Investigation of Machine Learning Using Satellite-Based Advanced Dvorak Technique Analysis Parameters to Estimate Tropical Cyclone Intensity

TIMOTHY OLANDER,[a] ANTHONY WIMMERS,[a] CHRISTOPHER VELDEN,[a] AND JAMES P. KOSSIN[b,a]

[a] *Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin–Madison, Madison, Wisconsin*
[b] *The Climate Service, Durham, North Carolina*

(Manuscript received 15 December 2020, in final form 27 September 2021)

ABSTRACT: Several simple and computationally inexpensive machine learning models are explored that can use advanced Dvorak technique (ADT)-retrieved features of tropical cyclones (TCs) from satellite imagery to provide improved maximum sustained surface wind speed (MSW) estimates. ADT (version 9.0) TC analysis parameters and operational TC forecast center best track datasets from 2005 to 2016 are used to train and validate the various models over all TC basins globally and select the best among them. Two independent test sets of TC cases from 2017 to 2018 are used to evaluate the intensity estimates produced by the final selected model called the "artificial intelligence (AI)" enhanced advanced Dvorak technique (AiDT). The 2017–18 MSW results demonstrate a global RMSE of 7.7 and 8.2 kt (1 kt ≈ 0.51 m s^{-1}), respectively. Basin-specific MSW RMSEs of 8.4, 6.8, 7.3, 8.0, and 7.5 kt were obtained with the 2017 dataset in the North Atlantic, east/central Pacific, northwest Pacific, South Pacific/south Indian, and north Indian Ocean basins, respectively, with MSW RMSE values of 8.9, 6.7, 7.1, 10.4, and 7.7 obtained with the 2018 dataset. These represent a 30% and 23% improvement over the corresponding ADT RMSE for the 2017–18 datasets, respectively, with the AiDT error reduction significant to 99% in both sets. The AiDT model represents a notable improvement over the ADT performance and also compares favorably to more computationally expensive and complex machine learning models that interrogate satellite images directly while still preserving the operational familiarity of the ADT.

KEYWORDS: Tropical cyclones; Satellite observations; Neural networks; Artificial intelligence

1. Motivation

Machine learning is a rapidly growing application of study being used to examine a wide variety of topics, especially in the environmental sciences. It can be employed to discern patterns in large datasets that are more difficult to examine using traditional methods due to its ability to decipher correlations in datasets objectively. Due to considerable advancements in computer hardware, such as new graphical processing units (GPUs), and software analysis packages such as Tensorflow (Abadi et al. 2016) and Keras (Chollet 2015), a greater number of researchers are able to access, learn, and utilize machine learning techniques than ever before.

The three most popular types of neural networks in the atmospheric sciences applications are multilayer perceptron, convolutional, and recurrent. Multilayer perceptron networks (MLP) are a type of "feed-forward" network where the data flows in one direction through the model. These are the most general type, containing varying numbers of layers and neurons. A MLP can be called "shallow" or "deep," depending on the number of hidden layers in the model. A typical MLP has three types of layers: input, output, and hidden layers. Input and output layers are self-explanatory and are equal in size to the vector size of the model input and output, respectively. By contrast, hidden layers exist between the input and output layers and can be any size according to their number of "neurons." Neurons are computational units that have weighted connections to the adjoining layers, with a layer being considered "fully connected" to the adjoining layers when each neuron in a layer is connected to each node (either neurons, input or output) of the adjacent layers. An "activation function" normally applies to each neuron: this can be as simple as a step function (pass the information on if the weighted sum is greater than a value, otherwise do not), and usually these functions are nonlinear to allow the network to learn nonlinear relationships.

Figure 1 provides a schematic diagram of a simple MLP with only three layers. The input layer contains the observed predictors (or features in ML literature) being passed forward to the next layer of the model containing 32 neurons. This layer is fully connected to a hidden layer since each of the 26 predictor values is connected to each of the 32 neurons. For each neuron in the hidden layer, a weighed sum (plus an offset) of each of the 26 values is calculated. The weights for each input value for each neuron and the corresponding offset are optimized during the training process of the model. An activation function is then applied to the weighted sum values of each neuron in the hidden layer to define how the information from the 32 neurons is passed to the next layer. This single node is the output layer and represents the final predicted value of the MLP.

Convolutional neural networks (CNN) are typically applied to computer vision analysis to perform abstract feature characterization. CNNs identify components within the images using convolutional filters and down-sampling processes that make up a method of hierarchical pattern-matching. The product of successive convolutions (called a "feature map") is then transformed, or "flattened," to a one-dimension array

Corresponding author: Timothy Olander, timo@ssec.wisc.edu

Publisher's Note: This article was revised on 22 November 2021 to include the complete affiliations of coauthor Kossin.

DOI: 10.1175/WAF-D-20-0234.1

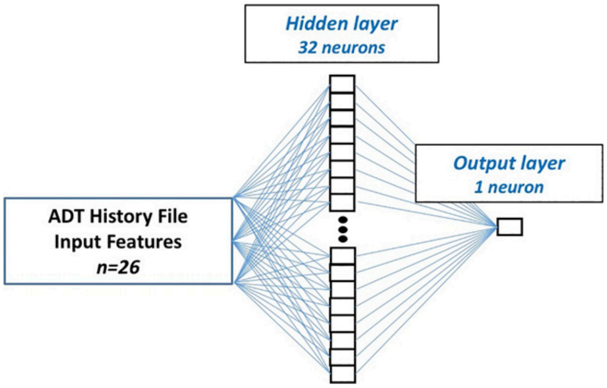


FIG. 1. Schematic diagram of final regression-based MLP model with one hidden layer of 32 neurons and one output layer with 1 neuron.

after all convolution and down-sampling iterations have been performed, which can then be analyzed with a MLP to produce a final predicted value. Finally, there are recurrent neural networks (RNN), which recursively feed the output of a layer back into itself. This allows the past iterations of a layer to influence the future iterations, which is advantageous for applications to sequential datasets.

Satellite-based tropical cyclone (TC) intensity analysis is one field where machine learning techniques have been explored a great deal in the past few years. Legacy techniques have achieved reasonable accuracies based upon empirical or statistics-based algorithms operating on geostationary satellite infrared (IR) imagery (Velden et al. 1998; Pineros et al. 2008; Ritchie et al. 2014; Olander and Velden 2019), polar-orbiting satellite passive microwave (PMW) imagery (Xiang et al. 2019; Jiang et al. 2019), or techniques using both (Velden and Herndon 2020).

Recent machine learning studies normally use the same datasets as the statistics-based algorithms but employ a much more powerful image analysis methodology to decipher patterns that may be missed or ignored in those techniques. These are CNN architectures, which is a process based on breaking down gridded data, such as satellite imagery, into a series of one or more progressively smaller grid layers using a series of image operations. A convolution operation is first performed, which highlights the larger, multi-element features in the image using filters. An element-wise, nonlinear operation can also be applied during the convolution process to remove data above or below a set threshold. A final down-sampling, or pooling, operation is performed to further highlight features and reduce the overall size of the image. The image is then flattened from a two-dimensional grid to a one-dimensional array of scalar values, which is then analyzed with a MLP to result in a final output layer of predicted value(s).

Highlighting select recent CNN studies, Pradhan et al. (2018), Combinido et al. (2018) and Maskey et al. (2020) focused on geostationary cloud top temperature information from a single IR channel, typically the longwave IR (LWIR) window channel (approximately 10.7 μm). Zhang et al. (2020) used a combination of LWIR and water vapor (WV, approximately 6.7 μm) imagery, while Lee et al. (2019) and Yu et al. (2020) employed several additional IR channels (3.9 and 12.0 μm) in addition to the LWIR and WV channels. Wimmers et al. (2019) focused exclusively on PMW data (37- and 85–92-GHz channels) from available polar-orbiting satellites. Chen et al. (2019) used both PMW rain rate and geostationary LWIR and WV imagery in their study. Utilization of a combination of channels and/or sensors over a single channel can provide more information to the model being derived, but doing so will increase the amount of time and computational power needed to incorporate each new set of data. Each of the previously listed studies balanced these requirements and availability of data when determining which datasets to use. All of these CNN studies showed promise for objectively estimating TC intensity, with some yielding superior results to the legacy methods.

CNNs have both the advantage and disadvantage of encapsulating spatial patterns in a neural network form. The advantage of this is the thoroughness and objectivity that comes from direct training on image inputs since all of the image data are utilized in the final determination of a predicted value. There are a few disadvantages to CNN models, however. It can be difficult to achieve a general optimization since many different methods can be used to subsample/filter the image in the convolution process. Also, CNN can have high computational costs, especially if the computer GPU is not sufficient and/or the model is very complex. In addition, CNN require a great deal of data to sufficiently train the model in order to avoid overfitting to the training data. Some of these drawbacks can be experienced with ANN, but are heightened with CNN due to their increased complexity.

By contrast, the statistical advanced Dvorak technique (ADT; Olander and Velden 2019) is a fully automated and objective algorithm that has been applied in real time for almost two decades by operational forecast centers worldwide as an aid to estimate TC intensity. The ADT has also been employed as the primary analysis tool in several TC climatological studies (Velden et al. 2017; Kossin et al. 2013, 2020; Courtney et al. 2020). It primarily examines geostationary satellite LWIR imagery to assess the intensity of TCs through pattern matching and explicit feature analysis techniques. PMW imagery (approximately 85–92 GHz) is also used in certain cases to provide indicators of early eyewall formation.

The basis of the ADT is the objective determination of a storm cloud pattern or ‘‘scene type,’’ which attempts to mimic the parent Dvorak technique (Dvorak 1975, 1984; Velden et al. 2006) methodology that requires a human analyst. Once the ADT scene type is derived the current storm intensity is estimated using statistical methods specific to that scene type. There are four primary scene-type categories used in the ADT, examples of which are shown in Fig. 2. ‘‘Eye’’ references when an eye feature is apparent and is a feature associated with stronger intensity TCs. A ‘‘CDO,’’ or central dense overcast, is a large, coherent cirrus cloud shield that covers the rotational storm center or forming eye feature. It typically occurs prior to the appearance of an eye. ‘‘Curved band’’ features typically occur in forming TCs when an arc of convection is wrapping around a storm circulation center as the storm is developing. As the storm increases in strength the

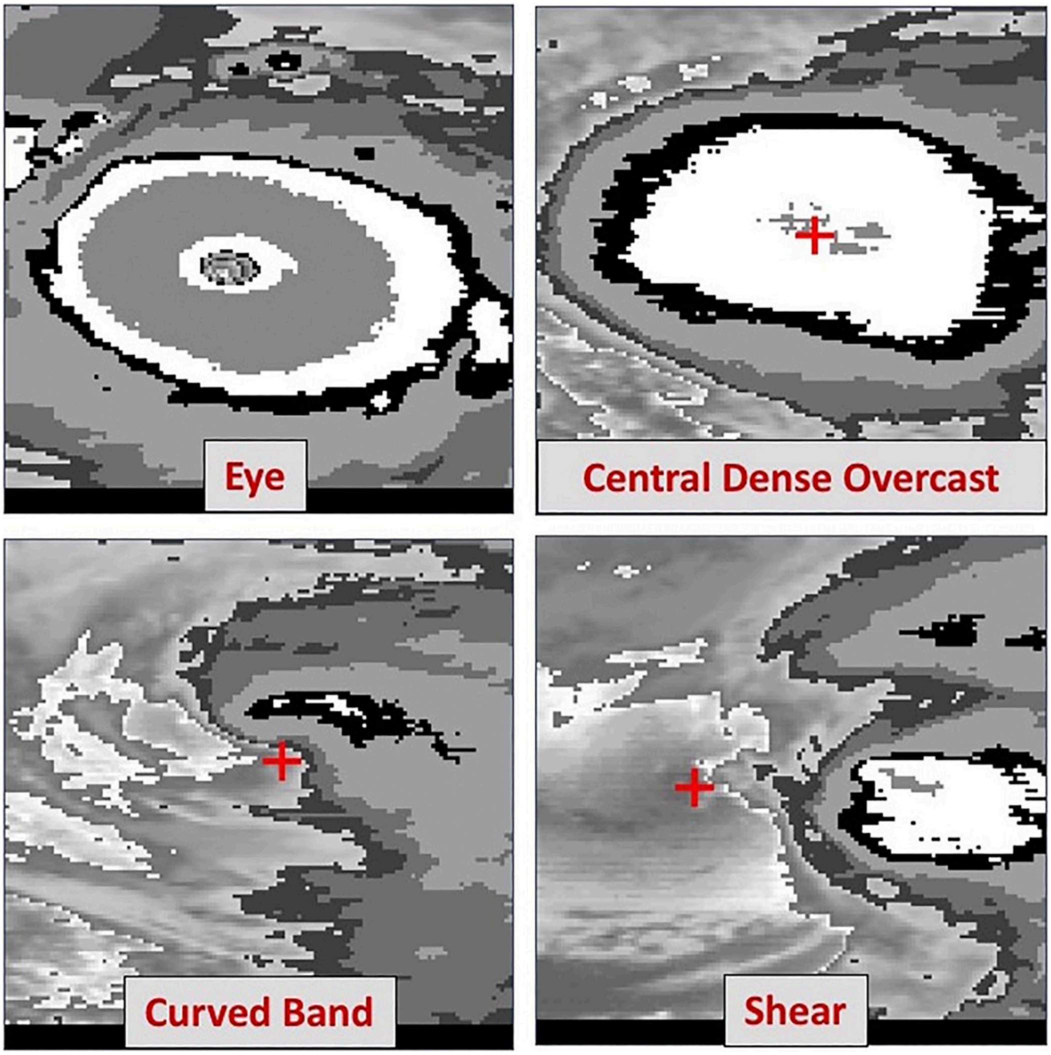


FIG. 2. Examples of ADT scene-type classifications. Imagery is LWIR imagery with the standard basic Dvorak "BD" enhancement to enhance ranges of cloud top temperatures. Black, white, and dark gray are colder temperatures, as in the eye and CDO examples, with several other shades of gray shown in the shear and curved band indicating warmer temperatures.

extent of the convective arc will also increase and move closer toward the storm circulation center. Once the convection wraps completely around the circulation center the clouds will tend to cover this location and form a CDO. Finally, the "shear" scene can occur at any time during the storm life cycle. This scene type occurs when a TC encounters strong environmental winds and the low level circulation center is exposed and separated from the convection. The more intense the atmospheric shear the larger the separation between the storm center and the convection (and typically the weaker the storm will be).

The scene type is calculated in the ADT using two separate score equations, one for the eye region (≤24 km from the storm center) and another for the cloud region (24–136 km from the storm center) of the storm. These equations use various TC parameters retrieved from the IR imagery to derive the scores and a series of threshold values to classify the final scene type. Depending upon the scene type classified, either a regression equation (eye and CDO scene types) or a linear relationship of select TC parameter values (shear and curved band scene types) can be used to derive the intensity estimate. The shear and curved band scene types are based upon the original analysis techniques outlined in the Dvorak technique and have not been investigated in depth and modified like the eye and CDO scene types.

The retrieved TC parameters are stored in an ADT "history file" for that storm and are used in the subsequent intensity analyses for the lifetime of the storm. Typically, the ADT is run every 30 min, providing real-time, objective estimates of TC intensity for all storms around the globe with accuracies commensurate with the legacy manual Dvorak technique. Further details on the ADT are provided in Olander and Velden (2019) and in Olander (2021).

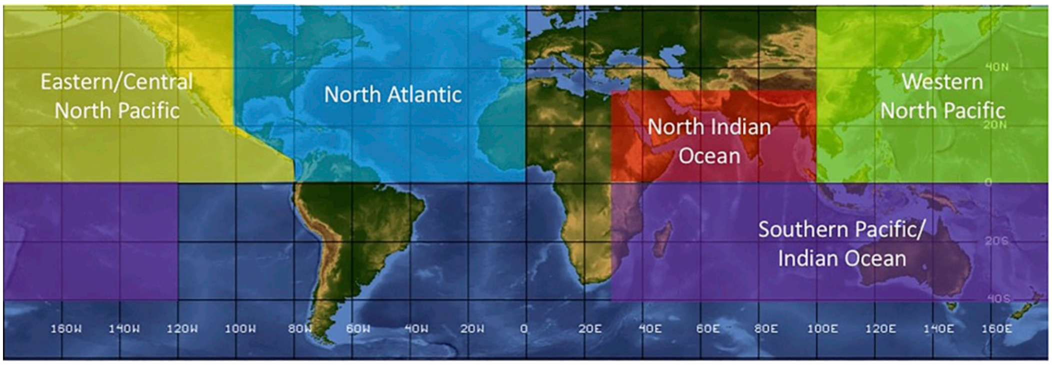


FIG. 3. World map outlining the tropical cyclone ocean region boundaries.

Since the ADT is a proven and mature algorithm that already objectively analyzes the IR imagery and determines the scene type and resultant intensity using methods that have been developed over two decades, why not utilize the output TC parameters stored within the ADT history file and apply machine learning techniques to assess value-added potential? Such a model would be relatively simple and computationally inexpensive to develop and deploy since the image interrogation is already done by the ADT. Development of a MLP takes minutes, not hours or days, to derive and can be done using relatively modest computer processing power available on a laptop or desktop computer. CNNs typically require much higher priced machines containing multiple and/or more expensive GPUs to derive a model efficiently. Repurposing the elements of the ADT, instead of replacing it with a CNN-style algorithm, allows operational users familiar with the ADT to understand the basis of the ADT intensity estimates while providing the user community with improved results, especially in areas where the ADT has struggled or has not been as thoroughly examined and improved.

This paper reports on the development of a MLP model to augment the ADT intensity estimation process. This "artificial intelligence (AI)" enhanced ADT (AiDT) model is executed after the real-time ADT processing sequence is completed for an active TC. It modifies the ADT intensity estimate by applying MLP techniques to the ADT analysis parameters familiar to operational TC users. Many different MLP networks and data inputs are explored to determine the best possible configuration. The final MLP configuration is independently validated and the TC intensity estimation performance is compared to several recent CNN algorithms to assess the competitiveness of the AiDT technique.

2. Data

ADT history files are collected globally from TCs during the period 2005–18 using the latest ADT-version 9.0 run at a 30-min temporal resolution for all storms with best track intensities of greater or equal to 30 kt (1 kt ≈ 0.51 m s^{-1}), including extratropical and subtropical cyclones. The satellite imagery is provided by the geostationary operational satellite with the lowest viewing angle at the time of the analysis for the storm in each TC basin: Geostationary Operational Environmental Satellites (*GOES-8–16*), *Himawari-8*, Multifunction Transport Satellites (*MTSAT-1R*, *MTSAT-2*), and *Meteosat-7–10*. The ADT does not require the IR imagery to be spatially homogeneous over the analysis period since it accounts for resolution changes within the algorithm itself. ADT estimates are derived when the storm center position is over open ocean. The ADT estimates are derived in terms of a "T-number" (tropical number, or T#) or current intensity (CI#), with each separate value defined from 1.0 (weakest TCs) to 8.5 (strongest TCs) in 0.1 increments. The CI# is calculated from the current and previous T# intensity values, employing time-dependent intensity change rules and a time averaging scheme, and represents the current storm intensity estimate. The CI# can be converted to a maximum wind speed (MSW) estimate using a standard conversion outlined in Velden et al. (2006).

Five oceanic regions (TC basins) are examined separately and also as a combined global dataset. The five regions are the North Atlantic, eastern/central North Pacific (east of the international date line, with central region defined from 140°W to the date line), western North Pacific (west of the international date line), northern Indian Ocean, and southern Pacific/Indian Oceans. These basins, shown in Fig. 3, will be referred to as the Atlantic, EastPac, WestPac, NIO, and SouthPac, respectively, with the combined set referred to as the global dataset.

MLP models are developed for each of the five separate basins as well as an "AllBasins" model using the combined global dataset. ADT history files from 2017 to 2018 TCs are set aside as independent "test" datasets. The test datasets from each year are examined separately for two reasons: 1) to provide a more direct comparison of results with the Chen et al. (2019) study which examined 2017 WestPac TCs, and 2) to discern the robustness of the results from one TC season to the next. Years 2007, 2010, and 2014 are designated as the "validation" dataset, with the remaining years between and including 2005 and 2016 serving as the "training" dataset. It must be noted that "validation" in machine learning terminology does not refer to the final independent evaluation process, but instead to the in-training check on model performance in order to tune the model design. The three years selected for the validation dataset are chosen to provide a representation of all

TC intensities, from tropical depression to category 5/Super Typhoon, for each of the five basins. These years were chosen before the model training and validation process were performed. The total number of ADT individual intensity analyses for each dataset and ocean basin are listed in Table 1.

The ground truth data, otherwise known as the ‘‘label’’ data in machine learning nomenclature, used in the training, validating, and testing of the model are the official final best track MSW estimates provided by the National Hurricane Center (NHC) for the Atlantic and EastPac storms, the Central Pacific Hurricane Center (CPHC) for EastPac storms in the Central Pacific region (west of 140W and east of the international date line), and the Joint Typhoon Warning Center (JTWC) for the WestPac, SouthPac, and NIO storms.[1] Both label datasets define MSW as the 1-min sustained wind at 10 m above the surface. The best track MSW estimates for each TC in the sample are provided every 6 h and linearly interpolated to each 30 min ADT history file record. The ADT current intensity numbers (CI#) are converted to MSW values using the standard Dvorak relationships (Dvorak 1984) to provide the baseline ADT MSW estimates shown in section 4. All MSW units are in knots (kt). The NHC and JTWC best track data are available from their respective websites, as listed in the data availability statement at the end of the article.

All training dataset feature values are normalized by removing the mean and scaling to variance (i.e., mean = 0 and standard deviation = 1) using the Keras StandardScaler function. This scalar transformation is then applied to the validation and test data to ensure all values are scaled in the same fashion.

3. Methodology

Several different neural networks configurations are explored in this study. The first is a regression-based network, outputting a single MSW intensity estimate value within a continuous range. This network is extensively examined to determine the best number of hidden layers to employ in the ANN. In addition, two multi-classification networks are explored. These types of networks result in an output expressed probabilistically over a range of 5-kt MSW bins instead of a single value. The main difference in the two networks is in the input label data. A single label (SL) bin is used for the first network, meaning the label data are assigned to a single bin, while in the second network the label data are assigned to several bins representing a multiple label (ML) bin distribution, such as a Gaussian distribution, of MSW. Within each of the two categorical classification networks two independent experiments are conducted to explore different methods of handing the input and/or output intensity bins.

TABLE 1. Total number of ADT history file records in the five ocean basin regions and combined global training, validation, and independent test datasets.

Basin	Training	Validation	Test 2017	Test 2018
Atlantic	36 087	10 846	5188	4944
EastPac	35 270	10 007	3677	5143
WestPac	38 636	9359	5475	4334
SouthPac	29 833	10 852	3766	3688
NIO	7076	1988	566	1227
Global	146 902	43 052	18 672	19 336

In addition to the different neural network configurations noted above, an additional experiment was conducted to explore the use of four scene-type specific models with their own set of model features versus using one single model with a set feature list.

To focus the scope of this article on the impact of the AiDT on the ADT MSW estimates, the experiment methodology details of each of the five networks variations and two scene-type experiments are presented in the appendixes at the end of this article. Detailed analysis of each network and experiment variations are examined using the training and validation datasets in appendix A. An independent analysis of the various network experiments is performed on the 2017 test dataset and is provided in appendix B, with a final best model selected in appendix C. A schematic diagram of the final model is presented in Fig. 1 to illustrate the structure of the final model selected.

4. Results

The selection of the best MLP is outlined in appendix C, with the regression network being chosen. This network is referred to as AiDT-SV, for ‘‘AiDT single value,’’ for the remainder of the article. The following sections will focus on the AiDT-SV and its performance in a multitude of analyses to highlight the impact of the MLP on the ADT. Analysis will focus not only on basin-specific and global statistical comparisons of the performance of the two techniques, but will also highlight specific situations during a TC life cycle where the MLP network aids the ADT the most. Significance testing between the MLP and ADT statistical comparison will be presented to demonstrate independence of the datasets. A final comparison between the regression network and other satellite-based TC intensity estimation neural network models and algorithms is provided at the end of this section.

a. *Time averaging of independently derived intensity estimates*

To smooth out some of the inherent noise associated with single, independently derived intensity estimates produced by the AiDT-SV, a weighted time-averaging technique is applied to the intensity values, similar to the time-averaging technique used within the ADT algorithm (Olander and Velden 2019). The methodology weights the records between the current analysis

[1] As noted in Olander and Velden (2019), ADT estimates can be used in the generation of NHC and JTWC best track intensity estimates, especially outside of the North Atlantic where in situ aircraft measurements are not available, thus the values may not be truly independent. However, given the availability of other intensity sources (e.g. in situ aircraft measurements, scatterometer winds, ship/buoy measurements, microwave imagery, subjective Dvorak, other objective intensity methods, etc.), the TC experts will account for the respective strengths and limitations of each value to formulate the best possible ‘‘educated’’ best track intensity estimate.

TABLE 2. Comparisons between the best regression network 3-h time-weighted average (AiDT), unaveraged single value estimate (AiDT-SV), and the original ADT MSW intensity estimates for the five ocean basins and the global dataset for the independent 2017 test dataset. MAE is mean absolute error. RMSE is root-mean-square error and is highlighted in bold text. Units are in knots. Negative bias indicates MSW estimates are generally weaker than the NHC/JTWC best track estimates.

Network	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE
		Atlantic			East Pacific			West Pacific	
ADT	−0.91	9.50	**12.33**	−0.15	7.38	**9.44**	−1.87	8.47	**10.88**
AiDT-SV	0.49	6.89	**8.76**	−0.13	5.50	**7.04**	−0.60	6.02	**7.56**
AiDT	0.33	6.59	**8.44**	−0.13	5.30	**6.77**	−0.86	5.89	**7.35**
No. of records	5188	5188	5188	3677	3677	3677	5475	5475	5475
		South Pacific			North Indian			All basins	
ADT	2.71	8.43	**10.70**	5.03	7.51	**9.96**	−0.13	8.50	**10.98**
AiDT-SV	0.80	6.52	**8.29**	1.50	5.90	**8.15**	−0.18	6.26	**7.98**
AiDT	−0.98	6.27	**7.99**	1.04	5.33	**7.49**	−0.35	6.03	**7.70**
No. of records	3766	3766	3766	566	566	566	18 672	18 672	18 672

(t_0) and 3.0 h prior using a weight of $(3.0 - \Delta T)$ where ΔT is the time difference in hours from the current analysis time [e.g., $t_0 -$ 30 min (0.5 h) prediction is weighted 2.5, $t_0 - 1.0$ h is weighted 2.0, and so forth]. To assess this application, comparisons of the weighted time-averaged intensity estimates versus the independently derived, non-time-averaged, single-value intensity estimates (AiDT-SV) are shown for each of the five ocean basins during 2017 in Table 2.

Use of the 3-h time-weighted average estimate results in slightly improved MAE and RMSE versus non-time-averaged, single-value intensity estimate values (AiDT-SV) in all five basins and the combined global dataset, with an improvement of about 0.3 kt noted in the RMSE using the time-weighted average over the unaveraged individual values. Therefore, the time-weighted version of the regression network is used for the remainder of the paper, and is designated as AiDT.

Closer examination of the AiDT improvements shows notable bias improvements in four of the five individual ocean basins. Reductions of large ADT positive biases in the South Pacific and northern Indian Ocean basins of 1.73 and 3.99, respectively, were noted with a lowering of the RMSE values around 2.5 kt in each basin. While smaller improvements to the negative ADT bias values (absolute value of the bias was reduced) were seen in the Atlantic and west Pacific using the AiDT, these corresponded to much larger RMSE reductions of over 4.1 and 3.5 kt in the two basins, respectively. The east Pacific RMSE was reduced by 2.67 kt with a very small negative bias obtained by both the ADT and AiDT. The AiDT MSW estimates demonstrated a 30% improvement over the ADT MSW estimates in RMSE for the 2017 global dataset, lowering from 10.98 to 7.70.

b. Tropical cyclone categorical analysis

Additional analysis of the 2017 independent test results is performed to determine where the AiDT impacts and improves upon the original ADT algorithm the most. The ADT and AiDT estimate errors versus NHC/JTWC best track for the global dataset are broken down by storm intensity using the Saffir–Simpson hurricane intensity classification categories. These categories are tropical depression (TD), tropical storm (TS), and five hurricane categories (H1–H5). Two additional groupings include weaker hurricanes (H1 and H2) and major hurricanes (H3–H5). The results are presented in Table 3.

Examination of the differences between the ADT and AiDT statistics shows the largest RMSE impact in the TS and H1 (and H1–H2 combined) categories, where the ADT RMSE values are reduced nearly 4 kt in each category. The H2 RMSE errors are also reduced 2.5 kt. This is a notable improvement since the ADT struggles (exhibiting a low bias) in these intensity ranges since a central dense overcast (CDO) obscuring an eye structure in geostationary IR imagery is usually apparent during these ranges. This will also be discussed in the following section. It must be noted that while the AiDT did reduce the bias in the H1 category by almost 2 kt, the bias worsened in the TS and H2 categories where the negative bias increased slightly.

Eye features normally appear in IR imagery in category H2 hurricanes and stronger, and ADT eye scene RMSE values are typically smaller here. This is noted in the smaller RSME reductions of the AiDT versus the ADT in the H3–H5 categories with limited reductions of the bias, especially noted in the H4 and H5 categories.

Unfortunately, the AiDT could not rectify the large overestimate and underestimate biases noted in the ADT for TD and H5 category storms, respectively. H5 cases are harder for the ADT and AiDT to analyze due to an insufficient number of training cases for this category, combined with a shortage of distinguishing features. H5 TCs are often characterized by very small eyes of less than 10 km (called ‘‘pinhole’’ eyes) that may not be fully resolved by the IR imagers, especially on older geostationary platforms,[2] leading to intensity underestimates

[2] Spatial resolution of the IR window channel (LWIR) imagery (approx. 10.7 μm) used by the ADT has varied between 2 and 5 km since 1994. Current operational geostationary satellites *GOES-16*, *GOES-17*, and *Himawari-8* possess resolutions of 2 km, with *Meteosat-8* and *Meteosat-11* (along with non-operational *Meteosat-9* and *Meteosat-10*) having 3-km resolutions. Prior GOES (*GOES-8–15*) and MTSAT imagers had a 4-km resolution, while *Meteosat-5–7* and *GMS-5* exhibited a 5-km resolution.

TABLE 3. Statistical comparisons between the AiDT 3-h time-weighted average (AiDT), and original ADT MSW intensity estimates for the independent 2017 global test dataset broken down into intensity bins based on Saffir–Simpson classifications. MAE is mean absolute error. RMSE is root-mean-square error and is highlighted in bold text. Units are in knots. Negative bias indicates MSW estimates are generally weaker than the NHC/JTWC best track estimates.

		ADT			AiDT		
Saffir–Simpson intensity category	Sample size	Bias	MAE	RMSE	Bias	MAE	RMSE
TD < 35.0 kt	3519	5.34	6.58	**9.27**	5.96	6.28	**7.83**
TS 35.0–63.9 kt	9016	−0.37	8.54	**10.72**	−1.19	5.30	**6.79**
H1 64.0–82.9 kt	3001	−3.99	9.90	**12.87**	−2.09	6.45	**8.15**
H2 83.0–95.9 kt	1445	−2.03	10.02	**12.43**	−3.50	8.01	**9.92**
H3 96.0–112.9 kt	845	2.44	8.35	**10.22**	−0.44	6.21	**7.86**
H4 113.0–136.9 kt	607	−4.18	7.83	**10.15**	−4.14	6.35	**8.24**
H5 > 137.0 kt	239	−10.34	10.84	**13.44**	−10.02	11.00	**12.82**
H1–H2 64.0–95.9 kt	4446	−3.35	9.94	**12.73**	−2.55	6.96	**8.77**
H3–H5 > 96.0 kt	1691	−2.95	8.52	**10.71**	−3.41	6.94	**8.88**

in these situations. TDs are difficult due to the lack of organization in the IR cloud features for these weaker systems. While it is common for statistical models to struggle with the extreme conditions, further research is required to explore and improve satellite-based intensity estimation procedures in both of these disparate classifications.

c. ADT scene-type analysis

An additional breakdown of the ADT and AiDT statistics is carried out to assess how the AiDT performs with regards to the four main ADT scene types. These results are presented in Table 4, and further highlight the situational impact of the AiDT.

While a notable reduction in error is evident for the CDO and eye scene-type classifications using the AiDT, larger reductions occur for the curved band and shear scene-types that are more difficult cloud patterns for the ADT to analyze empirically, so the improvements are a meaningful advancement over the current ADT techniques. Curved band scene types also typically occur during TC formation stages (i.e., during the TD and TS intensity classification stages) that can have higher cloud pattern analysis uncertainties than with more developed storms, so application of the AiDT will greatly help the ADT results during this important period of the storm life cycle when the convective structure is still organizing.

For shear scene types, the ADT uses this classification more often as a TC encounters stronger vertical wind shear in the midlatitudes as it is transitioning to an extratropical (ET) system or dissipating. An intensity adjustment scheme is implemented in the ADT to modify the estimates during and after ET transition (Manion et al. 2015). But the results in Table 4 indicate the AiDT provides further improvement to intensity estimates during periods when a TC is encountering stronger environmental shear.

d. Examples of AiDT performance and behavior

Figure 4 shows scatterplots of all the ADT and AiDT intensity estimates for each of the five basins during the 2017 test. The scatterplots show the MSW estimates versus the NHC/JTWC best track MSW estimates and highlights the reduction of spread between the AiDT estimates and the corresponding ADT estimates. The AiDT reduces the outliers present in the ADT estimates, most notably in the lower MSW ranges.

Figures 5–8 show example intensity time series displays for selected TCs in each of the different basins. The examples are selected to highlight storms that exhibit large intensity changes as well as some of the AiDT impacts demonstrated in sections 4b and 4c. Particular attention should be paid to those portions of the time series where the AiDT model deviates from the ADT estimates. For example, during the Atlantic storms Jose and Maria, EastPac storms Fernanda and Kenneth, and WestPac storms Sanvu and Talim, the AiDT improves the ADT estimates during the dissipation stage of the storm where shear scene types are primarily used to provide the MSW estimates as the storm moves into higher latitudes and encounters more atmospheric shear.

TABLE 4. Statistical comparisons between the AiDT and original ADT MSW intensity estimates for the 2017 global test dataset broken down by ADT scene types. MAE is mean absolute error. RMSE is root-mean-square error and is highlighted in bold text. Units are in knots. Negative bias indicates MSW estimates are generally weaker than the NHC/JTWC best track estimates.

		ADT			AiDT		
ADT scene type	Sample size	Bias	MAE	RMSE	Bias	MAE	RMSE
Eye	2590	0.10	8.66	**11.03**	−1.43	6.55	**8.30**
CDO	7246	2.20	8.92	**11.18**	−0.67	6.53	**8.30**
Curved band	5670	−1.50	8.54	**11.17**	0.57	5.75	**7.27**
Shear	3166	−3.21	7.36	**10.12**	−0.41	4.95	**6.35**

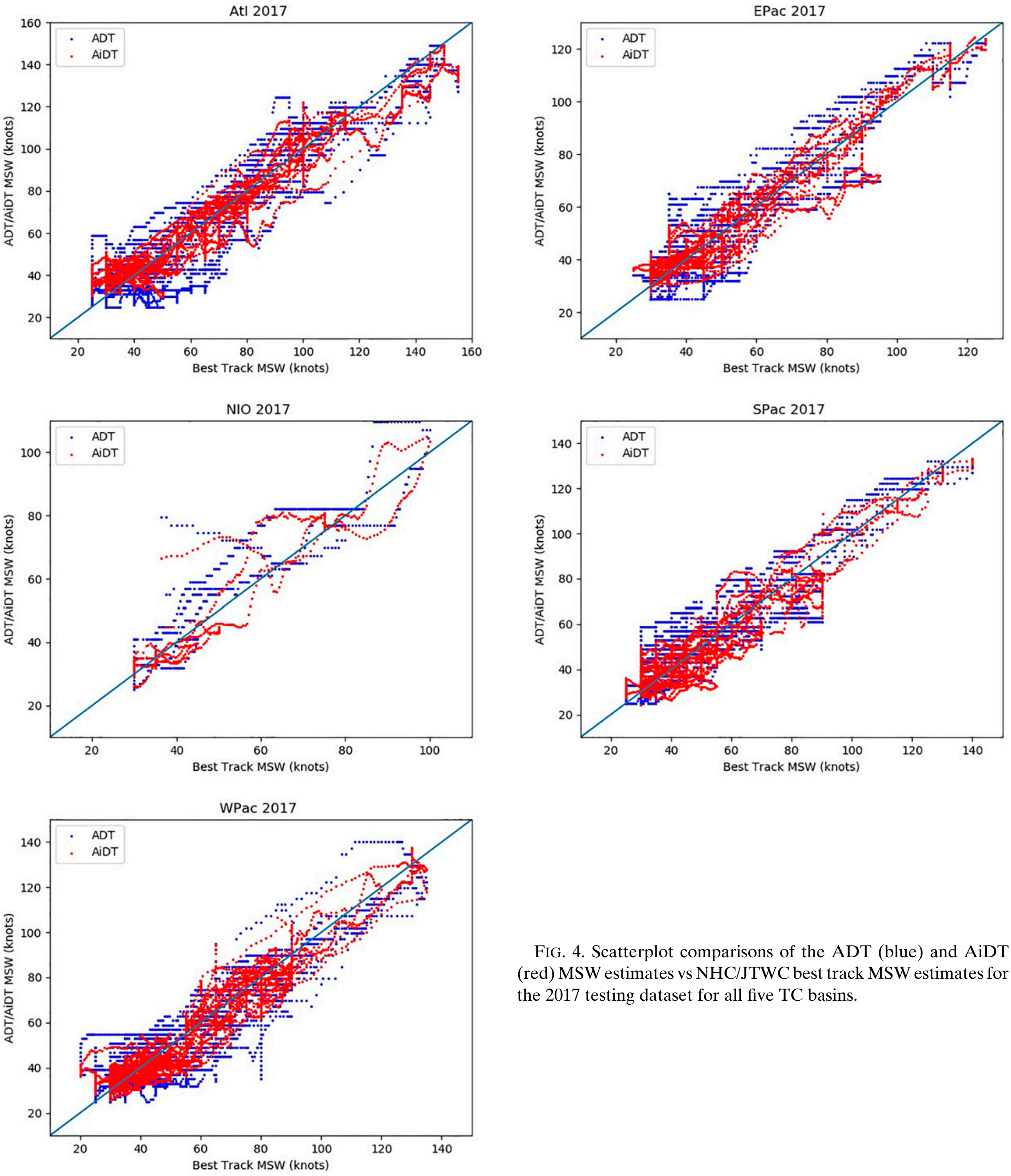


FIG. 4. Scatterplot comparisons of the ADT (blue) and AiDT (red) MSW estimates vs NHC/JTWC best track MSW estimates for the 2017 testing dataset for all five TC basins.

AiDT adjustments to the ADT MSW estimates during the formation stage are also highlighted in the examples where curved band scene types are primarily used. The Atlantic storms Harvey and Ophelia, EastPac storm Greg, and all four WestPac and SouthPac/NIO storms all demonstrate an AiDT deviation from the ADT toward the best track estimates during the early formation stage. These formation stage examples include weaker intensity periods when curved band scene types are primarily used by the ADT (in the TD and TS categories in section 4b) as well as H1 and H2 categories when the PMW adjustment is applied in conjunction with CDO scene types. As mentioned previously, the curved band and shear scene types have not been investigated in depth and still rely upon the original Dvorak technique techniques to provide a MSW

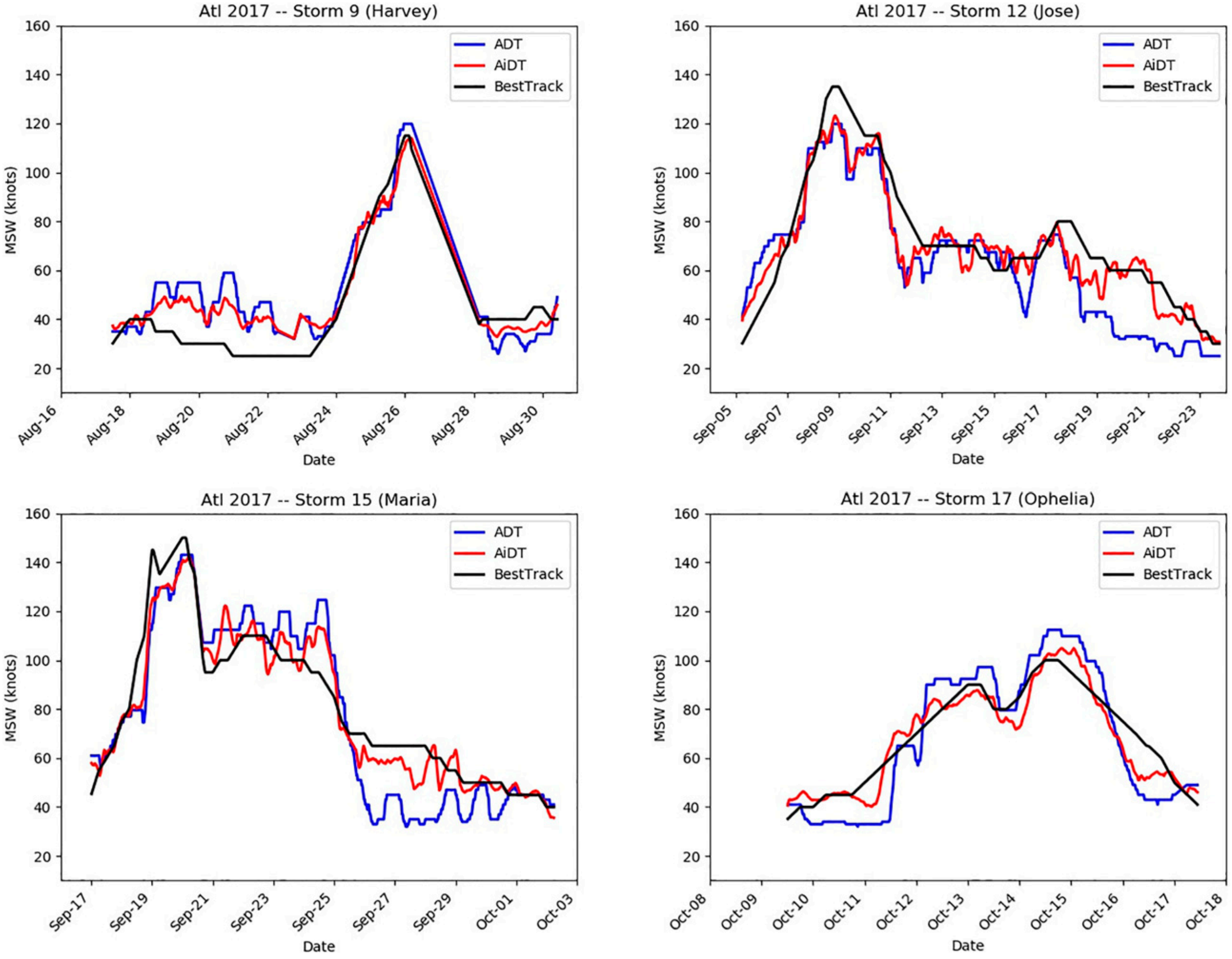


FIG. 5. Time series displays for four selected Atlantic storms in the 2017 independent test dataset. ADT (blue lines) and AiDT (red lines) are shown vs NHC/JTWC best track MSW (black lines). Units are in knots.

estimate in the ADT, so the impact of the AiDT during these scene types, as well as with CDO scene types, is quite promising.

In most cases only small deviations are noted between the AiDT and ADT estimates during the most mature TC stages of the storm life cycle (H3–H5 TC categories). These periods usually encompass TCs with well-defined eyes and convective structures, thus empirical methods to infer intensity (e.g., Dvorak, ADT) do quite well. However, in some cases, such as WestPac storms Talim and Lin and SouthPac storm Debbie, the AiDT estimates can adjust the ADT estimates upward or downward noticeably, demonstrating the power of machine learning models to identify and utilize additional information even in relatively well-behaved TC stages where the ADT MSW estimate methodology for eye scenes is typically reliable and stable. Some of these differences could be related to temporal changes associated with eyewall replacement cycles or spatial features that are not currently used in the eye analysis. While the AiDT can highlight periods to examine more closely to further improve the ADT algorithms in the future, it cannot state what the specific differences are. Performing machine learning feature selection analysis is one possible avenue to identify which specific features are important in each situation where the ADT and AiDT differ.

e. *Testing for AiDT robustness*

The robustness of the AiDT performance results in 2017 is tested on another independent sample of TC cases during 2018 for the five ocean basins analyzed previously. Table 5 presents the statistical analysis of the TC intensity estimates provided by the ADT and AiDT, and Fig. 9 illustrates a graphical comparison of the results between the two independent tests. Bias, RMSE and MAE are shown in Table 5 while only bias and RMSE are shown in Fig. 9 for clarity.

There are a similar total number of records (intensity analyses) examined in both years, with 600 more in 2018 than 2017. Only small deviations in AiDT RMSE are observed in four of the five TC basins, with changes of 0.46, −0.09, −0.29, and 0.24 kt in the Atlantic, east Pacific, west Pacific, and north Indian Ocean basins, respectively, from 2017 to 2018. Only the South Pacific basin results in 2018 deviate notably from 2017, with the RMSE increasing by 2.42 kt. There is

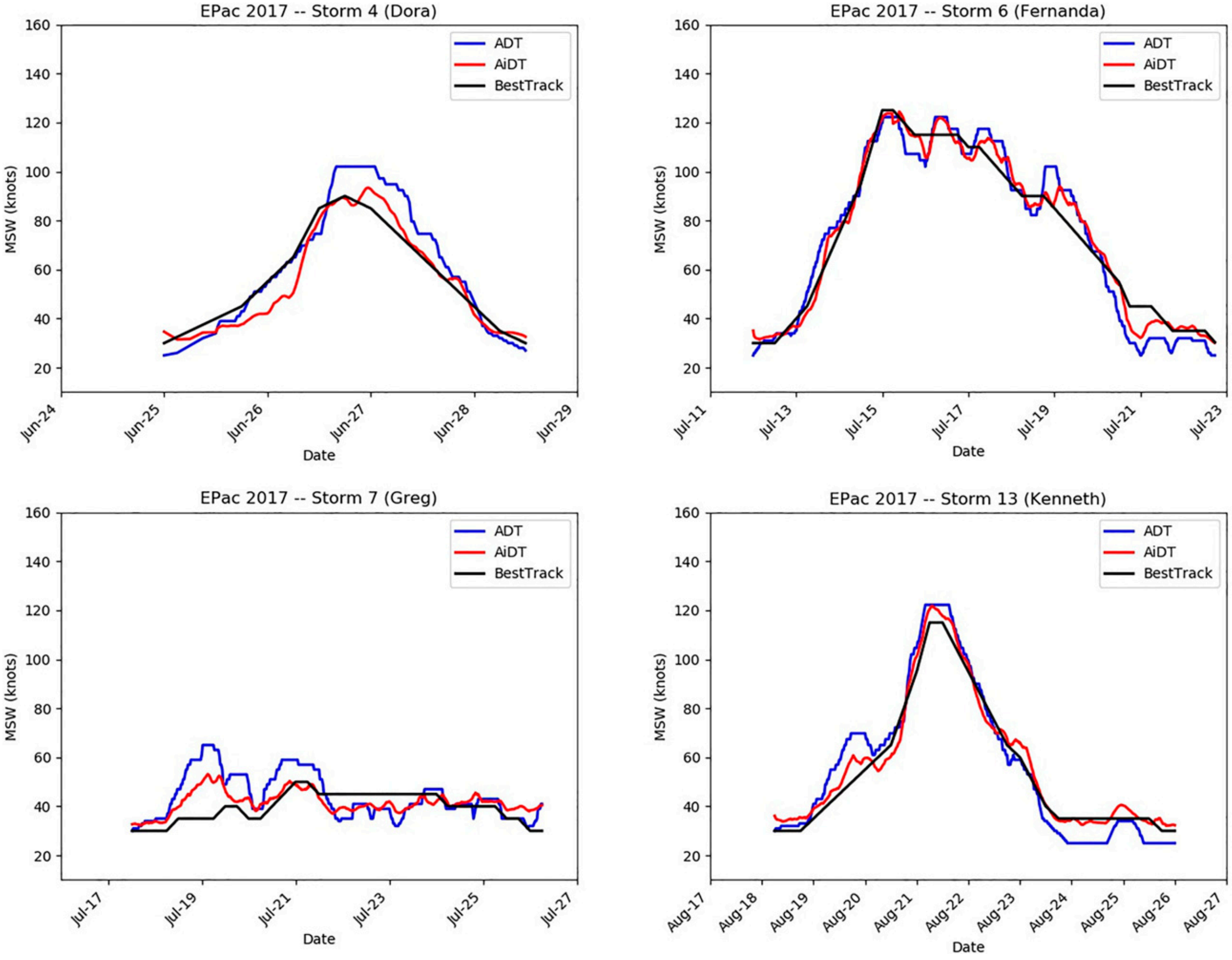


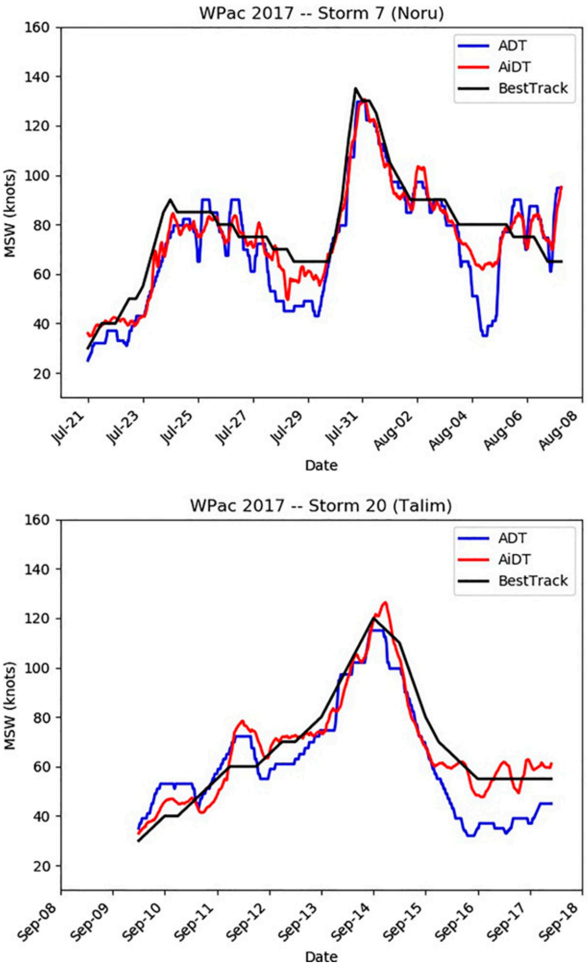
FIG. 6. As in Fig. 5, but for the 2017 east Pacific dataset.

also a corresponding increase in the ADT RMSE errors of 1.08 kt between the two years in the South Pacific along with a large AiDT bias shift from −0.98 in 2017 to −3.45 kt in 2018. A possible explanation of this increased negative bias is a significantly higher number of H3–H5 storms in the 2018 dataset than in the 2017 dataset (with all other categories being relatively the same). However, considering that the global RMSE only differs by 0.53 kt between the 2017 and 2018 independent sample tests, the robustness of the AiDT performance is clearly demonstrated. The AiDT MSW estimates represented a 23% improvement in RMSE over the corresponding ADT MSW estimates for the 2018 global dataset, lowering from 10.73 for the ADT to 8.23 for the AiDT.

f. Statistical significance testing

Significance testing is performed by examining the *p* value from a paired Student's *t* test using bootstrap sampling to determine whether the ADT and AiDT errors are significantly separated. This analysis was performed separately for the global and five individual basin datasets for 2017 (section 4a) and 2018 (section 4e) as well as the 2017 tropical cyclone category (section 4b) and ADT scene type (section 4c) datasets. The bootstrap sample sizes for each of these datasets were chosen to remove the temporal autocorrelation in the 30-min data. For this we used the decorrelation time for ADT estimates determined in Kossin et al. 2020, where it was found that the estimates decorrelate between 12 and 18 h. In that study, which used 6-hourly ADT estimates, the degrees of freedom were reduced by a conservative factor of 3 (the actual reduction factor was 2.7). Here, for the 30-min data, the degrees of freedom in the Student's *t* test would need to be reduced by a factor of 32 (i.e., 2.7 × 6 h/30 min). For each dataset described above, the bootstrap sample size was specified as *N*/32 where *N* is the size of the dataset. We then form 10 000 bootstrap samples and compute the bias and RMSE of each for the ADT and the AiDT data. The distributions of bias and RMSE are normal and represent independent errors.

Plots of the bias and RMSE probability density functions (PDF) are generated to visually compare the ADT and AiDT bootstrap datasets. Statistical significance for the 2017 and 2018 basin statistics as well as the 2017 categorical and scene-type statistics are determined by deriving the *p* value for the bias and RMSE distributions, with all *p* values found to be less than 0.01, thus the ADT and AiDT errors examined in sections 4a–d



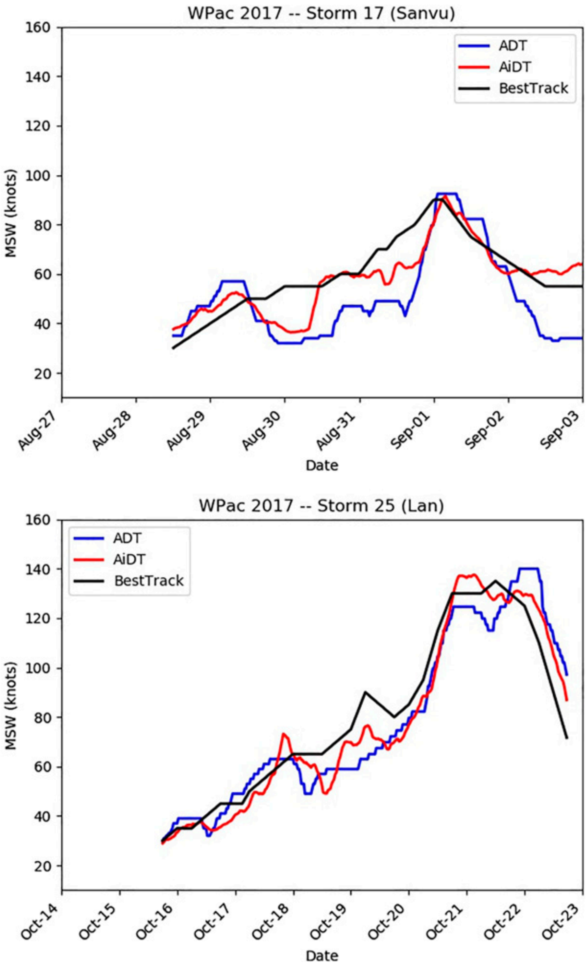


FIG. 7. As in Fig. 5, but for the 2017 west Pacific dataset.

are determined to be significantly separated with greater than 99% confidence.

Figure 10 displays the bias and RMSE PDF distributions for the 2017 ADT scene-type datasets, as discussed previously in section 4c. The 2017–18 basin PDF comparisons, as well as the categorical PDF comparisons, available at https://tropic.ssec.wisc.edu/real-time/adt/AiDT/pdf, show similar separation to those in Fig. 10. It is noted that the AiDT PDF bias and RMSE distributions have less overall bias/RMSE spread (x axis) and corresponding increase in PDF maxima (y axis) than the corresponding ADT distributions, meaning the AiDT estimates have greater overall accuracy than the ADT estimates. It is also noted again that the more notable PDF distribution changes occur in the curved band, shear, and CDO scene types (along with the previously discussed changed to the bulk bias and RMSE value reductions obtained with the AiDT over the corresponding ADT values).

g. *Comparisons with other satellite-based methods*

Table 6 lists a number of recently published satellite-based TC intensity estimation models and algorithms, including experimental deep learning/neural network methods as well as more traditional methods. The published performance accuracies are shown for comparison with the AiDT presented in this study.

As can be seen, the performance of the AiDT is very competitive with or superior to all of the methods, even many of the more sophisticated DL/CNN models which employ a variety of satellite data image sources. Most methods rely upon traditional geostationary imagery, primarily focusing on the infrared window channel cloud top temperature field around the TC center position. The Lee et al. (2019) study uses additional geostationary channels such as shortwave IR and water vapor, while the Wimmers et al. (2019) method uses passive microwave imagery sensitive to ice scattering below the cloud tops. Other techniques, like the Dvorak technique, ADT and DAV-T are not DL/CNN techniques, instead they rely on other methods to interrogate the satellite imagery. SATCON (Velden and Herndon 2020) is a weighted consensus of intensity estimates from several independent objective methods that include the ADT. This approach is being used operationally at several global TC analysis centers and provides accuracies that are better than its input members.

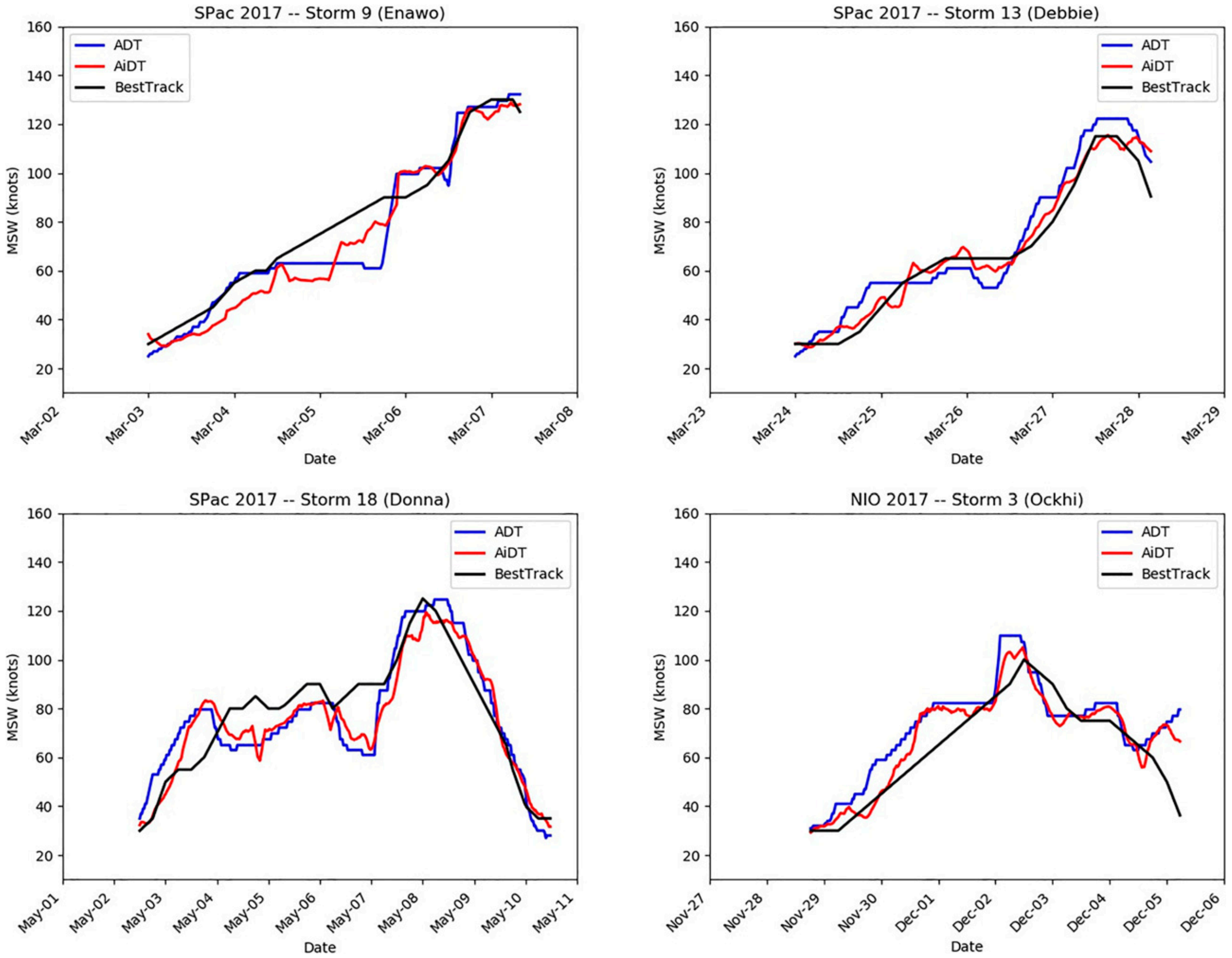


FIG. 8. As in Fig. 5, but for the 2017 South Pacific and north Indian Ocean datasets.

While it is difficult to directly compare all of these techniques since they are not tested on homogeneous TC samples, it can be inferred that the AiDT compares favorably with the two most accurate DL/CNN techniques, the Lee 2D3 model and Chen CNN-TC model. A specific comparison of the Chen results in the northwest Pacific during 2017 with the AiDT results obtained during the same period (Table 2) shows that Chen obtained an RMSE of 8.39 kt while the AiDT obtained a RMSE of 7.35 kt. While this specific comparison may not be statistically significant since the models

TABLE 5. Statistical comparisons between the AiDT and original ADT MSW intensity estimates for the five ocean basins and global dataset for the independent test sample of TC cases in 2018. MAE is mean absolute error. RMSE is root-mean-square error and is highlighted in bold text. Units are in knots. Negative bias indicates MSW estimates are generally weaker than the NHC/JTWC best track estimates.

Network	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE
		Atlantic			East Pacific			West Pacific	
ADT	−1.23	9.56	**12.16**	0.06	7.42	**9.64**	−0.21	7.17	**9.27**
AiDT	0.34	7.13	**8.90**	−0.80	5.30	**6.68**	0.34	5.68	**7.06**
No. of records	4944	4944	4944	5143	5143	5143	4334	4334	4334
		South Pacific			North Indian			All basins	
ADT	−0.58	9.44	**11.78**	3.56	8.14	**10.48**	−0.01	8.34	**10.73**
AiDT	−3.45	7.87	**10.41**	−1.23	5.84	**7.73**	−0.79	6.38	**8.23**
No. of records	3688	3688	3688	1227	1227	1227	19 336	19 336	19 336

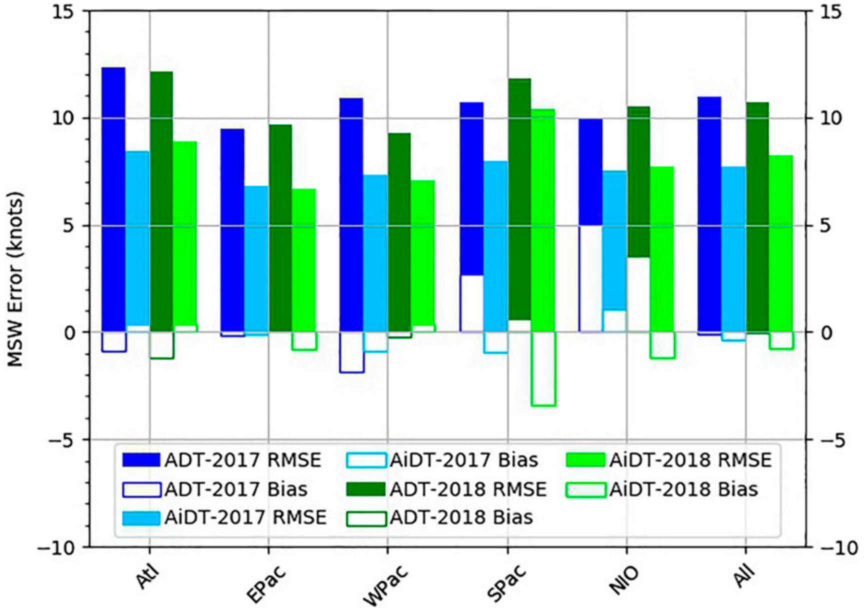


FIG. 9. Comparison of performance statistics for the ADT and AiDT covering the five different ocean basins and global dataset (All) for the two independent test data samples in 2017 and 2018. MSW intensity estimate bias and RMSE are presented. MSW units are in knots.

were not homogeneous, it demonstrates the potential of simple MLP model enhancements to existing methods (i.e., ADT) versus more computationally expensive and time-consuming full DL/CNN image analysis models. This result may be surprising but emphasizes the robustness of the ADT analysis techniques and related output features stored in the history file. As mentioned previously, many of the ADT analysis techniques have undergone extensive analysis over the years, but a simple MLP model was able to obtain additional information from these feature parameters not previously recognized and can help guide future research efforts to improve both the ADT and AiDT.

5. Summary and future directions

This study examines the potential to employ machine learning enhancements to an existing proven algorithm (ADT) that estimates the intensity of tropical cyclones from satellite

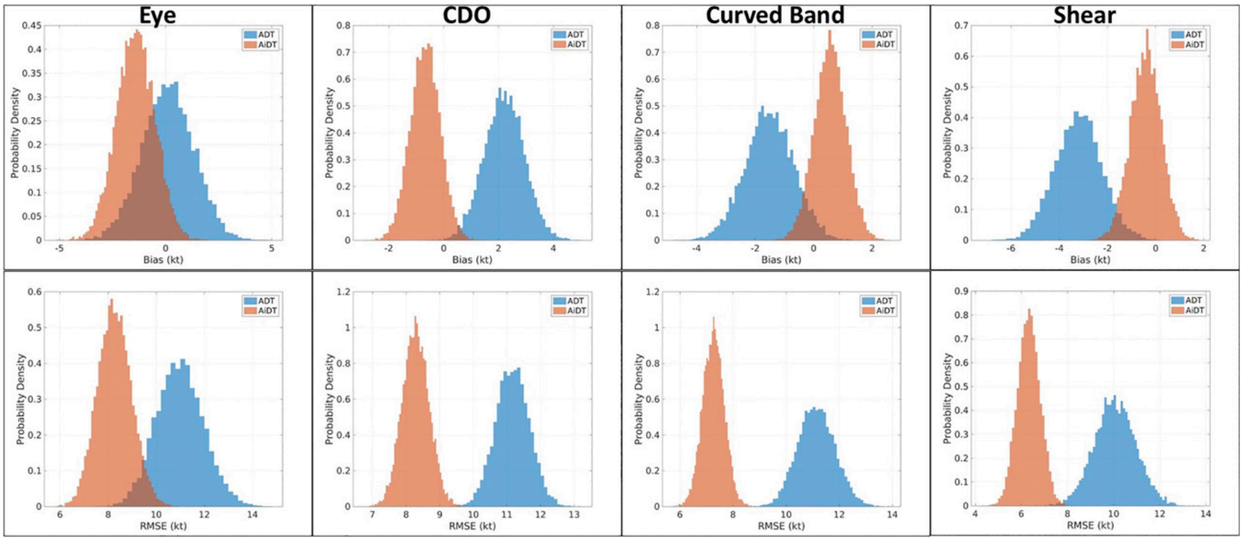


FIG. 10. Comparison of ADT (blue) and AiDT (orange) probability density function (PDF) bias and RMSE distributions for MSW intensity estimates for the 2017 global test dataset broken down by ADT scene types, as shown in Table 4. MSW units are in knots.

TABLE 6. Comparison of published TC intensity estimation models and algorithms.

Technique	Method	Data type	Inputs	Region	Dataset years	MSW RMSE (kt)
Dvorak (Dvorak 1975, 1984)	Empirical	Geo	IR, VIS	Global	1970s–80s	10–15
DAV-T (Ritchie et al. 2014)	Statistical	Geo	IR (10.7 μm)	North east/west Pacific	2007–11	12.9–13.4
SATCON (Velden and Herndon 2020)	Statistical Ensemble	Geo Leo	IR (10.7 μm) PMW (various, based on method)	Global	2006–14	9.0
ADT (Olander and Velden 2019)	Statistical Empirical	Geo Leo	IR (10.7 μm) PMW (eye score)	Global	2017	10.98
DeepMicroNet (Wimmers et al. 2019)	CNN	Leo	PMW (37 GHz, 85–92 GHz)	Global	2007, 2012	9.6–14.3
CNN-TC (Chen et al. 2019)	CNN	GeoLeo	IR (10.7 μm) WV (6.7 μm) PMW (rain rate)	Global	2017	8.39
Pradhan model (Pradhan et al. 2018)	CNN	Geo	IR	Global	1999–2014	10.18
2D3 (Lee et al. 2019)	CNN	Geo	IR1 (10.7 μm) IR2 (12.0 μm) WV (6.7 μm) SWIR (3.9 μm)	Northwest Pacific	2011–16/17	8.32
TCIENet (Zhang et al. 2020)	CNN	Geo	IR (10.7 μm) WV (6.7 μm)	Northwest Pacific	2017	9.97
VGG19 (Combinido et al. 2018)	CNN	Geo	IR (10.7 μm)	Northwest Pacific	1996–2016	13.23
AiDT	MLP	Geo Leo	IR (10.7 μm) PMW (eye score)	Global	2017, 2018	7.70–8.23

data, to assess whether superior performance can be achieved. It is found that various MLP models can augment the ADT by interrogating features that are output from the image processing, resulting in improvements in the accuracy of TC intensity estimates over the ADT itself. We found that a regression-based network, which derives an estimate in a continuous range of values, is slightly superior to multiclassification network models that derive estimate probabilities in a set series of range classes. The AiDT models are better when derived using the entire ADT feature list instead of a scene-specific subset of ADT features. In addition, the use of a network derived using the entire global combined dataset versus a set of individual basin-specific networks also produces superior results. The AiDT models, most importantly, improve TC intensity estimates for situations when the ADT (and Dvorak technique) struggles. Improvements of 30% and 23% were noted in the global AiDT MSW RMSE versus the ADT MSW estimates for the 2017 and 2018 independent tests, respectively, with the advantage being highly significant.

AiDT model TC intensity estimate accuracies and recently documented accuracies from other satellite-based neural network models (including CNN models) compare favorably. The AiDT is much easier and computationally cheaper to modify and run versus direct image interrogation models that can take considerable time to set up and execute, which makes it attractive for real-time application and potential operational implementation. Also, science upgrades and analysis modifications can be accomplished without requiring significant computational expense since adding new feature values to and recalculating the model is simple and not time consuming. Finally, the AiDT preserves much of the operational familiarity and heritage of the ADT, and Dvorak technique, while providing improved MSW estimates since it still relies upon the ADT analysis as input.

Future investigations will focus on optimizing the feature selection process that feeds the AiDT to determine which ADT history file features are most impactful and which can possibly be removed. Integration of other features, such as storm information from the ARCHER (Wimmers and Velden 2016) algorithm could be explored to augment the current ADT history file features utilized in this study. Finally, integration of the higher-precision AiDT results into the SATCON (Velden and Herndon 2020) model should improve the performance of that TC intensity estimate consensus algorithm, which currently relies upon the ADT.

Acknowledgments. We wish to thank the University of Wisconsin–Madison Space Science and Engineering Center for its support of this project. The authors also wish to thank the anonymous reviewers for their comments and suggestions during the review process. These have led to a much improved manuscript.

Data availability statement. Tropical cyclone best track datasets are available from The NOAA/National Hurricane Center (NHC) and Joint Typhoon Warning Center (JTWC) public webpages at https://www.nhc.noaa.gov/recon.php and

Table A1. ADT parameters (features) included in each of the five different scene models (ALL, eye, CDO, curved band, and shear). Checkmarks indicate whether the feature is used in the model. An asterisk indicates the feature is scene-type dependent (determined using the eye and cloud scene ID features) in the ALL scene-type model exclusively. C/W is coldest − warmest, PMW is passive microwave, CDO is central dense overcast, and FFT is fast Fourier transform.

Feature	ALL	Eye	CDO	Curved band	Shear
Raw T#	✓	✓	✓	✓	✓
Adjusted raw T#	✓	✓	✓	✓	✓
Final T#	✓	✓	✓	✓	✓
CI#	✓	✓	✓	✓	✓
Eye temperature	✓	✓	✓	✓	✓
Cloud temperature	✓	✓	✓	✓	✓
C/W temperature	✓	✓	✓	✓	✓
Latitude	✓	✓	✓	✓	✓
Sin of longitude	✓	✓	✓	✓	✓
Cos of longitude	✓	✓	✓	✓	✓
Viewing angle	✓	✓	✓	✓	✓
Eye FFT	✓	✓	✓	✓	✓
Cloud FFT	✓	✓	✓	✓	✓
Eye scene ID value	✓	✓			
Cloud scene ID value	✓		✓		
Eye std dev	✓	✓			
Cloud symmetry	✓	✓	✓		
Curved band value	✓			✓	
Curved band amount	✓			✓	
C/W temperature distance	✓			✓	
PMW eye score	✓		✓		
Extratropical flag	✓			✓	✓
Subtropical flag	✓			✓	
Eye size (2/eye size)	✓*	✓			
CDO size	✓*		✓		
Shear distance	✓*				✓
Total No.	26	17	17	18	15

https://www.metoc.navy.mil/jtwc/jtwc.html?best-tracks, respectively. The satellite data were obtained from the University of Wisconsin—Madison/Space Science and Engineering Center (SSEC) using the Man computer Interactive Data Access System (McIDAS) proprietary software which prohibits sharing the data publicly. The advanced Dvorak technique (ADT) version 9.0 history files may be obtained upon request from the lead author of this article. The ADT Users Guide may be obtained from https://tropic.ssec.wisc.edu/misc/adt.

APPENDIX A

Network Training and Validation

a. Scene-type experiments

Five separate models are first developed according to the ADT pattern (scene)-type categories of eye, central dense overcast (CDO), curved band, shear, and all. The first four models are applied to individual ADT scene types only (designated MIX), while the last "all" model uses all available ADT history file records regardless of the scene type, with scene type being a training input feature to the model (designated ALL). This is done to determine whether one single overall model or four separate scene-type models produce more accurate intensity estimates.

Different ADT history file parameters, sometimes referred to as predictors or "features" in machine learning nomenclature, are used in each of the five scene-type models (Table A1). Using different features in each model might seem unconventional, but it allows for more targeted information to be used in each scene-type model while removing as much unnecessary information as possible. One ADT history file parameter stores information specific to the scene type designated during the ADT processing and will change meaning based on the scene type. This single "shared" ADT history file parameter stores the CDO diameter for CDO scene types, the eye size for eye scene types, and the shear distance from center for shear scenes. In addition, for some scene types certain parameters are not measured and are designated with a missing value in the ADT history file. Such examples illustrate why select features are used for the different models. The list of features for the five different scene-type models is shown in Table A1. For detailed information regarding the features listed in Table A1, please refer to the ADT Users Guide (Olander 2021).

For the ALL scene-type model, the values of the three features in Table A1 designated with an asterisk following the check mark correspond to the single shared parameter outlined above. The model will assign the feature value based on the scene type of that history file record and will also fill the "missing" features with values that are reasonable and valid.

For the "eye size" feature during noneye situations or eye scenes with a missing ADT value, the feature is assigned a value of zero. For valid eye scene situations, this feature is set to 2.0/(eye size), where eye size is the diameter of the eye. This equation will result in values between 0 and 1 (assuming maximum IR imagery resolution is 2 km), with smaller eyes near 1 (meaning an eye diameter at or near 2 km, which are typically associated with more intense storms) and larger/no eye situations near/at 0 (meaning much larger eye diameters typically associated with less intense storms). This "normalized" value is also used in the eye scene model. "CDO size" feature values for non-CDO scenes are set to 170 km, which is the average CDO size for the training dataset. For nonshear scenes, the "shear distance" feature value is set to 0.

The longitude value is replaced in the scene-type models with two separate features, the sine and cosine of the longitude, in order to maintain meridional continuity. Finally, missing or negative PMW eye score values, derived during ADT processing when input PMW imagery is available, are given a value of zero. More information about these ADT history file parameters can be found in the ADT Users Guide (Olander 2021).

b. *Regression network*

The first network investigated is a regression-based network utilizing a varying number of hidden multineuron layers (including the single neuron output layer) producing a continuous range of output values. An "rmsprop" optimizer[A1] and "mean squared error" (MSE) loss function are used to compile the network.

An experiment is performed to focus on determining the best configuration of hidden layers for the model. Twelve different configurations of hidden layers are examined: from one to six hidden layers using 32 neurons, a three-layer network with 32/64/32 neurons, a five layer network with alternating 32 or 64 neurons, and four additional single layer networks with either 8, 16, 64, or 128 neurons. A final single-neuron output layer will contain the final MSW estimate for the regression network. A batch size of 150 with 250 epochs is used for all regression networks. A batch is a collection of training samples that are used for each iteration during the model derivation process before the model is tuned and another batch is examined. The number of iterations for one complete examination of the training set, called an epoch, will depend upon the size of the training set being utilized; the larger the batch size the smaller the number of iterations. The error of the model (model estimate versus the label data) is then derived for each epoch and the process is repeated for another pass through the training data, which are randomly shuffled for each epoch. The model is also applied to the validation dataset at the end of each epoch, as the model is being derived, to examine if the model behavior with the validation data are similar to the behavior exhibited with the training data. The errors obtained from each epoch during the training process will typically level out as less adjustment to the model is needed as more data are examined. However, if the model is tuned to this training set too much, the errors obtained during the validation may actually increase as more epochs are examined. This is known as model overfitting and must be avoided.

TABLE A2. Regression network training and validation mean-squared error (MSE) accuracy scores for the AllBasin model using the ALL scene-type model. Network configuration is listed as (number of hidden layers)–(number of neurons) in the network column.

Network (layers-neurons)	Training accuracy	Validation accuracy
1–8	76.85	85.98
1–16	72.47	84.85
1–32	70.24	82.74
1–64	67.28	84.78
1–128	62.95	89.43
2–32	64.26	84.61
3–32	56.07	92.11
3–mix	51.38	101.06
4–32	53.51	97.87
5–mix	36.69	115.98
6–32	48.06	105.25

Comparing the training and validation MSE values can identify overfitting. Table A2 shows the training and validation accuracy values obtained for the last (250th) epoch for the AllBasin/ALL scene-type models using a MSE model loss/accuracy metric. The AllBasin and ALL scene-type models are chosen for this analysis in order to reduce any basin or scene-type specific biases into the selection of the final model and make the data as homogenous as possible. Figure A1 shows a plot of the training and validation accuracy for 6 of the 12 models, with the final plotted point on each graph being the corresponding value in Table A2

As shown in Table A2 and Fig. A1, the networks with fewer hidden layers tend to be the most accurate, with networks more hidden layers having much higher MSE losses, typically due to overfitting of the networks. For the higher layer networks the training MSE values are quite low, indicating that the network has modeled the features to accurately match the label data. When this network is used on the validation data, however, the resulting intensity estimates are much different than the corresponding label data values, resulting in MSE values that are significantly higher. Thus the model is considered overfit to the training dataset. This behavior is noted in Fig. A1 with the lower layer network MSE errors (such as the three 1 layer models and 2 layer model) tending to flatten as the number of epochs increases, but higher layer models (the 4 and 6 layer models) have their validation errors minimize at lower epoch values and then increase as the epochs, and training data input into the derived model, increases.

The single hidden layer networks are typically best and display the least amount of overfitting, with the 1 layer, 32-neuron network being most accurate for the AllBasin, ALL scene-type models. It will represent the regression analysis in appendix C.

[A1] Unpublished learning rate model proposed by G. Hinton in Lecture 6e of his online Coursera class, which is available at https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf.

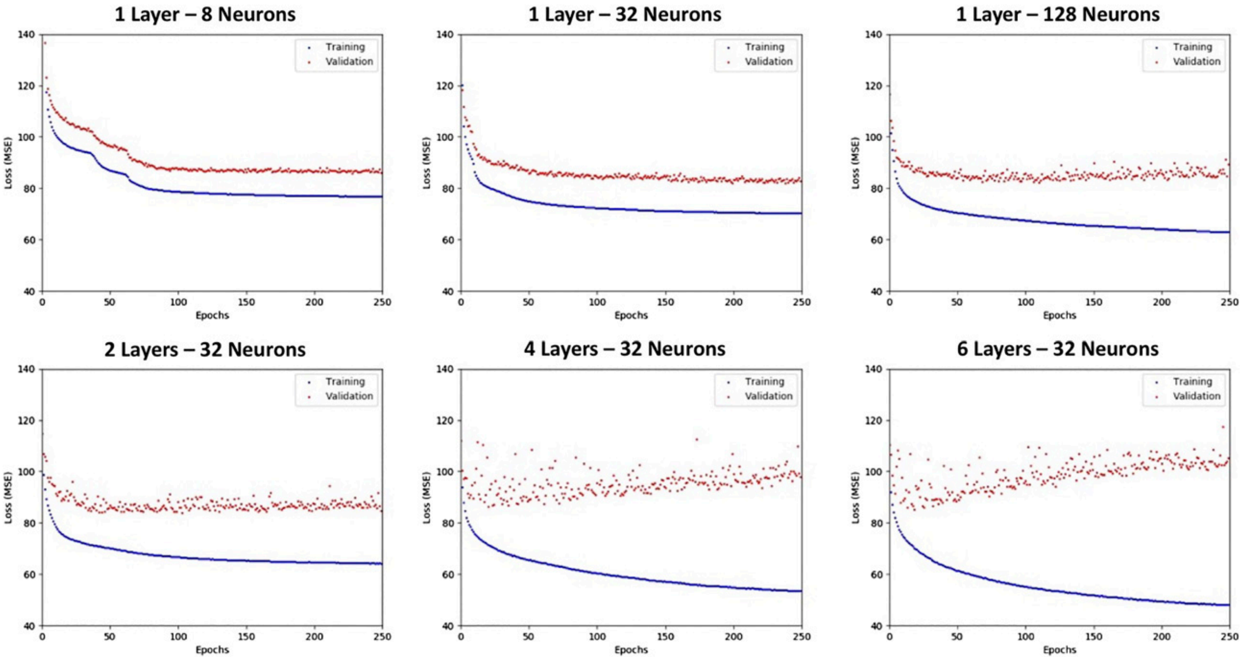


FIG. A1. Training (blue) and validation (red) loss plots for various regression models listed in Table A2. The number of layers and neurons are listed at the top of each plot.

It must be noted that other techniques to minimize overfitting exist, such as early stopping and learning rate reduction; however, the method in this study was employed for ease of model comparison between different network configurations (presented in the following sections). These methods can/will be explored in future experiments to investigate their impact on model performance.

c. *Multiclassification networks*

A second type of machine learning network is examined: multiclassification. Unlike the regression network outlined above which has a single neuron output layer containing a predicted value, these networks possess a multiple neuron output layer containing a probability distribution. The input labels are also restructured into bins of intensity value at 5-kt intervals, similar to Wimmers et al. (2019), with 32 bins defined, starting at 25–30 kt and ending at 180–185 kt. The input best track intensity label value is assigned to either a single intensity bin for the single label network, with a value of 1.0 in that single bin, or a range of intensity bins for the multilabel network, representing a probability distribution, summing to 1.0 for all 32 bins. The former, referred to as one-hot encoding, represents a single-label (SL), multiclass classification problem and uses a sparse categorical crossentropy loss function, while the latter is known as a multilabel (ML), multiclass classification problem and uses a categorical crossentropy loss-function. Experiments using both networks will be discussed in the following sections.

Both the SL and ML multiclassification networks will output a final layer of 32 neurons representing a probability distribution over the 32 output MSW bins. Each of the 32 neurons contains a likelihood, or percentage, that the final estimate is in that specific bin, from 0 to 1. One can either use the output bin with the maximum likelihood value as the predicted intensity, or a weighted-value using the specific likelihood values contained in each 5-kt output bin to derive a single value estimate. In this study two output layer weighting schemes will be explored for both the SL and ML networks to derive a single output MSW estimate from the 32 neuron distribution contained in the final layer. The first will derive a weighted intensity value using the bin with the maximum likelihood value and the two adjacent bin likelihood values (referred to as 3Bin), while the second will use an average of all of the likelihood bin values to calculate the intensity estimate value (AllBin). The equations for both are listed below, where i is the bin number, x is the maximum likelihood bin (for the 3Bin equation), W is the likelihood value for bin i, and M is the mean value for the 5-kt bin. This will produce a single MSW estimate for each method:

$$\text{3Bin} = \frac{\sum_{x-1}^{x+1} W_i \times M_i}{\sum_{x-1}^{x+1} W_i} \quad \text{or} \quad \text{AllBin} = \frac{\sum_{1}^{32} W_i \times M_i}{\sum_{1}^{32} W_i}.$$

Both the SL and ML networks utilize a softmax activation in the final layer to produce the probability distribution

As with the regression network, determination of the proper network configuration should be explored. This is more difficult using intensity bins for the verification label data since we are not comparing discrete values, but instead distributions of likelihood. Accuracy in categorical models is determined by

TABLE A3. SL and ML multiclassification network training and validation mean-squared error (MSE) accuracy scores for the AllBasin model using the ALL scene-type model. Network configuration is listed as (number of hidden layers)–(number of neurons) in the network column.

Network (layers-neurons)	SL train accuracy	SL validation accuracy	ML train accuracy	ML validation accuracy
1–8	0.280	0.269	0.278	0.270
1–16	0.291	0.266	0.286	0.262
1–32	0.308	0.256	0.302	0.260
1–64	0.325	0.258	0.323	0.252
1–128	0.354	0.252	0.350	0.237
2–32	0.341	0.246	0.324	0.241
3–32	0.333	0.228	0.349	0.239
4–32	0.358	0.239	0.355	0.238

deriving the difference between the SL input label data bin or ML input label data distribution with the output layer probability bins. Thus, any small deviation in the 32-bin output layer probability distribution can change a match to a nonmatch, thus affecting the "accuracy" of the model. That being said, we will examine the network validation accuracy for a limited set of network configurations consisting of the four single hidden layer networks with the five different number of neurons (8, 16, 32, 64, and 128) as well as the 2, 3, and 4 hidden layer networks using 32 neurons. Again, the number of hidden layers referred to in this this section does not include the final 32 neuron output layer, only the number of layers prior to that final output layer (which is uniform between all networks). Table A3 shows the training and validation accuracy values obtained for the last (250th) epoch for the AllBasin and ALL-Scenes model, as with the regression network shown in Table A2, using "sparse_categorical_accuracy" and "accuracy" metrics for the SL and ML networks, respectively.

Examination of the SL and ML network training and validation accuracy results in Table A3 shows that the single layer networks possess the highest validation accuracies, with the 8 and 16 neuron networks being the most accurate for the SL network, with the 32 neuron network very close to the 16 neuron accuracy in the ML network. It is interesting to note that the SL and ML networks both obtain similar accuracy characteristics.

Selection of the categorical network to be used is not as apparent as with the regression network. While the 1 layer/8 neuron network has the highest validation accuracy scores for both the SL and ML networks, the values are not notably higher than those obtained by the 16 and 32 neuron networks. Due to this fact, and in order to provide some consistency between the different networks in this paper, the 1 layer/32 neuron network will be utilized for the SL and ML networks also.

1) SINGLE-LABEL CATEGORICAL NETWORK

Two experiments are performed with the SL model; one using a label class-weighting scheme defined in the model fit (class_weight input parameter in Keras model.fit) and one without a label class-weighting scheme. To be clear, the label value being input is still assigned to single bin value. The purpose of a class weight is to normalize for the number of best track/label samples in each intensity bin across the entire training dataset. This is done in order to provide more/less weight to underrepresented/overrepresented bins so the distribution is not skewed heavily toward the overrepresented label bins For example, in the Atlantic training label dataset there are about 36 000 best track data points, with a majority of those intensities being tropical depression and tropical storm strengths (less than 65 kt). Bins 1–4 have over 4000 instances in each bin. However, higher intensity bins typically have less than 100, so this leads to a model that is pushed to err toward lower intensity estimates. A simple $1/N$ weighting scheme is used here (bins with zero records are given a 1.0 weight to avoid a divide by zero error). The non-class-weighting test will be referred to as "NoCW," with the $1/N$ class weighing scheme test being referred to as "1NCW."

2) MULTI-LABEL CATEGORICAL NETWORK

As with the SL network analysis, two experiments are performed using the ML network and focused on two types of weighting schemes when characterizing the best track label data. Unlike the SL analysis, which uses a one-hot, single-label classification identifying a single 5-kt bin where the best track data are placed, the ML network is a multilabel classification using a range of label values to characterize the label classification data. This is done in order to model any inherent errors associated with operational best track MSW datasets. One method, as discussed and utilized in Wimmers et al. (2019), characterizes the best track MSW label data using a Gaussian distribution centered on the best track value. This typically yields a center bin weight of 0.34, with values of 0.23 and 0.10 in the adjacent bins on either side of the center. The values are adjusted if the best track value lies within bins 0, 1, 30, and 31 in order to renormalize to a sum of 1. This method will be labeled "GaussD." A second experiment will give a value of 0.6 to the best track MSW label bin and 0.2 to the adjacent bins (again adjusted if the center bin is either bin 1 or 32), and will be labeled "262D."

APPENDIX B

Independent Network Testing

All TCs from 2017 are used as the primary independent testing dataset for the models under investigation. Performance statistics are presented by individual TC basin in order to

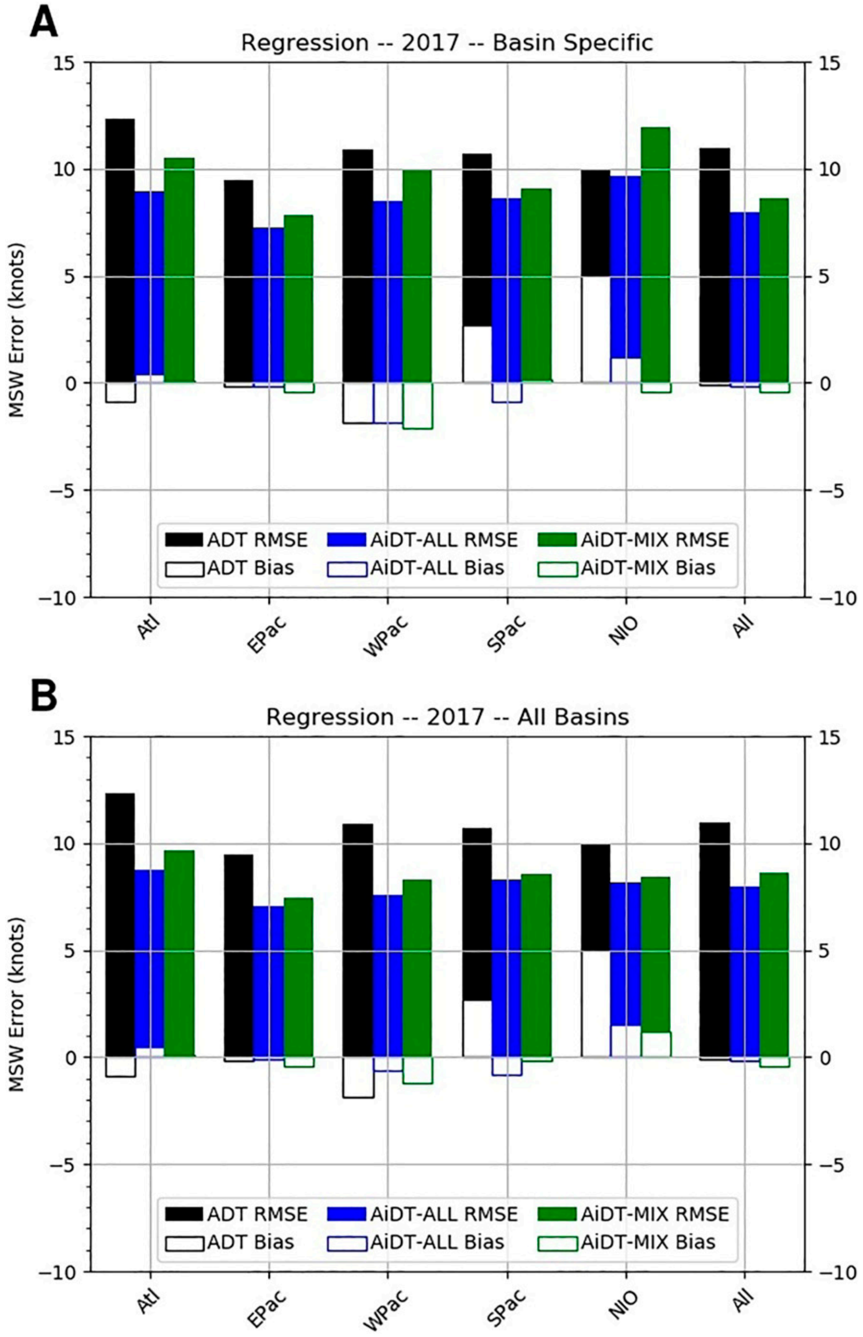


FIG. B1. Statistical comparisons between the regression network TC maximum sustained winds (MSW) estimates for the AiDT and ADT for the Atlantic (Atl), east Pacific (EPac), west Pacific (WPac), southern Pacific (SPac), and northern Indian Ocean (NIO), and combined global (All) basins for independent 2017 test dataset. Bias and root-mean-squared error (RMSE) are shown. (a) Five individual basin-specific models and (b) the globally derived AllBasin model. ALL indicates the single ''all-scene-type'' model and MIX the four ''combined scene-type'' models. MSW error units are in knots. A negative bias indicates the MSW estimate is weaker than the NHC/JTWC best track estimates.

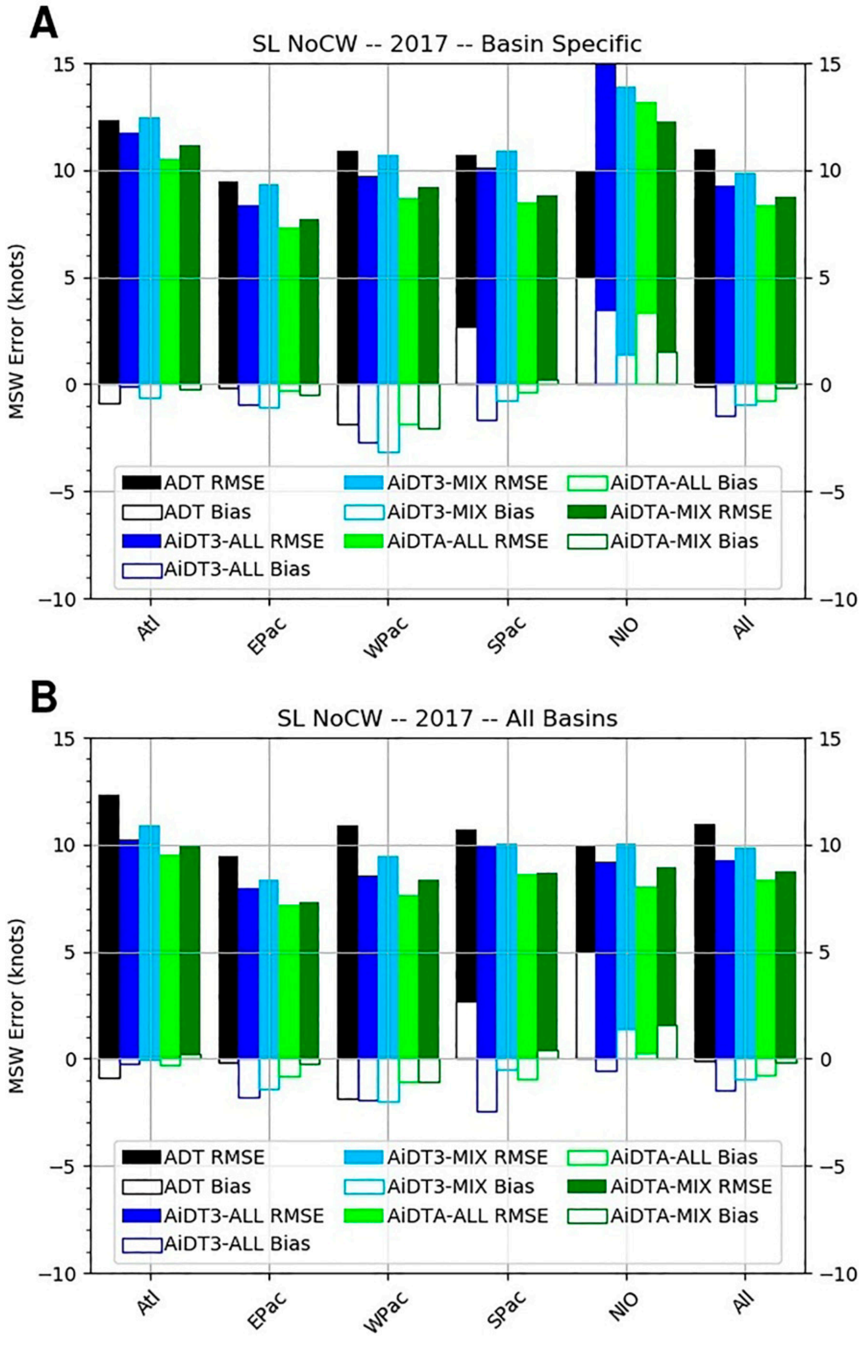


FIG. B2. As in Fig. B1, but for SL network, NoCW (no label weighing scheme utilized). In addition, the two output classification bin weighing schemes are displayed. AiDT3 and AiDTW designate the 3-bin and all-bin classification averaging schemes, respectively.

facilitate comparisons with other studies mentioned previously and to highlight basin-specific differences. Note again that ‘‘AllBasin’’ refers to the globally derived model using the combined ADT history file records for 2017 for all storm basins to determine if individual basin-specific models are better or worse than one all-encompassing global model. The AllBasin model is applied to the storms in each of the five individual basins, and results are tabulated by basin as well as overall global performance statistics.

a. Regression network

TC intensity estimates for the independent 2017 testing dataset are calculated with the regression network and

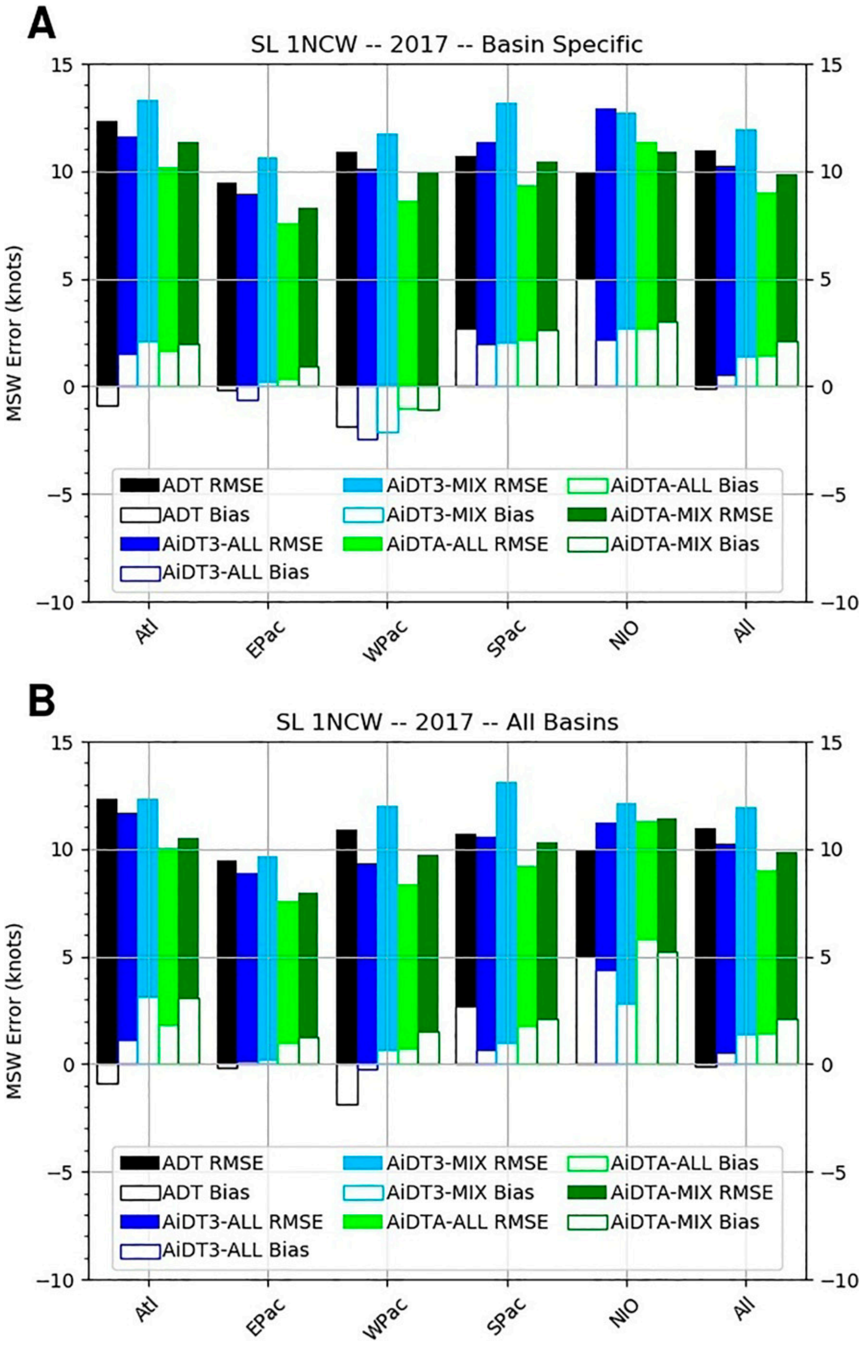


FIG. B3. As in Fig. B1, but for SL network, 1NCW (1/*N* label weighing scheme utilized).

compared to the original ADT estimates for that year. Figure B1 presents the regression network (labeled AiDT) and ADT MSW error statistics, including bias and root mean squared error (RMSE) (positive bias indicating a model overestimate versus NHC/JTWC best track MSW estimates) for each TC basin and the global dataset. Figure B1a presents the results from the five individual basin-specific models and Fig. B1b presents the results obtained using the global AllBasin model applied to each individual basin. The global result (labeled All in the figures to signify all combined global cases) is provided in both Figs. B1a and B1b for consistency but is equal in each figure. Two scene-specific variants are run for each of the five basin-specific models and the one AllBasin model, as described previously: one containing the combined four scene-type models (MIX) and one for the single, all-scene-type (ALL) model. The results from both scene-type model variants are presented in Figs. B1a and B1b.

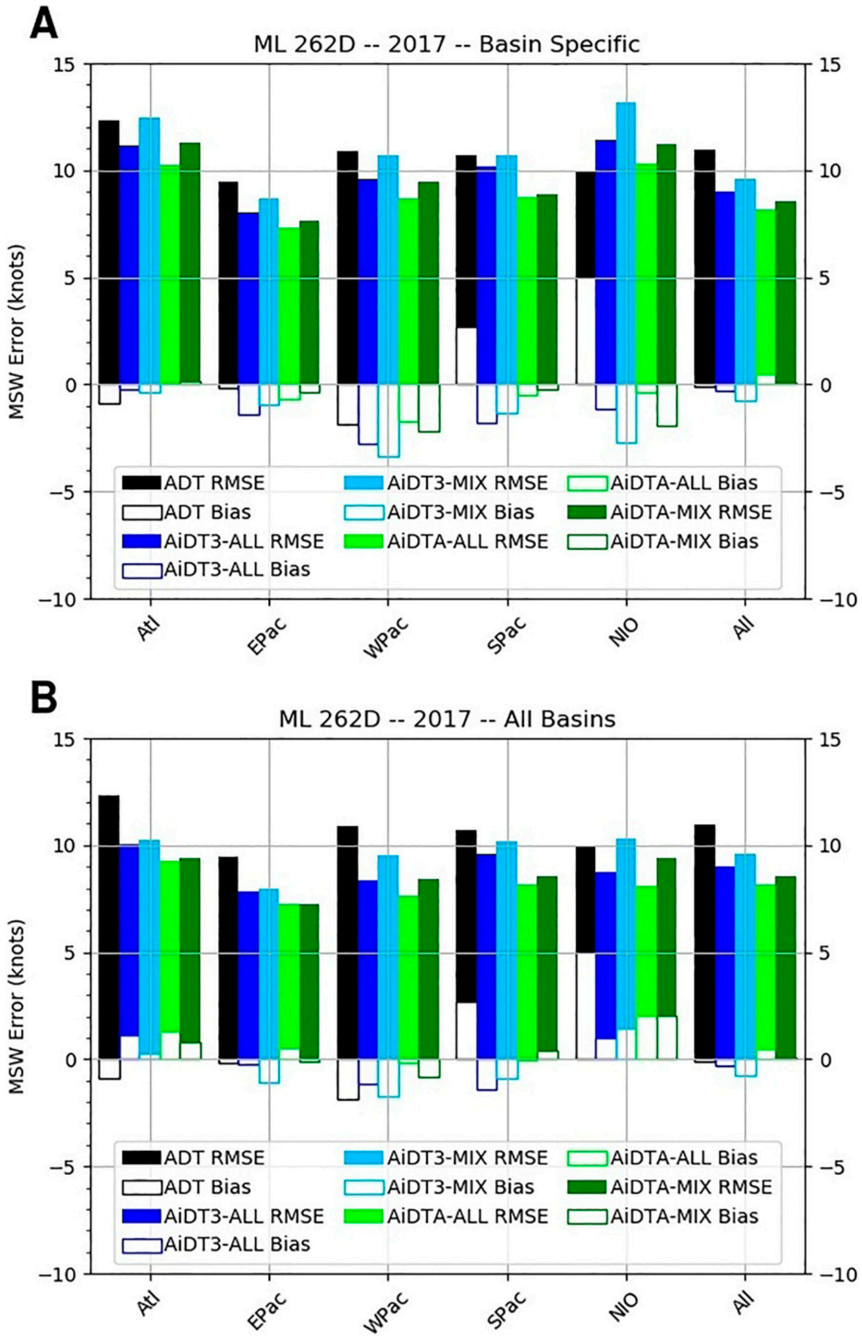


FIG. B4. As in Fig. B1, but for ML network, 262D (0.2/0.6/0.2 weighing distribution). In addition, the two output classification bin weighing schemes are displayed. AiDT3 and AiDTW designate the 3-bin and all-bin classification averaging schemes, respectively.

As shown in Fig. B1, both the ALL and MIX models are better than the original ADT, with the ALL scene-type model being superior in RMSE to the MIX scene-type in all basins for both the five basin-specific models (Fig. B1a) and the single AllBasin globally derived model (Fig. B1b) in RMSE. In addition, the AllBasin model results in Fig. B1b are superior to the corresponding basin-specific models in Fig. B1a for all five basins. Overall, the AllBasin model using the ALL scene-type model performed best for the regression network.

While it may be counterintuitive that a more homogenized network/model would be more accurate than the specialized basin/scene specific models, one can make the case that a more specific model might be ‘‘over-tuned’’ to that particular condition, especially if the sample size is small. For example, in the

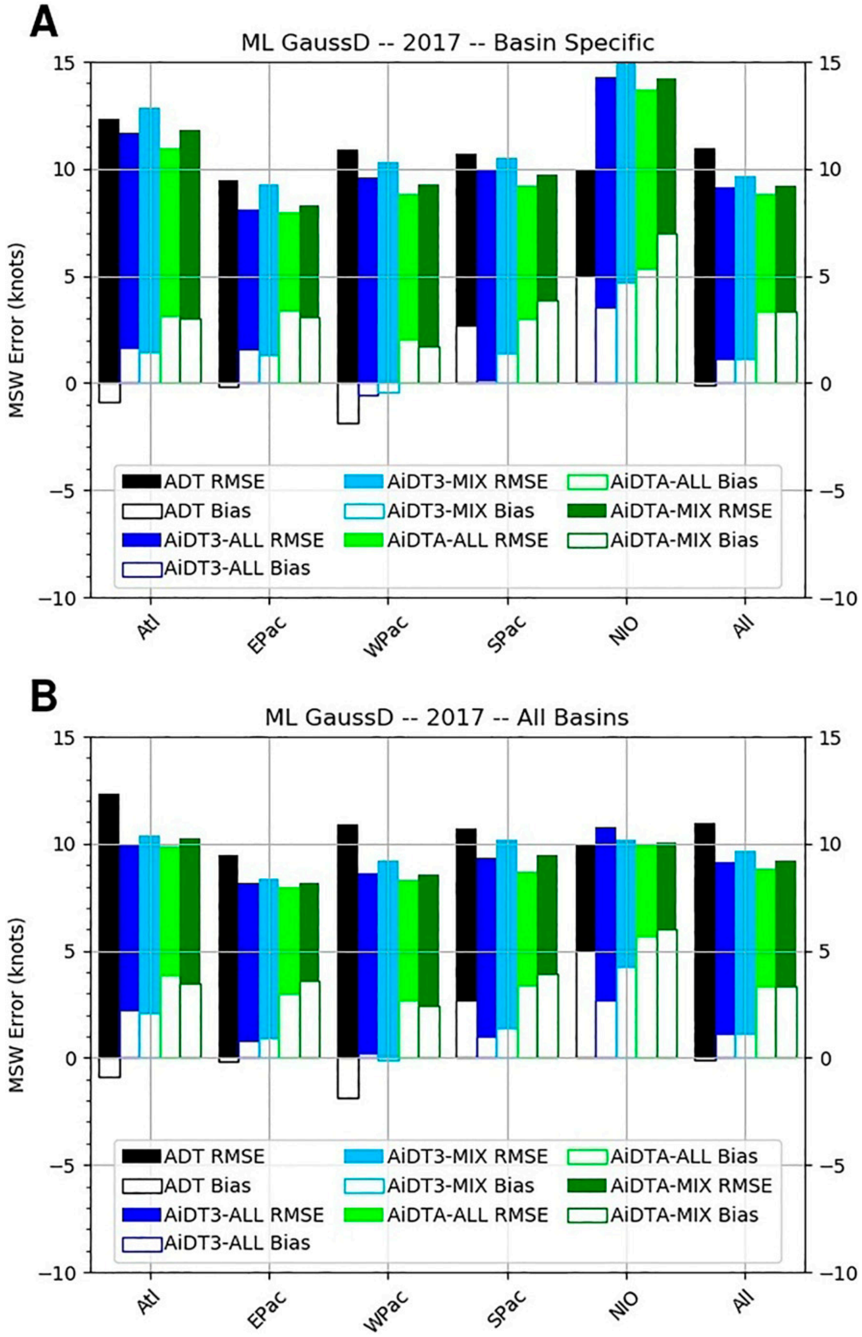


FIG. B5. As in Fig. B1, but for ML network, GaussD (Gaussian weighing distribution).

Atlantic model, there are almost 36 000 records in the training dataset, but of those about 21 500 (~60%) of those are CDO and curved band scene types. Only around 3000 (~8%) are eye scene types, so while a scene-specific model could provide a good relationship in the training data, when applied to the validation data the relationship might be overfit and provide a less accurate estimate than with a multi-feature model developed on a much larger dataset. This is highlighted in the basin specific plots where there are larger differences between the MIX and ALL models, illustrating the importance of having large data samples when training a machine learning model.

b. Single-label categorical networks

Figures B2 and B3 show the performance results obtained for both experiments using the SL networks. NoCW, shown in Fig. B2, uses no class_weight weighting scheme while 1NCW, shown in Fig. B3, utilizes the 1/*N* class_weight scheme for weighting the label data bins.

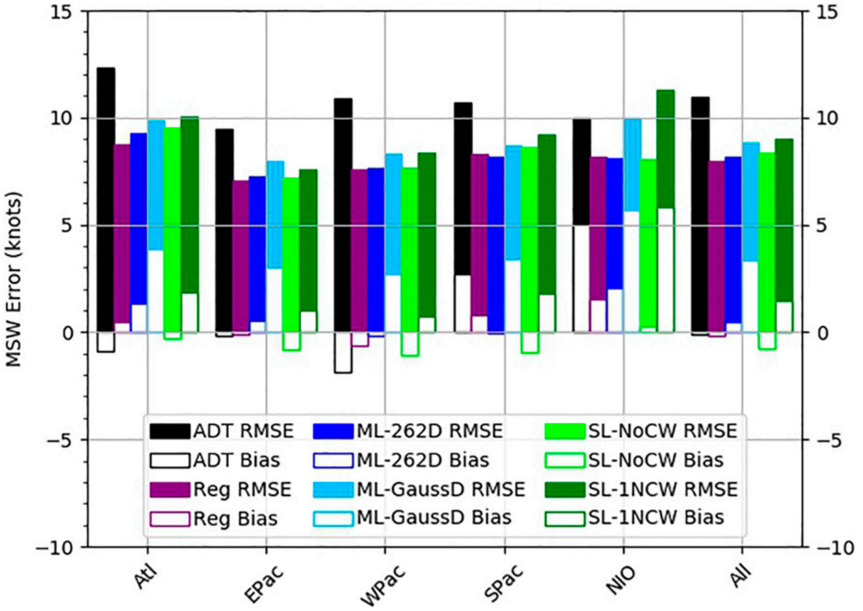


FIG. C1. As in Fig. B1, but a plot of the five best-performing networks using the AllBasin, ALL scene-type, and all-bin classification averaging scheme (AiDTA) for the regression (Reg), SL, and ML networks for five individual ocean basins and global total for independent 2017 test dataset. Units are in knots.

The same general results are obtained for both SL networks and are consistent with the results obtained with the regression network discussed in the previous section. The AllBasin model is, in general, superior to the five individual basin models, and the all-scene (ALL) model is superior to the combined four-scene (MIX) model. In addition, specific to the SL network experiments, the derivation of the final intensity value using a weighted average of all output classification bins (AiDTA) is more accurate than just using the maximum and two adjacent classification bins (AiDT3). Finally, the 1NCW network produces slightly better results overall than those obtained with the NoCW network.

It should be noted that the performance of the basin-specific NIO network NIO is poorer with respect to the other basin-specific models, especially in NoCW, where the RSME values are much larger than obtained with the ADT. This result is likely due to the smaller size of the samples used in the NIO when deriving the networks in the training stage (TCs are generally less common in this basin), leading to overfitting of the basin-specific networks. The RMSE values for the NIO 1NCW network are reduced but also result in an increased bias versus the NoCW network. Overall, use of the AllBasin model does improve the performance of the NIO storms, further illustrating the superiority of the AllBasin model against the basin-specific models.

A comparison of the SL NoCW and 1NCW network results using the AllBasin and ALL scene-type models and all-bin output classification weighting-scheme (AiDTA) against other networks is presented in appendix C.

c. *Multi-label categorical networks*

The ML multiclassification networks are investigated next. As described above, two different experiments are conducted and focus on the use a range of classification bins defining the input label data instead of a single label value. The first experiment, 262D, defines the input label bins using a simple 0.2/0.6/0.2 weighting distribution centered on the 5-kt bin where the label best track MSW value is contained, while the second experiment, GaussD, uses a Gaussian distribution weighting scheme for the label bins centered on the best track MSW label value classification bin.

Figures B4 and B5 present the ML network performance results obtained from both 262D and GaussD, respectively. The same output bin classification averaging schemes (AiDT3 and AiDTA) are used as explained in the SL discussion above.

As with the regression and SL results, the ML AllBasin model results are better than the basin-specific model results for the 262D and GaussD experiments, with the ALL scene-type model being superior to the MIX scene-type model in both experiments. GaussD typically performs worse than 262D, with much larger bias values noted with the GaussD experiments. In addition, the NIO dataset issues with the SL networks above are also noted in the ML network results, again highlighting the superiority of the AllBasin model over the individual basin-specific models.

In general, the AllBasin model using the ALL scene-type model and AiDTA output bin classification scheme yields the lowest errors for the ML networks, which is consistent with the SL network results presented previously.

APPENDIX C

Best Network Comparison

Figure C1 and Table C1 display results obtained from the regression, SL and ML networks using the AllBasin and ALL

TABLE C1. Comparison listing of various model performance values (MSW errors vs NHC/JTWC best track values) for all five individual ocean basins and overall global (All) set for the independent 2017 test data. MAE is mean absolute error. RMSE is root-mean-square error. Units are in knots. Negative bias indicates MSW estimates are generally weaker than the NHC/JTWC best track estimates. Minimum basin-specific and global MAE and RMSE values are highlighted in bold.

Network	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE
		Atlantic			East Pacific			West Pacific	
ADT	−0.91	9.50	12.33	−0.15	7.38	9.44	−1.87	8.47	10.88
ML-T1	1.35	7.26	9.24	0.57	5.71	7.27	−0.14	**5.93**	7.66
ML-T2	3.85	7.98	9.84	3.05	6.57	7.96	2.67	6.56	8.30
Regression	0.49	**6.89**	**8.76**	−0.13	**5.50**	**7.04**	−0.60	6.02	**7.56**
SL-T1	−0.31	7.41	9.54	−0.83	5.60	7.16	−1.07	6.02	7.65
SL-T2	1.84	7.39	10.08	0.99	5.94	7.55	0.76	6.47	8.37
No. of records	5188	5188	5188	3677	3677	3677	5475	5475	5475
		South Pacific			North Indian			Global (all)	
ADT	2.71	8.43	10.70	5.03	7.51	9.96	−0.13	8.50	10.98
ML-T1	−0.07	**6.30**	**8.19**	2.05	**5.63**	**8.08**	0.49	6.32	8.18
ML-T2	3.43	6.80	8.68	5.66	7.44	9.99	3.32	7.03	8.82
Regression	0.80	6.52	8.29	1.50	5.90	8.15	−0.18	**6.26**	**7.98**
SL-T1	−0.96	6.61	8.65	0.28	6.01	8.03	−0.75	6.44	8.34
SL-T2	1.81	7.06	9.22	5.85	8.39	11.29	1.47	6.80	9.00
No. of records	3766	3766	3766	566	566	566	18 672	18 672	18 672

scene-type models. Both experiments for the SL and ML networks are shown along with the single regression network results. Bias, RMSE, and mean absolute error (MAE) are shown in Table C1 while only bias and RMSE are shown in Fig. C1 for clarity.

Overall, the regression network produces the best statistical results for TC intensity estimates for models run on the 2017 independent test dataset. In three of the five individual basins, as well as the global set, the lowest RMSE values are achieved using the regression network, whereas the best results in the South Pacific and north Indian derive from the ML-262D and SL-NoCW networks, respectively (regression is second and third, only 0.10 and 0.12 higher RMSE in each respective basin). Given these findings, the regression network model is used for the network of choice in section 4. A schematic diagram of the final network is shown in Fig. 1.

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, https://dl.acm.org/doi/10.5555/3026877.3026899.

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

Chollet, F., 2015: Keras. GitHub, accessed 15 December 2020, https://github.com/fchollet/keras.

Combinido, J. S., J. R. Mendoza, and J. Aborot, 2018: A convolutional neural network approach for estimating tropical cyclone intensity using satellite-based infrared images. *24th Int. Conf. on Pattern Recognition (ICPR)*, Beijing, China, Institute of Electrical and Electronics Engineers, 1474–1480, https://doi.org/10.1109/ICPR.2018.8545593.

Courtney, J., A. Burton, C. Velden, T. Olander, L. Ritchie, C. Stark, and L. Majewski, 2020: Towards an objective historical tropical cyclone dataset for the Australian region. *Trop. Cyclone Res. Rev.*, **9**, 23–36, https://doi.org/10.1016/j.tcrr.2020.03.003.

Dvorak, V., 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., https://satepsanone.nesdis.noaa.gov/pub/Publications/Tropical/Dvorak_1984.pdf.

Jiang, H., C. Tao, and Y. Pei, 2019: Estimation of tropical cyclone intensity in the North Atlantic and northeastern Pacific basins using TRMM satellite passive microwave observations. *J. Appl. Meteor. Climatol.*, **58**, 185–197, https://doi.org/10.1175/JAMC-D-18-0094.1.

Kossin, J. P., T. L. Olander, and K. R. Knapp, 2013: Trend analysis with a new global record of tropical cyclone intensity. *J. Climate*, **26**, 9960–9976, https://doi.org/10.1175/JCLI-D-13-00262.1.

——, K. R. Knapp, T. L. Olander, and C. S. Velden, 2020: Global increase in major tropical cyclone exceedance probability over the past four decades. *Proc. Natl. Acad. Sci. USA*, **117**, 11 975–11 980, https://doi.org/10.1073/pnas.1920849117.

Lee, J., J. Im, D. Cha, H. Park, and S. Sim, 2019: Tropical cyclone intensity estimation using multi-dimensional convolutional neural networks from geostationary satellite data. *Remote Sens.*, **12**, 108, https://doi.org/10.3390/rs12010108.

Manion, A., C. Evans, T. L. Olander, C. S. Velden, and L. D. Grasso, 2015: An evaluation of advanced Dvorak technique–derived tropical cyclone intensity estimates during extratropical transition using synthetic satellite imagery. *Wea. Forecasting*, **30**, 984–1009, https://doi.org/10.1175/WAF-D-15-0019.1.

Maskey, M., and Coauthors, 2020: Deepti: Deep learning-based tropical cyclone intensity estimation system. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **13**, 4271–4281, https://doi.org/10.1109/JSTARS.2020.3011907.

Olander, T. L., 2021: The Advanced Dvorak Technique (ADT) Users Guide, version 9.0. Accessed 27 September 2021, http://tropic.ssec.wisc.edu/misc/adt/info.html.

——, and C. S. Velden, 2019: The advanced Dvorak technique (ADT) for estimating tropical cyclone intensity: Update and new capabilities. *Wea. Forecasting*, **34**, 905–922, https://doi.org/10.1175/WAF-D-19-0007.1.

Pineros, M., E. Ritchie, and J. Tyo, 2008: Objective measures of tropical cyclone structure and intensity change from remotely sensed infrared image data. *IEEE Trans. Geosci. Remote Sens.*, **46**, 3574–3580, https://doi.org/10.1109/TGRS.2008.2000819.

Pradhan, R., R. Aygun, M. Maskey, R. Ramachandran, and D. 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.

Ritchie, E. A., K. M. Wood, O. G. Rodriguez-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.

Velden, C. S., and D. Herndon, 2020: A consensus approach for estimating tropical cyclone intensity from meteorological satellites: SATCON. *Wea. Forecasting*, **35**, 1645–1662, https://doi.org/10.1175/WAF-D-20-0015.1.

——, 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.

——, T. L. Olander, D. Herndon, and J. P. Kossin, 2017: Reprocessing the most intense historical tropical cyclones in the satellite era using the advanced Dvorak technique. *Mon. Wea. Rev.*, **145**, 971–983, https://doi.org/10.1175/MWR-D-16-0312.1.

Wimmers, A., and C. Velden, 2016: Advancements in objective multisatellite tropical cyclone center fixing. *J. Appl. Meteor. Climatol.*, **55**, 197–212, https://doi.org/10.1175/JAMC-D-15-0098.1.

——, ——, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Mon. Wea. Rev.*, **147**, 2261–2282, https://doi.org/10.1175/MWR-D-18-0391.1.

Xiang, K., X. Yang, M. Zhang, Z. Li, and F. Kong, 2019: Objective estimation of tropical cyclone intensity from active and passive microwave remote sensing observations in the Northwestern Pacific Ocean. *Remote Sens.*, **11**, 627, https://doi.org/10.3390/rs11060627.

Yu, X., Z. Chen, H. Zhang, and Y. Zheng, 2020: A novel deep learning framework for tropical cyclone intensity estimation using FY-4 satellite imagery. *Proc. 2020 Fourth Int. Conf. on Innovation in Artificial Intelligence (ICIAI 2020)*, New York, NY, Association for Computing Machinery, 10–14, https://doi.org/10.1145/3390557.3394298.

Zhang, R., Q. Liu, and R. Hang, 2020: Tropical cyclone intensity estimation using two-branch convolutional neural network from infrared and water vapor images. *IEEE Trans. Geosci. Remote Sens.*, **58**, 586–597, https://doi.org/10.1109/TGRS.2019.2938204.