Predicting Rapid Intensification in North Atlantic and Eastern North Pacific Tropical Cyclones Using a Convolutional Neural Network

SARAH M. GRIFFIN,[a] ANTHONY WIMMERS,[a] AND CHRISTOPHER S. VELDEN[a]

[a] *Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin–Madison, Madison, Wisconsin*

(Manuscript received 24 November 2021, in final form 25 April 2022)

ABSTRACT: This study develops a probabilistic model based on a convolutional neural network to predict rapid intensification (RI) in both North Atlantic and eastern North Pacific tropical cyclones (TCs). Coined "I-RI," an advantage of using a convolutional neural network to predict RI is that it is designed to learn from spatial fields, like two-dimensional satellite imagery, as well as scalar features. The resulting model RI probability output is validated against two operational RI guidances—an empirical and a deterministic method—to assess skill at predicting RI over 12-, 24-, 36-, 48-, and 72-h lead times. Results indicate that in North Atlantic TCs, AI-RI is more skillful at predicting RI over 12- and 24-h lead times compared to both operational RI guidances. In eastern North Pacific TCs, AI-RI is more skillful than the empirical operational RI guidance at most RI thresholds, but less skillful than the deterministic RI guidance at all thresholds. For TCs north of 15°N, where the deterministic skill was lower, AI-RI was more skillful than the deterministic operational guidance for over half of the RI thresholds. It is also found that AI-RI struggles to reach the higher RI probabilities produced by both of the operational RI guidances in both basins. This work demonstrates that the two-dimensional structures within the satellite imagery of TCs and the evolution of these structures identified using the difference in satellite images, captured by a convolutional neural network, yield better 12–24-h indicators of RI than existing scalar assessments of satellite brightness temperature.

SIGNIFICANCE STATEMENT: The purpose of this study is to develop a method to predict tropical cyclone rapid intensification using artificial intelligence. The developed model uses a convolutional neural network, which can identify features in satellite imagery that are indicative of rapid intensification. The results suggest that, compared with current operational rapid intensification models, a convolutional neural network approach is generally more skillful at predicting rapid intensification.

KEYWORDS: Tropical cyclones; Forecasting; Neural networks

1. Introduction

Significant improvement has been achieved in the prediction of tropical cyclone (TC) rapid intensification (RI) over the last two decades (Cangialosi et al. 2020), thanks in part to the Statistical Hurricane Intensity Prediction Scheme (SHIPS) rapid intensification index (RII; Kaplan and DeMaria 2003; Kaplan et al. 2010; DeMaria et al. 2021). However, predicting RI remains a complex and challenging problem (Cangialosi and Franklin 2014), since multiscale processes, including environmental, inner-core, and oceanic, all likely contribute to the probability of RI occurring. Specifically, studies focused on inner-core processes contributing to RI have examined convective-scale bursts (e.g., Guimond et al. 2010; Wang and Wang 2014; Rogers et al. 2013, 2015), as vigorous convection with associated latent heat release through condensation processes is an essential ingredient for TC intensification (Adler and Rodgers 1977; Kuo 1965). However, the specific contribution to RI from this observed process remains difficult to quantify.

The purpose of this study is to expand upon existing statistical RI forecast methods by employing artificial intelligence (AI), namely, a convolutional neural network (CNN) coined AI-RI. CNNs are designed to learn from spatial patterns (Lagerquist et al. 2020), making them ideal for analyzing the convective organization of TCs as depicted in satellite infrared (IR) imagery. Identifying and quantifying active convection in the tropics has been attempted in a variety of ways and with varying success, mainly through the use of satellites (Steranka et al. 1986; Alcala and Dessler 2002; Liu and Zipser 2005; Romps and Kuang 2009; Olander and Velden 2009; Monette et al. 2012). SHIPS-RII initially accounted for TC convection by calculating the percent area within a 50–200-km radius with IR brightness temperatures (BTs) < -30°C (Kaplan et al. 2010, hereafter KDK10) and currently incorporates the value of a principal-component-analysis derived from IR imagery as well as the standard deviation of IR brightness temperature (Kaplan et al. 2015, hereafter K15). However, it is possible these methods could underestimate the potential growth of small-scale convection and any other relevant spatial structure associated with TC RI, whereas this structure could be better characterized with a CNN.

Recently, the TC research community has begun to use AI to analyze and understand TC characteristics. For example, to estimate current TC intensity, Wimmers et al. (2019) applied a CNN to passive microwave imagery, while Zhang et al. (2020) used a CNN model with inputs of satellite IR and water vapor features and Chen et al. (2019) implemented a

Corresponding author: Sarah M. Griffin, sarah.griffin@ssec.wisc.edu

DOI: 10.1175/WAF-D-21-0194.1

CNN model incorporating IR and microwave-derived rain rate. Mercer et al. (2021) employed unsupervised machine learning to identify differences between environments for RI and non-RI TCs. Shaiba and Hahsler (2016) used a random forest model on TC bulk statistics to predict TC RI and compare the skill to SHIPS-RII, and Xu et al. (2021) created a multilayer perceptron to predict TC intensity in the next 24 h, but to the authors' knowledge, there are no published studies of applying CNNs to predict TC RI.

This paper is organized as follows. Section 2 describes the TC and AI-RI input feature data used in this study. Sections 3 and 4 describe the CNN model and the method for determining which inputs to the CNN, known as features, are selected for RI prediction. Section 5 presents the findings and a discussion. Finally, a summary and conclusions are presented in section 6.

2. Data

a. *Tropical cyclone tracks and rapid intensification validation*

This study considers North Atlantic TCs, defined as west of 30°W, from 2003 to 2020 and eastern North Pacific TCs east of 140°W, from 1998 to 2020. TC intensity and location histories are obtained from the National Hurricane Center (NHC) best tracks and are used to analyze AI-RI forecasts at the synoptic hours of 0000, 0600, 1200, and 1800 UTC. Consistent with previous studies, we specify RI for eight predefined thresholds based on an increase in the maximum 1-min sustained 10-m winds of 20 kt (1 kt $\approx$ 0.51 m s^{-1}) over a 12-h period; 25, 30, 35, and 40 kt over a 24-h period; 45 kt over a 36-h period; 55 kt over a 48-h period; and 65 kt over a 72-h period. Cases where a TC center is over land within the RI lead time or the 12 h prior to the forecast time are not analyzed. Also, instances when a system is categorized as an open wave or invest are not analyzed (based on NHC best track data).

Current operational RI forecast guidance includes the SHIPS-RII (KDK10), as well as a logistic regression scheme and a Bayesian scheme (Rozoff and Kossin 2011**)**. The probability of RI from these three schemes are also averaged to create a consensus mean (K15), hereafter referred to as the "SHIPS Consensus" since the guidance is disseminated with the SHIPS graphical output package. Rozoff and Kossin (2011) and K15 reported that the consensus shows skill over the individual model members and is often used operationally. Therefore this consensus will be used as one way to assess the skill of AI-RI. Another operational RI forecast model is the Deterministic to Probabilistic Statistical Model (DTOPS; Onderlinde and DeMaria 2018; DeMaria et al. 2021), a logistic regression model using intensity change from five different numerical weather models as inputs, in addition to latitude and current intensity. DTOPS has been shown to be more skillful than SHIPS-Consensus for at least half of the RI thresholds (DeMaria et al. 2021). Therefore, the skill of AI-RI forecasts will be compared to both SHIPS Consensus and DTOPS in this study.

b. *Satellite features*

Inputs to AI-RI include features based on satellite imagery of TCs. The satellite features used in this analysis are derived from IR BTs obtained from the Geostationary Operational Environmental Satellite (GOES) series, from *GOES-10* through *GOES-17*. *GOES-10* became operational in 1998 over the eastern North Pacific Ocean while *GOES-12* was first operational over the North Atlantic Ocean in 2003. Since the spatial resolution of these GOES IR imagers varies from 2 to 4 km at nadir over the period of interest, the satellite data for each TC in this study is remapped to a 4-km spatial resolution encompassing 400 $\times$ 400 grid points centered on the TC using nearest-neighbor interpolation. This is done in order to homogenize the input data for the CNN. These images are not parallax corrected, as cloud height information is not available for the full dataset. However, since the average parallax error in the location of the lowest BTs from 2019 to 2020 TCs is similar to the precision of the TC center latitude and longitude, about 0.1° or 11.5 km, issues in patterns of BTs due to parallax with respect to the TC center will be negligible. A satellite scan over a TC must be within 1 h of a best track synoptic time to be included. To account for potential characterization discrepancies between North Atlantic and eastern North Pacific TCs, satellite feature data for each basin is normalized by subtracting the mean BT calculated from all grid points from every TC in the given basin and dividing by the standard deviation also calculated from all grid points from every TC in the given basin. Although the central IR window wavelengths are 10.3 μm for *GOES-16* and *GOES-17* and 10.7 μm for all other GOES satellites considered, this difference is deemed minimal and there is no attempt at normalization based on the IR wavelength.

In addition to the IR image BT feature, IR BT difference features are also calculated that include 1-, 3-, and 6-h BT difference within roughly 3° and 10° of the TC center from the NHC best track (corresponding to 82 $\times$ 82 and 276 $\times$ 276 grid sizes, respectively). Differences are calculated as previous satellite image minus current satellite image, so negative values indicate warming BTs. See Table 1 for a list of the satellite features.

c. *Scalar features*

The scalar features used in this analysis are available from the SHIPS developmental data provided by the Cooperative Institute for Research in the Atmosphere (CIRA) available at https://rammb.cira.colostate.edu/research/tropical_cyclones/ships/developmental_data.asp. A list of all the scalar and satellite features used in this analysis are shown in Table 1, which also serves as a reference for subsequent acronyms used in the text. These features include current and past information about the analyzed TC position and intensity, as well as the surrounding environment and oceanic characteristics. Some of these features are explicitly chosen because they are used in the SHIPS Consensus RI models, while others like TOD and the tangential wind predictors are chosen to investigate whether they add any skill at RI predictability in this new context. Like the satellite features, the scalar features are normalized for each oceanic basin. Data for the scalar features is either evaluated at time $t = 0$ or averaged over the entire RI lead time as noted in Table 1.

TABLE 1. List of all features considered when developing AI-RI.

Feature	Description (r = radius from TC center)
IR	4-km IR brightness temp (BT) data (400 × 400 grid points centered on TC)
1h-IRdiff-3deg	1-h IR BT difference (r = 3°)
1h-IRdiff-10deg	1-h IR BT difference (r = 10°)
3h-IRdiff-3deg	3-h IR BT difference (r = 3°)
3h-IRdiff-10deg	3-h IR BT difference (r = 10°)
6h-IRdiff-3deg	6-h IR BT difference (r = 3°)
6h-IRdiff-10deg	6-h IR BT difference (r = 10°)
POT	Difference between current and max intensity (time avg)
MPI	Maximum potential intensity (time avg)
TOD	Time of day (local solar time)
VMAX	Maximum wind
MSLP	Mean sea level pressure
HIST	The number of 6-h periods VMAX has been above 20 kt
DELV	−12- to 0-h intensity change
LAT	Latitude
LON	Longitude
RSST	Reynolds SST (time avg)
COHC	Climatological ocean heat content (time avg)
CD20	Climatological depth of 20°C isotherm from 2005 to 2010 NCODA analyses (time avg)
CD26	Climatological depth of 26°C isotherm from 2005 to 2010 NCODA analyses (time avg)
NC26	Depth of 26°C minus CD26 (time avg)
DTL	Distance to land (time avg)
U200	200-hPa zonal wind (r = 200–800 km) (time avg)
U20C	200-hPa zonal wind (r = 0–500 km) (time avg)
V20C	200-hPa meridional wind (r = 0–500 km) (time avg)
RHLO	850–700-hPa relative humidity (r = 200–800 km) (time avg)
RHMD	700–500-hPa relative humidity (r = 200–800 km) (time avg)
RHHI	500–300-hPa relative humidity (r = 200–800 km) (time avg)
Z850	850-hPa vorticity (r = 0–1000 km)
D200	200-hPa divergence (r = 0–1000 km)
V000	1000-hPa tangential wind azimuthally averaged at r = 500 km
V850	850-hPa tangential wind azimuthally averaged at r = 500 km
V500	500-hPa tangential wind azimuthally averaged at r = 500 km
V300	300-hPa tangential wind azimuthally averaged at r = 500 km
DIVC	200-hPa divergence centered at 85-hPa vortex location
SHDC	850–200-hPa shear with vortex removed (r = 0–500 km) (time avg)
SHRD	850–200-hPa shear (r = 200–800 km) (time avg)
SHRS	850–500-hPa shear (time avg)
MTPW01	r = 0–200-km average total precipitable water (TPW)
MTPW03	r = 200–400-km average TPW
MTPW05	r = 400–600-km average TPW
MTPW07	r = 600–800-km average TPW
MTPW09	r = 800–1000-km average TPW
MTPW11	r = 0–400-km average TPW
MTPW13	r = 0–600-km average TPW
MTPW15	r = 0–800-km average TPW
MTPW17	r = 0–1000-km average TPW
EPSS	Avg θ_e difference (only positive) between a parcel lifted from the surface compared with the saturated θ_e of the environment (r = 200–800 km) (time avg)
ENSS	Avg θ_e difference (only negative) between a parcel lifted from the surface compared with the saturated θ_e of the environment (r = 200–800 km) (time avg)

The scalar features used by AI-RI for comparison with SHIPS Consensus and DTOPS were gathered from real-time lsdiag files provided by the NHC. These files are available at https://ftp.nhc.noaa.gov/atcf/archive/MESSAGES/. The real-time lsdiag files are used instead of the SHIPS developmental data to provide a more homogeneous comparison between AI-RI and the real-time SHIPS Consensus and DTOPS, as the real-time files could contain potential errors in TC intensity and location and use the Global Forecast System (GFS) model when estimating the model-based scalar predictors. Since these real-time lsdiag files do not provide HIST or MSLP, these variables were calculated from working NHC best tracks, which are available at http://hurricanes.ral.ucar.edu/realtime/plots/.

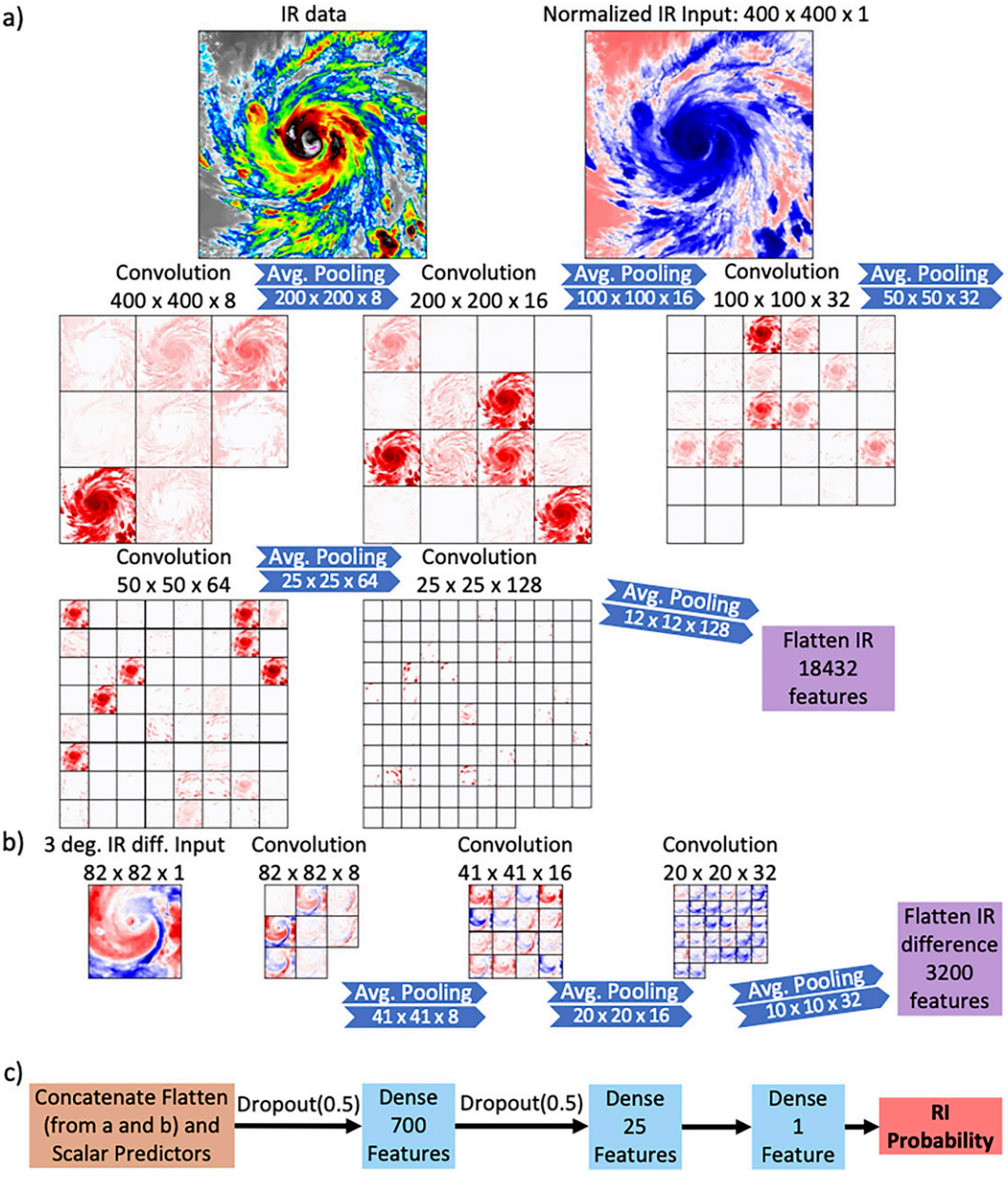

FIG. 1. Architecture of the AI-RI convolution neural network (CNN). The inputs used in this analysis are normalized (a) infrared (IR) BT data, (b) IR BT difference, and (c) scalar predictors. In the normalized input and feature maps produced by convolution and pooling layers, positive values are in red and negative values are in blue. Dropout (0.5) randomly sets half of the input units in the concatenated data to zero during training to mitigate against overfitting.

3. AI-RI

A schematic of the AI-RI CNN model can be seen in Fig. 1. This configuration uses IR and 3° IR difference satellite features as inputs. There are three main components to a CNN: convolution, pooling, and dense layers. AI-RI begins by passing the normalized IR and IR difference data through a convolution layer to produce an output map, or "feature map" (Fig. 1a). A useful definition of convolution in the context of deep learning is given by Eq. (4) in Lagerquist et al. (2019), though the standard meaning of image processing through filtering applies here as well. For each convolution layer, a 3×3 grid point convolution filter operates spatially on the combined input grids, encoding spatial patterns at higher levels of abstraction with each layer. Each convolutional filter in the CNN has a different set of weights, which are initialized randomly. In AI-RI, the first convolution layer produces a feature map that has dimensions of $400 \times 400 \times 8$. After

every convolution layer, a "leaky rectified linear unit" activation (ReLU; Maas et al. 2013) is applied, followed by average pooling layers. Activation is an elementwise nonlinear function applied to a given feature map; without it a CNN would only learn linear relationships. Average pooling layers downsample the feature map independently (e.g., Li et al. 2020), halving the spatial resolution at each pooling (which has become the fairly standard design). The process described above is repeated for 5 layers until the final IR feature map is $12 \times 12 \times 128$, which is then flattened into a 1D vector with 18 432 elements.

The process described above is repeated for the IR BT difference satellite features (Fig. 1b). For the 3° IR BT difference, which has an input dimension of $82 \times 82 \times 1$, three convolution, activation, and pooling layers are used until the feature map is flattened to a 1D vector with 3200 elements. The 10° IR BT difference goes through an additional set of layers before flattening to a 1D vector with 18 496 elements. Finally, the flattened IR feature map, flattened IR difference feature map, and additional input scalar features are concatenated together (Fig. 1c). The concatenated feature map is then sent through two dropout and three dense layers. Dropout (Hinton et al. 2012) randomly zeroes out a fraction of the layer's values, thereby forcing the weights in a given layer to evolve more independently and to reduce overfitting. Dense layers transform feature maps into predictions. The first two dense layers use the default ReLU activation function, which zeros out negative values, while the final dense layer uses the sigmoid activation function to squash the AI-RI output to range from 0 to 1 as a probability prediction (Chollet 2018).

Since TC RI prediction is a binary classification problem (RI either does or does not occur), we chose to minimize the binary cross-entropy loss function. This loss function identifies when the model should stop training. The equation for binary cross-entropy is

$$\varepsilon = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 + y_i)\log(1 - p_i)]. \qquad (1)$$

In Eq. (1), y_i is the RI label (1 for RI, 0 for no RI), p_i is the predicted probability of RI for each ith example, N is the number of examples, and ε is the binary cross-entropy, ranging from $[0, \infty]$.

4. Methodology

a. Data processing

The TCs used in this analysis are divided into three categories: training, validation, and testing. The testing dataset consists of all North Atlantic and eastern North Pacific TCs from 2019 to 2020, to compare to the most recent operational SHIPS Consensus RI forecasts. The remaining TCs are randomly divided into the training dataset (approximately 80% of the remaining TCs) and validation dataset (approximately 20% of the remaining TCs). Validation TCs are selected by binning TCs into basins and months and randomly selecting 20% of the TCs in each bin. This is done to ensure that any characteristics and environments of the TCs that vary with the time of year are proportional between the training and validation datasets. Since few TCs occur from January to June, TCs from these months are binned together when selecting the validation TCs. The same is done for November and December TCs. When developing AI-RI, training is done using the training dataset while the validation dataset is used to determine the optimal model configuration. Therefore, the testing dataset remains independent to any model training.

Depending on the RI threshold used, the training dataset consists of 3608–8442 TC images. To increase the size of the training dataset, the satellite feature data are augmented using image rotation. For each TC time in the training dataset, the image for each IR and IR-difference feature is rotated 0°, 90°, 180°, and 270°, technically quadrupling the size of the training dataset. However, the scalar data are not augmented for the training dataset, and therefore the four augmented TC satellite features match identical scalar features.

As mentioned in the previous section, each CNN begins with randomly initialized weights for each given feature. Since CNNs start with this random initialization, and due to the small number of RI events even in the augmented training dataset, it was found that training two different CNNs using the same input features would produce a significantly different probability of RI for a given TC. Therefore, it was decided to train five CNNs to create an ensemble and average the RI probabilities to provide greater forecast accuracy. Any reference to AI-RI hereafter refers to the ensemble of five CNNs. While employing an ensemble greatly increases the time necessary to develop the AI-RI, a given RI probability can still be produced within a few minutes.

b. Feature selection

To identify which features are optimal at predicting RI for each threshold (section 2a), RI prediction begins by training seven different types of model configurations. These model configurations are humorously coined the "kitchen sink" models as they include all scalar features listed in Table 1. One kitchen sink feature configuration includes just the IR feature, in addition to all scalar features, while the other six members include the IR feature and one IR difference feature from Table 1, in addition to all scalar features. After the seven different kitchen sink configurations are trained for each RI threshold, the optimal features for predicting RI for each of the seven different model configurations are selected using "permutation importance" applied to the validation dataset. In permutation importance, input data for a given feature are randomly shuffled among the validation TCs, while leaving the other features' input data consistent. Therefore, this method reveals the model's sensitivity to the permutated feature. The permutated feature's impact on overall model skill is measured with the Brier skill score (BSS; Wilks 2006). The BSS is calculated using the following equation:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{climo}}}, \qquad (2)$$

where

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2. \qquad (3)$$

TABLE 2. Climatological probability of RI for North Atlantic and eastern North Pacific TCs. These probabilities are available in SHIPS.

	20 kt (12 h)$^{-1}$	25 kt (24 h)$^{-1}$	30 kt (24 h)$^{-1}$	35 kt (24 h)$^{-1}$	40 kt (24 h)$^{-1}$	45 kt (35 h)$^{-1}$	55 kt (48 h)$^{-1}$	65 kt (72 h)$^{-1}$
North Atlantic								
2019	5.0%	10.9%	6.7%	3.8%	2.4%	4.5%	4.6%	5.2%
2020	5.2%	10.9%	6.9%	3.9%	2.5%	4.6%	4.6%	5.4%
Eastern North Pacific								
2019	6.1%	12.5%	8.4%	6.0%	4.0%	6.5%	5.9%	4.7%
2020	6.3%	12.6%	8.6%	6.2%	4.2%	6.7%	5.9%	4.8%

In Eq. (3), y represents the probability of RI, either from AI-RI or climatological, and o represents the actual occurrence of RI: 0 if no RI occurs and 1 if RI occurs. In Eq. (2), BS_{climo} refers to Eq. (3) where y represents a reference climatology. The reference climatologies for 2019 and 2020 from SHIPS can be seen in Table 2. A positive BSS indicates AI-RI is skillful compared to climatology. A feature is deemed optimal for predicting RI if the BSS of the permutated version is lower than that of the nonpermutated version. The BSS used in feature permutation here is an average of 1000 instances of bootstrap sampling with replacement from the validation dataset.

Once the optimal features for each RI threshold and combination of satellite features are identified, the AI-RI ensemble of models are then trained again using only the optimal features. As this still leaves seven different configurations for each RI threshold based on the different satellite features, the configuration with the highest BSS for the validation dataset is ultimately used to predict RI at that threshold. While the training and feature permutation importance was completed using TCs from both the North Atlantic and eastern North Pacific basins, the final AI-RI feature configuration is selected based on validation TCs from the given basin. Therefore, the feature configurations of AI-RI are different for each threshold and between the North Atlantic and eastern North Pacific basins. The optimal features for RI are shown in Table 3 (Table 4) for the North Atlantic (eastern North Pacific) basins. While some predictors are consistently optimal for all RI thresholds in both basins, other predictors vary based on threshold and basin. This is not uncommon, as the logistic regression and Bayesian models in SHIPS Consensus also include different predictors depending on RI threshold and basin (K15). Part of this difference in predictors between basins could also be due to aircraft reconnaissance, which is more common in the North Atlantic.

c. *Shapley additive explanation values*

While permutation importance described above indicates the impact of an individual feature on model skill, the impact of an individual feature on the probability of RI for any given TC is assessed using Shapley additive explanation (SHAP) values (Shapley 1953; Lundberg and Lee 2017; Lundberg et al. 2018, 2020). SHAP is an "explanation model" that uses a large combinatorial analysis to estimate the relative contribution of each input feature to the corresponding output of a predictive model. Although it relies on linear approximations of model performance, it has the advantage of supplying simple and interpretable solutions to questions of model performance, and the power of this approach increases with the number of examples supplied to it, as later analysis will demonstrate (section 5).

SHAP values (expressed as percentages) corresponding to an individual input feature indicate the extent to which that feature contributes to the probability of RI, and they can be positive or negative. For any given case, the SHAP values of all input features sum to about 6 to 13 percentage points lower than the RI probability, with this difference increasing with increasing RI probability. This difference is because the SHAP values cannot feasibly be calculated with the entire training dataset. In this analysis, SHAP values are calculated using the shap python library (https://github.com/slundberg/shap), with the already-trained AI-RI and inputs into AI-RI. For more information on the computation of SHAP values, please refer to Mangalathu et al. (2020).

5. Results and discussion

a. *North Atlantic basin*

The Brier skill score is used to assess forecast skill in this study as the AI-RI, SHIPS Consensus and DTOPS methods all produce RI forecasts in terms of probabilities. A higher BSS indicates greater skill. Figure 2 shows the results for 2019 and 2020 North Atlantic TCs where gray bars represent the BSS for AI-RI, the BSS for SHIPS Consensus is indicated by light blue bars, and dark blue bars represent the BSS when the AI-RI probability is included in the SHIPS Consensus average (hereafter "AI-RI in SHIPS Consensus"), with the AI-RI forecast receiving equal weight with the RII, logistic regression and Bayesian schemes. Similarly, lighter green bars indicate the BSS for DTOPS, and dark green bars represent the BSS when the AI-RI and DTOPS probabilities are averaged together (hereafter "AI-RI and DTOPS"). As seen for 25-, 30-, and 35-kt RI thresholds, AI-RI has a higher BSS than the SHIPS Consensus and DTOPS, indicating AI-RI is more skillful at predicting RI at these thresholds for the North Atlantic TCs. For RI thresholds of 12 and 24 h, the AI-RI in SHIPS Consensus has higher skill than the SHIPS Consensus alone. However, for RI thresholds of 36 h and above, AI-RI in SHIPS Consensus BSS is lower than SHIPS Consensus.

The DTOPS results show that for all RI thresholds, DTOPS has a lower BSS than SHIPS Consensus. However, adding

TABLE 3. Features used to predict RI in North Atlantic tropical cyclones.

Feature	20 kt (12 h)$^{-1}$	25 kt (24 h)$^{-1}$	30 kt (24 h)$^{-1}$	35 kt (24 h)$^{-1}$	40 kt (24 h)$^{-1}$	45 kt (36 h)$^{-1}$	55 kt (36 h)$^{-1}$	65 kt (72 h)$^{-1}$
IR	X	X	X	X	X	X	X	X
1h-IRdiff-3deg						X		
1h-IRdiff-10deg								
3h-IRdiff-3deg	X		X					
3h-IRdiff-10deg		X					X	
6h-IRdiff-3deg					X			
6h-IRdiff-10deg				X				
POT	X	X	X	X	X	X	X	X
MPI	X	X	X	X	X	X	X	X
TOD				X	X	X	X	X
VMAX	X	X	X	X	X	X	X	X
MSLP	X	X	X	X	X	X	X	X
HIST	X	X	X	X	X	X	X	X
DELV	X	X	X	X	X	X	X	X
LAT	X	X	X	X	X	X	X	X
LON		X	X	X	X	X	X	
RSST	X	X	X	X	X	X	X	X
COHC		X	X		X	X	X	X
CD20	X				X	X	X	X
CD26				X	X	X	X	X
NDML		X	X	X		X		X
DTL	X		X	X			X	
U200	X	X				X	X	X
U20C	X	X	X	X	X	X	X	X
V20C	X	X				X	X	X
RHLO				X			X	X
RHMD	X			X	X	X		X
RHHI	X		X	X	X	X		X
Z850	X	X	X	X	X			X
D200	X				X	X	X	
V000	X	X	X	X				X
V850	X	X	X	X				X
V500	X	X	X	X		X		X
V300	X	X	X	X		X	X	X
DIVC	X				X	X		
SHDC	X	X	X	X	X	X	X	X
SHRD	X	X	X	X	X	X	X	X
SHRS	X		X	X	X	X	X	
MTPW01						X	X	X
MTPW03				X	X	X		
MTPW05					X		X	
MTPW07				X			X	
MTPW09			X	X	X		X	
MTPW11					X			
MTPW13					X	X	X	
MTPW15				X	X	X	X	
MTPW17			X	X	X	X		
EPSS			X		X	X	X	X
ENSS		X	X	X	X	X	X	X

AI-RI leads to "AI-RI and DTOPS" performing the best for the 20-, 25-, 30-, and 55-kt RI categories.

Figure 3 shows another way to assess the relative improvement of AI-RI over the SHIPS Consensus (AI-RI is only compared to SHIPS Consensus in this figure, as well as the rest of this section, as SHIPS Consensus is more skillful than DTOPS in this basin based on the results in Fig. 2). During observed RI occurrences (red bars in Fig. 3a), an improved forecast is when the AI-RI probability is higher than the SHIPS Consensus. During non-RI occurrences (purple bars in Fig. 3a) an improved forecast is when the AI-RI probability of RI is lower than the SHIPS Consensus. Next, Fig. 3b displays the average difference between AI-RI and SHIPS Consensus probabilities over all RI/no RI verified forecasts for

TABLE 4. Features used to predict RI in eastern North Pacific tropical cyclones.

Feature	20 kt (12 h)$^{-1}$	25 kt (24 h)$^{-1}$	30 kt (24 h)$^{-1}$	35 kt (24 h)$^{-1}$	40 kt (24 h)$^{-1}$	45 kt (36 h)$^{-1}$	55 kt (36 h)$^{-1}$	65 kt (72 h)$^{-1}$
IR	X	X	X	X	X	X	X	X
1h-IRdiff-3deg							X	
1h-IRdiff-10deg								
3h-IRdiff-3deg								
3h-IRdiff-10deg								
6h-IRdiff-3deg	X	X	X	X	X	X		
6h-IRdiff-10deg								
POT	X	X	X	X	X	X	X	X
MPI		X	X	X	X	X	X	X
TOD			X	X	X	X		X
VMAX	X	X	X	X	X	X	X	X
MSLP	X	X	X	X	X	X	X	X
HIST	X	X	X	X	X	X	X	X
DELV	X	X	X	X	X	X	X	X
LAT	X	X	X	X	X	X	X	X
LON	X	X	X	X	X		X	
RSST	X	X	X	X	X	X	X	X
COHC				X	X	X	X	X
CD20	X			X		X		X
CD26		X	X	X	X	X	X	X
NDML	X		X	X		X		X
DTL	X					X	X	
U200	X					X	X	X
U20C	X	X			X	X	X	X
V20C	X	X				X	X	X
RHLO	X			X				X
RHMD	X			X	X	X	X	X
RHHI	X		X	X	X	X		X
Z850	X	X	X	X	X			X
D200	X			X	X		X	
V000	X	X	X	X				X
V850	X	X	X	X		X		X
V500	X	X	X	X		X		X
V300	X	X	X	X		X	X	X
DIVC				X		X	X	
SHDC	X	X	X	X	X	X	X	X
SHRD	X	X	X	X	X	X	X	X
SHRS		X	X	X	X	X		
MTPW01							X	X
MTPW03				X	X			
MTPW05				X	X	X	X	
MTPW07				X		X	X	
MTPW09		X	X	X	X			
MTPW11					X		X	
MTPW13					X			
MTPW15				X	X	X	X	
MTPW17			X	X	X	X		
EPSS			X	X	X		X	X
ENSS	X	X	X	X	X	X	X	X

each RI threshold. The AI-RI shows improvement over the SHIPS Consensus when this difference is positive for RI (red bars) and negative for non-RI (purple bars).

As seen in Fig. 3a, AI-RI forecasts of RI are generally an improvement over SHIPS Consensus forecasts for most RI thresholds concentrated at or below 24-h lead times. For these thresholds, the average difference between the AI-RI and SHIPS Consensus probabilities is positive (negative) when RI does (does not) occur (Fig. 3b), which helps to explain the higher skill of AI-RI in Fig. 2. Even though AI-RI forecasts improve less than 50% of RI occurrences compared to SHIPS Consensus for the 30-kt RI category, the AI-RI BSS (Fig. 2) is higher because of much higher forecast probabilities in certain cases within this category. One such case is the RI of

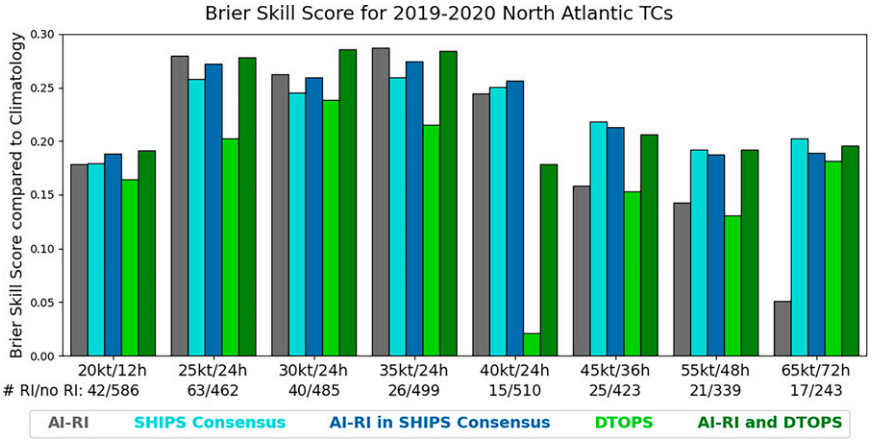

FIG. 2. The Brier skill score (BSS) compared to climatology for AI-RI (gray), SHIPS Consensus (light blue), AI-RI in SHIPS Consensus (dark blue), DTOPS (light green), and AI-RI and DTOPS (dark green) for the 2019–20 North Atlantic basin TCs. A higher BSS indicates a more skillful RI prediction. The maximum value for the BSS is 1.

Hurricane Laura (2020). An examination of SHAP values for Hurricane Laura (not shown) indicates that the highest contributing feature to the AI-RI probability is the 3h-IRdiff-3deg field (3-h image difference). If this feature is not used, the AI-RI forecast is less skillful than SHIPS Consensus. This demonstrates the potentially decisive role of the 3-h IR image trend, which is lacking in other existing objective methods in forecasting RI.

For RI at 36-h lead times and longer, the results are more mixed, possibly fueled by smaller sample sizes (especially for 65-kt RI). The average difference between AI-RI and SHIPS Consensus RI probabilities is negative when 45-kt RI occurs but positive for 55- and 65-kt RI thresholds (Fig. 3b). To investigate these RI probabilities, Fig. 4 presents a time series comparison between AI-RI and SHIPS Consensus RI probabilities during 45–65-kt RI events for the North Atlantic TCs in 2019 and 2020. As seen in Fig. 4, many late season (October and November) TCs underwent RI over 36-h and longer lead times. The 2020 North Atlantic Hurricane season had an unusually active October and November, with the highest number of major hurricanes since 1950 (Klotzbach et al. 2022). For these late season TCs, AI-RI has a lower probability of RI for the 45- and 65-kt thresholds compared to SHIPS Consensus, but a higher probability of RI for the 55-kt threshold. One hypothesis for the discrepancy between AI-RI and SHIPS Consensus differences for 45- and 55-kt RI is AI-RI uses an IR difference field within 3° of the TC center when predicting 45-kt RI and an IR difference field within 10° when predicting 55-kt RI. These thresholds were selected based on the North Atlantic validation dataset. In the North Atlantic validation dataset, the climatological probability of 45-kt RI in October and November TCs is lower compared to the entire validation dataset but higher for 55-kt RI. Since later season TCs tend to be larger in the North Atlantic basin, with a Pearson correlation coefficient of 0.23 between the radius of the outermost closed isobar and day of year, it is possible that AI-RI is not as well trained for RI in larger late-season TCs as lower RI probabilities result in a higher BSS when RI does not occur. If 45-kt RI is predicted with AI-RI trained using the 3h-IRdiff-10deg feature, similar to 55-kt RI, AI-RI does have a higher BSS than SHIPS Consensus.

In addition, when RI does not occur at 36-h lead times or longer, the AI-RI probability of RI is higher than SHIPS Consensus on average (Fig. 3b). This is especially noticeable for 65-kt RI. One hypothesis for these higher probabilities is that the satellite-derived features are less reliable at longer RI lead times. In agreement with this, K15 found that the importance of the satellite predictors decreases with increasing RI lead time. Another hypothesis is that the frequency of RI over longer lead times in the testing dataset is lower than in the validation dataset. For example, 65-kt RI occurs in 8.8% of North Atlantic TC validation times but only 6.4% in the testing dataset. Since a higher RI probability produces a higher BSS when RI occurs, selecting the optimal AI-RI features with the validation dataset where RI occurs more often could be leading to a general preference for higher AI-RI probabilities of RI.

Overall model confidence can also be assessed using reliability diagrams of forecast RI probabilities compared to observed occurrence (Fig. 5). In each reliability diagram, the 1:1 line represents perfect reliability for all forecast probabilities; i.e., the forecasted probabilities are the same as the observed probabilities. Points above the 1:1 line indicate forecast probabilities that are too low (underforecasted) and points below the 1:1 line indicate forecast probabilities that are too high (overforecasted). For 20-kt RI (Fig. 5a), AI-RI is generally well calibrated, as RI occurs about 25% of the time when the AI-RI forecast is between 20% and 30%, and is better calibrated than SHIPS Consensus at this RI threshold. By comparison, AI-RI overforecasts for 65-kt RI (Fig. 5d), especially at probabilities higher than 20%. This leads to an overconfident forecast, where the difference between predicted and observed probabilities is larger with increasing predicted probability. Also notable in Fig. 5 is that SHIPS Consensus

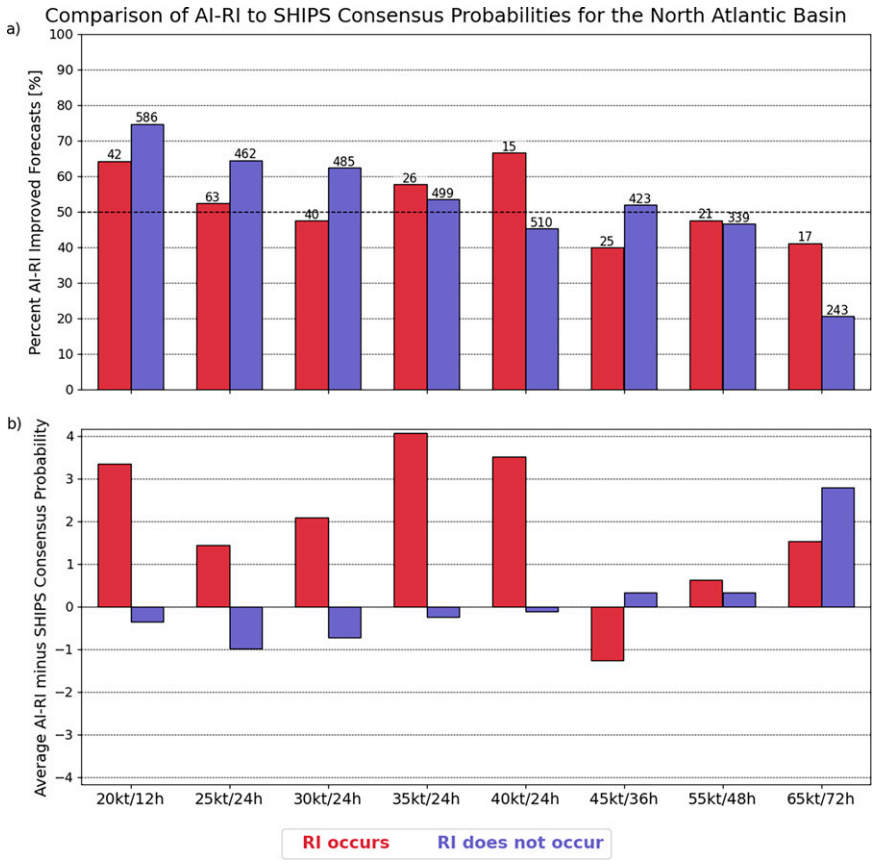

FIG. 3. (a) Percent of improved RI forecasts for AI-RI compared to the SHIPS Consensus for the 2019–20 North Atlantic basin TCs. The number above each bar indicates the total number (*N*) of valid RI (red bar) and non-RI (purple bar) forecasts for each threshold. (b) The average difference (in terms of %) between the AI-RI and SHIPS Consensus RI forecast probabilities over all *N* cases.

produces much higher probabilities for 45- and 65-kt RI (Figs. 6c and 6d) than AI-RI. These higher probabilities further contribute to the increased relative skill for SHIPS Consensus compared to AI-RI when RI occurs. Conversely, AI-RI produces the highest RI probabilities for 20-kt RI (Fig. 5a) and similarly high RI probabilities for 30-kt RI (Fig. 5b) and had a similar or higher BSS. Both AI-RI and SHIPS Consensus are underconfident for 30-kt RI, as the observed probability is much higher than the predicted probability for these higher probabilities.

b. *Eastern North Pacific basin*

The BSS for predicting RI in the 2019–20 eastern North Pacific can be seen in Fig. 6. Overall, for each RI threshold, AI-RI in SHIPS Consensus is more skillful or just as skillful as SHIPS Consensus. Furthermore, AI-RI itself is more skillful than SHIPS Consensus for 25-, 40-, 45-, and 65-kt RI, while the difference for 30-kt RI is negligible. Overall, AI-RI is the most skillful empirical method for predicting RI in this basin. However, with the curious exception of 65-kt RI, DTOPS (which has a deterministic element) is more skillful than both AI-RI and SHIPS Consensus and even AI-RI and DTOPS. The increased AI-RI skill for 65-kt RI is in contrast to the North Atlantic basin, even though the SHAP values (which indicate relative importance) of the IR feature decrease with longer RI time in both basins. Again, there is relatively small sample size for 65-kt RI due to the extended lead time.

The percent of improved forecasts and average difference between the AI-RI and DTOPS RI probabilities for eastern North Pacific TCs is shown in Fig. 7. Here, AI-RI is only compared to DTOPS because it is the most skillful operational model in this basin as indicated in Fig. 6. DTOPS is probably more accurate than SHIPS Consensus in the eastern North Pacific because it has higher (lower) average RI probabilities when RI did (did not) occur compared to SHIPS Consensus, but lower average probability of RI regardless of RI occurring in the North Atlantic. Figure 7 indicates that the number of improved RI forecasts from AI-RI are less than DTOPS for all RI thresholds when RI occurs, except for 65-kt RI. This is largely due to AI-RI assigning lower probabilities to RI events than DTOPS. As seen in the reliability diagram in Fig. 8, DTOPS more frequently produces higher RI probabilities. In

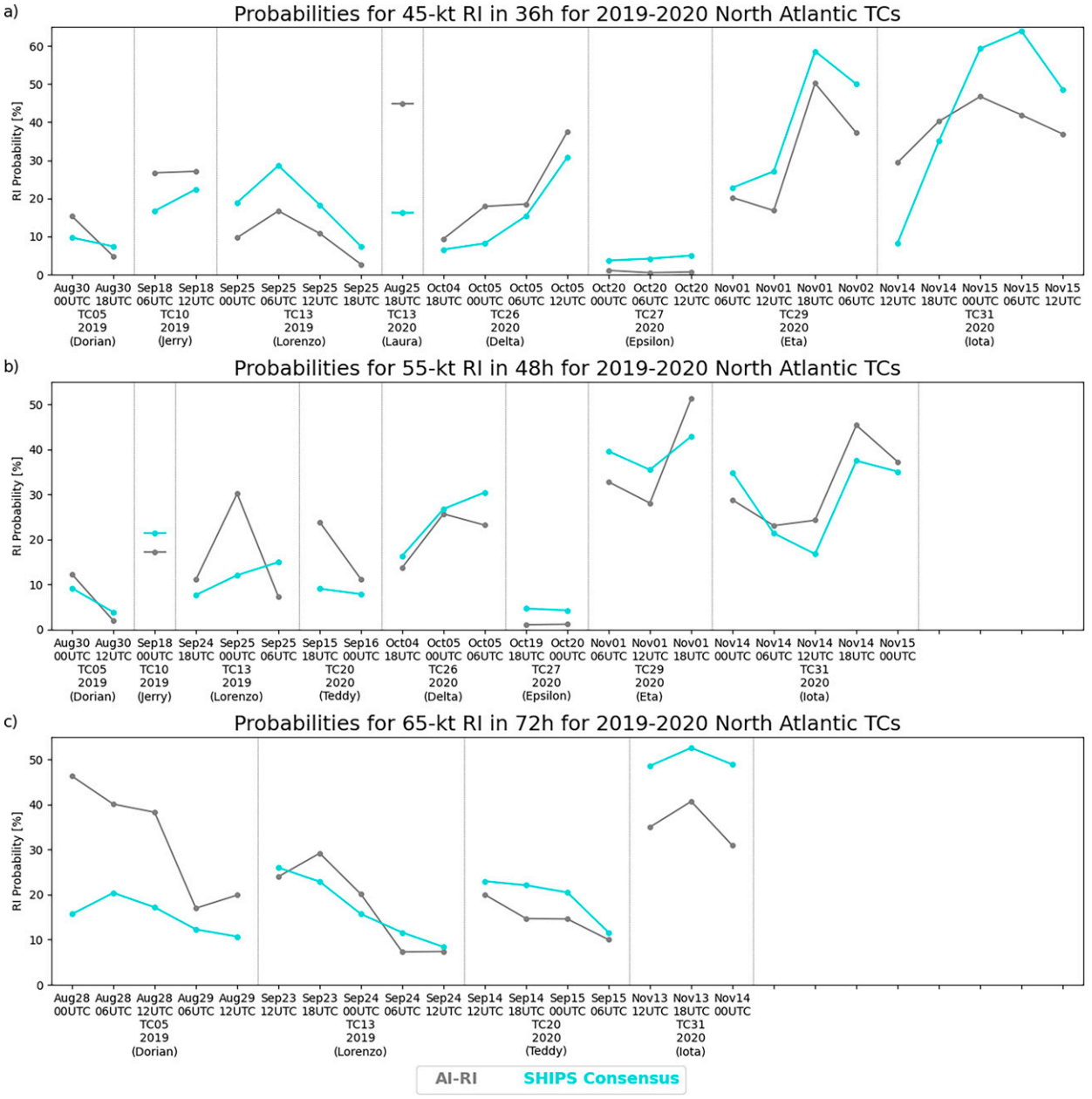

FIG. 4. A comparison of RI probability for AI-RI (gray) and SHIPS Consensus (light blue) for all verifying RI forecasts of (a) 45-, (b) 55-, and (c) 65-kt RI for 2019–20 North Atlantic TCs.

all but 65-kt RI (Fig. 8d), RI occurs at a higher probability than AI-RI predicts. It then follows that the overforecasting of 65-kt RI by DTOPS is potentially why the AI-RI is more skillful at this threshold. When RI does not occur, less than half of the RI forecasts have an AI-RI probability lower than DTOPS.

In spite of the overall better performance of DTOPS in this basin, a closer look at the testing cases reveals that AI-RI can improve on DTOPS in higher latitude cases. In DTOPS, one of the highest contributing predictors is cos(LAT) × VMAX (Onderlinde and DeMaria 2018), with lower values of latitude increasing the probability of RI. Therefore, DTOPS may be less skillful for TCs at higher latitudes. AI-RI also includes latitude as a feature, but the relative impact is much lower than in DTOPS. Figure 9a displays the BSS for all TCs, and Fig. 9b displays the BSS for only TCs with a latitude above 15°N. DTOPS is less skillful than AI-RI for 5 RI thresholds, and this is due to AI-RI producing a higher average RI probability than DTOPS when 30-, 40-, 55-, and 65-kt RI occur as well as a lower average RI probability when 35-, 40-, and 55-kt RI does not occur (not shown). Therefore, the increased skill for AI-RI compared to DTOPS for these higher latitude eastern North Pacific TCs is not only due to increasing the RI

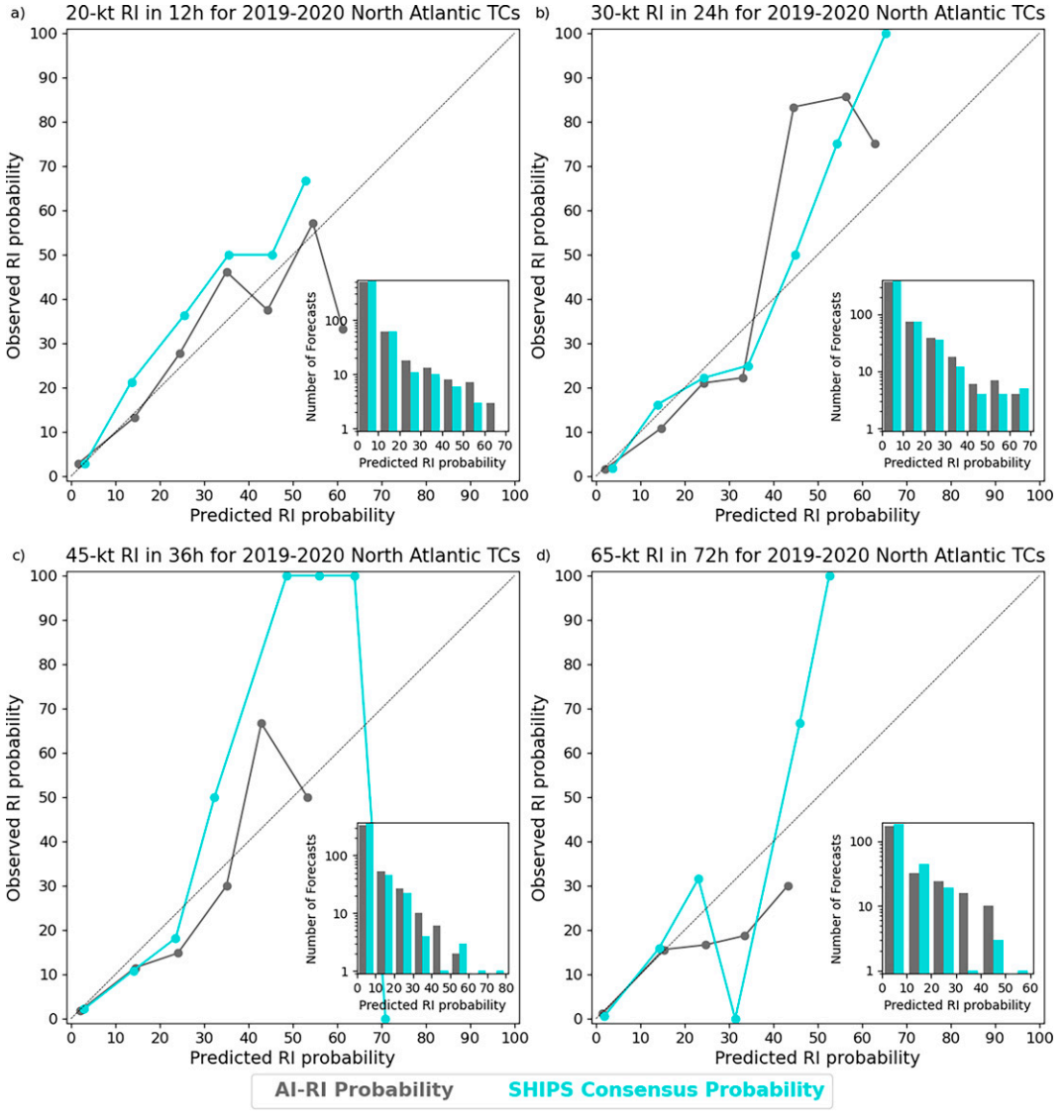

FIG. 5. Reliability diagram for the AI-RI (gray) and SHIPS Consensus (light blue) of (a) 20-, (b) 30-, (c) 45-, and (d) 65-kt RI probabilities for 2019–20 North Atlantic TCs. The inset depicts the corresponding number of forecasts for each predicting RI probability bin.

probability when RI does occur, but also having a lower probability of RI than DTOPS when RI does not occur.

c. *Feature contributions to AI-RI probabilities*

As noted earlier, SHAP values can be used to identify which characteristics in the satellite IR images (features) are contributing, positively or negatively, toward RI probabilities. Figure 10 depicts IR features and SHAP values from five different TC times when 30-kt RI is occurring. In Fig. 10, the left column displays the IR BTs while the middle column displays the normalized BTs from the left column after subtracting the mean and dividing by the standard deviation of all BTs for the North Atlantic TCs in the training dataset. Therefore, blue (red) in Fig. 10 indicate BTs that are colder (warmer) than the mean BT. These normalized BTs are the inputs for AI-RI. The resulting total SHAP values are depicted in the right column. The cases displayed in Fig. 10 were chosen as they all have similar total SHAP values (contributing approximately seven to nine percentage points to the RI probability), but the total satellite feature value (normalize BT summed over all grid points) is much different. One contributing characteristic to the increased probability of RI is the extent of the central dense overcast (left column). These positive SHAP values (right column) are similar to the SHIPS-RII predictor of the percentage of area with −30°C GOES-IR BT within a 50–200-km range and are especially evident in all but Fig. 10b (Delta). In Fig. 10b, the higher BTs near Delta's center

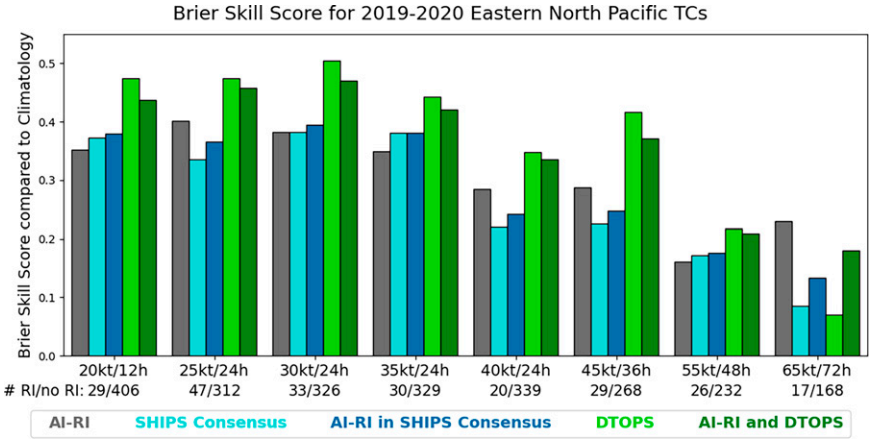

FIG. 6. As in Fig. 2, but for eastern North Pacific TCs.

contribute negatively to RI, possibly because AI-RI is interpreting these BTs as a dry slot. However, the highest SHAP values in Fig. 10 are associated with isolated cold BTs in the satellite image feature, with texture clearly indicating convection. Convection positively contributing to RI is consistent with previous studies that associate RI with intense convection (DeMaria et al. 2012; Monette et al. 2012).

Figure 11 depicts the SHAP values for the 3h-IRdiff-3deg feature, which relates convective tendencies to RI probabilities. Again, the cases in Fig. 11 were chosen as they all have a

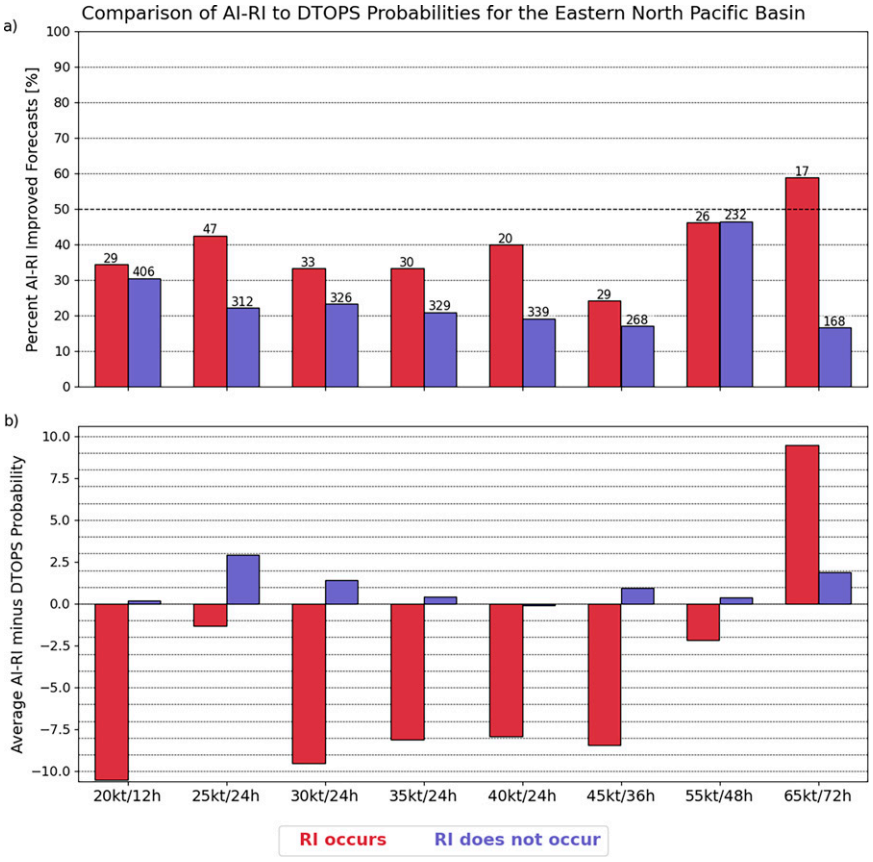

FIG. 7. (a) Percent of improved RI forecasts for AI-RI compared to DTOPS for the 2019–20 eastern North Pacific basin TCs. The number above each bar indicates the total number (N) of valid RI (red bar) and non-RI (purple bar) forecasts for each threshold. (b) The average difference (in terms of %) between the AI-RI and DTOPS RI forecast probabilities over all N cases.

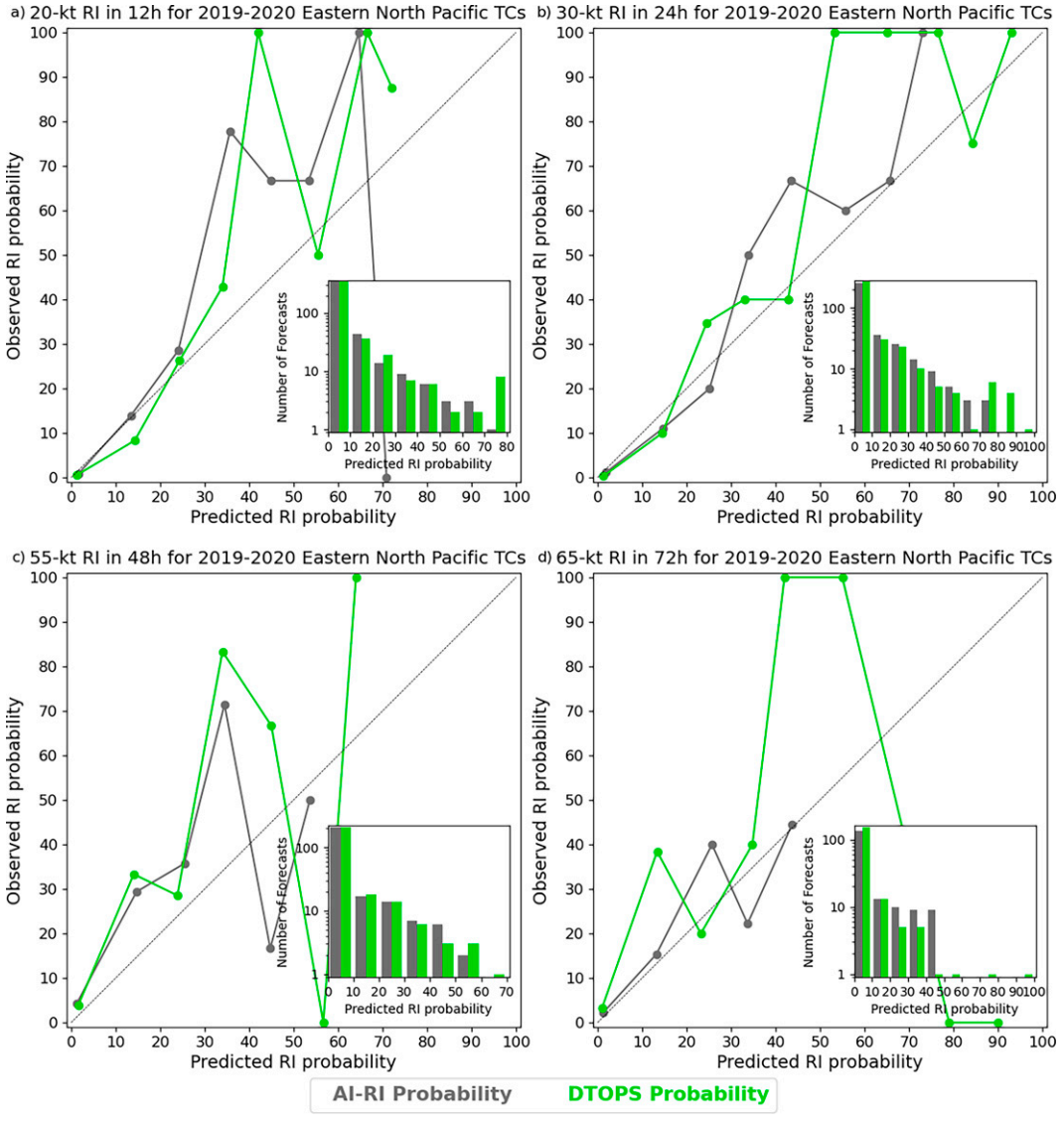

FIG. 8. Reliability diagram for the AI-RI (gray) and DTOPS (light green) of (a) 20-, (b) 30-, (c) 40-, and (d) 65-kt RI probabilities for 2019–20 eastern North Pacific TCs. The inset depicts the corresponding number of forecasts for each predicting RI probability bin.

similar total SHAP values (contributing approximately ten to thirteen percentage points to the RI probability) but the total satellite feature value is much different. In Fig. 11, the leftmost columns display the IR BTs for the two satellite images contributing to the normalized IR difference value in the rightmost center column. In Fig. 11, blue (red) indicate BT differences that are larger (smaller) than the mean BT different. SHAP values are displayed in the right column. It is evident based on the SHAP values that a warming or neutral change near the TC center positively contributes to the probability of RI. The warming signal is likely coincident with the beginning of an eye formation in the IR, which has been associated with rapid intensification (Vigh et al. 2012). Indeed, an eye is subsequently present in the IR images in the majority of these cases 24 h later.

To shed more light on the contributing features to the AI-RI probabilities for this RI category, the SHAP values for all instances of 30-kt RI in 24 h occurring in the 2019 and 2020 North Atlantic TCs are shown in Fig. 12. Features are ordered from the highest contribution to the total SHAP value at the top to the lowest at the bottom. Each RI forecast is represented by a dot, with the color of each dot indicating the normalized value of a given feature when 30-kt RI occurred in 2019 and 2020 North Atlantic TCs. Based on Fig. 12, the highest individual contributing features to the overall RI probabilities are the satellite features and then the scalar features. While there are no distinguishable trends between satellite features and SHAP values, some trends can be identified between scalar features and SHAP values. Some of these trends are intuitive. The highest contributing scalar feature is DELV,

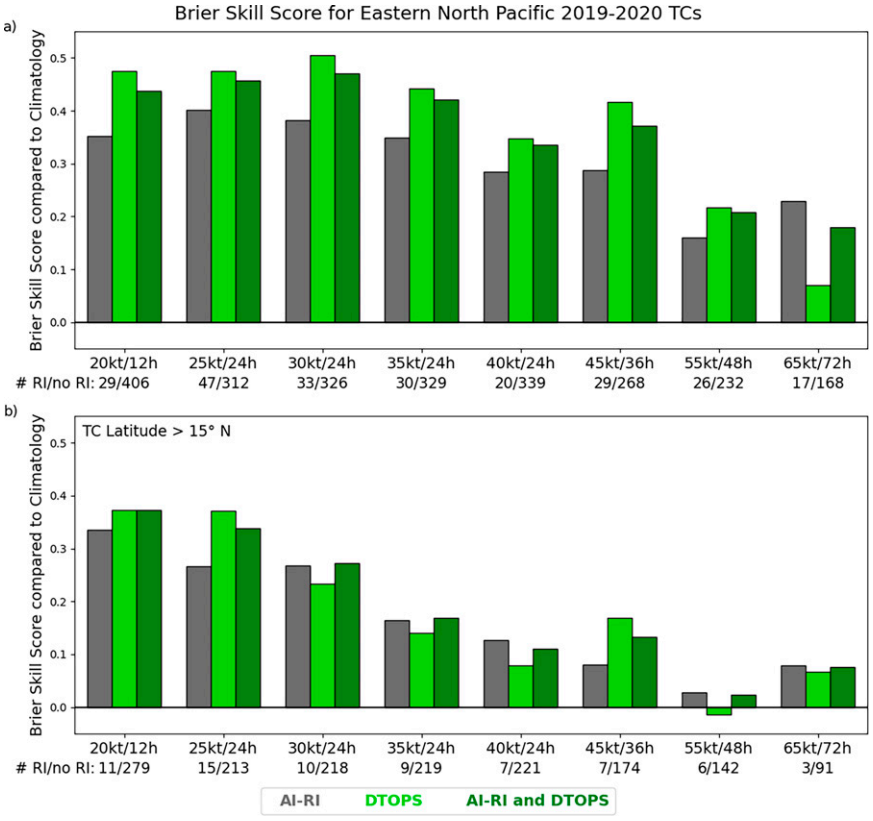

FIG. 9. (a) The Brier skill score (BSS) compared to climatology for AI-RI (gray), DTOPS (light green), and AI-RI and DTOPS (dark green) for the 2019–20 eastern North Pacific basin TCs. A higher BSS indicates a more skillful RI prediction. (b) BSS comparison of RI forecasts for TCs north of 15°N.

with the red dots associated with larger SHAP values indicating that a higher rate of intensification in the previous 12 h is correlated with increasing the probability of RI. Both 850–200-hPa vertical wind shear (VWS) features (SHDC and SHRD) also contribute highly to the probability of RI, with higher values being less favorable for RI (it should be noted that the lack of negative contribution instances by VWS features is symptomatic of this subsample of cases being limited to only when RI actually occurs). Higher MPI and a greater difference between current and maximum potential intensity (POT) also increase the probability of RI, as well as warmer underlying ocean values (RSST, COHC). A less intuitive feature that highly contributes to the 30-kt RI probability is the 850-hPa tangential wind. V850 is possibly the highest contributing tangential wind feature because the TC maximum tangential wind usually occurs between 800 and 900 hPa (Doyle et al. 2017). Figure 12 shows that lower values of V850 are mainly associated with more positive SHAP values. Weaker TCs have lower tangential wind (Doyle et al. 2017; Stern and Nolan 2011). Therefore, this correlation is likely due to already strong TCs being less likely to undergo RI. Lower values of V850 are also associated with smaller TCs, as the V850 is highly correlated with the radius of outer closed isobar (ROCI). Smaller TCs are also more favorable for rapid deepening (Knaff et al. 2018); however, Carrasco et al. (2014) indicated the ROCI appears to have little to no relationship with subsequent intensification. Therefore, it is unclear if the lower V850 is indicative of smaller TCs, which are more likely to rapidly intensify.

The correlation between all feature and SHAP values, ranked by contribution to the total SHAP value, for all RI thresholds for all 2019 and 2020 TCs is shown in Fig. 13. Features at the bottom of the panels with no plotted correlations are not used when predicting RI at any threshold in the basin. As in Fig. 12, the satellite features have the highest individual contributions to the total SHAP values. In the eastern North Pacific basin, the contribution of the IR feature is higher than most IR difference features, possibly due to the disparity in convection between intensifying and weakening TCs being greater in the eastern North Pacific basin than North Atlantic basin (DeMaria et al. 2012). For the scalar features, many of the correlations between feature and SHAP value are consistent with the SHIPS Consensus RI models. For example, increased POT, DELV (known as PER in SHIPS-RII), ocean temperatures (COHC, RSST, CD26), and total precipitable water (TPW), as well as low VWS, are correlated with high SHAP values and increased probability of RI (K15). VWS between 850 and 200 hPa (SHDC, SHRD) contributes more to the total SHAP value than VWS between 850 and 500 hPa

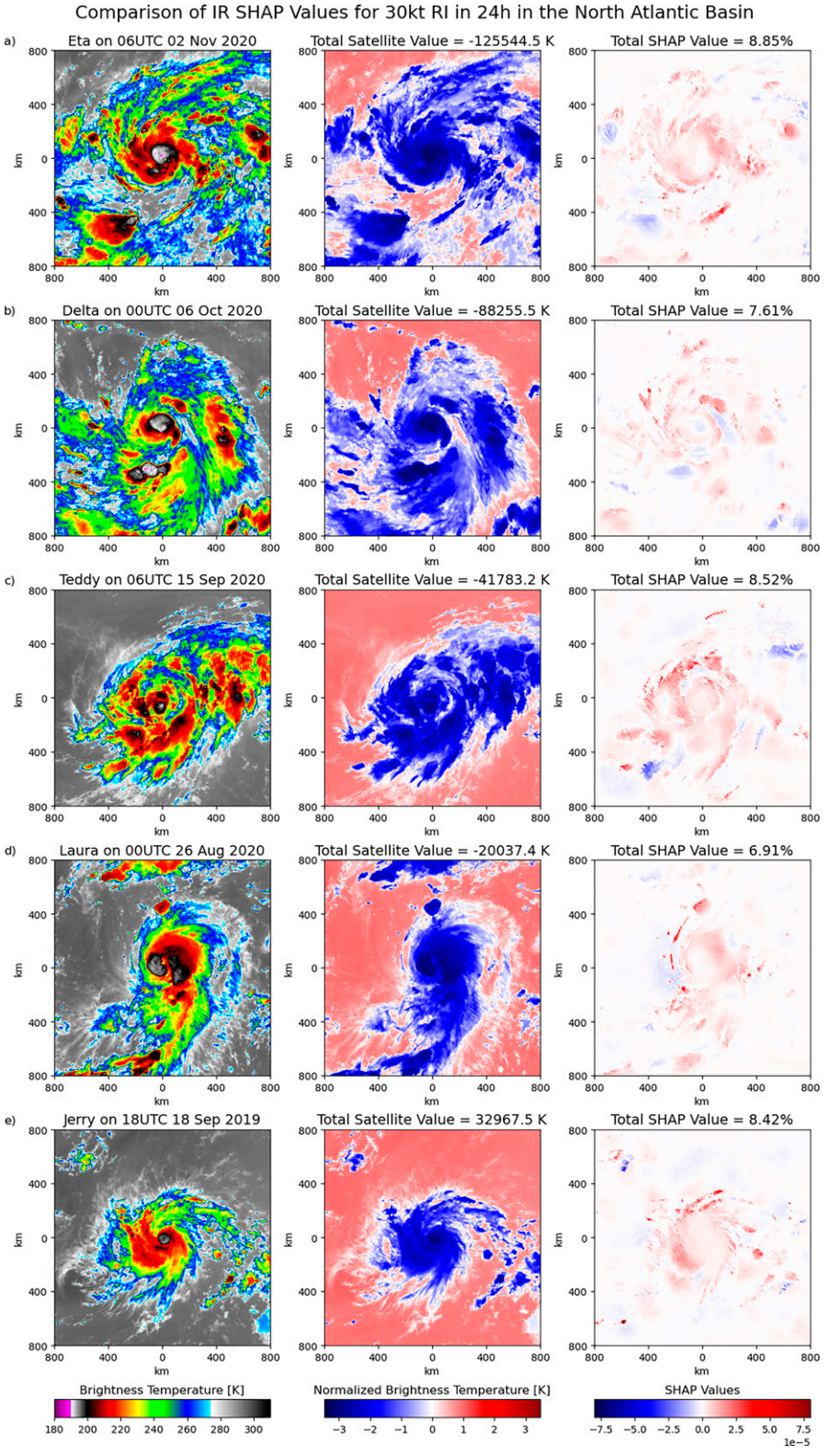

FIG. 10. IR data for a given RI forecast, normalized IR data for AI-RI, and corresponding SHAP values for five TCs that underwent 30-kt RI in the North Atlantic basin.

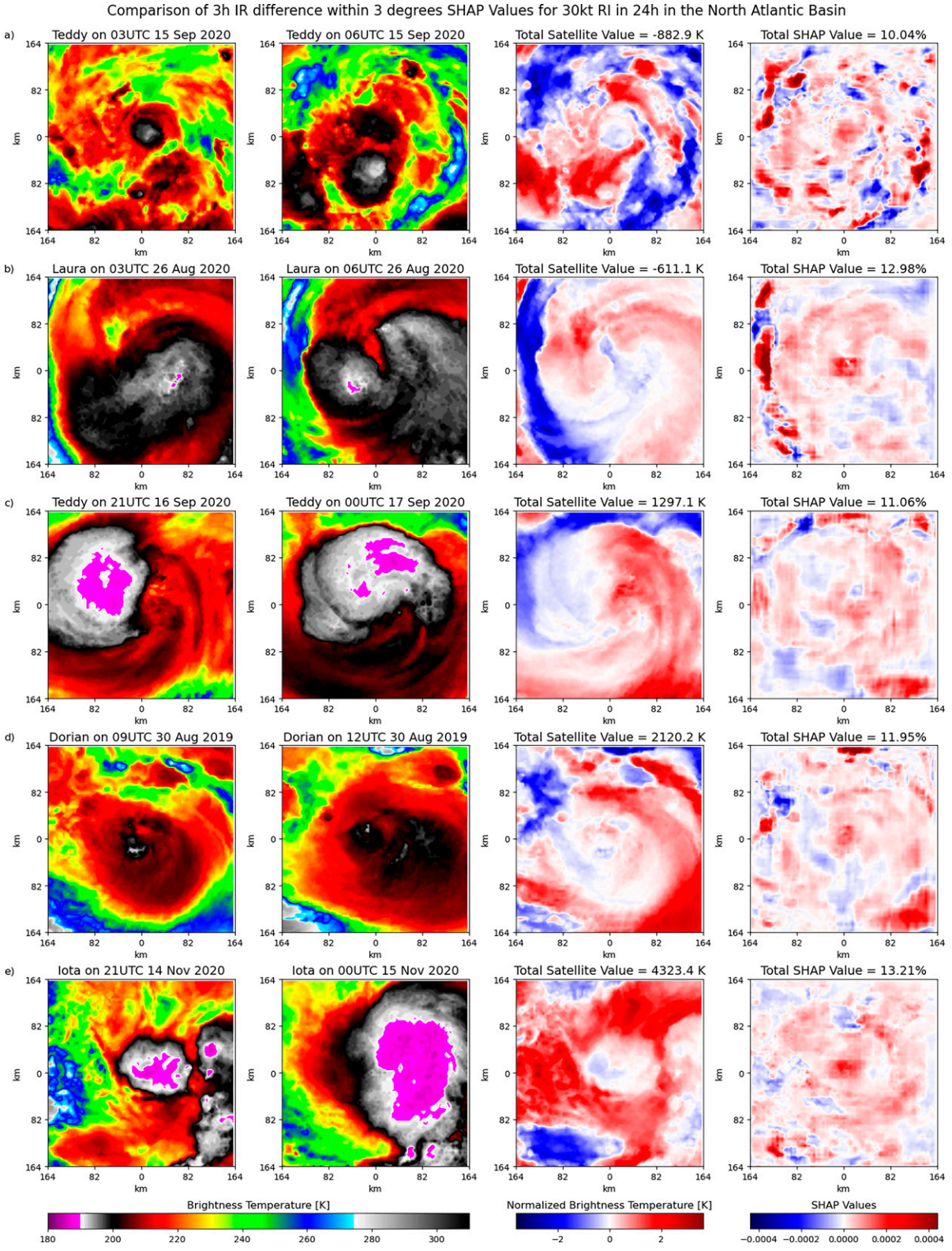

FIG. 11. IR data within 3° of the TC center from 3 h prior to RI forecast, current IR data, normalized difference between the 3-h prior IR data and the current IR data for AI-RI, and corresponding SHAP values for five TCs that underwent 30-kt RI in the North Atlantic basin.

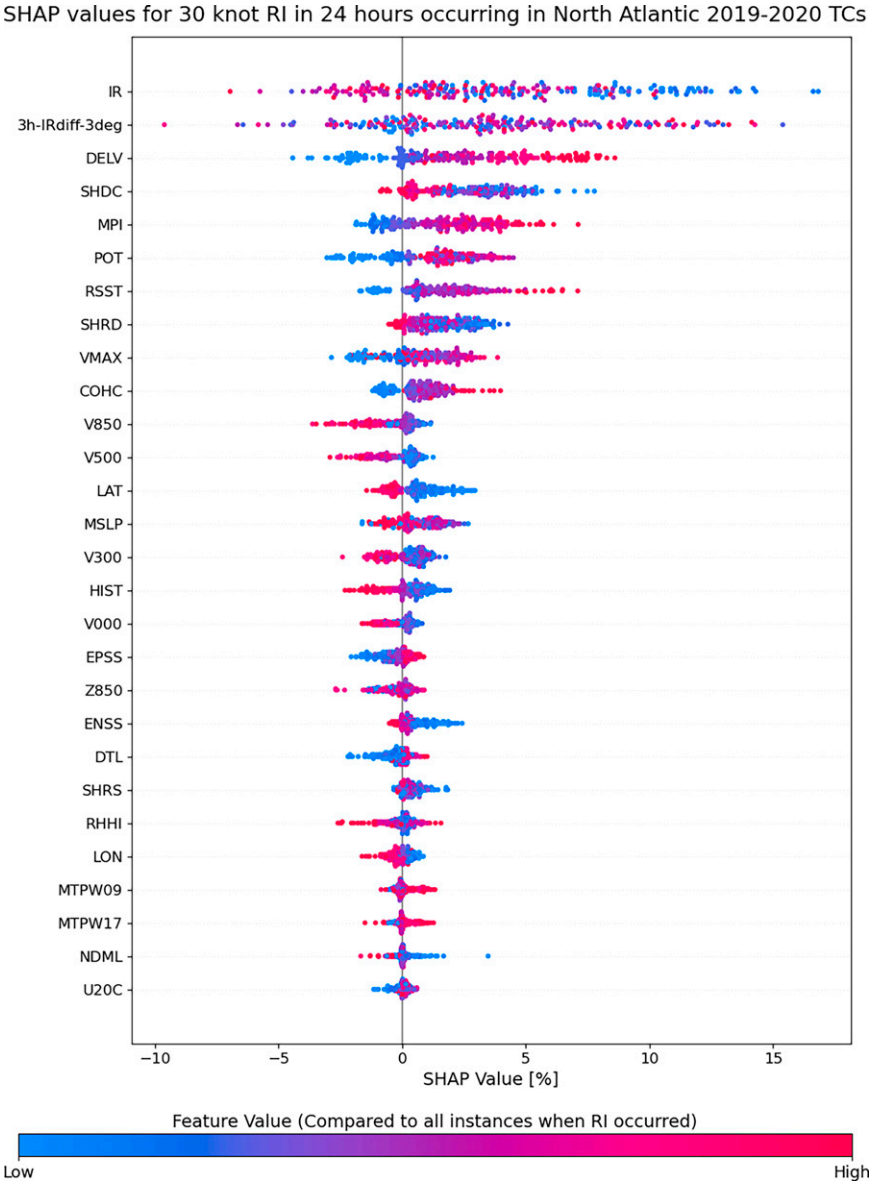

FIG. 12. Feature SHAP values for all instances of 30-kt RI occurring in the 2019–20 TCs in the North Atlantic basin (N = 40 RI cases). Features are sorted from highest to lowest based on their contribution to the overall SHAP value total. Each RI forecast is represented by a dot, with the color of each individual dot indicating the value of a given feature is low (blue) or high (red) compared to all values of that given feature when 30-kt RI occurred in 2019 and 2020 North Atlantic TCs.

(SHRS). Also consistent with K15 is the correlation between VMAX and SHAP values. Decreasing VMAX is correlated with increasing SHAP values as the RI time period lengthens, similar to K15 finding that RI occurs at lower VMAX values as the RI lead time increases. High EPSS and low ENSS is correlated with increased SHAP values, also as seen in K15. For relative humidity (RH) predictors, increased RHLO is correlated with higher SHAP values (consistent with KDK10), while increased RHMD is also mostly correlated with higher SHAP values. For RHHI, decreased RH is associated with increased SHAP values, except for 65-kt RI. These correlations are potentially due to the relationship between intensity and 300–500-hPa RH, as weaker TCs tend to have higher values of RH at these levels (Wu et al. 2012). Conversely, no consistent correlation between D200 and SHAP values exists, though K15 suggests D200 is higher when RI occurs.

While many of the scalar features are consistent between the AI-RI and SHIPS Consensus models, other scalar features in AI-RI are not considered by the SHIPS Consensus RI models. Most noticeable is LAT, one of the larger

FIG. 13. Correlation between feature and SHAP values for all 2019–20 TCs. Features are sorted from highest to lowest based on their contribution to the overall SHAP value total.

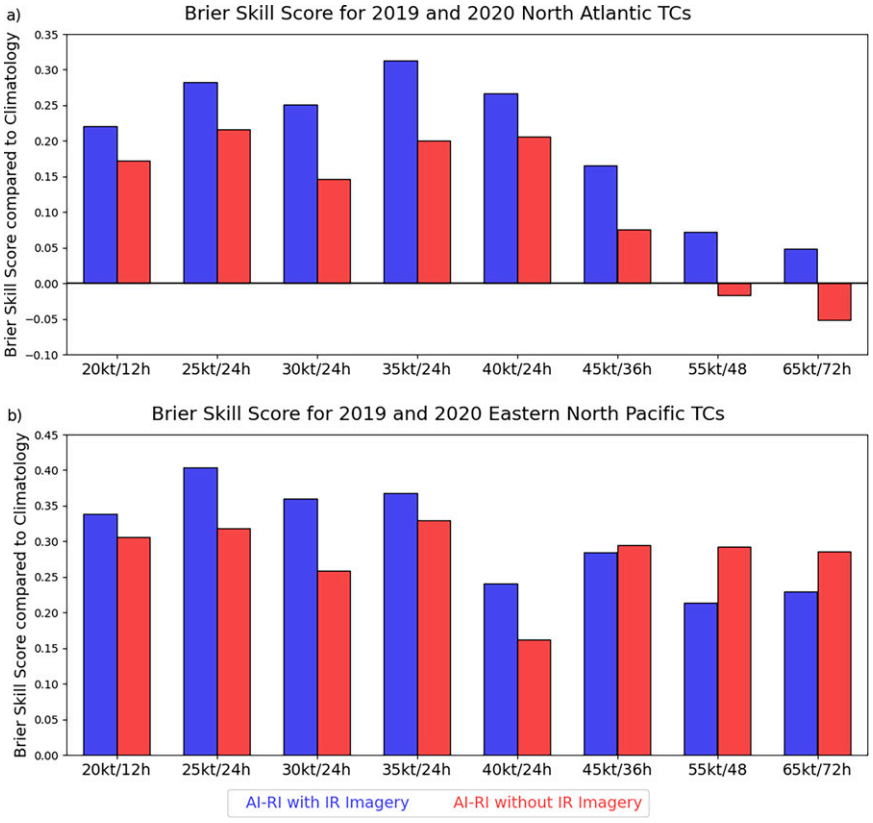

FIG. 14. The Brier skill score (BSS) compared to climatology for AI-RI trained with (blue) and without (red) the IR satellite features for 2019–20 (a) North Atlantic basin and (b) eastern North Pacific TCs.

contributors to total SHAP values. Latitude is negatively correlated with SHAP values. Therefore, lower latitude TCs are more likely to undergo RI, which is consistent with a study of western North Pacific TCs by Wang and Zhou (2008). Latitude is included in DTOPS and is also negatively correlated with RI probability (Onderlinde and DeMaria 2018). Other scalar features not considered by the SHIPS Consensus models that are a part of AI-RI include the tangential wind features (previously described) as well as HIST, Z850, and TOD. V850 is in the upper half of contribution features for AI-RI, and greater differences between the AI-RI and DTOPS probabilities of RI are moderately correlated with increased SHAP values from V850. HIST is also a moderate contributor to total SHAP values; the probability of RI is increased for TCs early in their life cycle. Z850 and TOD are generally lower contributors to the total SHAP values, and do not have consistent correlations between SHAP and feature values for most RI thresholds. Higher Z850 SHAP values, though, are correlated with a greater difference between the AI-RI and DTOPS probabilities for RI lead times of 24 h or less.

The changing of various features in this model across each of the RI thresholds and basins remains an interesting but unresolved issue, just as in previous models for RI (Rozoff and Kossin 2011; K15; Shaiba and Hahsler 2016). Any attempt to account for these differences would exceed the precision of the limited training dataset. As it stands, the varying feature differences across threshold and basin suggest an intriguingly complex interplay between these factors and the varieties of RI.

The importance of satellite features to the prediction of RI can be further observed in Fig. 14. Figure 14 compares the BSS for AI-RI trained with IR imagery (blue bars) to a CNN trained without IR imagery (red bars). For all RI thresholds, except those over 36 h or longer in the eastern North Pacific basin, AI-RI with IR imagery is more skillful than without IR imagery, further highlighting the importance of IR imagery.

6. Summary and conclusions

This study develops a convolutional neural network, named AI-RI, to predict the probability of rapid intensification (RI) for North Atlantic and eastern North Pacific tropical cyclones (TCs) using satellite infrared (IR) imagery, as well as scalar features. A selection of optimal features for AI-RI are determined by first developing a "kitchen sink" version of the model, which uses all scalar features listed in Table 1 in addition to the satellite features, and then randomizing the data for one feature at a time and recalculating the probability of

RI to determine the feature's impact on RI prediction skill. A feature is considered optimal for RI prediction if AI-RI skill, denoted by the Brier skill score, decreases with the feature's randomization. RI is predicted for 8 different intensity increase (VMAX) thresholds: 20 kt in 12 h; 25, 30, 35, and 40 kt in 24 h; 45 kt in 36 h; 55 kt in 48 h, and 65 kt in 72 h.

AI-RI is tested on an independent dataset consisting of North Atlantic and eastern North Pacific TCs in 2019 and 2020. RI probabilities are compared with two existing objective methods used operationally at the National Hurricane Center: SHIPS Consensus (which includes the SHIPS-RII model) and DTOPS (which includes some deterministic model predictors). In the North Atlantic basin, results show that AI-RI is more skillful for seven of the eight RI thresholds compared to DTOPS, and three RI thresholds over the 24-h time period compared to SHIPS Consensus. Adding AI-RI into the SHIPS Consensus is also more skillful than SHIPS Consensus for all of the 12- and 24-h RI thresholds. For these five RI thresholds, on average AI-RI has higher RI probabilities for cases when RI occurs compared to SHIPS Consensus, and lower probabilities when RI does not occur. For RI thresholds covering 36-h lead times and longer, AI-RI is less skillful than the SHIPS Consensus due to higher average RI probabilities when RI does and does not occur. One hypothesis for these higher probabilities is that the satellite-derived features are less relevant to longer RI lead times. An exception occurs for the 45-kt RI category, where the AI-RI probability of RI is lower when RI occurs. This is possibly due to 45-kt RI occurring in many larger late-season October and November North Atlantic TCs in 2019 and 2020, and the IR difference feature within 3° feature was inadequate for these larger systems.

In the eastern North Pacific basin, AI-RI is more skillful than SHIPS Consensus for four of the eight RI thresholds, and including the AI-RI in the SHIPS Consensus has a higher or similar skill than SHIPS Consensus for all eight RI thresholds. Therefore, AI-RI is a skillful empirical method for predicting RI in this TC basin. However, the quasi-deterministic DTOPS model is more skillful than AI-RI for all RI thresholds except 65 kt. However, for TCs north of 15°N, AI-RI is more skillful than DTOPS for five of the RI thresholds in this basin.

In addition to skillfully predicting TC RI, AI-RI highlights the importance of vigorous organized convection in initiating RI, with strongly positive SHAP values corresponding to convective features in IR satellite imagery. Though it can be difficult to "see" RI in a given IR satellite image, the highest contributing features to the AI-RI probabilities are IR-based. By comparison, the relative weight of the IR predictors ranks third or fourth in SHIPS-RII, and is lower than vertical wind shear (K15). Perhaps more importantly, the convective trend (IRdiff feature) supplies a strong signal to the AI-RI probabilities, whereas other existing RI methods do not include this predictor. Since convective behavior is difficult to quantify as a scalar input to algorithms such as SHIPS-RII, the convolutional neural network used here is more appropriate for the task. SHAP values indicate the importance of also including short-term trends in the IR pixels (and cloud organization). These trends are missed by the existing operational RI models that only input and analyze the current cloud pattern.

Since AI-RI identifies convection as an important contributor to RI probabilities, future work will include training AI-RI with a short series of rapid-scan IR images to better identify active convection, as convective overshooting tops can have a lifespan as short as 10 min (Gettelman et al. 2002). In addition, AI-RI will be trained with additional two-dimensional environmental features, such as vertical wind shear and layered precipitable water, to identify if the location/orientation of these features with respect to the TC core can improve RI prediction as suggested by K15.

Acknowledgments. The authors of this paper would like to thank Tim Olander of UW–CIMSS for creating the TC-centered infrared satellite dataset used in this analysis and Mark DeMaria and the SHIPS team at the Regional and Mesoscale Meteorology Branch for providing the archived SHIPS predictor files used to obtain the scalar features. Also, thank you to Ryan Lagerquist, John Cintineo, and Charles White for their assistance in optimizing the process to develop AI-RI. This research was funded by the Office of Naval Research Award N00014-20-1-2149: A Deep Learning Approach to Examining and Predicting TC Rapid Intensification.

Data availability statement. Tropical cyclone best tracks can be found at https://ftp.nhc.noaa.gov/atcf/archive/and the scalar features used in this analysis are available from the SHIPS developmental data at https://rammb.cira.colostate.edu/research/tropical_cyclones/ships/developmental_data.asp. Archived satellite data can be found via the Space Science and Engineering online archive. More information can be found at https://www.ssec.wisc.edu/datacenter/goes-archive/.

REFERENCES

Adler, R. F., and E. B. Rodgers, 1977: Satellite-observed latent heat release in a tropical cyclone. *Mon. Wea. Rev.*, **105**, 956–963, https://doi.org/10.1175/1520-0493(1977)105<0956:SOLHRI>2.0.CO;2.

Alcala, C. M., and A. E. Dessler, 2002: Observations of deep convection in the tropics using the Tropical Rainfall Measuring Mission (TRMM) precipitation radar. *J. Geophys. Res.*, **107**, 4792, https://doi.org/10.1029/2002JD002457.

Cangialosi, J. P., and J. L. Franklin, 2014: 2013 National Hurricane Center verification report. NHC, 84 pp., http://www.nhc.noaa.gov/verification/pdfs/Verification_2013.pdf.

——, E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Wea. Forecasting*, **35**, 1913–1922, https://doi.org/10.1175/WAF-D-20-0059.1.

Carrasco, C. A., C. W. Landsea, and Y. Lin, 2014: The influence of tropical cyclone size on its intensification. *Wea. Forecasting*, **29**, 582–590, https://doi.org/10.1175/WAF-D-13-00092.1.

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., 2018: *Deep Learning with Python*. Manning Publications Co., 361 pp.

DeMaria, M., R. T. DeMaria, J. A. Knaff, and D. Molenar, 2012: Tropical cyclone lightning and rapid intensity change. *Mon. Wea. Rev.*, **140**, 1828–1842, https://doi.org/10.1175/MWR-D-11-00236.1.

——, J. L. Franklin, M. J. Onderlinde, and J. Kaplan, 2021: Operational forecasting of tropical cyclone rapid intensification at the National Hurricane Center. *Atmosphere*, **12**, 683, https://doi.org/10.3390/atmos12060683.

Doyle, J., and Coauthors, 2017: A view of tropical cyclones from above: The Tropical Cyclone Intensity Experiment. *Bull. Amer. Meteor. Soc.*, **98**, 2113–2134, https://doi.org/10.1175/BAMS-D-16-0055.1.

Gettelman, A., M. L. Salby, and F. Sassi, 2002: Distribution and influence of convection in the tropical tropopause region. *J. Geophys. Res.*, **107**, 4080, https://doi.org/10.1029/2001JD001048.

Guimond, S. R., G. M., Heymsfield, and F. J., Turk, 2010: Multiscale observations of Hurricane Dennis (2005): The effects of hot towers on rapid intensification. *J. Atmos. Sci.*, **67**, 633–654, https://doi.org/10.1175/2009JAS3119.1.

Hinton, G., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2012: Improving neural networks by preventing co-adaptation of feature detectors. arXiv, 12070580, https://arxiv.org/pdf/1207.0580.pdf.

Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108, https://doi.org/10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2.

——, ——, and J. A. Knaff, 2010: A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **25**, 220–241, https://doi.org/10.1175/2009WAF2222280.1.

——, and Coauthors, 2015: Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Wea. Forecasting*, **30**, 1374–1396, https://doi.org/10.1175/WAF-D-15-0032.1.

Klotzbach, P. J., and Coauthors, 2022: A hyperactive end to the Atlantic hurricane season: October–November 2020. *Bull. Amer. Meteor. Soc.*, **103**, E110–E128, https://doi.org/10.1175/BAMS-D-20-0312.1.

Knaff, J. A., C. R. Sampson, and K. D. Musgrave, 2018: An operational rapid intensification prediction aid for the western North Pacific. *Wea. Forecasting*, **33**, 799–811, https://doi.org/10.1175/WAF-D-18-0012.1.

Kuo, H. L., 1965: On formation and intensification of tropical cyclones through latent heat release by cumulus convection. *J. Atmos. Sci.*, **22**, 40–63, https://doi.org/10.1175/1520-0469(1965)022<0040:OFAIOT>2.0.CO;2.

Lagerquist, R., A. McGovern, and D. Gagne, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Wea. Forecasting*, **34**, 1137–1160, https://doi.org/10.1175/WAF-D-18-0183.1.

——, ——, C. R. Homeyer, D. J. Gange II, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

Li, F., J. Johnson, and S. Yeung, 2020: CS231n convolutional neural networks for visual recognition. GitHub, accessed 7 July 2021, http://cs231n.github.io/convolutional-networks/.

Liu, C., and E. J. Zipser, 2005: Global distribution of convection penetrating the tropical tropopause. *J. Geophys. Res.*, **110**, D23104, https://doi.org/10.1029/2005JD006063.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions, arXiv, 1705.07874, https://arxiv.org/abs/1705.07874.

——, G. G. Erion, and S.-I. Lee, 2018: Consistent individualized feature attribution for tree ensembles. arXiv, 1802.03888, https://arxiv.org/abs/1802.03888.

——, and Coauthors, 2020: From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 56–67, https://doi.org/10.1038/s42256-019-0138-9.

Maas, A., A. Hannun, and A. Ng, 2013: Rectifier nonlinearities improve neural network acoustic models. *Proc. 30th Int. Conf. on Machine Learning*, Atlanta, GA, International Machine Learning Society, 6 pp., https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.

Mangalathu, S., S.-H. Hwang, and J.-S. Jeon, 2020: Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Eng. Struct.*, **219**, 110927, https://doi.org/10.1016/j.engstruct.2020.110927.

Mercer, A. E., A. D. Grimes, and K. M. Wood, 2021: Application of unsupervised learning techniques to identify Atlantic tropical cyclone rapid intensification environments. *J. Appl. Meteor. Climatol.*, **60**, 119–138, https://doi.org/10.1175/JAMC-D-20-0105.1.

Monette, S. A., C. S. Velden, and K. S. Griffin, 2012: Examining trends in satellite-derived tropical overshooting tops as a potential predictor of tropical cyclone rapid intensification. *J. Appl. Meteor. Climatol.*, **51**, 1917–1930, https://doi.org/10.1175/JAMC-D-11-0230.1.

Olander, T. L., and C. S. Velden, 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.

Onderlinde, M., and M. DeMaria, 2018: Deterministic to probabilistic statistical rapid intensification index (DTOPS): A new method for forecasting RI probability. *33rd Conf. on Hurricanes and Tropical Meteorology*, Ponte Vedra, FL, Amer. Meteor. Soc., 16C.3, https://ams.confex.com/ams/33HURRICANE/webprogram/Paper339346.html.

Rogers, R. F., P. D. Reasor, and S. Lorsolo, 2013: Airborne Doppler observations of the inner-core structural differences between intensifying and steady-state tropical cyclones. *Mon. Wea. Rev.*, **141**, 2970–2991, https://doi.org/10.1175/MWR-D-12-00357.1.

——, ——, and J. Zhang, 2015: Multiscale structure and evolution of Hurricane Earl (2010) during rapid intensification. *Mon. Wea. Rev.*, **143**, 536–562, https://doi.org/10.1175/MWR-D-14-00175.1.

Romps, D. M., and Z. Kuang, 2009: Overshooting convection in tropical cyclones. *Geophys. Res. Lett.*, **36**, L09804, https://doi.org/10.1029/2009GL037396.

Rozoff, C. M., and J. P. Kossin, 2011: New probabilistic forecast models for the prediction of tropical cyclone rapid intensification. *Wea. Forecasting*, **26**, 677–689, https://doi.org/10.1175/WAF-D-10-05059.1.

Shaiba, H., and M. Hahsler, 2016: Applying machine learning methods for predicting tropical cyclone rapid intensification events. *Res. J. Appl. Sci. Eng. Technol.*, **13**, 638–651, https://doi.org/10.19026/rjaset.13.3050.

Shapley, L., 1953: A value for n-Person games. *Contributions to the Theory of Games (AM-28)*, Vol. II, Princeton University Press, 307–318, https://doi.org/10.1515/9781400881970-018.

Steranka, J., E. B. Rodgers, and R. C. Gentry, 1986: The relationship between satellite measured convective bursts and tropical cyclone intensification. *Mon. Wea. Rev.*, **114**, 1539–1546, https://doi.org/10.1175/1520-0493(1986)114<1539:TRBSMC>2.0.CO;2.

Stern, D. P., and D. S. Nolan, 2011: On the vertical decay of the maximum tangential winds in tropical cyclones. *J. Atmos. Sci.*, **68**, 2073–2094, https://doi.org/10.1175/2011JAS3682.1.

Vigh, J. L., J. A. Knaff, and W. H. Schubert, 2012: A climatology of hurricane eye formation. *Mon. Wea. Rev.*, **140**, 1405–1426, https://doi.org/10.1175/MWR-D-11-00108.1.

Wang, B., and X. Zhou, 2008: Climate variation and prediction of rapid intensification in tropical cyclones in the western North Pacific. *Meteor. Atmos. Phys.*, **99**, 1–16, https://doi.org/10.1007/s00703-006-0238-z.

Wang, H., and Y., Wang, 2014: A numerical study of Typhoon Megi (2010). Part I: Rapid intensification. *Mon. Wea. Rev.*, **142**, 29–48, https://doi.org/10.1175/MWR-D-13-00070.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Wimmers, A., C. Velden, 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.

Wu, L., and Coauthors, 2012: Relationship of environmental relative humidity with North Atlantic tropical cyclone intensity and intensification rate. *Geophys. Res. Lett.*, **39**, L20809, https://doi.org/10.1029/2012GL053546.

Xu, W., K. Balaguru, A. August, N. Lalo, N. Hodas, M. DeMaria, and D. Judi, 2021: Deep learning experiments for tropical cyclone intensity forecasts. *Wea. Forecasting*, **36**, 1453–1470, https://doi.org/10.1175/WAF-D-20-0104.1.

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.