC 17 Oct 2015, and (d)–(f) corresponding images five days later at 0600 UTC 22 Oct 2015. The best track intensities are also indicated.

image (Fig. 1b) both have an asymmetric cold cloud shield associated with a sheared anvil to the south. By contrast, the CMORPH PMW reveals the eye and eyewall of Champi and a separate convection area to the southwest (Fig. 1c). Five days later (Figs. 1d–f), the IR1 and WV both have a large eye, but the PMW has the capability to resolve asymmetries in the rain rate of the eyewall and in the rainbands. Although the PMW better observes the TC inner core, the IR1 reveals more cloud features in the outer region, including shallow rainbands and nonprecipitating anvil clouds.

In this study, IR1, WV, and PMW satellite observations from a total of 1379 TCs during 2003–17 were examined every 3 h. Table 1 shows the sample sizes of TCs and 3-hourly images for the western North Pacific (WP), eastern North Pacific (EP), Atlantic (AL), Southern Hemisphere (SH, including South Pacific and south

Indian Ocean), central North Pacific (CP), and north Indian Ocean (IO). Note that the data were categorized into three groups: training, validation, and testing (Table 1). The training dataset was used to fit the CNN weights, and the validation dataset was used to find the best hyperparameters (parameters selected a priori, rather than learned during training). Finally, the best-performing model on the validation dataset is applied to the testing data, which provides an independent assessment of the model's performance.

Postseason analyzed best track TC intensities were used as the "ground truth" for developing the CNN model. These intensities were provided by the JTWC for TCs in the WP, IO, and SH basins, and from the revised hurricane database (HURDAT2) for TCs in the EP, CP, and AL basins. Although these best track intensities were considered ground truth, most of them are based

TABLE 1. Sample sizes of TCs and 3-hourly frames of training data (left portion), validation data (center portion), and testing data (right portion) for various basins. Each frame comprises an IR1, a WV, and a PMW image.

	For training (2003–14)		For validation (2015–16)		For testing (2017)	
	TCs	Frames	TCs	Frames	TCs	Frames
WP	320	17 104	59	2955	33	1367
EP	203	11 910	44	2826	20	972
AL	207	11 921	28	1786	18	1154
SH	285	16 131	45	2303	19	969
CP	17	2771	2	358	0	0
IO	65	1121	10	434	4	118
Global	1097	60 958	188	10 662	94	4580

on satellite, rather than in situ reconnaisance or research observations that are available in the EP and AL basins. Nevertheless, best track intensities are currently the most suitable intensity datasets to develop a worldwide TC intensity estimation technique. As the best track datasets are generally available only at the 6-hourly synoptic times, TC locations and intensities at 0300, 0900, 1500, and 2100 UTC in this study were generated by a simple linear interpolation to match the times of the satellite datasets. Although this interpolation might introduce some small errors, it is necessary to train the model with satellite images that are not at the 6-hourly synoptic times.

Because a large majority of best track intensities have been generated using the subjective Dvorak technique by forecasters in various agencies, some considerable uncertainty (i.e., error relative to the real truth) exists in the best track data. Thus, a separate evaluation of the model performance will be relative to a subset of recon-aided best track data that includes only the intensities within ±3 h of aircraft reconnaissance observations[4] in the AL and EP basins during 2015–17.

b. Brief introduction to convolutional neural networks

CNNs, a class of deep-learning model, are the state of the art in many computer vision problems, such as digit identification (e.g., LeCun et al. 1998) and object recognition (e.g., Krizhevsky et al. 2012). A CNN consists of several processing layers to extract a progressively more abstract representation of the input data, called "features," and fits these features to some target categories for a classification task or a target value for a regression task.

Each layer consists of a number of units (or neurons), which compute a weighted linear combination of the input and are followed by an element-wise nonlinearity (activation function). The weights (or model parameters) are optimized by "training" the model on a dataset. A common choice of the element-wise nonlinear function is linear rectification [$f(x) = \max(x, 0)$], which gives rise to rectified linear units (ReLUs; Nair and Hinton 2010). This application of ReLUs is critical for learning nonlinear relationships between the input and output variables and effectively reduces the "vanishing gradient" problem in training neural networks (Nair and Hinton 2010; Glorot et al. 2011). This ReLU application also makes other pretraining procedures (e.g., Bengio 2007) unnecessary in most cases.

To train the CNN to fit TC intensities that are represented as single values by updating the weights of neurons with many learning iterations (i.e., epochs), the model calculates the error (or loss function) at the end of each epoch and propagates it backward. As the errors are back-propagated in the network, the CNN "learns" the task by updating the weights of each layer to minimize the loss function. As shown in Fig. 2, our CNN model is consisted of convolution layers and fully connected layers. The CNN components responsible for feature extraction are the convolutional layers, as the weights in the convolutional filters are learned during training. Note that the transformation through convolution layers is useful to extract features when the input data exhibit some kind of topological structure, such as the ordering of pixels in a grid or the temporal structure of an audio signal. Specifically, a learnable convolution filter (Fig. 2, red cuboids), which has a specific window size and a depth equivalent to the depth of the input layer, scans through the data with restricted connectivity to the next layer, and transforms the multidimensional array scanned by the filter into one high-level output feature at every step of scanning. As a set of convolution filters is applied to the input layer, a subsequent convolution layer is produced with a depth equivalent to the number of the filters. After passing through several convolution layers, the feature maps (Fig. 2, the 3D array of the last convolution layer) are flattened to a 1D array in the CNN. In this study, fully connected layers transform these features to the target value, which is an estimate of TC intensity. Similar to a conventional artificial neural network, neurons in a fully connected layer have full connections to all neurons in the previous layer, and their output can hence be computed with a weight and a bias offset.

Pooling techniques (or pooling layers), which are usually applied in CNNs, aim to reduce computational complexity by storing the statistics of a group of features instead of their original values (Krizhevsky et al. 2012).

---

[4] The aircraft reconnaissance data can be downloaded at the NHC website (https://www.nhc.noaa.gov/recon.php).
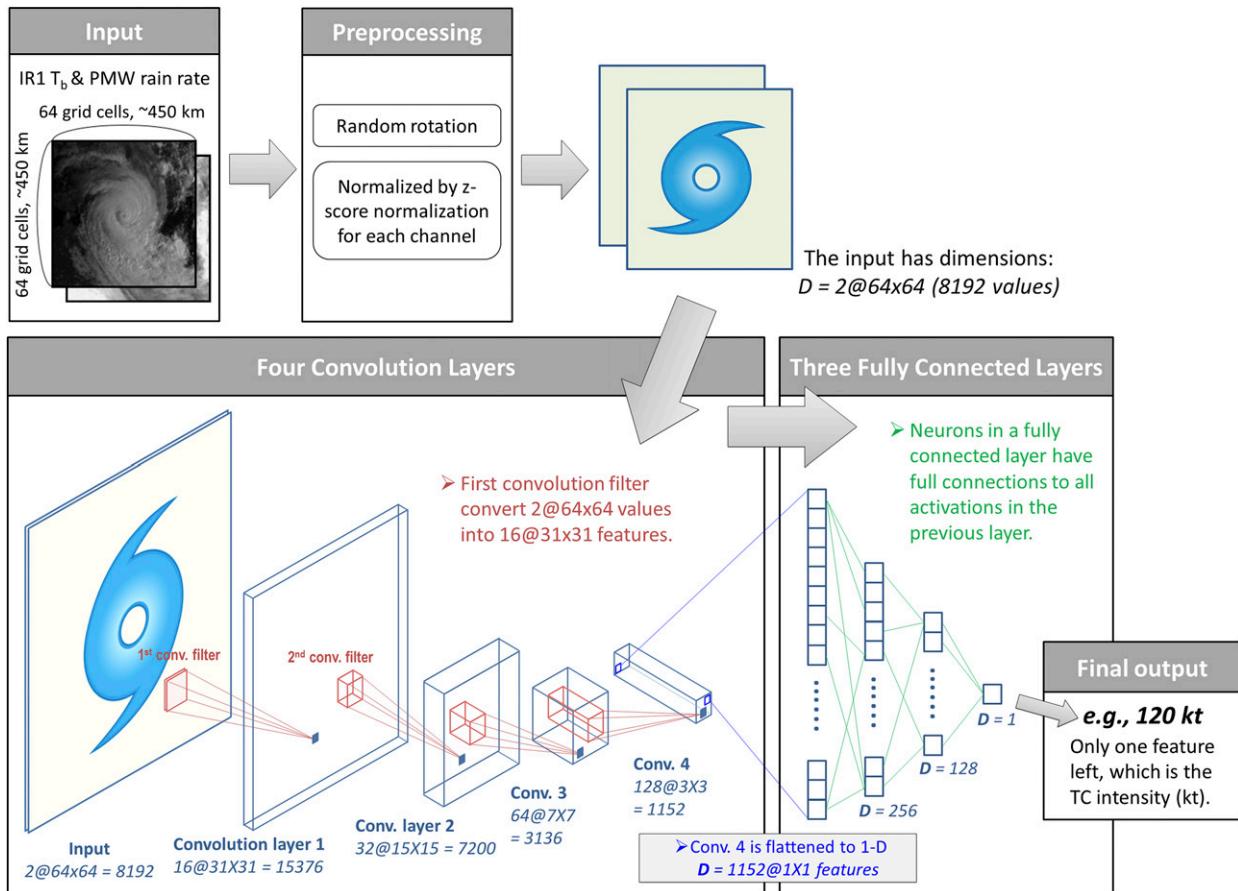
FIG. 2. Flowchart and architecture of the CNN-TC model (see description in section 2c).

However, the characteristics of TC satellite images lead to the decision of omitting pooling layers in our model, because pooling degrades the resolution of the detected features, such as small-scale gradients (e.g., a clear hurricane eye with large gradients on the eyewall), which are important in determining TC intensity (B. Chen et al. 2018, their section 4.3).

Another common procedure applied in standard CNNs is dropout, which is a regularization technique where randomly selected neurons are ignored (the output is temporarily set to zero) during the training to prevent overfitting (Srivastava et al. 2014). However, our previous study (B. Chen et al. 2018, their section 4.3 and Fig. 4) suggested that the application of dropout led to a negative bias in the CNN model, which did not have a softmax layer (Krizhevsky et al. 2012) in order to solve intensity estimation as a regression task. Thus, no dropout is applied to the CNN model in this study. As will be described later in section 2c, an alternative regularization technique, which is the random rotation of input images, is applied to prevent overfitting and improve the model performance.

c. Description of the CNN model

The proposed CNN model was developed based on the structure of AlexNet (Krizhevsky et al. 2012), instead of a very deep CNN (e.g., VGGNet; Simonyan and Zisserman 2014) because our training set is smaller (60 000 images) than those typically required for a very deep model. A deeper CNN contains more weights and therefore requires more data to properly adjust these weights; it could easily overfit without a sufficient data amount. Because a relatively deep CNN usually needs to be pretrained, we reduce the number of layers and use fewer weights on each layer, so that no pretraining is needed. The main procedures and the architecture of the CNN model (hereafter referred to as CNN-TC) are described in Fig. 2. However, some optimizations will be described in section 3 with additional experiments. For example, an experiment to be described in section 3b, excluding the WV datasets, leads to a later decision to only include the IR1 brightness temperatures and PMW rain rates as the model inputs.

Three preprocessing steps of middle-cropping, random rotation during the learning phase, and normalization

TABLE 2. The shape of the input array and hyperparameters of four convolution layers (Conv. 1–4) and three fully connected layers (FC 1–3). See also Fig. 2 and the description in section 2c.

Layer	Input shape	Number of filters	Shape of the filter	Stride	Nonlinearity	Output (features learned)
Conv. 1	$2 \times 64 \times 64$	16	$2 \times 4 \times 4$	2	ReLU	$16 \times 31 \times 31 = 15\,376$
Conv. 2	$16 \times 31 \times 31$	32	$16 \times 3 \times 3$	2	ReLU	$32 \times 15 \times 15 = 7200$
Conv. 3	$32 \times 15 \times 15$	64	$32 \times 3 \times 3$	2	ReLU	$64 \times 7 \times 7 = 3136$
Conv. 4	$64 \times 7 \times 7$	128	$64 \times 3 \times 3$	2	ReLU	$128 \times 3 \times 3 = 1152$
FC 1	1152	—	—	—	ReLU	256
FC 2	256	—	—	—	ReLU	128
FC 3	128	—	—	—	—	1

were applied to the satellite observations (Fig. 2). As the TC intensity may correlate better with inner-core features than with cloud patterns in the outer region, an area of 64×64 grid points (approximately 2.25° latitude–longitude) with respect to the TC center was cropped. The main reason for focusing on this inner-core region was to omit less important and more complicated features in order to avoid overfitting. Second, as dropout layers were omitted in the CNN-TC (section 2b), data augmentation is performed by randomly rotating the input fields with respect to the TC center before using them to train the model. This augmentation not only efficiently increased the amount of training data, but also guarded against overfitting. The third preprocessing step was to separately normalize IR1 and PMW values by *z*-score normalization. For instance, a brightness temperature value of an IR1 image was normalized by subtracting the mean of all brightness temperatures of all IR1 images in the training data and dividing the value by the standard deviation. This step significantly improved the computational efficiency because most of the values would range between -3 and 3 after the normalization. Finally, an additional preprocessing step was utilized with the images of SH TCs. As the Southern Hemisphere TCs rotate clockwise while Northern Hemisphere TCs rotate counterclockwise, the images of SH TCs were flipped horizontally to make them suitable to be simultaneously trained with NH TCs.

After preprocessing, the model had $64 \times 64 \times 2 = 8192$ values (Fig. 2). Subsequently, four convolution layers were applied to extract/transform these 8192 features into 1152 high-level features. The convolution window size, number of convolution filters, stride (i.e., the grid points by which the convolution window shifts), and other configurations are shown in Table 2. Taking the first convolution layer as an example, scanning through the input array ($D = 64 \times 64 \times 2$) by a total of 16 filters with a $4 \times 4 \times 2$ window size and with a stride of 2 transformed the input to $31 \times 31 \times 16 = 15\,376$ features. Similarly, the second, third, and fourth convolution layers extracted $15 \times 15 \times 32 = 7200$, $7 \times 7 \times 64 = 3136$, and $3 \times 3 \times 128 = 1152$ features, respectively.

Also, ReLUs were applied to account for nonlinearity on every layer.

Finally, three fully connected layers were applied to transform these 1152 features to one predicted TC intensity. At the end of the learning process, the model calculates the loss function and propagates it backward to update the weights. Because the objective here is to predict the actual magnitude of the TC intensity, the mean squared error (MSE; kt^2) is selected as the loss function. Note that the CNN hyperparameters as shown in Table 2 (e.g., layer numbers and filter sizes) were determined by testing various CNN structures based on the training and validation datasets.

The CNN-TC was trained with randomly rotated TC images from 2003 to 2014 (Table 1). Each image is horizontally rotated with a random angle at each epoch during the training process to enhance the variability of the input data and prevent overfitting (Dieleman et al. 2015). The model was trained on GeForce GTX 1080 8GB GPU, using the TensorFlow[5] framework (in Python), which supports CUDA.[6] It took approximately 2 h to complete 100 epochs, and 100–300 epochs are generally sufficient for learning this task with the aforementioned CNN architecture.

The MSEs relative to the best track intensity for each epoch in the CNN-TC are shown in Fig. 3 for the training data (red line) and the validation data (green line). For the training data, the MSEs decrease as the epoch number increases. This continuous MSE decrease is because the CNN-TC focuses more and more on detailed and minor features of each image in the training data after about 200 epochs, so that smaller and smaller MSE can be achieved (i.e., overfitting the data). In contrast, the MSE for the validation data stops decreasing after about 200 epochs, and then increases slightly because of the
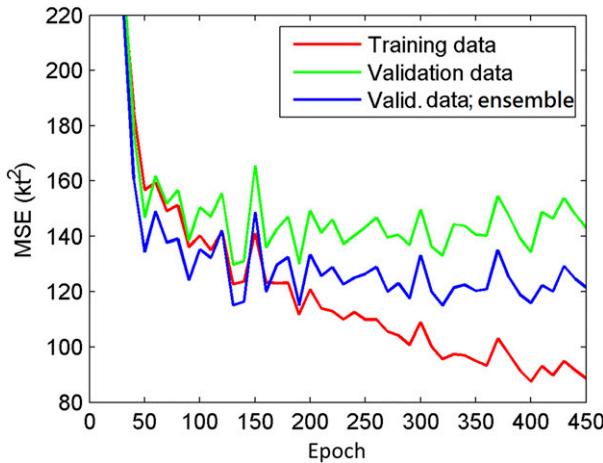
FIG. 3. Learning curves (i.e., MSE for each epoch) for the CNN model using training data (red line) and validation data for both the non-ensemble-averaging estimation (green line) and the rotational-ensemble-averaging estimation (blue line).

CNN overfitting. Consequently, the decision was to stop training the model at around 200 epochs and to use this CNN configuration for estimating TC intensity. Further optimizations for TC intensity estimation across the globe are described in the next section.

3. Model optimization and preliminary results

a. Rotational ensemble averaging

Rotational ensemble averaging is a special procedure only applied in the prediction (validation and testing) phase to enhance the performance of the CNN-TC. It is somewhat similar to the concept of the ensemble mean, in which ensemble members are created with initial perturbations as the model input and an average of all members is taken as the final output. In this study, the CNN-TC estimates TC intensity by averaging multiple estimates based on images rotated by evenly distributed angles from $0°$ to $360°$. For example, if an ensemble number of 4 is selected, four estimates would be obtained by using the unrotated image and those rotated by $90°$, $180°$, and $270°$. Note again that this procedure is not applied in the training phase, in which each image was randomly rotated before fed into the model at each epoch.

The MSE with an ensemble number of 10 for the validation data (Fig. 3, blue line) is lower than the MSE without rotational ensemble averaging (Fig. 3, green line). In addition, a series of experiments was conducted to test the effect of various ensemble numbers on the model performance (Fig. 4a). As the ensemble number increases from 1 to 12, the MSEs decrease but tend to equilibrate with ensemble numbers larger than 6. Therefore, an ensemble number of 10 was adopted for the CNN-TC model.
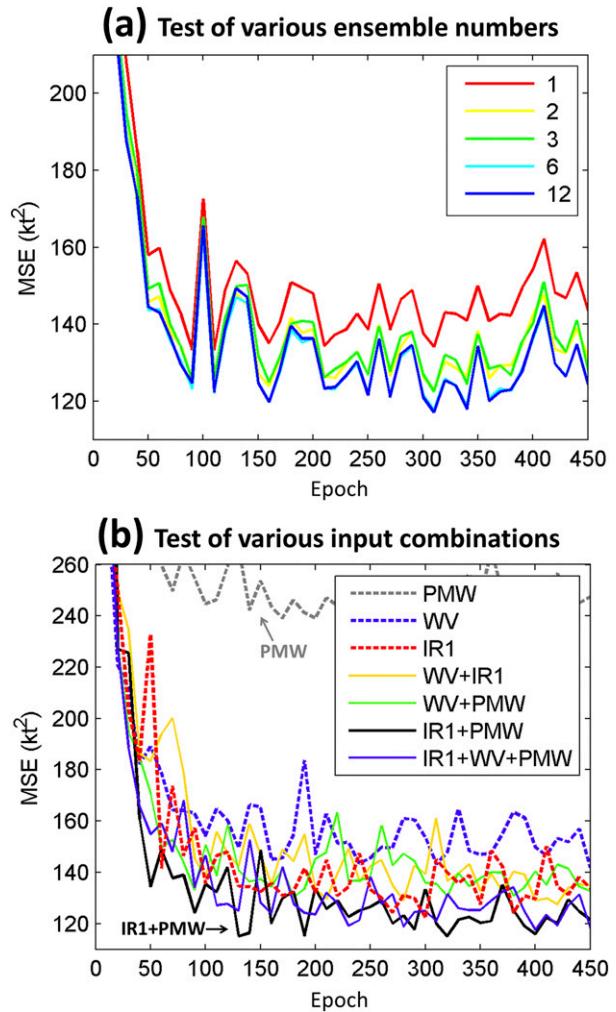

FIG. 4. Learning curves (i.e., MSE for each epoch) based on the validation data for (a) CNN models with various ensemble numbers and (b) CNN models trained with various combinations of satellite images.

b. Test of the combinations of different channels

With the selection of an ensemble number of 10 (section 3a), experiments evaluating the model performance with various combinations of satellite images (i.e., IR1, WV, and PMW) as the model input were conducted (Fig. 4b). To examine the seven combinations of the satellite channels, the filter size of the first convolution layer was changed to fit the depth of the input array.

The first test is the impact of using a single channel among the PMW, WV, and IR1 channels. An MSE of around $250\,kt^2$ could be achieved if only the PMW is used (Fig. 4b, gray dotted line). If only the WV is used, an MSE of around $160\,kt^2$ is achieved (Fig. 4b, blue dotted line). Note that the IR1-only experiment achieves an MSE of around $130\,kt^2$ (Fig. 4b, red dotted line), and thus the IR1
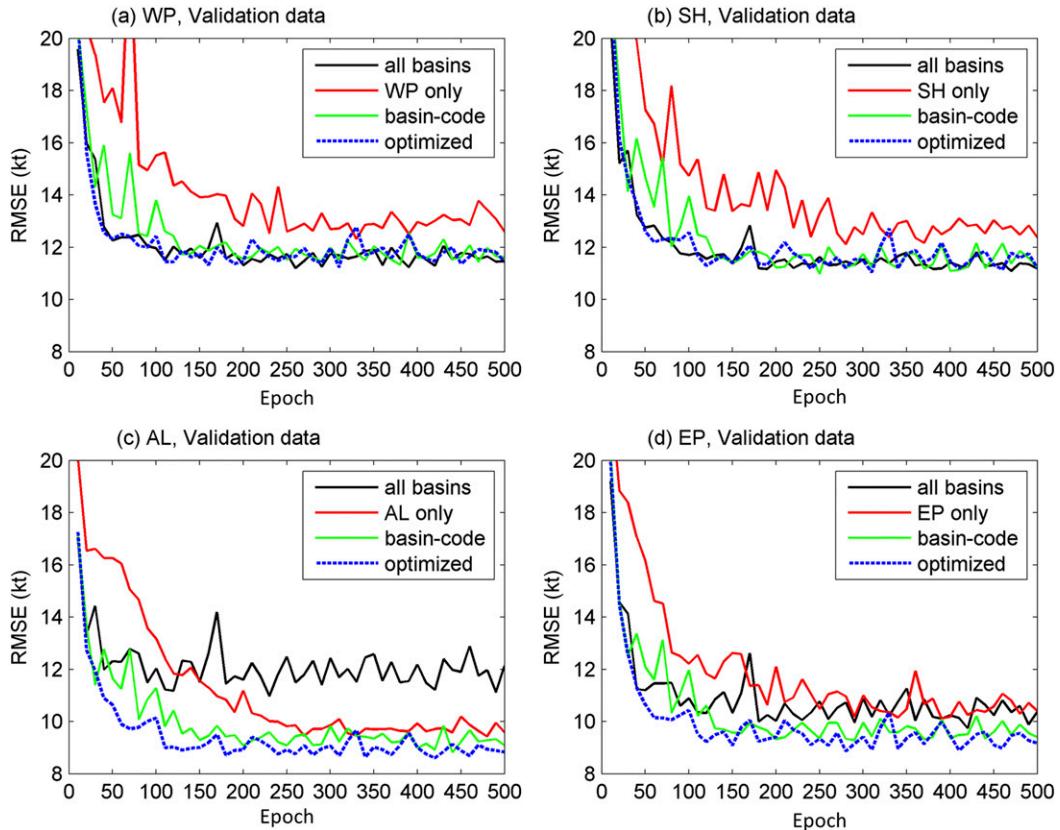
FIG. 5. Learning curves of intensity RMSE for (a) WP TCs for the CNN models trained with all data (black line), data from the specific basin (WP in this case, red line), and the other two optimized versions of the CNN model (green and blue lines), and corresponding graphs for TCs from (b) SH, (c) AL, and (d) EP basins.

is concluded to have the most useful features for TC intensity estimation. However, the optimum combination of two channels is found to be IR1 and PMW (Fig. 4b, black solid line). The reason the combination of the IR1 and WV channels has higher MSEs (Fig. 4b, yellow solid line) is that these channels have similar features (Fig. 1), while the PMW resolves the convective features under the cold cloud shield and thus provides additional information for estimating the TC intensity. Furthermore, the MSE of the three-channel combination (Fig. 4b, blue solid line) is slightly higher than that of IR1 + PMW combination (Fig. 4b, black solid line) presumably because the current CNN may be not deep enough to effectively extract features from three channels at the same time. Because the combination of IR1 and PMW is more effective than the three-channel combination, the WV channel was omitted in the final CNN-TC configuration.

c. Optimized versions of CNN-TC for TCs from all basins

In the previous subsections, TCs from all basins during 2003–14 were used for training the model. However,

the model was found to have different performance in different basins, presumably because the TCs are affected by different environments, which contribute to different TC structures and cloud features. Learning curves of the root-mean-square errors (RMSEs) of the estimated intensity relative to the best track intensity are shown in Fig. 5 for TCs from various basins in the verification data. For the all-basins model (black lines in Fig. 5), the RMSEs for the WP (Fig. 5a), SH (Fig. 5b), and AL (Fig. 5c) basins are around 12 kt, but the RMSE for the EP is close to 10 kt (Fig. 5d). Except for the AL basin (Fig. 5c), the all-basins model has smaller RMSEs than the basin-specific models (red lines in Fig. 5), which were trained with data from each specific basin. Although the all-basin model performance is generally better because of the larger training dataset, the lower RMSEs of the AL-specific model (Fig. 5c, red line) suggest that some special cloud features in the AL basin are not learned by the all-basins CNN.

Another modification of the CNN architecture to develop an upgraded version of CNN-TC for estimating TC intensity in all basins was to include other TC
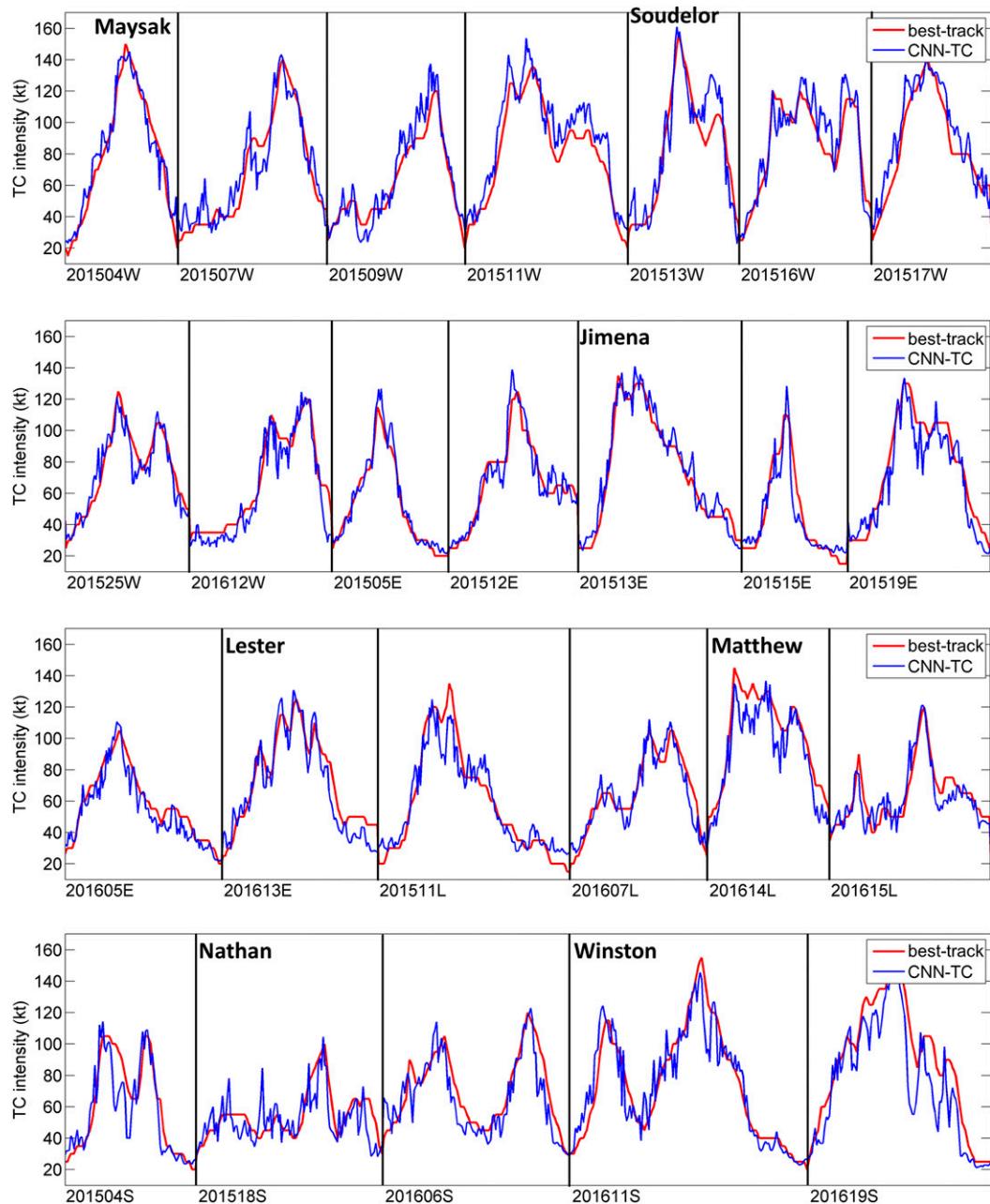
FIG. 6. Best track intensities every 6 h (red line) vs the intensity estimates every 3 h by the optimized CNN-TC (blue line) for the 25 global TCs with maximum intensities greater than 96 kt and lifetimes longer than 10 days during 2015–16 (validation dataset). The year, storm number, and basin indicator (W for western North Pacific, E for eastern North Pacific, L for Atlantic, and S for Southern Hemisphere) of each TC is indicated at the bottom, and names of specific TCs mentioned in the text are provided at the top.

information into the model. This modification was made by adding nonsatellite TC parameters into the 1152 features extracted by the four convolution layers before feeding them into the fully connected layers (Fig. 2). The first modification was simply including six basin codes (1 or 0) representing the WP, EP, AL, SH, CP, and IO basins. For example, a WP TC has the six basin codes of [1, 0, 0, 0, 0, 0]. Therefore, a total of $1152 + 6 = 1158$ features were fed into the fully connected layers. This "basin-code" version (Fig. 5, green lines) has comparable RMSEs to the all-basin model in the WP and SH basins and has substantially lower RMSEs than both all-basin and basin-specific models in the EP and AL basins.
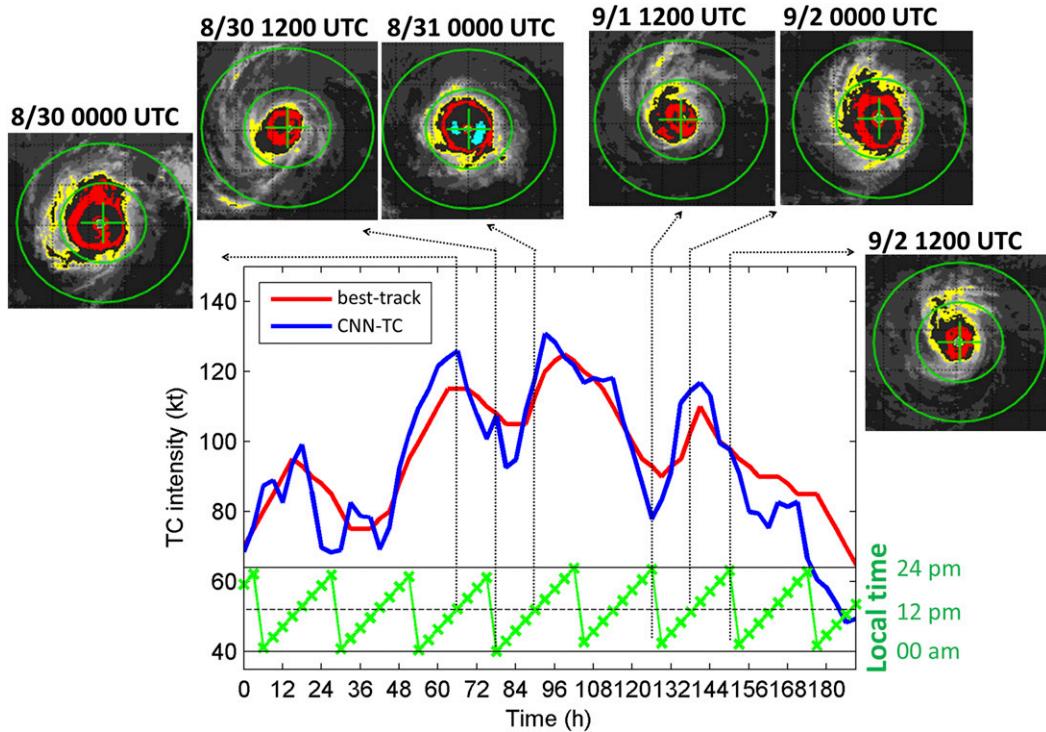
FIG. 7. Best track intensities (red line) and CNN-TC estimates (blue line) for Hurricane Lester (201613E) and the local time at the TC location (green line, right axis). Diurnal variations of the TC IR1 cloud features are illustrated by the six IR1 images. The green circles on the IR1 image indicate the 200- and 400-km radii from the TC center.

An upgraded CNN-TC version (hereafter referred to as "optimized version") was achieved by further adding day of year (represented by sine and cosine), local time (represented by sine and cosine), and TC location (longitude and latitude) into the basin-code version. Therefore, a total of $1152 + 12 = 1164$ features were fed into the fully connected layers. This optimized version (Fig. 5, blue dotted lines) has RMSEs around 11.5 kt for the WP and the SH and 9.5 kt for the AL and the EP. Finally, the weights at epoch 160 were selected as the final CNN configuration, and the performance of this optimized CNN-TC is evaluated in the next section.

4. Performance of the CNN-TC and comparison to other techniques

a. Performance of the optimized CNN-TC based on the validation dataset

This subsection aims to further evaluate how well the optimized CNN-TC fit the validation dataset and introduces the application of a smoothing procedure to improve the estimation accuracy.

The 3-hourly intensity estimates from the optimized version of the CNN-TC are compared in Fig. 6 with the 6-hourly best track intensities for all 25 global TCs during 2015–16 that reached major-hurricane intensity (>96 kt) and had lifetimes longer than 10 days. Generally, CNN-TC exhibits good performance for estimating TC intensity based on images at a single time. For example, the CNN-TC estimated intensity closely follows the best track intensity during the rapid intensifications of Typhoon Maysak (201504W), Typhoon Soudelor (201513W), Hurricane Jimena (201513E), and Hurricane Matthew (201614L). Note that the peak intensity of 155 kt of Typhoon Soudelor (201513W) is actually captured by the CNN-TC.

Although the CNN-TC estimates generally follow the best track intensities, 3-hourly CNN-TC estimates have larger fluctuations. For example, the CNN-TC estimated intensity of Severe Tropical Cyclone Nathan (201518S) has oscillations of 20 kt when Nathan moved along the northern coast of Australia. For Severe Tropical Cyclone Winston (201611S), which was the most intense tropical cyclone on record in the SH, the CNN-TC also has short-term intensity fluctuations even though it follows the general intensity evolution rather well.

Hurricane Lester (Fig. 6, 201613E) was a case with a pronounced diurnal intensity evolution in the verification data. Examination of Lester's intensity oscillations relative to the local time and a series of IR1 images
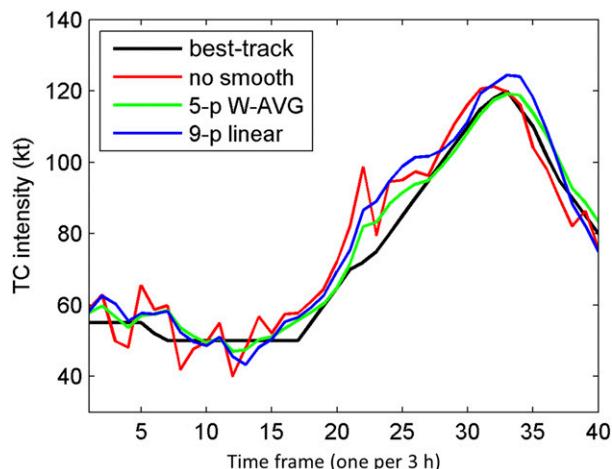
FIG. 8. Intensity evolutions of Typhoon Namtheun (201615W) from the optimized CNN-TC model estimations with different smoothing methods: no smoothing (red line); 5-point weighted average (green line); and 9-point linear regression (blue line). The black line indicates the best track intensity.
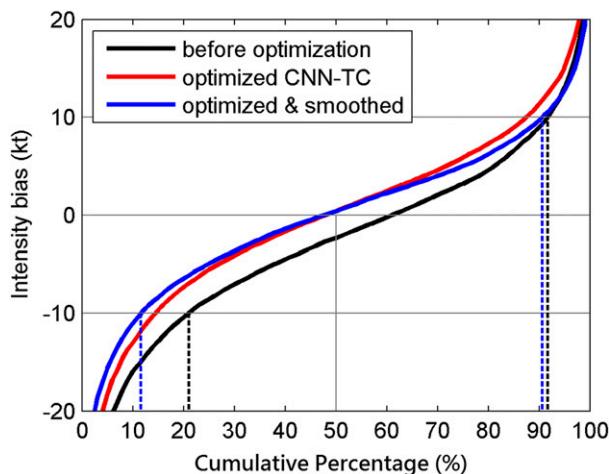

FIG. 9. Cumulative distribution functions of intensity biases relative to the best track intensity for the original version of CNN-TC model (black line), the optimized version (red line), and the smoothed-optimized version (blue line).

(Fig. 7) reveal that Lester had a relatively large cold cloud shield and a clear eye near local noon and a small cold cloud shield near midnight. While the larger cold cloud shields and clear eye appear to be correlated with large CNN-TC intensity estimates, the amplitudes of diurnal intensity variations in CNN-TC estimates are larger than the amplitudes of best track intensities (recall that the best track values are every 6 h and are interpolated to 3-h values). Thus, oscillations of the CNN-TC intensity estimates may in part be due to overfitting the diurnal variation of cloud features.

Another contribution to the large short-term intensity fluctuations of the CNN-TC may be that each 3-hourly estimate by the CNN-TC is time-independent. Therefore, several smoothing methods were tested for the CNN-TC estimates, which are illustrated for Typhoon Namtheun (201615W) as an example (Fig. 8). The smoothing methods tested include: (i) five-point weighted average, which averages the five estimates during the previous 12 h with weights of 1, 2, 3, 4, and 5 (i.e., the latest time, $t = 0$, has a weight of 5; Fig. 8, green line); and (ii) nine-point linear fitting, in which the final estimates is based on a linear fit of estimates during the previous 24 h (Fig. 8, blue line). These smoothing methods are most effective in reducing differences relative to best track intensities during the steady-state stage (Fig. 8, time frame ~0–20) and for a short time scale oscillation during the intensification period (Fig. 8, time frame = 23).

In this study, the RMSE relative to best track intensities for the optimized (but nonsmoothed) CNN-TC version based on all samples in the validation dataset (Table 1) is 10.38 kt. This RMSE can be reduced to 8.74

and 9.80 kt if five-point weighted average and nine-point linear fitting are applied, respectively. Consequently, postsmoothing by the five-point weighted average is applied to the optimized version of CNN-TC, which is hereafter referred to as the "optimized and smoothed" version. In this version, only 11% of the estimations in the validation dataset have a negative bias larger than 10 kt and only 9% of the estimations have an intensity overestimate larger than 10 kt. Thus, the optimized and smoothed version has better performance [79% of errors < 10 kt (Fig. 9, blue line)] than the optimized and unsmoothed version [73% of errors < 10 kt (Fig. 9, red line)]. By contrast, the original version, which is trained without additional TC information, substantially underestimates the intensity (Fig. 9, black line).

The performance of the optimized and smoothed CNN-TC relative the best track intensities is also evaluated by stratifying the validation dataset by various TC intensities, environmental vertical wind shear (VWS) magnitudes, and latitudes (Fig. 10). As shown in Fig. 10d, CNN-TC has relatively small biases for the large sample (Fig. 10a) of TCs within the 30–60-kt intensity bin, which may be due to the training of the CNN having focused on those cases in reducing the MSE during the learning process. Similar to previous studies (e.g., Ritchie et al. 2012, their Fig. 11; and Fetanat et al. 2013, their Fig. 8), the CNN-TC tends to have positive biases (overforecasts) for tropical depressions and negative biases (underforecasts) for hurricanes and typhoons (Fig. 10d). The underestimations for hurricanes and typhoons (intensity > 65 kt) contribute to the relatively high RMSEs (10–15 kt) and mean absolute errors (MAEs, 8–12 kt), compared to the RMSE and MAE for
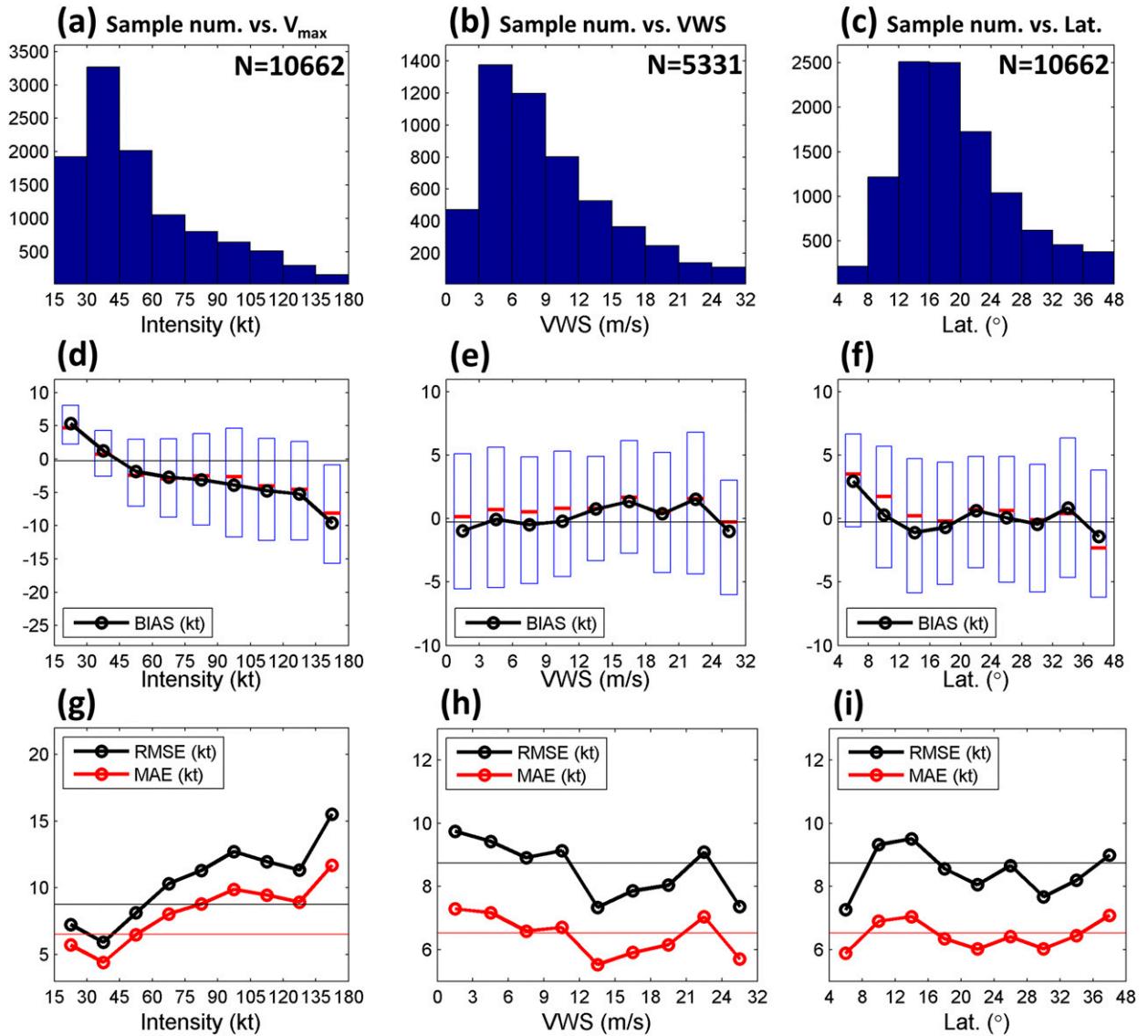
FIG. 10. (a) Sample sizes in 15-kt intensity bins, (d) biases, and (g) RMSEs and MAEs of the optimized and smoothed CNN-TC estimations. Note that the box with the red horizontal bar in (d) indicates the lower, middle, and upper quartiles of the biases, while the black line indicates the mean. (b),(e),(h) As in (a), (d), and (g), but for various 3 m s^{-1} VWS bins. (c),(f),(i) As in (a), (d), and (g), but for various 4° latitude bins.

the 30–60-kt TCs (Fig. 10g). A future modification of the CNN-TC to address this issue will be to randomly drop out the samples with tropical storm intensities in the training data for reducing the overall sampling bias.

Following B.-F. Chen et al. (2018) and Galarneau and Davis (2013), VWS is defined as the difference between 200- and 850-hPa mean flows and was calculated based on the National Centers for Environmental Prediction final operational global analyses. As applied here, both the rotational and divergent winds associated with a TC vortex have been removed within 5° from the TC center, and the mean flow at each level could be calculated by

averaging the wind within this 5° area (Galarneau and Davis 2013). Note also that only half of all samples ($n =$ 5331) are available for this analysis because the global analysis data are produced every 6 h. As shown in Fig. 10e, the CNN-TC intensity biases are fairly stable for various VWS bins, except that the biases are slightly larger for larger VWS magnitudes. This result suggests that the CNN-TC is capable of detecting asymmetric cloud features associated with the TC–VWS interaction, which is important as VWS is considered to be the most critical factor contributing to asymmetric TC convection. Furthermore, relatively small intensity RMSEs
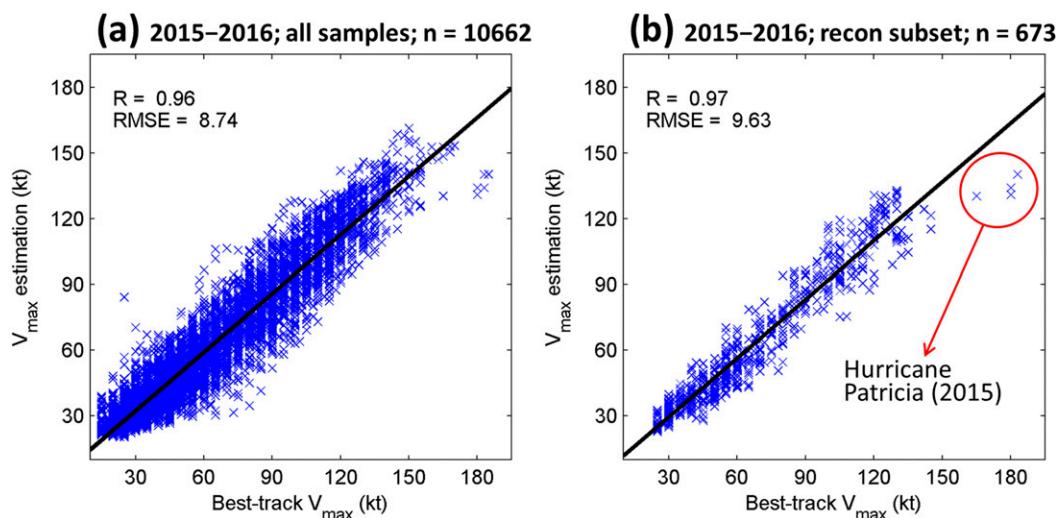
FIG. 11. (a) Scatterplot of CNN-TC estimations and best track intensities in the validation dataset of 188 global TCs in 2015–16. The sample number *n*, least squares regression line (black line), correlation coefficient *R*, and RMSE are also shown. (b) As in (a), except for samples with recon-aided best track for which aircraft reconnaissance observations were available within a ± 3-h period.

and MAEs are found for VWS $> 12\,\mathrm{m\,s^{-1}}$ (Fig. 10h). The interpretation is that a TC under strong VWS ($>12\,\mathrm{m\,s^{-1}}$) usually cannot maintain a hurricane/typhoon intensity (i.e., >64 kt), while TCs under moderate VWS ($<10\,\mathrm{m\,s^{-1}}$) are more likely to reach hurricane/typhoon intensities. It is also possible that the intensity estimation errors of TCs under strong VWS are inherently constrained to be smaller because of the smaller intensities (see also Fig. 10g).

For various latitude bins (Figs. 10c,f,i), the performance of CNN-TC is also stable, except the positive bias for low-latitude TCs (Fig. 10f, 4°–8° bin) and negative bias for high-latitude TCs (Fig. 10f, 36°–48° bin). For low-latitude TCs, the higher bias (Fig. 10f, 4°–8° bin) but lower RMSE (Fig. 10i, 4°–8° bin) may be attributed to generally weak TC intensities. For high-latitude TCs, CNN-TC might underestimate the intensity, as the TC cloud pattern begins to disperse as the extratropical transition begins.

Recall that a large majority of best track intensities were generated using the subjective Dvorak technique and thus might have considerable uncertainty. Following the practice at Cooperative Institute for Meteorological Satellite Studies (CIMSS), the CNN-TC performance is also evaluated with respect to recon-aided best tracks, which is a subset of best track data that includes only the intensities within ± 3 h of aircraft reconnaissance observations in the AL and EP basins. Whereas the intensity RMSE of the optimized and smoothed CNN-TC model for all 10 662 samples from global TCs is 8.74 kt with a correlation coefficient of 0.96 (Fig. 11a), the RMSE for

the 673 recon-only samples is 9.63 kt, with a correlation coefficient of 0.97 (Fig. 11b). Again, a significant underestimation of the extraordinarily intense Hurricane Patricia (indicated by the red circle in Fig. 11b) indicates that the current CNN-TC model lacks skill for intensities above 140 kt. Although the recon-only RMSE (9.63 kt) is larger than the all-sample RMSE (8.74 kt), a two-sample Student's *t* test indicates that the mean absolute error of these two groups are not significantly different at the 99% confidence level. These results suggest that although the CNN-TC model is trained to fit the global best track data, a statistically consistent performance could be also evaluated based on just the recon-aided best tracks for EP and AL TCs.

b. Final verification based on the testing dataset and comparing the optimized and smoothed CNN-TC to ADT, AMSU, and SATCON

Recall that the final CNN configuration and the option of smoothing CNN-TC estimates have been determined based on the evaluation of the validation dataset. Thus, the independent testing dataset (TCs in 2017) is used as a documentation of the CNN-TC performance as it might be operationally used in the future. For the optimized and smoothed CNN-TC, the RMSE relative to all 4580 best track intensities from global TCs is 8.39 kt (Fig. 12a), which is slightly lower than but not significantly different to the RMSE (8.74 kt) calculated based on the validation dataset (Figs. 11a). Furthermore, the RMSE relative to 482 recon-aided best track samples from EP and AL TCs is 8.79 kt, with a correlation coefficient of 0.97 (Fig. 12b).
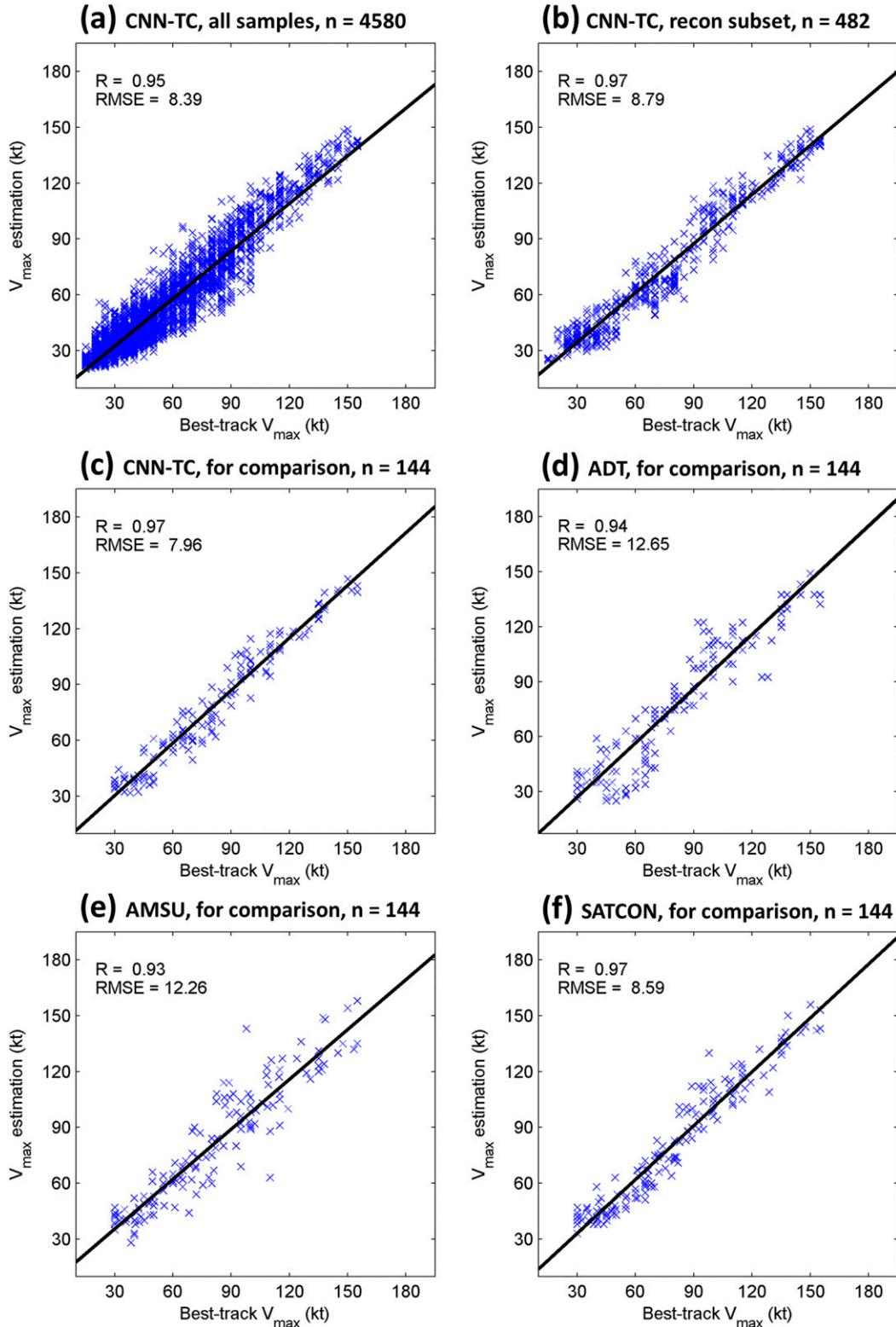
FIG. 12. Scatterplots of CNN-TC estimations and best track intensities for (a) all samples from the testing dataset (94 TCs in 2017), (b) samples with recon-aided best track intensity, and (c) a subset of recon-aided samples for homogenous comparison with other techniques. The sample number *n*, least squares regression line (black line), correlation coefficient *R*, and RMSE are also shown. As in (c), except for (d) ADT, (e) AMSU, and (f) SATCON.

This RMSE of 8.79 kt is considered the most robust and independent in the current study.

The optimized and smoothed CNN-TC intensity estimates were compared with three objective operational techniques for intensity estimation: ADT (Olander and Velden 2007), AMSU (Demuth et al. 2004; Herndon and Velden 2014), and the SATCON (Velden and Herndon 2014; Herndon and Velden 2018). These intensity estimates were downloaded from the CIMSS website.[7]

Different rules were necessary to select samples for homogenous comparisons among the four techniques. As the CNN-TC estimates were available every 3 h, the closest ADT estimate within ±20 min of a CNN-TC estimate was selected. As AMSU estimates were available only at irregular times when the polar-orbiting satellites overpassed the TC, the closest AMSU estimate within ±60 min of each CNN-TC estimate was selected. As SATCON estimates were also only available when a polar-orbiter with a microwave sounder (e.g., AMSU) passed over a TC, the analysis only included SATCON estimates in which a CIMSS AMSU estimate was available in the past 60 min. Note also that, to ensure a fair comparison, these three techniques were validated with corresponding best track values interpolated to their observation times. Finally, only CNN-TC estimates that were able to be matched with all three other estimates were included in the homogenous comparisons.

As ADT, AMSU, and SATCON have been designed to minimize error with respect to the recon-aided best track intensities, scatterplots of the 144 intensity estimates and the corresponding recon-aided best tracks during 2017 are shown for the four techniques (Figs. 12c–f). For this homogeneous comparison, the RMSE and the correlation coefficient R for the CNN-TC are 7.96 kt and 0.97 (Fig. 12c). The CNN-TC generally has a better performance than the ADT (Fig. 12d, RMSE = 12.65 kt, R = 0.94) and AMSU (Fig. 12e, RMSE = 12.26 kt, R = 0.93), and a comparable performance to the SATCON (Fig. 12f, RMSE = 8.59 kt, R = 0.97). In addition, the distributions of absolute errors of the four techniques are also compared (Fig. 13). The mean absolute error of CNN-TC is 6.5 kt. According to two-sample Student's t tests, this mean absolute error is significantly smaller than that of ADT (9.6 kt) and AMSU (9.5 kt) at the 99% confidence level, but is not significantly smaller than for SATCON (6.9 kt).
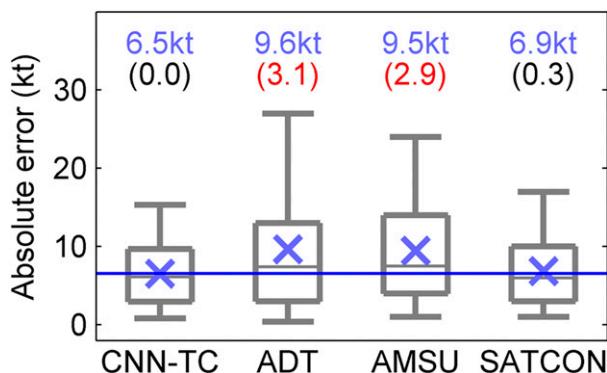
FIG. 13. Boxplots of the absolute errors relative to the recon-aided best track intensties for CNN-TC, ADT, AMSU, and SATCON in a homogenous comparison with n = 144 samples from the testing datasets (2017 TCs). On each box, the median, 25th percentile, and 75th percentile are indicated; the whiskers extend to the 5th and 95th percentiles; the blue "X" and the blue text above the box indicate the mean absolute error. The red (black) text indicates the significant (not significant) mean difference of each techniques to the mean of CNN-TC (blue horizontal line) at the 99% confidence level.

5. Conclusions

This study proposes a satellite imagery-based CNN for TC intensity estimation that is objective and addresses some deficiencies in other techniques, such as inherent subjectivity leading to inconsistent intensity estimates within various basins. Several modifications to the general CNN, such as omitting the pooling and dropout, allow the CNN-TC to directly estimate TC intensity. To our knowledge, CNN-TC is the first CNN model that can estimate TC intensity as a regression task.

The CNN-TC was trained with datasets of 1097 TCs during 2003–14 and optimized with data from 188 TCs during 2015–16. An optimized version was developed that uses both satellite infrared brightness temperatures and microwave rain rates as inputs and incorporates other TC information (i.e., basin, day of year, local time, longitude, and latitude) into the model to estimate TC intensity all over the globe. A postanalysis smoothing of short time-scale intensity fluctuations based on a five-point weighted average further reduces the model RMSE to 8.74 kt based on all 10 662 global samples in the validation dataset (2015–16 TCs). Furthermore, 79% of the CNN-TC estimations have "errors" (differences relative to the best track intensities) less than 10 kt. An independent testing dataset of 94 global TCs during 2017 was used to evaluate the model performance. The optimized and smoothed CNN-TC has RMSEs of 8.39 and 8.79 kt for all 4580 testing samples and 482 recon-aided best track intensities, respectively. Although the CNN-TC model is trained to fit the global best track data, a statistically consistent performance

TABLE 3. RMSEs of several intensity estimation techniques. Descriptions for the CNN-TC in this study are provided in rows 5–6. Three operational intensity techniques that were compared in Figs. 12 and 13 with the CNN-TC are described in rows 7–9.

	Technique	Dataset for verification	RMSE (kt)
Kossin et al. (2007)	Linear regression of IR features	1988−2006, 165 hurricanes (with reconnaissance)	13.20
Fetanat et al. (2013)	k-nearest-neighbors algorithm identifying analog TCs		12.70
Ritchie et al. (2012)	DAV techniques	2004−10, 95 hurricanes	12.70
Liu et al. (2015)	DAV techniques	Only 5 typhoons in 2011	9.34
CNN-TC	Optimized and smoothed CNN	482 samples with recon-aided best track intensity for 2017 TCs in the EP and AL basins	8.79
CNN-TC	Optimized and smoothed CNN	144 homogenous samples with recon-aided best track intensity for 2017 TCs in the EP and AL basins	7.96
ADT	CIMSS-ADT		12.65
AMSU	CIMSS-AMSU		12.26
SATCON	CIMSS-SATCON		8.59

could be also evaluated based on just the recon-aided best tracks for EP and AL TCs. Moreover, the RMSE of 8.79 kt for the 482 recon-aided best track intensities (Table 3, row 5) is considered the most robust and independent in the current study. One of the limitations of the current model is that it would operate with a 24-h latency because the CMORPH data are a blend of microwave imager data before and after the valid time. A future modification to be tested for a real-time operational use is to utilize the nearest PMW rain rate observation within the past 1.5-h time window.

Based on a homogenous verification of 144 recon-aided samples in the testing dataset, the optimized and smoothed CNN-TC has an RMSE of 7.96 kt (Table 3, row 6), which is significantly lower than the RMSE of the operational ADT (12.65) and AMSU (12.26), and statistically comparable to the SATCON (8.59 kt). Although homogeneous comparisons were not possible with the techniques of Kossin et al. (2007), Fetanat et al. (2013), Ritchie et al. (2012), and Liu et al. (2015), the RMSE of the CNN-TC is smaller than those of these techniques (Table 3, rows 1–4). In summary, the results of this study suggest that the CNN approach is competitive with existing methods to estimate TC intensity from satellites.

The successful application of CNN-TC for intensity estimation demonstrates the potential of applying data science for TC analyses. For example, future studies may apply advanced applications for tasks such as estimating TC size and structure, which are important parameters related to forecasting the potential societal impacts of a TC (Powell and Reinhold 2007; Irish et al. 2008; Done et al. 2018; B.-F. Chen et al. 2018). The demonstration that CNN-TC was capable of incorporating external TC information into the neural network suggests it may be possible to develop CNN models for predicting the TC formation based on both the satellite images and environmental factors from external sources, such as the SHIPS parameters (DeMaria and Kaplan 1994; Knaff

et al. 2005; Jones et al. 2006; Kaplan et al. 2015). Finally, the challenging task of predicting the probability of rapid intensification of a TC might be approached with a CNN model.

REFERENCES

Abadi, M., and Coauthors, 2016: TensorFlow: A system for large-scale machine learning. *Proc. 12th USENIX Conf. on Operating Systems Design and Implementation (OSDI'16),* Savannah, GA, USENIX, 265–283, http://dl.acm.org/citation.cfm?id=3026877.3026899.

Bengio, Y., 2007: Learning deep architectures for AI. Tech. Rep. 1312, Dept. IRO, Universit'e de Montreal, 56 pp., https://www.iro.umontreal.ca/~lisa/pointeurs/TR1312.pdf.

Cecil, D. J., and E. J. Zipser, 1999: Relationships between tropical cyclone intensity and satellite-based indicators of inner core convection: 85-GHz ice-scattering signature and lightning. *Mon. Wea. Rev.,* **127**, 103–123, https://doi.org/10.1175/1520-0493(1999)127<0103:RBTCIA>2.0.CO;2.

Chen, B., B.-F. Chen, and H.-T. Lin, 2018: Rotation-blended CNNs on a new open dataset for tropical cyclone image-to-intensity regression. *KDD'18 Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, 90–99, London, United Kingdom, ACM, https://doi.org/10.1145/3219819.3219926.

Chen, B.-F., C. A. Davis, and Y. Kuo, 2018: Effects of low-level flow orientation and vertical shear on the structure and intensity of tropical cyclones. *Mon. Wea. Rev.,* **146**, 2447–2467, https://doi.org/10.1175/MWR-D-17-0379.1.

DeMaria, M., 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.

——, 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.

Demuth, J. L., M. DeMaria, J. A. Knaff, and T. H. Vonder Haar, 2004: Evaluation of Advanced Microwave Sounding Unit tropical-cyclone intensity and size estimation algorithms. *J. Appl. Meteor.*, **43**, 282–296, https://doi.org/10.1175/1520-0450(2004)043<0282:EOAMSU>2.0.CO;2.

Dieleman, S., K. W. Willett, and J. Dambre, 2015: Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Mon. Not. Roy. Astron. Soc.*, **450**, 1441–1459, https://doi.org/10.1093/mnras/stv632.

Done, J. M., D. PaiMazumder, E. Towler, and C. M. Kishtawal, 2018: Estimating impacts of North Atlantic tropical cyclones using an index of damage potential. *Climatic Change*, **146**, 561–573, https://doi.org/10.1007/s10584-015-1513-0.

Dvorak, V. F., 1975: Tropical cyclone intensity analysis and forecasting from satellite imagery. *Mon. Wea. Rev.*, **103**, 420–430, https://doi.org/10.1175/1520-0493(1975)103<0420:TCIAAF>2.0.CO;2.

——, 1984: Tropical cyclone intensity analysis using satellite data. NOAA Tech. Rep. NESDIS 11, 47 pp. [Available from NOAA/NESDIS, 5200 Auth Rd., Washington, DC 20233.]

Fetanat, G., A. Homaifar, and K. R. Knapp, 2013: Objective tropical cyclone intensity estimation using analogs of spatial features in satellite data. *Wea. Forecasting*, **28**, 1446–1459, https://doi.org/10.1175/WAF-D-13-00006.1.

Galarneau, T. J., and C. A. Davis, 2013: Diagnosing forecast errors in tropical cyclone motion. *Mon. Wea. Rev.*, **141**, 405–430, https://doi.org/10.1175/MWR-D-12-00071.1.

Glorot, X., A. Bordes, and Y. Bengio, 2011: Deep sparse rectifier neural networks. *Proc. Machine Learn. Res.*, **15**, 315–323 .

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, IEEE, https://doi.org/10.1109/CVPR.2016.90.

Herndon, D. C., and C. S. Velden, 2014: An update on tropical cyclone intensity estimation from satellite microwave sounders. *31st Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., 34, https://ams.confex.com/ams/31Hurr/webprogram/Paper244770.html.

——, and ——, 2018: An update on the CIMSS SATellite CONsensus (SATCON) tropical cyclone intensity algorithm. *33rd Conf. on Hurricanes and Tropical Meteorology*, Jacksonville, FL, Amer. Meteor. Soc., 284, https://ams.confex.com/ams/33HURRICANE/webprogram/Paper340235.html.

——, ——, J. Hawkins, T. L. Olander, and A. Wimmers, 2010: The CIMSS SATellite CONsensus (SATCON) tropical cyclone intensity algorithm. *29th Conf. on Hurricanes and Tropical Meteorology*, Tucson, AZ, Amer. Meteor. Soc., 4D.4, https://ams.confex.com/ams/29Hurricanes/techprogram/paper_167959.htm.

Inamdar, A. K., and K. R. Knapp, 2015: Intercomparison of independent calibration techniques applied to the visible channel of the ISCCP B1 data. *J. Atmos. Oceanic Technol.*, **32**, 1225–1240, https://doi.org/10.1175/JTECH-D-14-00040.1.

Irish, J. L., D. T. Resio, and J. J. Ratcliff, 2008: The influence of storm size on hurricane surge. *J. Phys. Oceanogr.*, **38**, 2003–2013, https://doi.org/10.1175/2008JPO3727.1.

Jones, T. A., D. J. Cecil, and M. DeMaria, 2006: Passive microwave-enhanced Statistical Hurricane Intensity Prediction Scheme.

Wea. Forecasting, **21**, 613–635, https://doi.org/10.1175/WAF941.1.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeor.*, **5**, 487–503, https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.

Kaplan, J., 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.

Knaff, J. A., C. R. Sampson, and M. DeMaria, 2005: An operational statistical typhoon intensity prediction scheme for the western North Pacific. *Wea. Forecasting*, **20**, 688–699, https://doi.org/10.1175/WAF863.1.

Knapp, K. R., and Coauthors, 2011: Globally gridded satellite observations for climate studies. *Bull. Amer. Meteor. Soc.*, **92**, 893–907, https://doi.org/10.1175/2011BAMS3039.1.

Kossin, J. P., K. R. Knupp, D. J. Vimont, R. J. Murnane, and B. A. Harper, 2007: A globally consistent reanalysis of hurricane variability and trends. *Geophys. Res. Lett.*, **34**, L04815, https://doi.org/10.1029/2006GL028836.

Krizhevsky, A., and G. E. Hinton, 2009: Learning multiple layers of features from tiny images. Tech. Rep., University of Toronto, 7 pp., https://www.semanticscholar.org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086.

——, I. Sutskever, and G. E. Hinton, 2012: ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS2012),* 1097–1105, https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, 1998: Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2324, https://doi.org/10.1109/5.726791.

Liu, C.-C., C.-Y. Liu, T.-H. Lin, and L.-D. Chen, 2015: A satellite-derived typhoon intensity index using a deviation angle technique. *Int. J. Remote Sens.*, **36** (4), 1216–1234.

Maskey, M., D. J. Cecil, R. Ramachandran, and J. Miller, 2018: Tropical cyclone intensity estimation using deep convolutional neural networks. *33rd Conf. on Hurricanes and Tropical Meteorology*, Jacksonville FL, Amer. Meteor. Soc., 15C.8, https://ams.confex.com/ams/33HURRICANE/webprogram/Paper340019.html.

Nair, V., and G. E. Hinton, 2010: Rectified linear units improve restricted Boltzmann machines. *Proc. 27th Int. Conf. on Machine Learning (ICML-10)*, Haifa, Israel, Omnipress, 807–814.

Nakazawa, T., and S. Hoshino, 2009: Intercomparison of Dvorak parameters in the tropical cyclone datasets over the western North Pacific. *SOLA*, **5**, 33–36.

Olander, T. L., and C. S. Velden, 2007: The advanced Dvorak technique: Continued development of an objective scheme to estimate tropical cyclone intensity using geostationary infrared satellite imagery. *Wea. Forecasting*, **22**, 287–298, https://doi.org/10.1175/WAF975.1.

——, and ——, 2009: Tropical cyclone convection and intensity analysis using differenced infrared and water vapor imagery. *Wea. Forecasting*, **24**, 1558–1572, https://doi.org/10.1175/2009WAF2222284.1.

Powell, M. D., and T. A. Reinhold, 2007: Tropical cyclone destructive potential by integrated kinetic energy. *Bull. Amer. Meteor. Soc.*, **88**, 513–526, https://doi.org/10.1175/BAMS-88-4-513.

Pradhan, R., R. S. Aygun, M. Maskey, R. Ramachandran, and D. J. Cecil, 2018: Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Trans. Image Process.*, **27**, 692–702, https://doi.org/10.1109/TIP.2017.2766358.

Rappaport, E. N., J.-G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The Joint Hurricane Test Bed: Its first decade of tropical cyclone research-to-operations activities reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371–380, https://doi.org/10.1175/BAMS-D-11-00037.1.

Ritchie, E. A., G. Valliere-Kelley, M. F. Piñeros, and J. S. Tyo, 2012: Tropical cyclone intensity estimation in the North Atlantic basin using an improved deviation angle variance technique. *Wea. Forecasting*, **27**, 1264–1277, https://doi.org/10.1175/WAF-D-11-00156.1.

——, K. M. Wood, O. G. Rodríguez-Herrera, M. F. Piñeros, and J. S. Tyo, 2014: Satellite-derived tropical cyclone intensity in the North Pacific Ocean using the deviation-angle variance technique. *Wea. Forecasting*, **29**, 505–516, https://doi.org/10.1175/WAF-D-13-00133.1.

Sanabia, E. R., B. S. Barrett, and C. M. Fine, 2014: Relationships between tropical cyclone intensity and eyewall structure as determined by radial profiles of inner-core infrared brightness temperature. *Mon. Wea. Rev.*, **142**, 4581–4599, https://doi.org/10.1175/MWR-D-13-00336.1.

Simonyan, K., and A. Zisserman, 2014: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, https://arxiv.org/abs/1409.1556.

Spencer, R. W., and W. D. Braswell, 2001: Atlantic tropical cyclone monitoring with AMSU-A: Estimation of maximum sustained wind speeds. *Mon. Wea. Rev.*, **129**, 1518–1532, https://doi.org/10.1175/1520-0493(2001)129<1518:ATCMWA>2.0.CO;2.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958, http://jmlr.org/papers/v15/srivastava14a.html.

Velden, C. S., and D. C. Herndon, 2014: Update on the SATellite CONsensus (SATCON) algorithm for estimating TC intensity. *31st Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., 30, https://ams.confex.com/ams/31Hurr/webprogram/Paper243759.html.

——, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from 89-GHz band microwave satellite imagery. *18th Conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 2B.3, https://ams.confex.com/ams/2019Annual/meetingapp.cgi/Paper/352708.

——, T. L. Olander, and R. M. Zehr, 1998: Development of an objective scheme to estimate tropical cyclone intensity from digital geostationary satellite infrared imagery. *Wea. Forecasting*, **13**, 172–186, https://doi.org/10.1175/1520-0434(1998)013<0172:DOAOST>2.0.CO;2.

——, and Coauthors, 2006: The Dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bull. Amer. Meteor. Soc.*, **87**, 1195–1210, https://doi.org/10.1175/BAMS-87-9-1195.

Zhai, A. R., and J. H. Jiang, 2014: Dependence of US hurricane economic loss on maximum wind speed and storm size. *Environ. Res. Lett.*, **9** (6), https://doi.org/10.1088/1748-9326/9/6/064019.

Zhang, C., J. Qian, L. Ma, and X. Lu, 2016: Tropical cyclone intensity estimation using RVM and DADI based on infrared brightness temperature. *Wea. Forecasting*, **31**, 1643–1654, https://doi.org/10.1175/WAF-D-15-0100.1.

Zhao, Y., C. Zhao, R. Sun, and Z. Wang, 2016: A multiple linear regression model for tropical cyclone intensity estimation from satellite infrared images. *Atmosphere*, **7**, 40, https://doi.org/10.3390/atmos7030040.