

Identifying Data Sources and Physical Strategies Used by Neural Networks to Predict TC Rapid Intensification

RYAN LAGERQUIST^{a,b}, JOHN A. KNAFF^d, CHRISTOPHER J. SLOCUM^d, KATE MUSGRAVE^a,
AND IMME EBERT-UPHOFF^{a,c}

^a Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado

^b NOAA/Global Systems Laboratory, Boulder, Colorado

^c Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado

^d NOAA Center for Satellite Applications and Research, Fort Collins, Colorado

(Manuscript received 15 August 2024, in final form 5 March 2025, accepted 23 March 2025)

ABSTRACT: The rapid intensification (RI) of tropical cyclones (TCs), defined here as an intensity increase of ≥ 30 kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$) in 24 h, is a difficult but important forecasting problem. Operational RI forecasts have considerably improved since the late 2000s, largely thanks to better statistical models, including machine learning (ML). Most ML applications use scalars from the Statistical Hurricane Intensity Prediction Scheme (SHIPS) development dataset as predictors, describing the TC history, near-TC environment, and satellite presentation of the TC. More recent ML applications use convolutional neural networks (CNNs), which can ingest full satellite images (or time series of images) and freely “decide” which spatiotemporal features are important for RI. However, two questions remain unanswered: 1) Does image convolution significantly improve RI skill? 2) What strategies do CNNs use for RI prediction—and can we gain new insights from these strategies? We use an ablation experiment to answer the first question and explainable artificial intelligence (XAI) to answer the second. Convolution leads to only a small performance gain, likely because, as revealed by XAI, the CNN’s main strategy uses image features already well described in scalar predictors used by preexisting RI models. This work makes three additional contributions to the literature: 1) Neural networks (NNs) with SHIPS data outperform preexisting models in some aspects; 2) NNs provide well calibrated uncertainty quantification (UQ), while preexisting models have no UQ; 3) the NN without SHIPS data performs surprisingly well and is fairly independent of preexisting models, suggesting its potential value in an operational ensemble.

SIGNIFICANCE STATEMENT: Rapid intensification (RI) is a hazardous event and is difficult to forecast. RI forecasting could potentially be improved by machine learning (ML) models that ingest full satellite images—or even time series of images—centered on a tropical cyclone (TC). We investigate whether this improvement is actually realized and whether explainable ML can teach us anything new about RI precursors in satellite imagery.

KEYWORDS: Hurricanes/typhoons; Tropical cyclones; Empirical orthogonal functions; Machine learning; Model interpretation and visualization; Neural networks

1. Introduction

Tropical cyclone (TC) warning centers around the globe issue 6-hourly forecasts for all active TCs, including two important quantities: track, defined as the geographic location of the TC center, and intensity, defined as the spatial maximum of sustained wind¹ inside the TC. While forecasts of both quantities have steadily improved (DeMaria et al. 2014, 2021), the National Hurricane Center (NHC) and Joint Typhoon Warning Center (JTWC) note that rapid intensification (RI) remains a difficult problem (Cangialosi et al. 2020). RI is also

a problem of great societal importance, because this rapid change often overwhelms the capabilities of the public and emergency management, leaving insufficient time for preparedness activities. This is especially true when RI—typically defined as an intensity increase of 30 kt (15.4 m s^{-1}) within 24 h (Rappaport et al. 2012)²—occurs shortly before landfall. While RI is not a special physical process, it represents the extreme end of the intensity change distribution, with $\sim 5\%$ of 24-h intensity changes meeting the 30-kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$) threshold (Kaplan and DeMaria 2003; Kaplan et al. 2015). Also, most major TCs—with lifetime maximum intensity > 96 kt (49.4 m s^{-1})—undergo RI at least once (Lee et al. 2016). Recent improvements in RI forecasting have been afforded by better understanding of the processes responsible for RI; better dynamical models, with sufficient resolution to capture these processes; and better statistical models, which often bias

¹ The 1-min-average wind speed at 10 m above ground level.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-24-0166.s1>.

Corresponding author: Ryan Lagerquist, ralager@colostate.edu

² There are other definitions of RI, typically defined for longer time periods, such as 65 kt in 72 h (DeMaria et al. 2021). However, in this paper we focus on 30 kt (24 h)^{−1}.

correct the systematic underforecasting of RI by dynamical models (DeMaria et al. 2021).

The rest of this introduction reviews the literature on TC dynamics (because we seek to connect our models' RI-forecasting strategies to TC dynamics), followed by the literature on statistical RI forecasting, and then concludes with the focus of the current study.

a. Background on TC dynamics

Processes leading to RI can be divided into three categories: environmental, oceanic, and inner core. The first two categories represent most predictor variables used in statistical RI models. For example, two seminal studies (Kaplan and DeMaria 2003; Kaplan et al. 2010) found that RI is favored by small values of vertical wind shear; large differences between the TC's current intensity and maximum potential intensity (MPI; Emanuel 1986), the latter determined by the environment; and large values of sea surface temperature (SST), ocean heat content (OHC), low-to-midlevel humidity, upper-level divergence, and recent intensity increase. Kaplan et al. (2015) also identified current intensity I_0 as important for RI. They found that RI is favored by intermediate values of I_0 ; weak TCs (with small I_0) are too disorganized to undergo RI, and strong TCs (with large I_0) are too close to their MPI. Recent machine learning studies (Xu et al. 2021; Griffin et al. 2022) have also found that RI is favored by weaker 200-hPa zonal wind, which may indicate less vertical wind shear or less ventilation of the warm core by environmental flows.

The inner core of a TC is dominated by two circulations: the primary circulation, characterized by horizontal rotation around a low pressure center; and the secondary circulation, characterized by vertical and radial motions which act to warm and moisten the TC core. TC intensity is determined largely by the strength of the coupling between these two circulations, which is often influenced by convective bursts (CBs; Shapiro and Willoughby 1982; Heymsfield et al. 2001), defined as convection with an updraft speed exceeding a certain threshold, often 5 m s^{-1} (Black et al. 1996; Reasor et al. 2009; Rogers 2010). However, recent studies suggest that the pathway from CBs to RI is indirect and involves weaker convection as well (Rogers 2010; Chen and Zhang 2013; Tang et al. 2018). RI usually coincides with the axisymmetrization of convection in the inner core—i.e., coverage evolves from one or two quadrants to a full ring around the TC center (Rogers 2010; Chen and Zhang 2013; Tang et al. 2018).

Other inner-core processes may sometimes be responsible for RI, including vortex Rossby waves and mesovortices caused by barotropic-convective instability (Montgomery and Kallenbach 1997; Rogers 2010; Alford et al. 2023). However, these processes are not reliably identifiable in the infrared satellite imagery used in this study; rather, they are typically identified by numerical simulations or targeted field observations (Cotto et al. 2015; Alford et al. 2023).

b. Background on statistical RI forecasting

Statistical models are often used to operationally forecast RI, because dynamical models suffer from an extreme

underforecasting bias (DeMaria et al. 2021, henceforth D21). Three popular statistical models are the Statistical Hurricane Intensity Prediction Scheme (SHIPS)–RI index (RII), SHIPS consensus, and deterministic-to-probabilistic statistical (DTOPS) model. The SHIPS-RII has been used by the NHC since 2001, undergoing several modifications during that time (D21). For example, Kaplan et al. (2010) replaced the original thresholding algorithm of Kaplan and DeMaria (2003) with linear discriminant analysis (LDA). The SHIPS consensus averages output from three models: the preexisting LDA, logistic regression, and naïve Bayes (Kaplan et al. 2015, henceforth K15). DTOPS, which was developed in 2018 and introduced globally in 2023 (D21; Sampson et al. 2023), uses logistic regression to combine deterministic intensity forecasts from dynamical models into one probabilistic RI forecast. Although DTOPS uses no predictors other than intensity forecasts and the TC's initial latitude/intensity, it outperforms all versions of SHIPS-RII (Fig. 10 of D21), showing that the deterministic models have skill that can be “unlocked” with simple ensemble and bias correction.

Some of these statistical models have recently been expanded to a global domain (Knaff et al. 2018, 2020) or to include novel observations, such as microwave satellite imagery and lightning (DeMaria et al. 2012; Su et al. 2020; Slocum et al. 2023). However, statistical RI models generally rely on the same data sources, so their predictions are highly interdependent. While the spread is large enough that these models can be profitably blended in a consensus (Sampson et al. 2023), some questions remain on how to gain additional independent skill with new models.

Many statistical RI models are examples of ML, a field characterized by complex, nonlinear, and data-driven methods. There are also more recent ML applications in RI research (McNeely et al. 2020; Mercer et al. 2021), including deep neural networks (NNs). The deep NNs of Griffin et al. (2022) show some advantages over operational models in predicting 24-h RI, while the deep NNs of Xu et al. (2021) beat most operational models in predicting 24-h intensity change (not specifically RI).

c. Focus of this study

A key advantage of convolutional NNs (CNNs) is that they can ingest a full image—or even a video (time series of images)—instead of summary statistics describing the image. This allows a CNN to freely “decide” which spatial features are important for RI prediction—features that could be missed by summary statistics. However, two science questions remain unanswered:

- 1) What is the influence of different predictor sources on RI skill? For example, does the use of full satellite images or videos (which requires more computationally expensive models such as CNNs) lead to a significant improvement in RI skill in practice?
- 2) When ML models are applied to full satellite images, what strategies do they use for RI prediction? Can we learn anything new, especially *vis-à-vis* convection and CBs in the inner core?

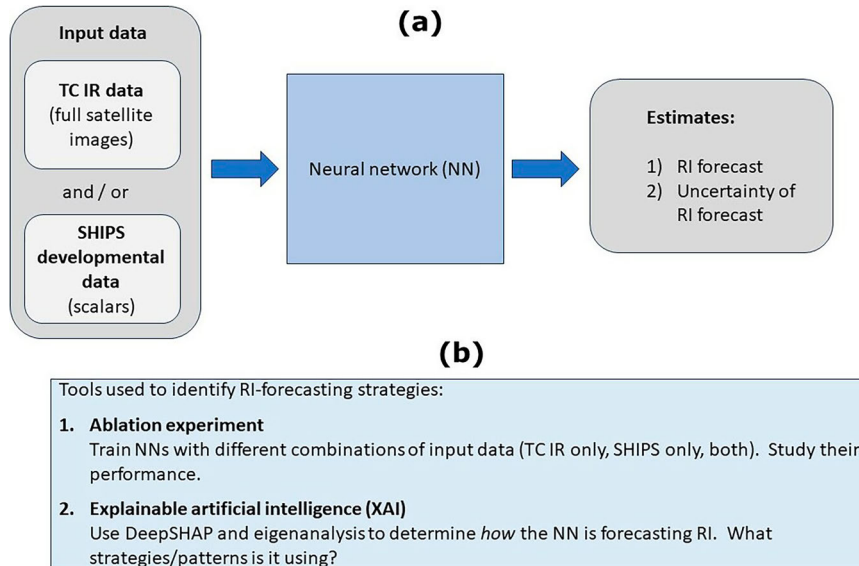


FIG. 1. Overview of (a) our approach to forecasting RI with NNs and (b) our approach to understanding what the NNs have learned.

To answer the first question, we perform an ablation experiment, training NNs with and without different data sources such as full satellite images (Fig. 1a). To answer the second question, we use explainable artificial intelligence (XAI). Furthermore, we incorporate uncertainty quantification (UQ) into our NNs, training them to output a well-calibrated ensemble of RI predictions. UQ is crucial for high-impact decision-making and can be used to assess the trustworthiness of ML models on a case-by-case basis (Haynes et al. 2023); it also provides forecasters with an often-requested objective measure of uncertainty.

The rest of this article is organized as follows. Sections 2 and 3 describe our data and methodology; section 4 shows the results of our ablation experiment, including a detailed evaluation of the selected NN models and comparison of these with baseline models used operationally; section 5 shows the results of our XAI analysis, including a detailed discussion of what the NNs have learned; and section 6 summarizes our findings.

2. Input data

This section describes the three datasets used to create output (target) variables and input (predictor) variables for the NNs, as well as preprocessing for these variables. We use TCs from around the globe: specifically the North Atlantic (AL), eastern North Pacific (EP), western North Pacific (WP), central North Pacific (CP), Southern Hemisphere (SH), and Indian Ocean (IO) Basins.

a. Best track intensity data (ground truth)

We use the final best tracks—provided by the NHC, JTWC, and Central Pacific Hurricane Center—from the archives of the Automated Tropical Cyclone Forecast (ATCF) system (Sampson and Schrader 2000). Best tracks are updated at

synoptic times (0000, 0600, 1200, and 1800 UTC)—and occasionally other important times, such as landfall—throughout the lifetime of each TC. This postevent dataset (reanalyzed after every TC season) contains a smoothed representation of the best subjective analysis for many TC properties, including location, storm motion, intensity, etc. For this study, we use only one best track variable—intensity—to provide correct answers (labels) for the NN models. Specifically, for each TC object, the binary label is created as follows:

$$y = \mathcal{I} \left\{ \max_{t=t_0}^{t=t_0+24\text{h}} (I_t) - I_{t_0} \geq 30 \text{ kt} \right\}, \quad (1)$$

where I_t is the best track intensity at time t and $\mathcal{I}$ is the indicator function, evaluating to 1 if the condition is true and 0 if false. Thus, $y = 1$ if, at any time between t_0 and $t_0 + 24$ h, the intensity exceeds its t_0 value by at least 30 kt.

b. TC IR satellite data (image predictors)

This study uses a global archive of infrared images, henceforth the TC IR archive. The archive is maintained and updated by the NOAA Center for Satellite Applications and Research, collocated with the Cooperative Institute for Research in the Atmosphere (CIIRA) at Colorado State University. The IR images are mapped to a Mercator projection on a 480×640 pixel ($1920 \text{ km} \times 2560 \text{ km}$) grid with 4-km spacing. The images contain one variable—brightness temperature at a central wavelength of $\sim 11 \mu\text{m}$, henceforth T_b —created using the global constellation of geostationary satellites. For each available TC³ from 1993 through 2021, the dataset contains images at ~ 30 -min time intervals. For additional details on the TC IR

³ TCs in the AL, EP, WP, CP, SH, and IO are available starting in 1993, 2002, 2002, 2005, 2018, and 2020, respectively.

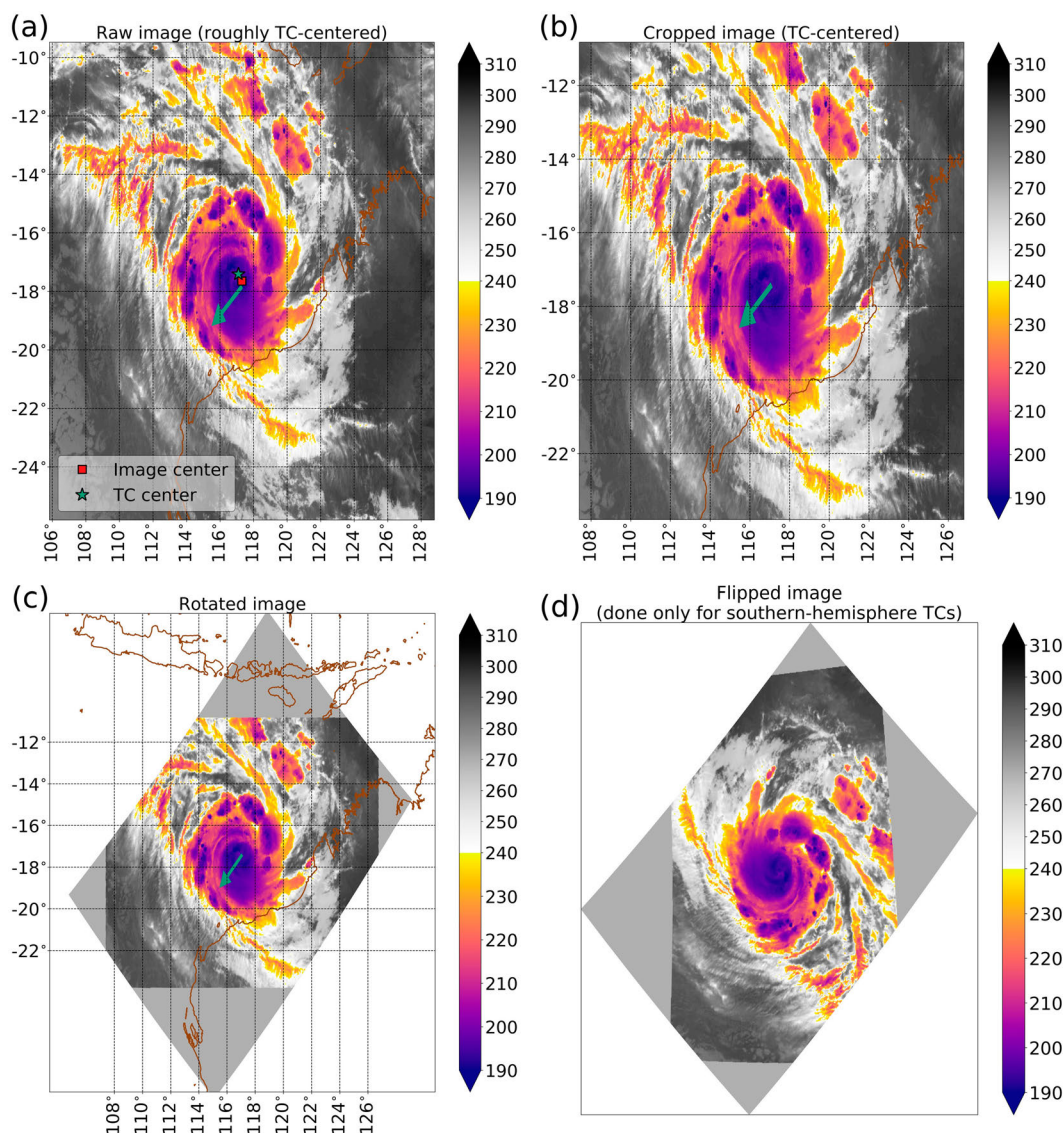


FIG. 2. Overview of TC IR data and the preprocessing thereof. The green arrow is the direction of storm motion. The color map in each panel is T_b (K). (a) The raw TC IR image for Cyclone Veronica at 1810 UTC 21 Mar 2019. (b) Cropped version of (a), which is now TC-centered. (c) Rotated version of (b). When plotted in latitude–longitude coordinates, the rotated image is a rhombus; but in row–column coordinates (the way the image is fed to an NN), it is rectangular, with the direction of storm motion being the right edge. In the particular case shown, this is the southwestern edge. (d) Flipped version of (c). The TC now appears to rotate counterclockwise, i.e., cyclonically if it were in the Northern Hemisphere. This flipping operation is done only for SH TCs.

archive, see [Mueller et al. \(2006\)](#), [Zehr and Knaff \(2007\)](#), and [Knaff et al. \(2014\)](#).

We preprocess the TC IR data in five steps. First, we align image centers with TC centers. The original images are only roughly TC-centered ([Fig. 2a](#)), but for this study—especially the XAI analyses, where we composite images from many TCs—more exact centering is useful. Specifically, we crop each image from its original size (480×640) to 380×540 , aligning the image center with the center fix stored in the TC IR metadata. TC IR center fixes are based on the best track archive, available at 6-hourly synoptic times; for off-synoptic

times, TC IR applies cubic spline interpolation to the synoptic times. The cropped image is smaller than the original image because, when recentering at a different point, the 480×640 window runs off the edge of the original image.

Second, we rotate images to align them with TC motion, which is provided in the TC IR files and based on a 6-h backward difference from points in the abovementioned spline interpolation ([Fig. 2c](#)). Arbitrarily, we make TC motion point toward the right, i.e., the positive x direction. Accounting for differences in TC motion is crucial for the XAI analyses—which involve compositing over many TCs—as some important RI precursors

TABLE 1. Partitioning of data into independent subsets. One “TC object” is one TC at one 6-h time step. Although the TC IR archive has 30-min time steps, the other two datasets have 6-h time steps, so our NN models make one forecast every 6 h. An “RI object” is a TC object that undergoes rapid intensification in the next 24 h. The first three datasets use postevent SHIPS data, while the last uses real-time SHIPS data.

Data subset	Years	Number of unique TCs	Number of TC objects	Number of RI objects
Training	1993–2004, 2015–19	673	16 329	1217
Validation	2005–09	310	6287	518
Testing	2010–14, 2020–21	448	8948	640
Real-time testing	2010–14, 2020–21	260	5000	290

in IR imagery have preferred orientations with respect to TC motion (K15). TC motion also leads to asymmetric frictionally induced convergence (Chow 1971; Shapiro and Willoughby 1982) and, to the first order, determines asymmetries in the surface wind (Klotz and Jiang 2017). Note that rotation leaves the image corners with no data (see light gray pixels in Fig. 2c). These grid points need a sentinel value, for which we choose the average T_b over the training data, namely, 268.9 K.

Third, because TCs in the two hemispheres rotate in opposite directions, training the same model with data from around the globe requires special care. Specifically, we flip TC IR images in the Southern Hemisphere upside down (with respect to the x axis), so that IR features appear as they would if the TC rotated counterclockwise (Fig. 2d). Thus, all the IR images appear as they would in the Northern Hemisphere, allowing the NN models to learn relevant IR features in a common framework.

Fourth, we split the data into training, validation, and testing sets. Training data are used to optimize NN parameters (i.e., weights and biases); validation data are used to optimize NN hyperparameters (settings that cannot change during training, e.g., number of layers); and testing data are used for an independent assessment of the selected models. The three datasets are independent, each containing TCs from a different set of years (Table 1). Supplemental section 1 in the online supplemental material discusses the geographic distribution of each dataset.

Fifth, we scale T_b via the following procedure, which is repeated for every atomic T_b value (an “atomic value” being one grid point in one image).

- 1) Quantile normalization. Transform T_b to its quantile over the training data.⁴ This results in a transformed variable T'_b with a uniform distribution over $[0, 1]$.
- 2) z -score normalization. Transform T'_b to a standard normal distribution (mean = 0 and variance = 1), using the inverse cumulative density function.

The z -score normalization is common practice for NNs (section 3.6.2 of Chollet 2018; Shanker et al. 1996). However, the common approach does not involve step 1, because it assumes that the raw data follow a normal distribution (just not the standard normal distribution) already. This is not true for T_b in our dataset, which is heavily skewed to the left.

⁴ Ideally, we would use all values in the training data. However, finding a quantile over 42 billion values (204 658 images with 380×540 pixels each) is too computationally expensive, so we randomly select 100 000 training values for this procedure.

c. SHIPS developmental data (scalar predictors)

The SHIPS dataset (NOAA 2020) contains three types of predictors: environmental, describing near-TC atmospheric and oceanic conditions; satellite-based, describing TC-centered geostationary satellite imagery; and storm-based metrics, including the TC’s location, current intensity, and recent intensity changes. All SHIPS predictors are scalars, and most are environmental. Thus, while the TC IR archive quantifies the current state of the TC, the SHIPS data mainly quantify the near-TC environment. The SHIPS data are updated at synoptic times throughout the lifetime of each TC.

In the real-time SHIPS dataset, most of the predictors are forecasts, provided at lead times up to 5 days.⁵ However, we train with the postevent SHIPS dataset,⁶ where these predictors are perfect prognoses (Kalnay 2002) instead of forecasts. In other words, what appears to be an N -hour forecast initialized at time t_0 is actually a 0-h analysis initialized at time $t_0 + N$. Thus, using any lead time beyond zero in the NN predictors is tantamount to leaking information from the future, which is not possible in an eventual real-time application of the NNs. For environmental predictors, we use only the first five lead times from the postevent dataset—0, 6, 12, 18, and 24 h—at which true forecasts and perfect prognoses are quite close. We quantify the effects of this choice in section 4c. For predictors involving TC intensity, we use only the 0-h lead time, because using any perfect prognosis for these variables makes RI prediction too easy.

There are many environmental and storm-based predictors to choose from, and our choice (Table 2 and Fig. 3b) is based on two factors. The first is a predictor-importance analysis from a “kitchen sink” model trained with all SHIPS predictors (not shown). We identified top predictors according to the multipass backward version of the permutation test, which screens out correlated predictors (section 7.1.1 of Lagerquist 2020). The second factor is data quality and

⁵ Storm-based predictors (location and intensity) are forecasts from the SHIPS model; satellite-based predictors are analyses (based on data available at the current time); within the environmental predictors, atmospheric variables are forecasts from the Global Forecast System (GFS), while oceanic variables are a hybrid forecast/analysis. Specifically, oceanic variables—such as sea-surface temperature and ocean heat content—come from the Navy Coupled Operational Data Assimilation (NCODA; Cummings 2005) at the forecast TC location (Sampson et al. 2022).

⁶ Because postevent data are available over a longer time period than real-time data.

TABLE 2. Environmental and storm-based SHIPS variables used as NN predictors. “Low level” means averaged from 850 to 700 mb (1 mb = 1 hPa); “low-to-midlevel” means over the 850–500-mb layer; “deep layer” means over the 850–200-mb layer; “full column” means averaged from 0 to 15 km above ground level; “inner ring” means averaged over a 500-km radius around the TC center; and “outer ring” means averaged over a 200–800-km annulus around the TC center. Low-to-midlevel eastward shear is computed from two SHIPS variables: SHRS (the magnitude of the shear vector) and SHTS (its direction). As discussed in section 2b, LAT and TWXC are multiplied by -1 for SH cases.

Variable	Name in SHIPS files	Units	Forecast leads used (h)	Category
Current TC intensity	VMAX	m s^{-1}	0	Storm-based
TC-intensity change over last 6 h	INCV	m s^{-1}	0	Storm-based
Low-to-midlevel eastward shear	SHRS and SHTS	m s^{-1}	0, 6, 12, 18, 24	Environmental
Deep-layer, inner-ring, vortex-removed shear	SHGC	m s^{-1}	0, 6, 12, 18, 24	Environmental
200-mb outer-ring temperature	T200	K	0, 6, 12, 18, 24	Environmental
Low-level, inner-ring temperature gradient	TGRD	K m^{-1}	0, 6, 12, 18, 24	Environmental
Maximum 850-mb symmetric tangential wind	TWXC	m s^{-1}	0, 6, 12, 18, 24	Environmental
Full-column, inner-ring, vortex-removed vertical velocity	VVAC	m s^{-1}	0, 6, 12, 18, 24	Environmental
TC-center latitude	LAT	$^{\circ}\text{N}$	0, 6, 12, 18, 24	Environmental
MPI	VMPI	m s^{-1}	0, 6, 12, 18, 24	Environmental
Ocean age	OAGE	s	0, 6, 12, 18, 24	Environmental
Best available SST	NSST, DSST, RSST, or CSST	K	0, 6, 12, 18, 24	Environmental
Best available OHC	NOHC, RHCN, PHCN, XOHC, or COHC	J m^{-2}	0, 6, 12, 18, 24	Environmental

homogeneity—i.e., we choose variables that have better overall quality and better consistency over the time period (1993–2021). In other words, our choice of predictors is based on both objective methods and subjective domain expertise. For variables with “best available” in the name, the second column of Table 2 gives the priority list. For best available SST, we use the first option available from NSST (NCODA), DSST (daily Reynolds analysis), RSST (Reynolds analysis), and CSST (climatology). For best available OHC, we use the first available out of NOHC (at the 26°C isotherm from NCODA), RHCN (satellite-based), PHCN (based on climatology and SST), XOHC (climatology), and COHC (climatology).

For the satellite-based predictors (Table 3 and Fig. 3a), each SHIPS record contains four versions of the predictors: in lines labeled IR00 (based on satellite imagery at the present time), IRM1 (based on imagery 1.5 h ago), IRM3 (3 h ago), and IRXX (climatology). For each variable, we use the best available from the following priority list: IR00, IRM1, IRM3, and IRXX. We use all satellite-based predictors in SHIPS except for “time of the GOES image” and “variables needed for storm-size estimation.” Some of these variables were identified by the kitchen-sink model as important, and some were not. Nonetheless, we use the full set of variables because 1) they include all summary statistics of the full satellite image, and 2) our first science question involves determining whether the full image adds predictive value beyond this summary.

We preprocess the SHIPS data in two steps. First, we normalize SHIPS predictors, using the same procedure as for TC IR data (section 2b). In that procedure, replace “ T_b ” with “ x_j ” (the j th SHIPS predictor) and replace atomic value with just “value” (since the SHIPS predictors are scalars and not images). The procedure is repeated for every SHIPS predictor; the result is that all predictors (SHIPS scalars and T_b) have mean 0 and variance 1. Second, to align Southern

Hemisphere TCs with their Northern Hemisphere counterparts, we multiply LAT and TWXC (Table 2) by -1 . Hence, LAT can be interpreted as meridional distance from the equator, and TWXC can be interpreted as maximum tangential wind in the cyclonic direction. Before (after) this preprocessing, the positive direction for TWXC is counterclockwise (cyclonic) regardless of hemisphere.

3. Methods

This section discusses our NN architecture, uncertainty-quantification methods coupled with the NNs, XAI methods, and the ablation experiment used to find the best NN model with/without certain data sources.

a. NNs: Fully connected and convolutional

CNNs can ingest full images as predictors, whereas fully connected NNs can use only scalar data. Since we train some NNs with full satellite images and some without, we use both CNNs and fully connected NNs. See Fig. 4 for a diagram of one of our CNN architectures,⁷ supplemental section 2 for deeper background on CNNs, and supplemental section 3 for our hyperparameter choices.

b. UQ

Our methods for machine-learned UQ (ML-UQ) are described below with frequent reference to Haynes et al. (2023, henceforth H23). There are two fundamental types of uncertainty—aleatory and epistemic (section 2b of H23)—and

⁷ Including all SHIPS variables and two TC IR images as predictors. Note that different predictor sets lead to slightly different architectures.

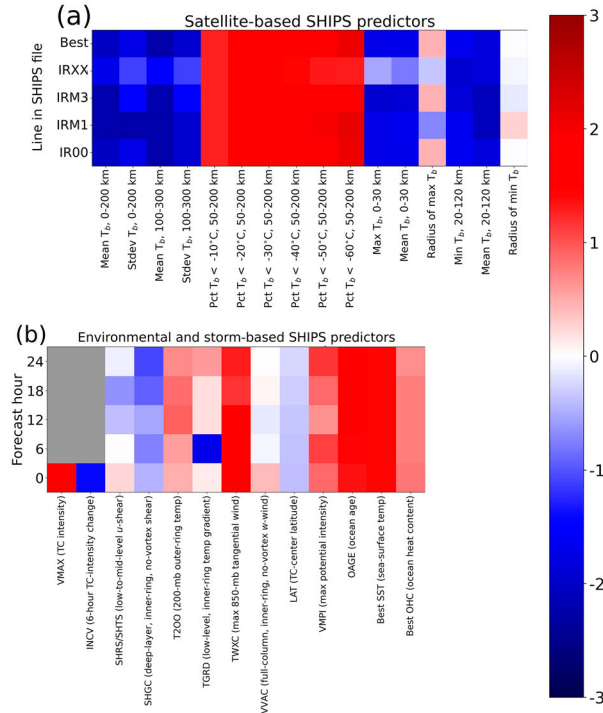


FIG. 3. Example of SHIPS data for Cyclone Veronica at 1800 UTC 21 Mar 2019. Variables are plotted in normalized units; otherwise, given the different scales of the variables, they would be difficult to plot with a single color bar. (a) The 16 satellite-based predictors we use. “Best” is the best available from the following priority list: IR00, IRM1, IRM3, and IRXX. In this case, all variables are available in the IR00 line, so the “IR00” and “Best” rows of the plot are identical. (b) Environmental and storm-based predictors. Note that while the SHIPS dataset contains “forecasts” (actually perfect prognoses) of VMAX and INCV beyond 0 h, we do not use these values. This is why, for both variables, values at nonzero lead times are greyed out.

most UQ methods can capture only one type (section 3 of H23). Thus, we combine two UQ methods. For aleatory uncertainty, we have found that the continuous ranked probability score (CRPS) method performs best (e.g., sections 5 and 6 of H23). This method involves training the NN to output an ensemble of predictions, optimizing the ensemble with the CRPS loss function. For epistemic uncertainty, we have found in other studies that Bayesian NNs (BNNs) perform best, followed closely by the multimodel ensemble (MME). However, for the RI problem, even after trying hundreds of hyperparameter settings for BNNs, the resulting models barely outperformed climatology. Thus, we use the MME method instead, which involves training many different NNs with the same hyperparameters but different random seeds. The random seed determines how NN weights are initialized, and different random seeds lead to different solutions, i.e., final weights. To combine the two UQ methods, we train 50 NN models, each with the CRPS loss function and 100 output neurons. Thus, the inner (outer) ensemble size is 100 (50), leading to a total ensemble size of 5000.

TABLE 3. Satellite-based SHIPS variables used as NN predictors. Any distance range (e.g., 0–200 km) is a range of radii outside the TC center.

Variable	Units
Mean 0–200-km T_b	K
Standard deviation of 0–200-km T_b	K
Mean 100–300-km T_b	K
Standard deviation of 100–300-km T_b	K
Fractional area from 50 to 200 km with $T_b < -10^\circ\text{C}$	Unitless
Fractional area from 50 to 200 km with $T_b < -20^\circ\text{C}$	Unitless
Fractional area from 50 to 200 km with $T_b < -30^\circ\text{C}$	Unitless
Fractional area from 50 to 200 km with $T_b < -40^\circ\text{C}$	Unitless
Fractional area from 50 to 200 km with $T_b < -50^\circ\text{C}$	Unitless
Fractional area from 50 to 200 km with $T_b < -60^\circ\text{C}$	Unitless
Maximum 0–30-km T_b	K
Mean 0–30-km T_b	K
Radius of maximum T_b	m
Minimum 20–120-km T_b	K
Mean 20–120-km T_b	K
Radius of minimum T_b	m

c. Probability calibration

To calibrate RI probabilities and produce a better attributes diagram, we apply isotonic regression (Niculescu-Mizil and Caruana 2005) to every NN. Isotonic regression is commonly used for this purpose, and an explanation geared toward atmospheric scientists is found in Lagerquist et al. (2017). The sole input (predictor) variable for isotonic regression is the NN-forecast RI probability ($\hat{y}_{\text{raw}}$), and the target variable is the true RI label (0 or 1). The output of isotonic regression is a calibrated probability $\hat{y}$, and we use only the training data (Table 1) to tune this calibration. More specifically, we train isotonic regression only to correct the NN ensemble mean. At inference time, to adjust the ensemble mean of a forecast by $\Delta = \hat{y} - \hat{y}_{\text{raw}}$, we adjust every ensemble member by Δ .

d. DeepSHAP

Shapley values, which come from the field of cooperative game theory (Shapley 1953), represent the expected marginal contribution of each player in a cooperative game. Shapley additive explanation (SHAP; Lundberg and Lee 2017, henceforth LL17) is an XAI method based on Shapley values. In ML, the definition of Shapley value is the same, except that “player” is replaced with “predictor variable” and the “marginal contribution” is to the ML model’s prediction $\hat{y}$.⁸ The expected marginal contribution of each predictor x to $\hat{y}$ is a measure of x ’s influence, often called a “feature attribution” in the XAI literature. The summed attribution over all predictors equals $\hat{y}$, so SHAP is considered an additive feature-attribution method. SHAP is a local method, meaning that it provides a different explanation for each data sample (here, TC object). Furthermore, SHAP unifies and combines the best properties of six

⁸ Although our models output an ensemble, we compute Shapley values only for the ensemble mean. Thus, $\hat{y}$ in this discussion should be interpreted as the model’s mean prediction.

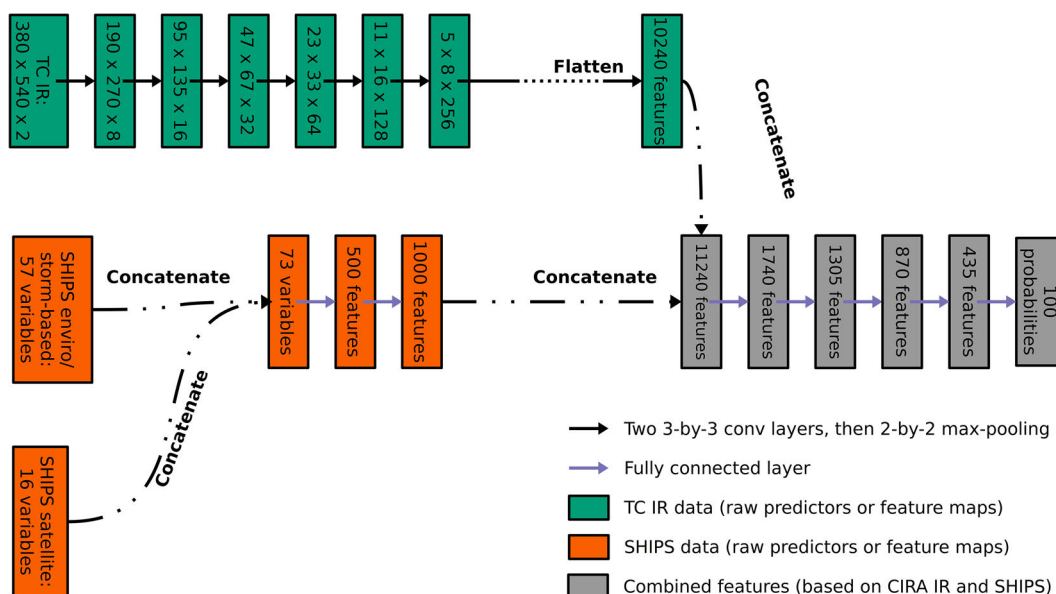


FIG. 4. The CNN architecture shown here is used when the predictors include all environmental/storm-based SHIPS variables, all satellite-based SHIPS variables, and two lag times for TC IR images. There are six convolutional blocks (solid black arrows), each containing two convolutional layers with 3×3 pixel filters, followed by a maximum pooling layer with a 2×2 pixel pooling window. There are two fully connected layers (FCLs) for extracting SHIPS-only features (purple arrows in the left half of the figure) and five FCLs for extracting combined features, which include both SHIPS and TC IR data (purple arrows in the right half of the figure). After concatenating SHIPS-only and TC-IR-only features into the same vector, we exponentially decrease the number of features from its initial value (11 240) to its final value (100), which is the ensemble size. In other words, the final output is an ensemble of 100 RI probabilities.

preexisting XAI methods, as explained in LL17. We specifically use the DeepSHAP algorithm, which is a computationally efficient approximation of SHAP for NNs.

e. Eigenanalysis of DeepSHAP output

Our goal for XAI is to learn more about RI precursors in satellite imagery. Thus, although DeepSHAP produces a Shapley value for every predictor—including the TC IR image pixels and SHIPS scalars—we focus on the TC IR image. However, like most XAI methods (Nielsen et al. 2022), DeepSHAP produces a lot of pixel-level noise when applied to image data, as in supplemental Fig. S24. This makes it difficult to interpret Shapley values for an individual data sample. Thus, we use eigenanalysis to find patterns in large composites of Shapley maps, i.e., over many data samples. Specifically, we wish to find the leading modes of covariability between two fields— T_b and Shapley value—interpreting each mode as a strategy used by the NN to predict RI.

When one spatial field is to be eigenanalyzed, the most common strategy is principal component analysis (PCA; Wilks 2006, chapter 11), which eigenanalyzes the temporal covariance matrix of the field to find leading modes of variability. When two fields must be considered—e.g., T_b and Shapley values—the analogous method is maximum covariance analysis (MCA; section 12.4 of Wilks 2006), which is often used in climate studies (Frankignoul et al. 2011; Barreto et al. 2017; Pangaluru et al. 2019). MCA detects modes of covariability by eigenanalyzing the temporal

covariance matrix of the two fields. For details of our MCA procedure, see supplemental section 4.

f. Ablation experiment: Finding the best model with/without certain data sources

We conduct a hyperparameter experiment, where most of the hyperparameters control which data sources are used to create predictors. The goal of this experiment is to find the best NN models with and without certain data sources. Specifically, we aim to find the best model overall, the best without TC IR data (i.e., with only SHIPS data), and the best without SHIPS data (i.e., with only TC IR data). The best model overall—which will likely use most or all data sources and thus be computationally expensive—would be useful in a stand-alone setting (where no other models are available). The best model without TC IR data—which will include no images and thus no convolution—would be useful in a setting with limited computing resources. The best model without SHIPS data (let this be $\mathcal{M}$) would be useful in a setting where it is combined with information from other models. This is because most preexisting RI models use SHIPS data but not full satellite images, so $\mathcal{M}$ would be fairly independent of the preexisting models.

In addition to data sources, we experiment with two other hyperparameters. The first is dropout rate (let this be d) for dropout regularization (Hinton et al. 2012), a method that mitigates overfitting by randomly dropping $d \cdot 100\%$ of neurons when making a prediction. Supplemental section 3 contains

TABLE 4. Experimental hyperparameters. When N_{lag} (the number of TC IR lag times) = 0, the predictors include no TC IR data; when $N_{\text{lag}} = 1$, we use data at the forecast issue time (t_0); when $N_{\text{lag}} = 2$, we use data from t_0 and $t_0 - 4$ h; and when $N_{\text{lag}} = 3$, we use data from t_0 , $t_0 - 2$ h, and $t_0 - 4$ h. When N_{lag} in the table is followed by an asterisk, this means that we use temporal differencing for nonzero lag times. “Count” is the number of values attempted for the given hyperparameter.

Hyperparameter	Values attempted	Count	Explanation
Dropout rate	0.5, 0.6, 0.7, 0.8, 0.9	5	
Number of TC IR lag times	0, 1, 2, 2*, 3, 3*	6	
Use SHIPS environmental predictors?	Yes, no	2	If “yes,” we use the last 11 columns of Fig. 3b.
Use SHIPS storm-based predictors?	Yes, no	2	If yes, we use the first two columns of Fig. 3b.
Use SHIPS satellite-based predictors?	Yes, no	2	If yes, we use the top row of Fig. 3a.

more details on dropout. The second additional hyperparameter is temporal differencing of TC IR images, which can be either “on” or “off.” When turned on, TC IR images at nonzero lag times contain T_b differences instead of absolute values, as in supplemental Fig. S1.

A full list of attempted hyperparameter values is shown in Table 4. We try all possible combinations of these values. The total number of NN models would then be $5 \times 6 \times 2 \times 2 \times 2 = 240$, except that five of these models have no predictors at all. Hence, the total number of models is 235. For a full listing and justification of all other hyperparameter values (those not optimized by the experiment), see supplemental section 3. To choose the best models, we compute eight evaluation metrics (defined below) on the validation data. The models in this experiment are trained with the CRPS-only method, as opposed to the hybrid MME/CRPS method, for UQ. Using MME for every hyperparameter combination would involve training $235 \times 50 = 11\,750$ NNs, which is too computationally expensive. For each of the best CRPS-only models, we then train an MME/CRPS model with the same hyperparameters and evaluate this final model on the testing data.

The eight evaluation metrics include five concerning the ensemble-mean prediction $\bar{y}_{\text{pred}}$ and three UQ-based metrics, which concern the full ensemble. The first five are area under the receiver operating characteristic curve (AUC; Metz 1978), area under the performance diagram (AUPD; Roebber 2009), Brier skill score with respect to a climatological forecast (BSS; Hsu and Murphy 1986), critical success index (CSI) maximized over all probability thresholds, and frequency bias (FB) at the CSI-maximizing probability threshold. See Lagerquist et al. (2021) for a succinct explanation of these metrics. The three UQ-based metrics are spread-skill reliability (SSREL) and spread-skill ratio (SSRAT), based on the spread-skill plot, and monotonicity fraction (MF), based on the discard test. See H23 for details. AUC, AUPD, BSS, CSI, and MF are positively oriented with an optimal value of 1.0; SSREL is negatively oriented with an optimal value of 0.0; FB has an optimum of 1.0, with higher (lower) values indicating overprediction (underprediction); and SSRAT has an optimum of 1.0, with higher (lower) values indicating underconfidence (overconfidence).

4. Results of ablation experiment

In the validation phase, we use validation data to select the three best NN models: the best overall, without TC IR data, and without SHIPS data. These models are trained with the

CRPS-only UQ method. In the testing phase—after retraining the best models on training data with the same hyperparameters but with the hybrid MME/CRPS UQ method—we use testing data for a final assessment of the best models.

a. Validation phase

Detailed results—i.e., all eight evaluation metrics for all 235 models—are shown in Fig. 5 and supplemental Figs. S2–S8. UQ-based evaluation metrics are shown in supplemental Figs. S6–S8, while mean-prediction-based metrics are shown in Fig. 5 and supplemental Figs. S2–S5. We summarize the results below.

- 1) All five mean-prediction-based metrics improve as more SHIPS predictors are added (i.e., as the panel letter increases from Figs. 5a to h).
- 2) The mean-prediction-based metrics do not have a clear relationship with the amount (number of lag times) or format (temporal differences vs raw values) of TC IR data used.
- 3) However, mean-prediction-based metrics improve slightly as the number of TC IR lag times increases from 0 to 1 (the left two columns of each panel). In other words, the models are better when they have access to some TC IR data instead of none.
- 4) The mean-prediction-based metrics have no clear relationship with dropout rate.
- 5) The UQ-based metrics also improve as more SHIPS predictors are added, but not as markedly as the mean-prediction-based metrics.
- 6) The UQ-based metrics do not have a clear relationship with any hyperparameter not involving SHIPS predictors, except that SSREL improves slightly as the number of TC IR lag times increases from 0 to 1. In other words, the SSREL is better with some TC IR data instead of none.

Overall, the most important hyperparameter is the inclusion of SHIPS data (all three types—environmental, storm-based, and satellite-based—are useful), followed by the inclusion of TC IR data (one lag time—i.e., an image rather than a video—is sufficient). The dropout rate and format of TC IR data (raw values vs temporal differences) are unimportant. Based on all eight metrics, we choose the following as the best models, marked by circles in Fig. 5. Supplemental section 5a contains more details on the best models.

- Best model overall: all SHIPS predictors, 1 TC IR lag time, dropout rate of 0.8

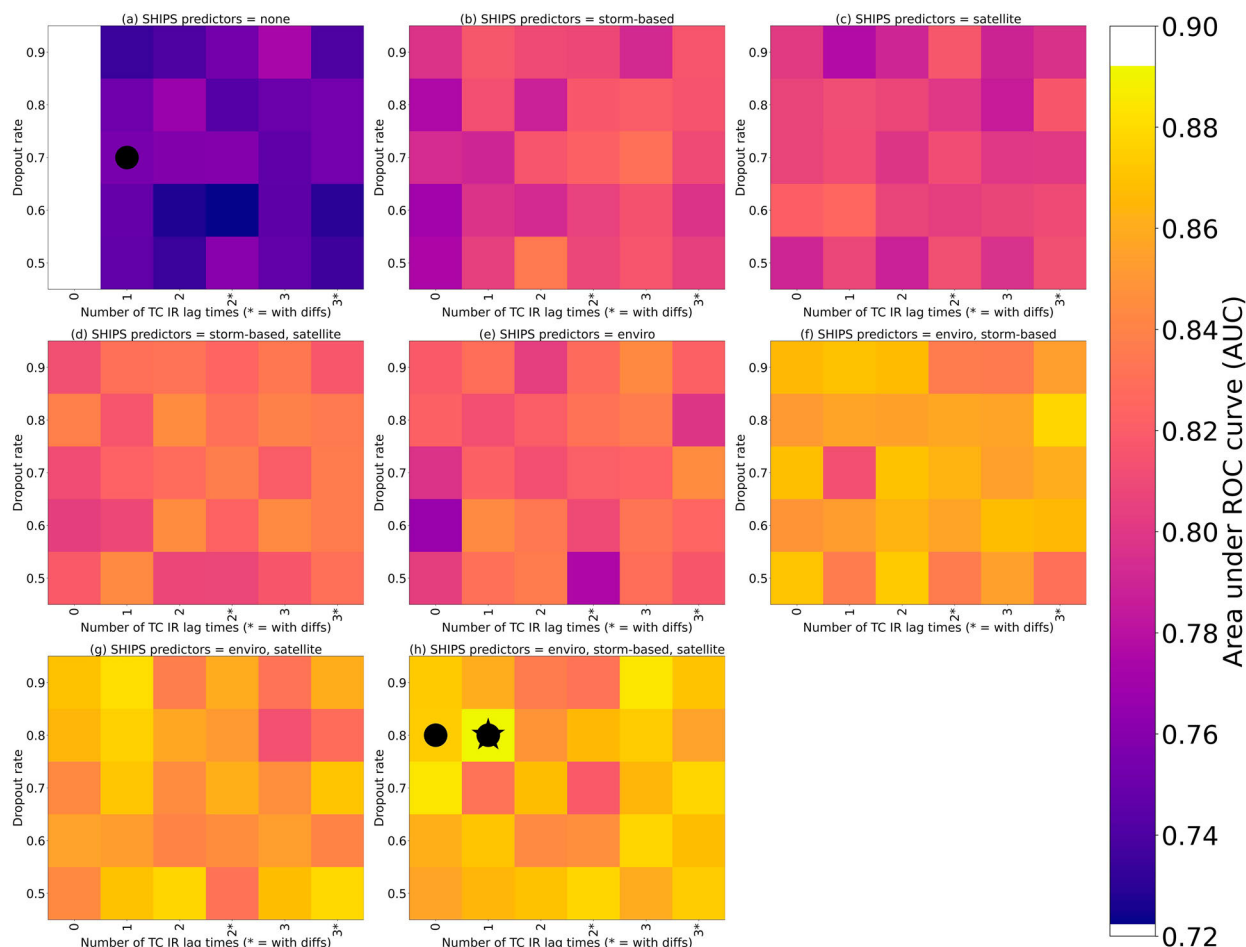


FIG. 5. AUC on validation data for each NN model. Each panel represents models trained with a different set of SHIPS predictors; within one panel, each row represents models trained with a different dropout rate, and each column represents models trained with a different amount and format of TC IR data. If the number of lag times (N_{lag}) is 0, the predictors include no TC IR data; if $N_{\text{lag}} = 1$, the predictors include a single TC IR image; if $N_{\text{lag}} = 2$ or 3, the predictors include a TC IR video; and if $N_{\text{lag}} = 2^*$ or 3^* , the predictors include a TC IR video with raw T_b values at the zero lag time and temporal differences at the nonzero lag times. The three circles mark the best models, and the star represents the model with the best value for this metric (AUC). Five boxes in (a) are missing; these correspond to models with no predictors, so we do not actually train a model.

- Best model without TC IR: all SHIPS predictors and dropout rate of 0.8
- Best model without SHIPS: 1 TC IR lag time and dropout rate of 0.7

Henceforth, we call these the “NN with everything,” “NN without TC IR,” and “NN without SHIPS,” respectively.

b. Testing phase with postevent data

Figure 6 shows testing results for the NN with everything. Figures 6a–c and 6f–g consider only the ensemble-mean prediction, while Figs. 6d and 6e consider the full predictive distribution, including UQ. The receiver operating characteristic (ROC) curve (Fig. 6a) shows an AUC > 0.9, considered excellent performance (Luna-Herrera et al. 2003; Muller et al. 2005). However, the performance diagram (Fig. 6b) is much lower quality, with an AUPD of only 0.483. At the CSI–

maximizing probability threshold (p^*) for the testing data (the “T” in Fig. 6b), the CSI is 0.369 and the FB is 0.981, so the model does not substantially over- or underpredict RI. However, the testing data are meant to simulate an operational environment, where p^* has already been tuned on the validation data and cannot be retuned. In other words, one can think of p^* as a hyperparameter. Using p^* chosen from the validation data (marked as “V” in Fig. 6b), the CSI is 0.272 and the FB is 2.741, so the model substantially overpredicts RI. This contrast illustrates that great care must be taken in choosing decision thresholds.

The attributes diagram (Fig. 6c) shows that probabilistic forecasts are often poorly calibrated, with overprediction at all probabilities below 0.8 and negative skill at all probabilities from 0.1 to 0.6. Calibration is also an issue for preexisting RI models (D21), and this will be discussed in section 4c. The spread–skill plot (Fig. 6d) shows that the model is underconfident at all

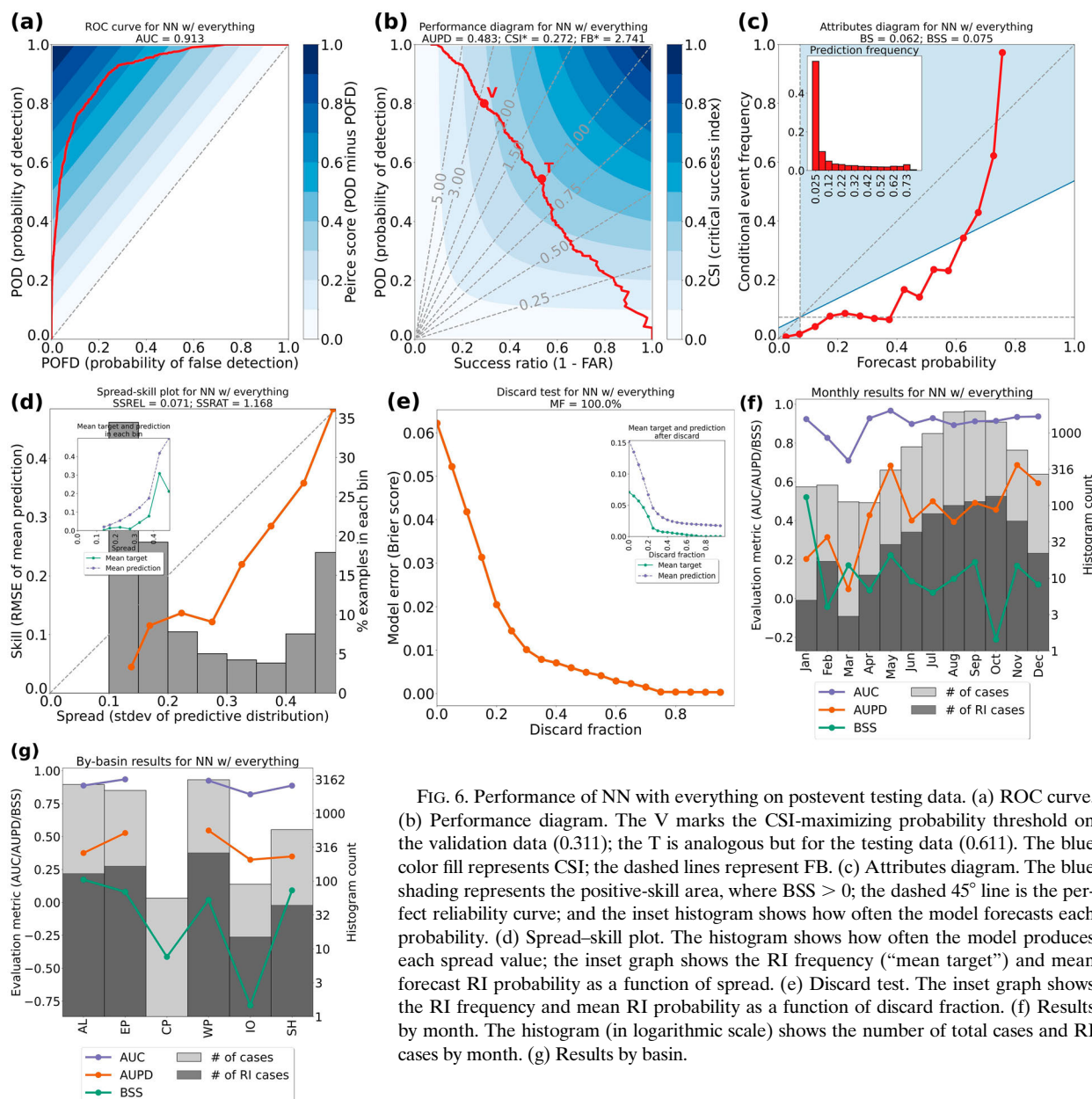


FIG. 6. Performance of NN with everything on postevent testing data. (a) ROC curve. (b) Performance diagram. The V marks the CSI-maximizing probability threshold on the validation data (0.311); the T is analogous but for the testing data (0.611). The blue color fill represents CSI; the dashed lines represent FB. (c) Attributes diagram. The blue shading represents the positive-skill area, where BSS > 0; the dashed 45° line is the perfect reliability curve; and the inset histogram shows how often the model forecasts each probability. (d) Spread-skill plot. The histogram shows how often the model produces each spread value; the inset graph shows the RI frequency ("mean target") and mean forecast RI probability as a function of spread. (e) Discard test. The inset graph shows the RI frequency and mean RI probability as a function of discard fraction. (f) Results by month. The histogram (in logarithmic scale) shows the number of total cases and RI cases by month. (g) Results by basin.

spread values except the highest; however, the underconfidence is most minor (i.e., the curve is closest to the 45° line) at the most common spread values (0.10–0.20 and 0.45–0.50). Also, the overall SSRAT is quite close to one, at 1.168. The discard test (Fig. 6e) shows that the model's estimate of its own uncertainty is strongly correlated with its Brier score, yielding an MF of 100%. In other words, the expected Brier score decreases whenever the discard fraction is increased (i.e., whenever the maximum acceptable uncertainty is decreased). This property could be useful as an "override switch" in operations. For example, suppose that the maximum acceptable Brier score is 0.042, which corresponds to a discard fraction of 0.1 (shown in Fig. 6e) and model spread

$\sigma_{y_{\text{pred}}}$ of 0.480 (not shown in Fig. 6e). Then, forecasters could choose to override the model whenever $\sigma_{y_{\text{pred}}} > 0.480$.

Results split by month and basin (Figs. 6f and g) lead to three main takeaways. First, AUC and AUPD—which measure the quality of binary forecasts—are largely controlled by the number of cases in each subset, especially the number of RI cases. AUC and AUPD are worst in January–March, which are among the 4 months with the fewest RI cases,⁹ and in the

⁹ In this paragraph, the phrase "fewest RI cases" is based only on the testing data used to create Fig. 6. We make no claims about the climatological frequency of RI in any month/basin.

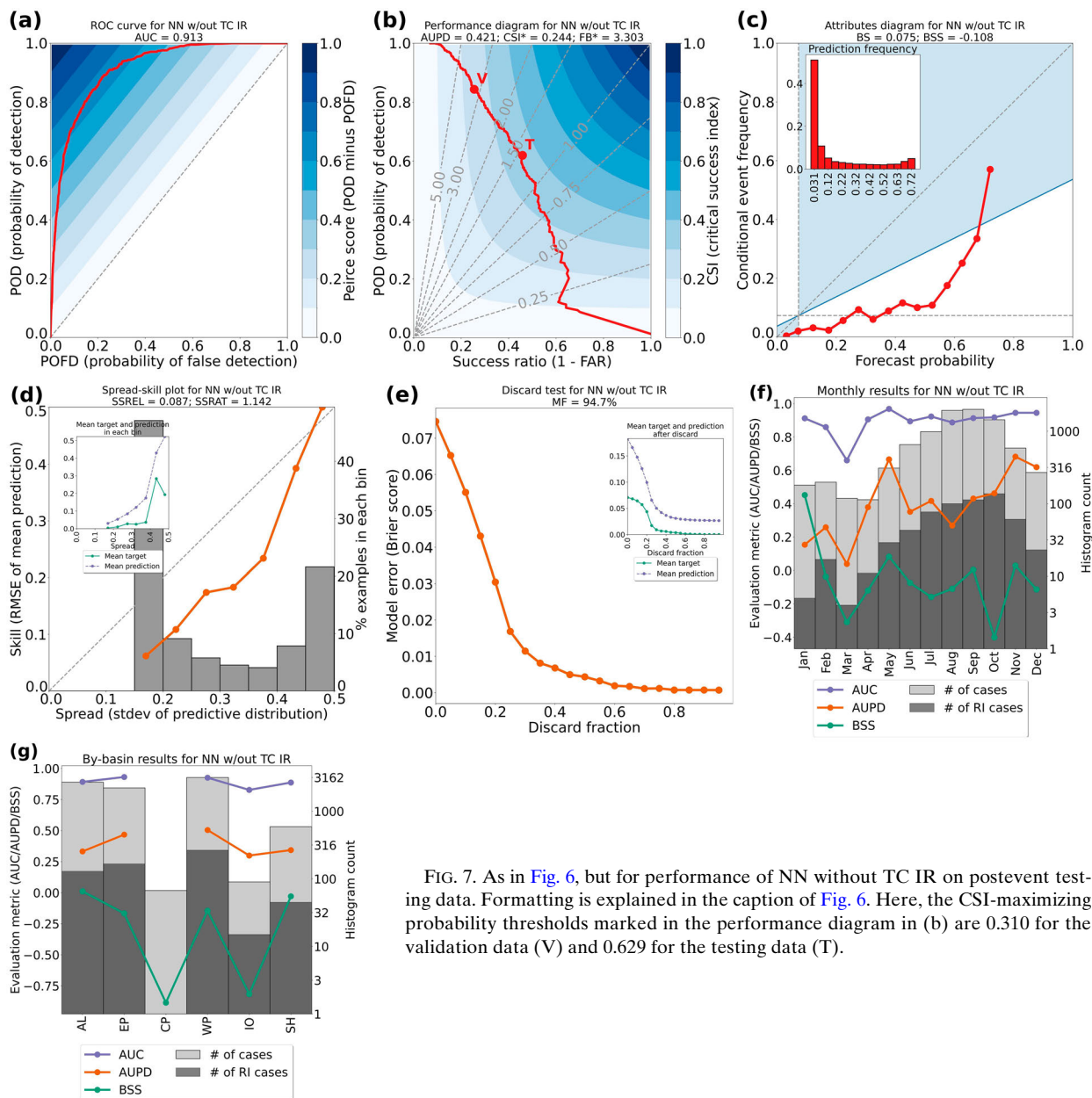


FIG. 7. As in Fig. 6, but for performance of NN without TC IR on postevent testing data. Formatting is explained in the caption of Fig. 6. Here, the CSI-maximizing probability thresholds marked in the performance diagram in (b) are 0.310 for the validation data (V) and 0.629 for the testing data (T).

IO and SH, which are among the three basins with the fewest RI cases. Second, in the monthly distribution of BSS (Fig. 6f)—which measures the quality of probabilistic forecasts—there is no clear seasonal cycle or relationship to the number of cases. However, third, there is a weak relationship between BSS and the number of RI cases in a given basin (Fig. 6g). For example, BSS is worst in the CP and IO, which are the two basins with the fewest RI cases. In all other basins, BSS is positive, indicating that the NN's probabilistic forecasts are better than climatology.

Figure 7 shows testing results for the NN without TC IR. In the following, we highlight differences between this model and the NN with everything (Fig. 6). The ROC curves (Fig. 7a) are

nearly identical, with an identical AUC. However, the performance diagram (Fig. 7b) is worse without TC IR, showing a 13% decrease in AUPD, a 10% decrease in CSI, and a 21% increase in FB. The attributes diagram (Fig. 7c) is also worse without TC IR, showing a negative BSS and overprediction at all probabilities from 0.1 to 0.7 (vs 0.1–0.6 for the NN with everything). Between Figs. 7b and 7c, we conclude that the NN without TC IR has a worse overprediction problem for both binary and probabilistic forecasts. The spread-skill plots (Fig. 7d) are similar, with both models showing underconfidence. The discard tests (Fig. 7e) are similar, except that the NN without TC IR has a slightly worse MF at 94.7% (18/19). Nonetheless, this large MF shows good correlation between

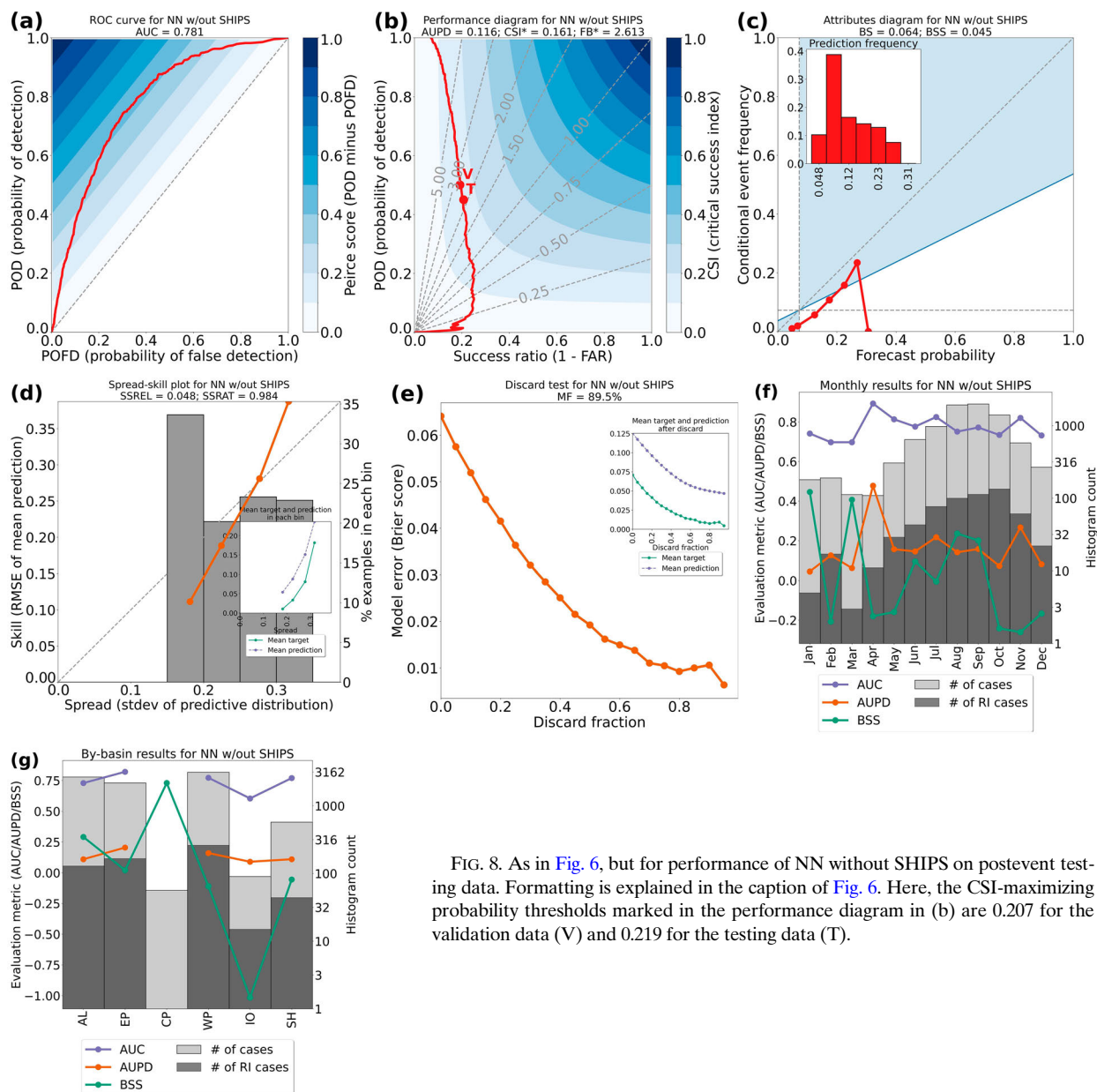


FIG. 8. As in Fig. 6, but for performance of NN without SHIPS on postevent testing data. Formatting is explained in the caption of Fig. 6. Here, the CSI-maximizing probability thresholds marked in the performance diagram in (b) are 0.207 for the validation data (V) and 0.219 for the testing data (T).

model spread and skill. Last, results by month and basin (Figs. 7f and g) are similar between the two models, leading to the same main takeaways.

Figure 8 shows testing results for the NN without SHIPS. In the following, we highlight differences between this model and the NN with everything (Fig. 6). The ROC curve (Fig. 8a) and performance diagram (Fig. 8b) are much worse without SHIPS data, showing a decrease of 14% in AUC, 76% in AUPD, and 41% in CSI. The huge decrease in AUPD occurs because, at a given POD (y axis in the performance diagram), the NN without SHIPS produces many more false alarms than the NN with everything, leading to a much smaller success ratio (x axis in the performance diagram). This suggests that SHIPS data are crucial for reducing false alarms. The attributes

diagram (Fig. 8c) is overall worse without SHIPS data, showing a slight decrease in BSS and a major decrease in sharpness. The NN with everything forecasts all probabilities up to ~ 0.8 with nonnegligible frequency, while the NN without SHIPS never forecasts a probability > 0.35 . However, the NN without SHIPS is inside the positive-skill area at most probabilities it forecasts, unlike the NN with everything. We note that without probability calibration (section 3c), the NN without SHIPS has greater sharpness (can forecast probabilities up to 45%) but worse calibration, manifesting in a BSS of -0.126 . This trade-off between sharpness and calibration was also observed by Burke et al. (2020); see their Fig. 6. The spread-skill plot (Fig. 8d) is surprisingly better without SHIPS data, showing a 32% decrease in SSREL and SSRAT very close to 1.0. In other words,

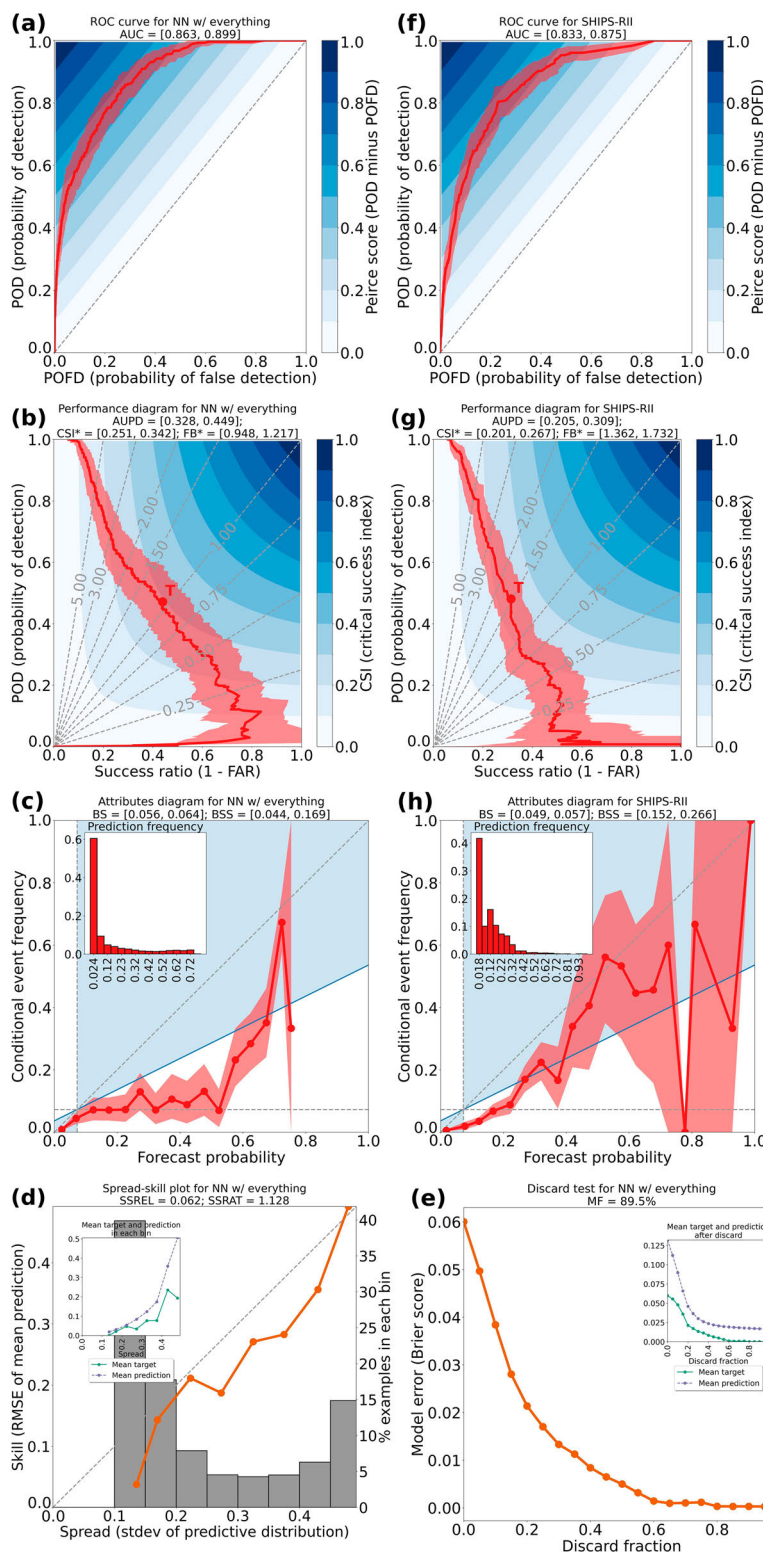


FIG. 9. Comparison of NN with everything vs SHIPS-R11 baseline on real-time testing data. (a)–(e) NN results; (f)–(h) SHIPS-R11 results. (a),(f) ROC curves for the two models. (b),(g) Performance diagrams for the two models. The T marks the CSI-maximizing probability threshold (0.590 for NN and 0.290 for SHIPS-R11) on the testing data used to make the figure. We do not include a V for validation data

the NN without SHIPS does not have the across-the-board underconfidence exhibited by the NN with everything. The discard test (Fig. 8e) is slightly worse without SHIPS data, exhibiting an MF of 89.5% (17/19). But again, this MF is still large, showing good correlation between spread and skill. Last, results by month and basin (Figs. 8f and g) are similar between the two models, except that for the NN without SHIPS, the relationship between BSS and the number of RI cases in a basin (Fig. 8g) is not apparent. Overall, the removal of SHIPS data severely impacts model performance, but it impacts binary forecasts more than probabilistic forecasts or UQ.

Our main conclusions from postevent testing are summarized below:

- The NNs with access to SHIPS data produce skillful binary forecasts (e.g., $AUC > 0.9$).
- All NNs produce a better ROC curve than performance diagram, resulting in $AUC \gg AUPD$. This is because RI is a rare event and is overpredicted by the NNs, leading to many false alarms (the same is true for operational models; see section 4c).
- All NNs produce poorly calibrated probabilistic forecasts ($BSS < 0.1$).
- All NNs produce useful uncertainty estimates; results from the discard test could be used to create an override switch for operations.
- The NNs perform better in months/basins where RI is more common and worse in months/basins where RI is less common.
- Removing TC IR data hurts performance a little, but removing SHIPS data hurts performance a lot. In other words, the SHIPS predictors are collectively much more important to our NNs than IR satellite imagery. This makes sense, given that SHIPS predictors 1) include summary statistics of IR imagery and 2) have been identified and continually improved over several decades.

c. Testing phase with real-time data

This section compares our NN models with three baselines: the SHIPS-RII, SHIPS consensus, and DTOPS. Because the baseline models are operational, they use real-time data rather than postevent; to ensure fairness, our NNs also use real-time data for this comparison. Also, the baseline models have incomplete forecast archives; to ensure fairness, when comparing our NNs with baseline $\mathcal{B}$, we use only TC objects (samples) in the archive for $\mathcal{B}$. The complete real-time testing dataset includes 5000 samples, covering all seven testing years and three basins: the AL, EP, and CP. The SHIPS-RII archive includes 4588 samples, also covering the seven years and three basins; the SHIPS-consensus archive includes 1705 samples, covering only

the AL and EP basins in 2020–21; and the DTOPS archive includes 1583 samples, covering the same two basins and two years.

The overall performance of the NN with everything is worse on real-time data (Figs. 9a–e) than on postevent data (Fig. 6), as expected since postevent data have the benefit of hindsight. However, the decline in performance is not catastrophic—e.g., the AUC decreases from 0.913 to 0.881, while the AUPD decreases from 0.483 to 0.395. Also, note that some aspects of NN performance actually improve from the postevent to real-time testing data. For example, the BSS increases from 0.075 to 0.105, indicating better probabilistic forecasts despite worse binary forecasts, and both the SSREL and SSRAT improve, indicating that uncertainty estimates become better calibrated. This mixture of results reemphasizes the trade-off among the quality of binary forecasts, probabilistic forecasts, and uncertainty estimates. The NN has a higher AUC (cf. Figs. 9a,f) and AUPD (cf. Figs. 9b and g) than SHIPS-RII, indicating better binary forecasts, but a worse BSS (cf. Figs. 9c and h), indicating worse probabilistic forecasts. All three differences are statistically significant at the 95% level, as determined by a paired, two-sided bootstrapping test with 1000 replicates (henceforth, we use “SS” as a shorthand). The NN’s uncertainty estimates cannot be compared with SHIPS-RII or any other baseline, since the baseline models do not perform UQ. However, we note that the NN’s uncertainty estimates could be useful in operations, as the SSRAT and MF are both near their optimal value of 1.0.

Figure 10 compares the best-performing NN (NN with everything) to the best-performing baseline (DTOPS). Again, the NN produces better binary forecasts, indicated by a higher AUC (cf. Figs. 10a and 10f) and AUPD (cf. Figs. 10b and 10g), but worse probabilistic forecasts, indicated by a lower BSS (cf. Figs. 10c and 10h). The AUC and BSS differences are SS, but the AUPD differences are not. Comparison of the NN with everything to the remaining baseline, namely, the SHIPS consensus, is relegated to supplemental Fig. S9. The above conclusions from Fig. 10 also apply to supplemental Fig. S9.

For the other two NNs—i.e., that without TC IR data and that without SHIPS data—Figs. 11 and 12 show the comparison to DTOPS. Comparison between these NNs and the other two baselines is relegated to supplemental section 5b. The NN without TC IR performs similarly to the NN with everything on samples in the DTOPS archive (cf. Figs. 11a–e and 10a–e), with the most notable difference being a much lower BSS (cf. Figs. 11c and 10c). This difference between the two NNs was also noted for postevent testing (cf. Figs. 7c and 6c). Like the NN with everything, the NN without TC IR produces better binary forecasts (with SS differences for AUC and non-SS differences for AUPD) and worse probabilistic forecasts (with SS differences for BSS) than the DTOPS baseline. Meanwhile,

←

here, because we do not have SHIPS-RII forecasts for the validation period and thus cannot properly tune this threshold for SHIPS-RII. (c),(h) Attributes diagrams for the two models. (d),(e) Spread–skill plot and discard test for the NN. These graphics cannot be created for the SHIPS-RII model, which does not perform UQ. In (a)–(c) and (f)–(h), the solid line represents the mean over 1000 bootstrap replicates of the testing samples, while the shaded envelope represents the 95% confidence interval.

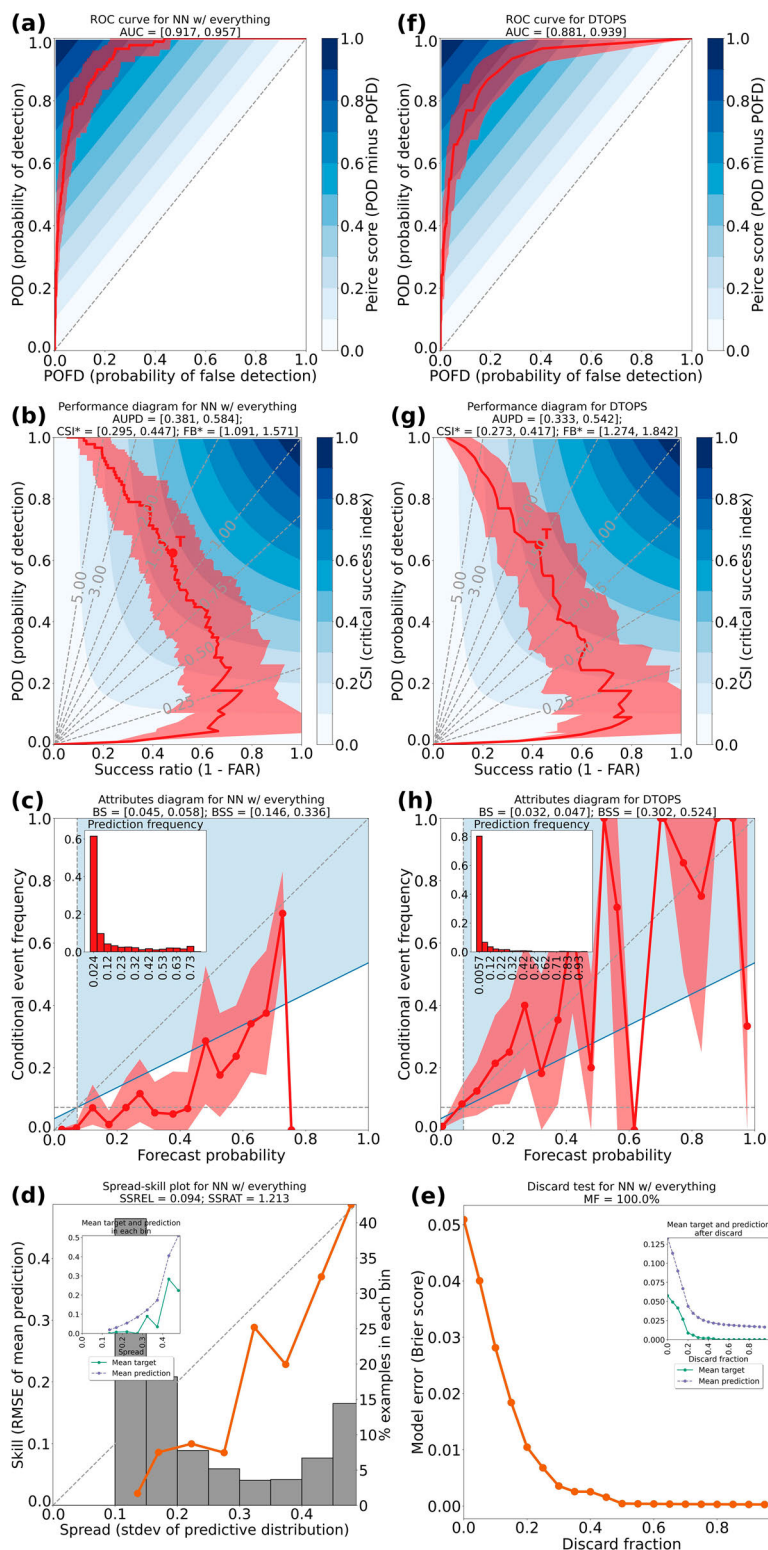


FIG. 10. As in Fig. 9, but comparing NN with everything vs DTOPS. The T in each performance diagram marks the CSI-maximizing probability threshold (0.583 for NN and 0.171 for DTOPS) on the testing data.

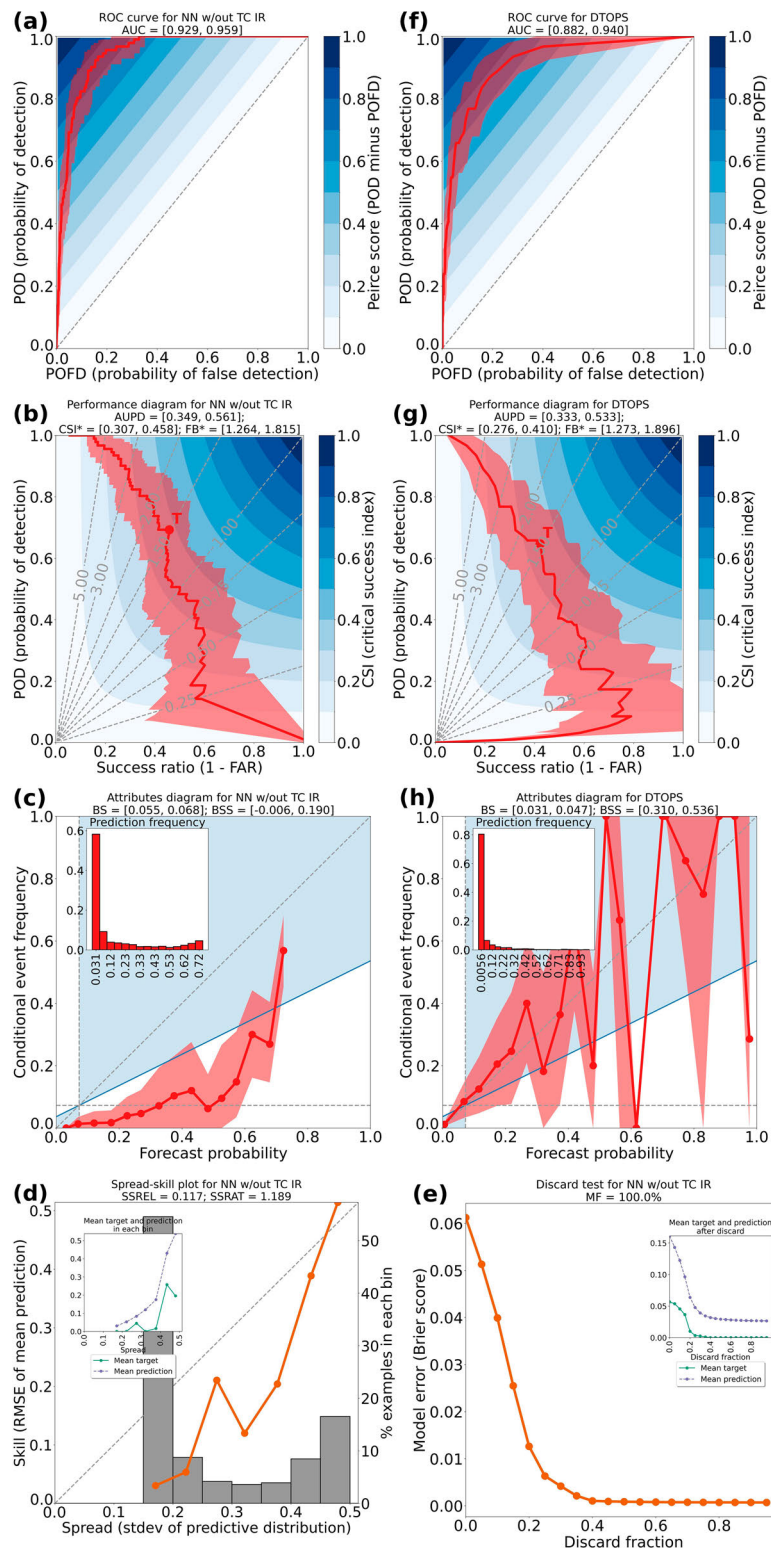


FIG. 11. As in Fig. 9, but comparing NN without TC IR vs DTOPS. The T in each performance diagram marks the CSI-maximizing probability threshold (0.634 for NN and 0.171 for DTOPS) on the testing data.

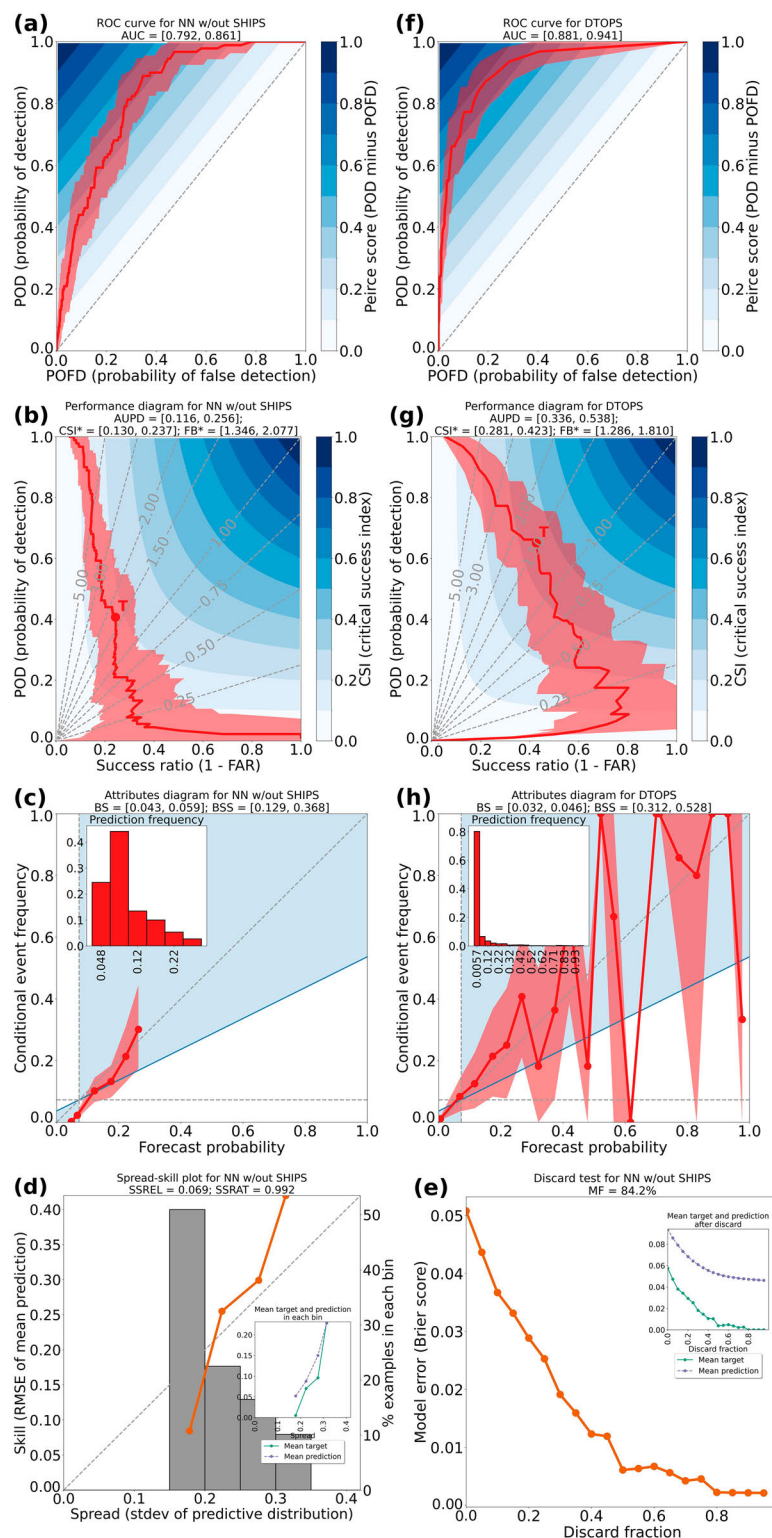


FIG. 12. As in Fig. 9, but comparing NN without SHIPS vs DTOPS. The T in each performance diagram marks the CSI-maximizing probability threshold (0.193 for NN and 0.171 for DTOPS) on the testing data.

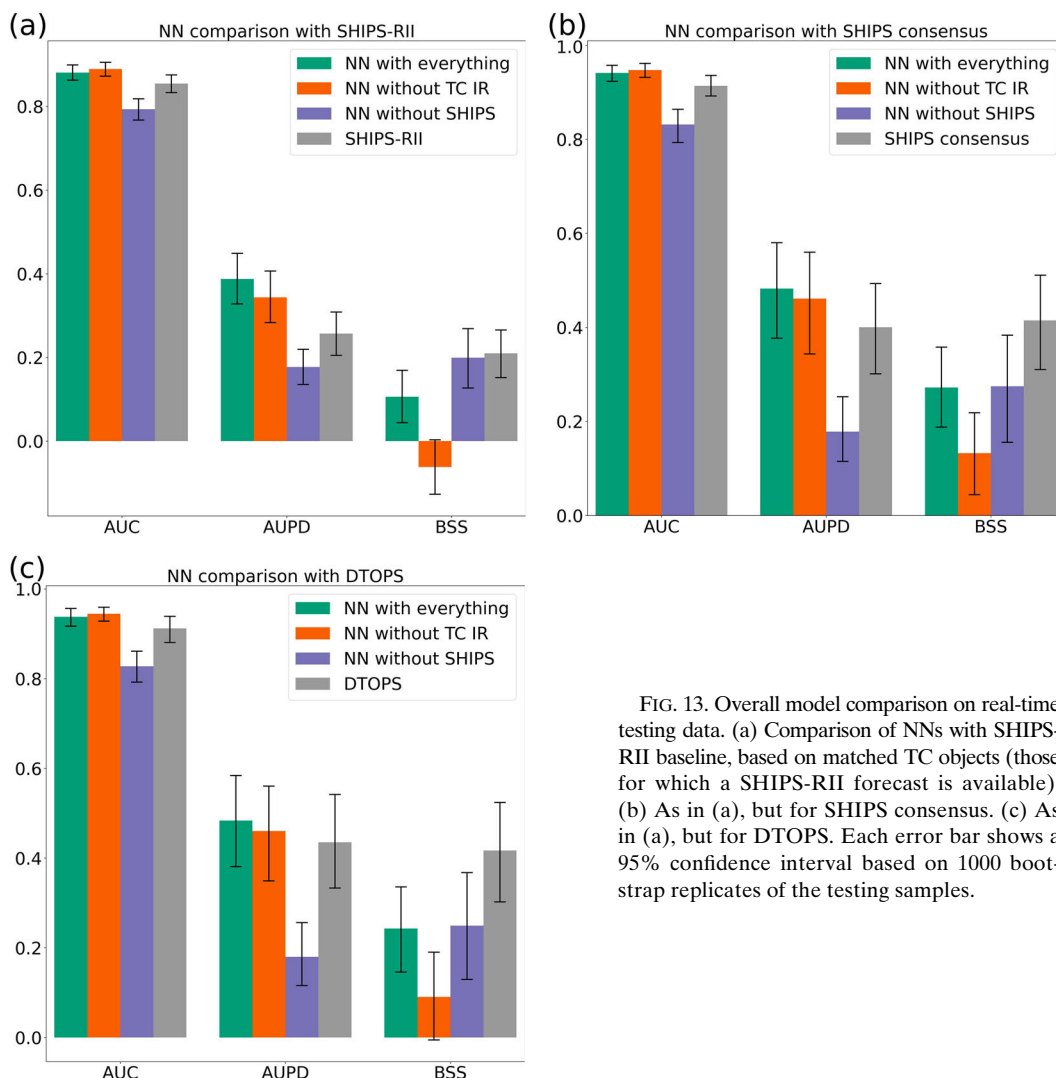


FIG. 13. Overall model comparison on real-time testing data. (a) Comparison of NNs with SHIPS-RII baseline, based on matched TC objects (those for which a SHIPS-RII forecast is available). (b) As in (a), but for SHIPS consensus. (c) As in (a), but for DTOPS. Each error bar shows a 95% confidence interval based on 1000 bootstrap replicates of the testing samples.

the NN without SHIPS produces much worse binary forecasts than DTOPS, with a SS AUC difference (cf. Figs. 12a and 12f) and AUPD difference (cf. Figs. 12b and 12g), and worse probabilistic forecasts than DTOPS, with a SS BSS difference (cf. Figs. 12c and 12h). However, the NN without SHIPS still produces surprisingly good probabilistic forecasts (much higher BSS than the NN without TC IR; cf. Figs. 12c and 11c) and uncertainty estimates (better SSREL and SSRAT than any other NN; cf. Figs. 12d, 11d, and 10d).

Our main conclusions from real-time testing are summarized below:

- All NNs perform slightly worse on real-time data than post-event data, as expected.
- The NNs produce worse probabilistic forecasts than the baselines (shown by the “BSS” bars in Fig. 13).
- The NN with everything and the NN without TC IR produce better binary forecasts than the baselines, while the NN without SHIPS produces worse binary forecasts (shown by the “AUC” and “AUPD” bars in Fig. 13).

- All NNs provide useful uncertainty estimates; the baselines provide none.
- Though outperformed by the baselines, the NN without SHIPS still provides useful uncertainty and surprisingly good probabilistic forecasts. Furthermore, the NN without SHIPS is the least correlated with (most independent of) the other models; see Fig. 14. Thus, although the NN without SHIPS does not have competitive skill on its own, it could be a useful addition to operational RI ensembles like the SHIPS consensus and that discussed in Sampson et al. (2023).¹⁰

Supplemental section 5c shows five case studies from the real-time testing data: one randomly selected, the best true

¹⁰ Creating this new ensemble would involve retuning ensemble weights; this would require additional SHIPS-consensus data, i.e., a tuning set in addition to the testing set used in this section. Thus, we have left a potential new ensemble for future work.

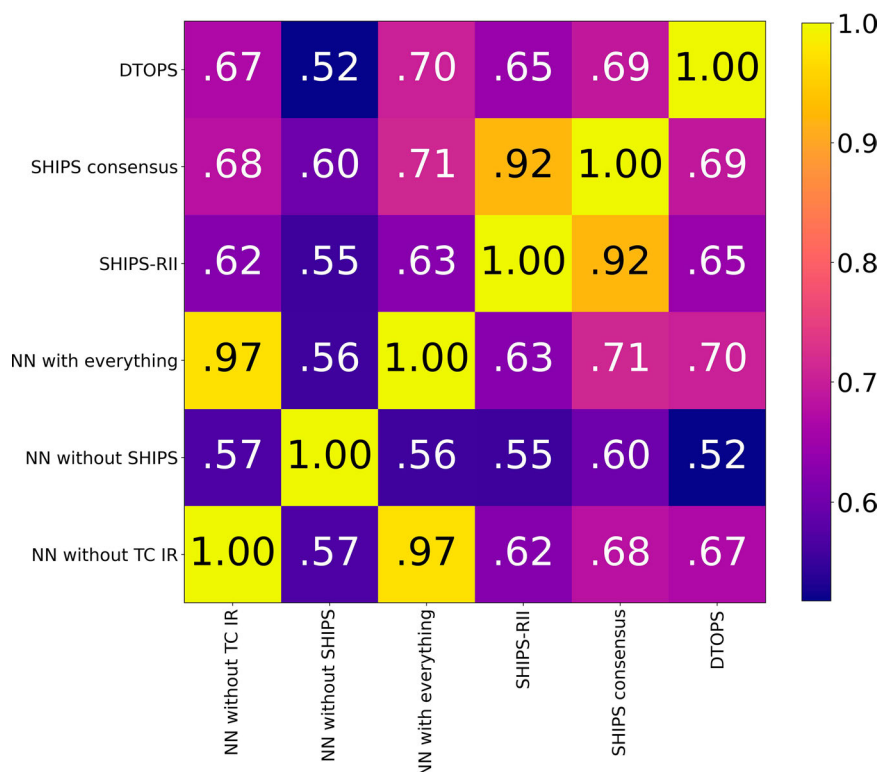


FIG. 14. Correlation matrix for RI probabilities on real-time testing data. The (i, j) entry is the Pearson correlation between forecast probabilities from the i th and j th models, based on matched TC objects for the two models.

positive, the best true negative, the worst false positive, and the worst false negative.

5. Results of XAI

We apply the DeepSHAP XAI method (section 3d) to every TC object in the postevent testing data, then eigenanalyze the outputs via MCA (section 3e), interpreting the leading modes of covariability between T_b and Shapley value as the leading strategies used by an NN to predict RI. Because we are interested in the satellite imagery and not the SHIPS scalars, we use the NN without SHIPS. Also, because DeepSHAP is designed for a single NN and not a multimodel ensemble, we use the model from the validation phase rather than the testing phase. This decision should have no impact on XAI results; the main difference between the two models is their UQ method, but DeepSHAP considers a model's mean prediction, not the full ensemble (section 3d). DeepSHAP requires a set of baseline data samples, for which we use 50 randomly chosen TC objects from the training data, each from a different TC.¹¹

Supplemental Fig. S25 shows eigenvalues for the top 10 modes of covariability from the MCA. All 10 eigenvalues have nonoverlapping error bars, suggesting that at least 10 modes are significant. However, the top 5 modes collectively

explain over 99% of the covariance between T_b and Shapley value, so we focus on the top 5 modes. Homogeneous-regression maps¹² for these modes are shown in Fig. 15.

Mode 1 (Fig. 15a) explains 93.2% of the covariance and therefore represents by far the leading strategy used by the NN to predict RI. The TC center is surrounded by positive Shapley values (solid contours; evidence for RI) associated with colder cloud tops (i.e., deeper convection), and both fields are nearly axisymmetric. Also, large Shapley values in mode 1 are concentrated within $\sim 2^\circ$ of the TC center. Thus, taken alone, our results for mode 1 justify current operational strategies that use pixel counts (i.e., number of pixels with T_b colder than a threshold value) within 200–300 km of the TC center for RI prediction.

Meanwhile, mode 2 (Fig. 15b) features a dipole, with convection biased toward the right-front quadrant of the TC and warmer cloud tops in the left-rear quadrant. Mode 3 (Fig. 15c) is essentially mode 2 but rotated 90° counterclockwise, with convection biased toward the left-front quadrant. K15 found a similar pattern in univariate PCA of IR imagery (their Fig. 2) and concluded that this pattern, with convection concentrated

¹¹ Using more TC objects in the baseline leads to out-of-memory errors.

¹² Homogeneous-regression maps are explained in supplemental section 4. For readers more familiar with PCA, the equivalent is a regression map (the word “homogeneous” is not needed when there is only one field), such as those in Fig. 4 of Knaff et al. (2015).

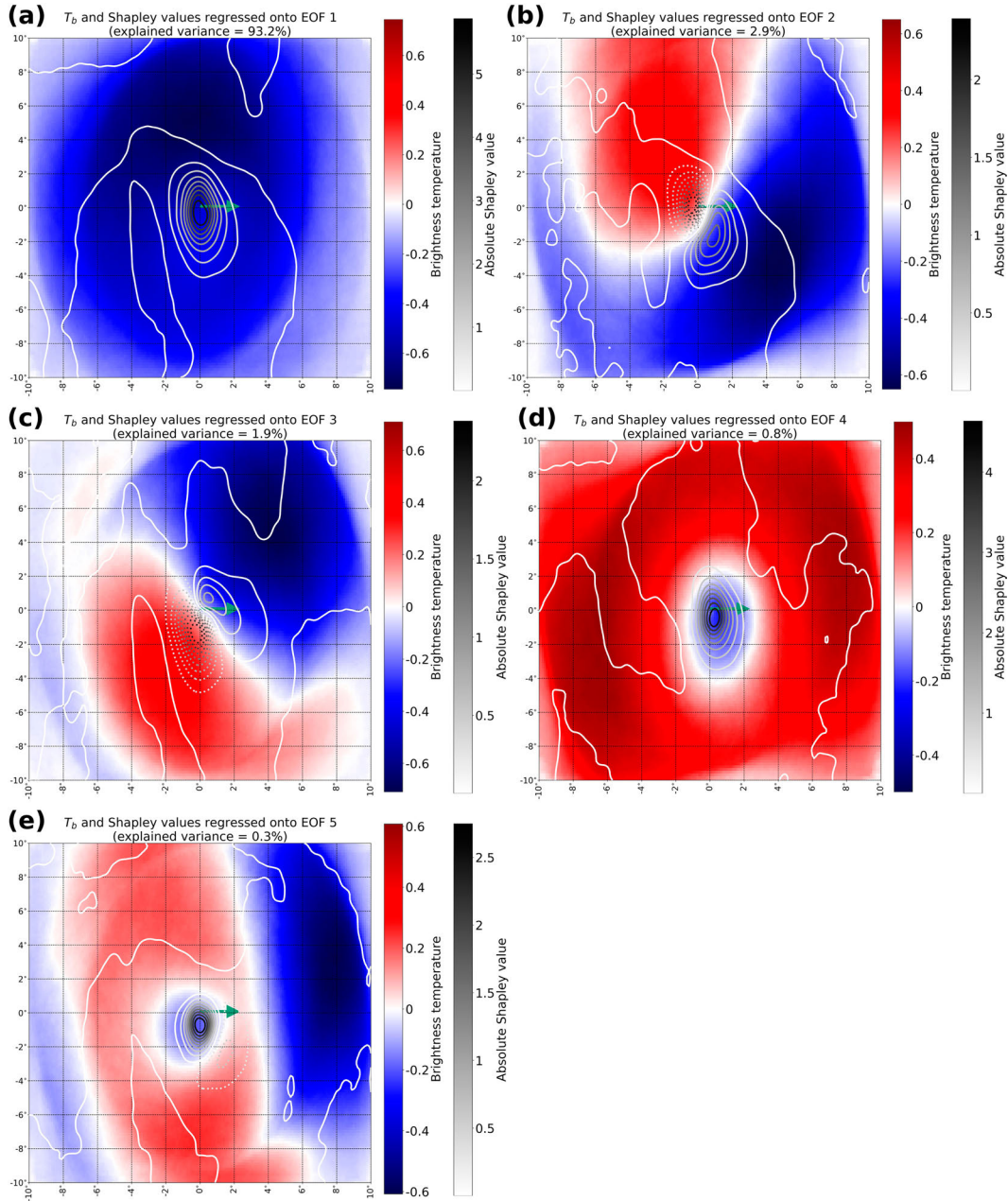


FIG. 15. Homogeneous regression maps for top five modes of MCA. All inputs to the MCA (TC objects) are centered at dummy coordinates (0° , 0°) with TC motion pointing to the right; thus, (0° , 0°) is the composite TC center, while the green arrow is the direction of composite TC motion. The red–blue color fill shows double-normalized T_b (see supplemental section 4 for the meaning of “double-normalized”), and the grayscale line contours show Shapley values, one for every T_b pixel. The color scale represents the magnitude of a Shapley value, while the line style (dashed or solid) represents the sign of a Shapley value. Dashed contours represent negative Shapley values—i.e., negative contributions to RI probability, or evidence against RI. Solid contours represent positive Shapley values—i.e., positive contributions to RI probability, or evidence for RI.

in the left-front quadrant, is favorable for RI. Collectively, T_b dipoles in modes 2 and 3 suggest that the NN can leverage nonaxisymmetric patterns of convection. From the literature (section 1), we know that these nonaxisymmetric patterns typically occur earlier—before and at the onset of RI—evolving to

become axisymmetric during RI. Thus, modes 2 and 3 could represent strategies used by the NN earlier during RI, allowing for a longer lead time. K15 supports this view, stating that enhanced convection in the left-front quadrant often precedes axisymmetrization of the overall convective pattern. However, in

both modes 2 and 3, Shapley values show that the NN attributes positive RI evidence to the cold cloud tops and negative evidence to the warm cloud tops. In other words, only the cold side of the dipole is seen as conducive to RI. Thus, modes 1–3 collectively suggest that the NN prefers cold cloud tops everywhere near the TC center and does not prefer warm cloud tops anywhere. In other words, the NN's top three strategies for RI prediction leverage mostly axisymmetric patterns (although modes 2 and 3 suggest a slight preference for cold cloud tops ahead of the TC). The axisymmetric areal coverage of cold cloud tops around the TC center is already described in several satellite-based SHIPS predictors; this is likely a key reason that NNs without full satellite images perform nearly as well as those with the images (section 4).

Last, modes 4 and 5 are similar to mode 1, but with cold cloud tops confined to a much smaller ring around the TC center. In both modes 4 and 5, positive Shapley values collocated with cold cloud tops indicate higher RI probabilities. We hypothesize that mode 4, which is nearly symmetric, is related to the diurnal oscillations often seen in TCs (Gray and Jacobson 1977; Kossin 2002; Dunion et al. 2014). Knaff et al. (2019) found a similar mode with univariate PCA that oscillates diurnally, with a maximum of cold cloud tops occurring in the early morning. These oscillations occur in {45%, 54%, and 61%} of TCs with at least {tropical storm, hurricane, and major hurricane} strength. Ditchek et al. (2019) came to similar conclusions, using different methods. Also, we hypothesize that mode 5 is related to variations in the size of the convective core; both overall TC size and inner-core size contribute weakly to predicting RI (Knaff et al. 2018, 2020). However, our exploration of these two hypotheses yields very weak correlations (supplemental section 6). Thus, while diurnal pulsing and TC size may contribute weakly to improving RI forecasts, our analysis suggests that these phenomena are not correct interpretations of modes 4 and 5, respectively.

6. Summary and future work

RI is a difficult forecasting problem of great societal importance. Recent improvements in RI forecasting have been largely fueled by statistical models, including machine learning (ML) and, more recently, deep learning such as convolutional neural networks (CNNs). However, two questions have still not been answered. First, what is the influence of different predictor sources (e.g., full satellite images versus scalars in the SHIPS developmental data, which include summaries of the images) on RI skill? Does the use of full images or videos significantly improve RI skill? Second, what can NNs learn from full satellite images?

To answer the first question, we perform an ablation experiment, training NNs with different combinations of full satellite images and scalar summaries of the images, TC history, and near-TC environment. We found that the SHIPS data are collectively much more important than the TC IR image data and that all three types of SHIPS data—satellite-based, environmental, and storm-based—are important. We also found that, while the inclusion of full satellite images marginally improves NN performance, this performance gain is achieved by a single

image (one time step), with no additional performance gain achieved by a video. Thus, NNs can successfully leverage some spatial information in the satellite images that is not subsumed by the SHIPS predictors, but they cannot do the same for spatiotemporal information. However, these conclusions are tempered by the facts that 1) SHIPS predictors and IR imagery contain redundant information; 2) we have not exhausted all possible NN or deep learning approaches; and 3) higher-resolution IR imagery or microwave imagery (although the latter is available at much lower temporal frequency) could provide more skill. Also, the problem investigated here is 24-h RI; for shorter time windows such as 12-h RI, the initial TC-state information contained in IR imagery might be more important. We selected three best NN models: the best overall (called the “NN with everything” because it uses data from all sources), without TC IR data, and without SHIPS data. We selected three operational models as baselines: SHIPS-RII, SHIPS consensus, and DTOPS. The two NNs including SHIPS data produce better binary (yes/no) forecasts, but worse probabilistic forecasts, than the baselines; the NN without SHIPS is worse than the baselines in both categories. However, the NNs provide well-calibrated uncertainty estimates, while the baselines provide no uncertainty. Also, the NN without SHIPS performs surprisingly well, and its forecasts are fairly independent of pre-existing RI models, suggesting that it could be a valuable addition to an operational ensemble (e.g., Sampson et al. 2023). Nonetheless, the main conclusion from the ablation experiment is that full satellite images add little skill to RI prediction beyond the SHIPS scalars.

To answer the second question, we used the DeepSHAP XAI method. For every TC object, DeepSHAP produces an attribution map, quantifying the RI evidence at each satellite-image pixel. Because individual attribution maps are noisy, we used maximum covariance analysis (MCA) to find the leading modes of covariability—between satellite brightness temperature and RI evidence. The leading modes suggest that the NN's main strategy for RI prediction is to look for an axisymmetric ring of cold cloud tops (deep convection) around the TC center. Although we know from the literature that nonaxisymmetric patterns are also prevalent—occurring just before or at the onset of RI—we did not find evidence that the NN leverages nonaxisymmetric patterns. This could indicate that nonaxisymmetric patterns are overall not good predictors of RI or that the NN simply fails to use them. Nonetheless, the NN's main strategy—looking for cold cloud tops in an axisymmetric ring—is also the main strategy used to derive most of the satellite-based SHIPS predictors, which helps explain why adding the full TC IR image data to SHIPS data yields such a small performance gain.

In future work, we will incorporate the NN without SHIPS into an ensemble with other RI models and find the best ways to estimate uncertainty from such an ad hoc ensemble, i.e., to ensure a good spread–skill relationship. Also, we will explore improved ways to leverage spatiotemporal information in satellite videos; leveraging spatiotemporal information remains a vexing problem in ML applications to environmental science (Singh Bansal et al. 2022).

Acknowledgments. This work was partially supported by the NOAA Global Systems Laboratory, Cooperative Institute for Research in the Atmosphere, and NOAA Award NA19OAR4320073. Author Ebert-Uphoff's work was partially supported by NSF AI Institute Grant 2019758.

Data availability statement. We used version 1.0.0 of Machine Learning for Tropical Cyclones (ML4TCs; <https://zenodo.org/doi/10.5281/zenodo.10268620>)—a Python library managed by author Lagerquist—for all tasks in this study. A subset of the real-time testing data (87 of 260 TCs, totaling ~100 GB)—fully processed and normalized, following the methods in section 2—can be found at <https://zenodo.org/records/13272829> and <https://zenodo.org/records/13272877>. The trained NNs—i.e., the 50-NN ensemble with everything, the 50-NN ensemble without TC IR data, and the 50-NN ensemble without SHIPS data—can be found at <https://zenodo.org/records/13274366>. Raw ATCF and best track files for the North Atlantic (AL), eastern North Pacific (EP), and central North Pacific (CP) basins are available from the NHC at <https://ftp.nhc.noaa.gov/atcf/>. Best track files for the western North Pacific (WP), Indian Ocean (IO), and Southern Hemisphere (SH) are available from the JTWC at <https://www.metoc.navy.mil/jtwc/jtwc.html?best-tracks>. ATCF files for these basins are not publicly available.

REFERENCES

- Alford, A. A., M. I. Biggerstaff, and G. Carrie, 2023: Using ground-based radar observations to evaluate asymmetric convection and eyewall dynamics during the landfall of Hurricane Harvey (2017). *J. Atmos. Sci.*, **80**, 1867–1889, <https://doi.org/10.1175/JAS-D-22-0053.1>.
- Barreto, N. J. C., M. S. Mesquita, D. Mendes, M. H. C. Spyrides, G. U. Pedra, and P. S. Lucio, 2017: Maximum covariance analysis to identify intraseasonal oscillations over tropical Brazil. *Climate Dyn.*, **49**, 1583–1596, <https://doi.org/10.1007/s00382-016-3401-3>.
- Black, M. L., R. W. Burpee, and F. D. Marks Jr., 1996: Vertical motion characteristics of tropical cyclones determined with airborne Doppler radial velocities. *J. Atmos. Sci.*, **53**, 1887–1909, [https://doi.org/10.1175/1520-0469\(1996\)053<1887:VMCOTC>2.0.CO;2](https://doi.org/10.1175/1520-0469(1996)053<1887:VMCOTC>2.0.CO;2).
- Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning-based probabilistic hail predictions for operational forecasting. *Wea. Forecasting*, **35**, 149–168, <https://doi.org/10.1175/WAF-D-19-0105.1>.
- Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latta, 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>.
- Chen, H., and D.-L. Zhang, 2013: On the rapid intensification of Hurricane Wilma (2005). Part II: Convective bursts and the upper-level warm core. *J. Atmos. Sci.*, **70**, 146–162, <https://doi.org/10.1175/JAS-D-12-062.1>.
- Chollet, F., 2018: *Deep Learning with Python*. Manning, 361 pp.
- Chow, S., 1971: A study of the wind field in the planetary boundary layer of a moving tropical cyclone. M.S. thesis, Dept. Meteorology and Oceanography, New York University, 59 pp.
- Cotto, A., I. Gonzalez III, and H. E. Willoughby, 2015: Synthesis of vortex Rossby waves. Part I: Episodically forced waves in the inner waveguide. *J. Atmos. Sci.*, **72**, 3940–3957, <https://doi.org/10.1175/JAS-D-15-0004.1>.
- Cummings, J. A., 2005: Operational multivariate ocean data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131**, 3583–3604, <https://doi.org/10.1256/qj.05.105>.
- DeMaria, M., R. T. DeMaria, J. A. Knaff, and D. Molnar, 2012: Tropical cyclone lightning and rapid intensity change. *Mon. Wea. Rev.*, **140**, 1828–1842, <https://doi.org/10.1175/MWR-D-11-00236.1>.
- , C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, <https://doi.org/10.1175/BAMS-D-12-00240.1>.
- , 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>.
- Ditchev, S. D., J. Molinari, K. L. Corbosiero, and R. G. Fovell, 2019: An objective climatology of tropical cyclone diurnal pulses in the Atlantic Basin. *Mon. Wea. Rev.*, **147**, 591–605, <https://doi.org/10.1175/MWR-D-18-0368.1>.
- Dunion, J. P., C. D. Thorncroft, and C. S. Velden, 2014: The tropical cyclone diurnal cycle of mature hurricanes. *Mon. Wea. Rev.*, **142**, 3900–3919, <https://doi.org/10.1175/MWR-D-13-00191.1>.
- Emanuel, K. A., 1986: An air-sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. *J. Atmos. Sci.*, **43**, 585–605, [https://doi.org/10.1175/1520-0469\(1986\)043<0585:AASITF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1986)043<0585:AASITF>2.0.CO;2).
- Frankignoul, C., N. Chouaib, and Z. Liu, 2011: Estimating the observed atmospheric response to SST anomalies: Maximum covariance analysis, generalized equilibrium feedback assessment, and maximum response estimation. *J. Climate*, **24**, 2523–2539, <https://doi.org/10.1175/2010JCLI3696.1>.
- Gray, W. M., and R. W. Jacobson Jr., 1977: Diurnal variation of deep cumulus convection. *Mon. Wea. Rev.*, **105**, 1171–1188, [https://doi.org/10.1175/1520-0493\(1977\)105<1171:DVODCC>2.0.CO;2](https://doi.org/10.1175/1520-0493(1977)105<1171:DVODCC>2.0.CO;2).
- Griffin, S. M., A. Wimmers, and C. S. Velden, 2022: Predicting rapid intensification in North Atlantic and eastern North Pacific tropical cyclones using a convolutional neural network. *Wea. Forecasting*, **37**, 1333–1355, <https://doi.org/10.1175/WAF-D-21-0194.1>.
- Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artif. Intell. Earth Syst.*, **2**, e220061, <https://doi.org/10.1175/AIES-D-22-0061.1>.
- Heymsfield, G. M., J. B. Halverson, J. Simpson, L. Tian, and T. P. Bui, 2001: ER-2 Doppler radar investigations of the eyewall of Hurricane Bonnie during the convection and moisture experiment-3. *J. Appl. Meteor.*, **40**, 1310–1330, [https://doi.org/10.1175/1520-0450\(2001\)040<1310:EDRIOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(2001)040<1310:EDRIOT>2.0.CO;2).
- Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, 2012: Improving neural networks by preventing co-adaptation of feature detectors. arXiv, 1207.0580v1, <https://doi.org/10.48550/arXiv.1207.0580>.
- Hsu, W., and A. H. Murphy, 1986: The attributes diagram A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, [https://doi.org/10.1016/0169-2070\(86\)90048-8](https://doi.org/10.1016/0169-2070(86)90048-8).

- Kalnay, E., 2002: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 368 pp., <https://doi.org/10.1017/CBO9780511802270>.
- Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108, [https://doi.org/10.1175/1520-0434\(2003\)018<1093:LCORIT>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<1093:LCORIT>2.0.CO;2).
- , —, and J. A. Knaff, 2010: A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **25**, 220–241, <https://doi.org/10.1175/2009WAF2222280.1>.
- , and Coauthors, 2015: Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Wea. Forecasting*, **30**, 1374–1396, <https://doi.org/10.1175/WAF-D-15-0032.1>.
- Klotz, B. W., and H. Jiang, 2017: Examination of surface wind asymmetries in tropical cyclones. Part I: General structure and wind shear impacts. *Mon. Wea. Rev.*, **145**, 3989–4009, <https://doi.org/10.1175/MWR-D-17-0019.1>.
- Knaff, J. A., S. P. Longmore, and D. A. Molenaar, 2014: An objective satellite-based tropical cyclone size climatology. *J. Climate*, **27**, 455–476, <https://doi.org/10.1175/JCLI-D-13-00096.1>.
- , —, R. T. DeMaria, and D. A. Molenaar, 2015: Improved tropical-cyclone flight-level wind estimates using routine infrared satellite reconnaissance. *J. Appl. Meteor. Climatol.*, **54**, 463–478, <https://doi.org/10.1175/JAMC-D-14-0112.1>.
- , C. R. Sampson, and K. D. Musgrave, 2018: Statistical tropical cyclone wind radii prediction using climatology and persistence: Updates for the western North Pacific. *Wea. Forecasting*, **33**, 1093–1098, <https://doi.org/10.1175/WAF-D-18-0027.1>.
- , C. J. Slocum, and K. D. Musgrave, 2019: Quantification and exploration of diurnal oscillations in tropical cyclones. *Mon. Wea. Rev.*, **147**, 2105–2121, <https://doi.org/10.1175/MWR-D-18-0379.1>.
- , C. R. Sampson, and B. R. Strahl, 2020: A tropical cyclone rapid intensification prediction aid for the Joint Typhoon Warning Center's areas of responsibility. *Wea. Forecasting*, **35**, 1173–1185, <https://doi.org/10.1175/WAF-D-19-0228.1>.
- Kossin, J. P., 2002: Daily hurricane variability inferred from GOES infrared imagery. *Mon. Wea. Rev.*, **130**, 2260–2270, [https://doi.org/10.1175/1520-0493\(2002\)130<2260:DHVIFG>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<2260:DHVIFG>2.0.CO;2).
- Lagerquist, R., 2020: Using deep learning to improve prediction and understanding of high-impact weather. Ph.D. dissertation, University of Oklahoma, 290 pp., <https://shareok.org/handle/11244/324145>.
- , A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, <https://doi.org/10.1175/WAF-D-17-0038.1>.
- , J. Q. Stewart, I. Ebert-Uphoff, and C. Kumler, 2021: Using deep learning to nowcast the spatial coverage of convection from Himawari-8 satellite data. *Mon. Wea. Rev.*, **149**, 3897–3921, <https://doi.org/10.1175/MWR-D-21-0096.1>.
- Lee, C.-Y., M. K. Tippett, A. H. Sobel, and S. J. Camargo, 2016: Rapid intensification and the bimodal distribution of tropical cyclone intensity. *Nat. Commun.*, **7**, 10625, <https://doi.org/10.1038/ncomms10625>.
- Luna-Herrera, J., G. Martínez-Cabrera, R. Parra-Maldonado, J. A. Enciso-Moreno, J. Torres-López, F. Quesada-Pascual, R. Delgadillo-Polanco, and S. G. Franzblau, 2003: Use of receiver operating characteristic curves to assess the performance of a microdilution assay for determination of drug susceptibility of clinical isolates of *Mycobacterium tuberculosis*. *Eur. J. Clin. Microbiol. Infect. Dis.*, **22**, 21–27, <https://doi.org/10.1007/s10096-002-0855-5>.
- Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., 4768–4777, <https://dl.acm.org/doi/10.5555/3295222.3295230>.
- McNeely, T., A. B. Lee, K. M. Wood, and D. Hammerling, 2020: Unlocking GOES: A statistical framework for quantifying the evolution of convective structure in tropical cyclones. *J. Appl. Meteor. Climatol.*, **59**, 1671–1689, <https://doi.org/10.1175/JAMC-D-19-0286.1>.
- 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>.
- Metz, C. E., 1978: Basic principles of ROC analysis. *Semin. Nucl. Med.*, **8**, 283–298, [https://doi.org/10.1016/s0001-2998\(78\)80014-2](https://doi.org/10.1016/s0001-2998(78)80014-2).
- Montgomery, M. T., and R. J. Kallenbach, 1997: A theory for vortex Rossby-waves and its application to spiral bands and intensity changes in hurricanes. *Quart. J. Roy. Meteor. Soc.*, **123**, 435–465, <https://doi.org/10.1002/qj.49712353810>.
- Mueller, K. J., M. DeMaria, J. Knaff, J. P. Kossin, and T. H. Vonder Haar, 2006: Objective estimation of tropical cyclone wind structure from infrared satellite data. *Wea. Forecasting*, **21**, 990–1005, <https://doi.org/10.1175/WAF955.1>.
- Muller, M. P., G. Tomlinson, T. J. Marrie, P. Tang, A. McGeer, D. E. Low, A. S. Detsky, and W. L. Gold, 2005: Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia? *Clin. Infect. Dis.*, **40**, 1079–1086, <https://doi.org/10.1086/428577>.
- Niculescu-Mizil, A., and R. Caruana, 2005: Predicting good probabilities with supervised learning. *ICML'05: Proc. 22nd Int. Conf. on Machine Learning*, Bonn, Germany, International Machine Learning Society, 625–632, <https://doi.org/10.1145/1102351.1102430>.
- Nielsen, I. E., D. Dera, G. Rasool, R. P. Ramachandran, and N. Carla, 2022: Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Process. Mag.*, **39**, 73–84, <https://doi.org/10.1109/MSP.2022.3142719>.
- NOAA, 2020: Statistical tropical cyclone intensity forecast technique development: Developmental data. https://rammb.cira.colostate.edu/research/tropical_cyclones/ships/developmental_data.asp.
- Pangaluru, K., I. Velicogna, G. A. Y. Mohajerani, E. Ciraci, S. Charakola, G. Basha, and S. V. B. Rao, 2019: Soil moisture variability in India: Relationship of land surface–atmosphere fields using maximum covariance analysis. *Remote Sens.*, **11**, 335, <https://doi.org/10.3390/rs11030335>.
- Rappaport, E. N., J.-G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The Joint Hurricane Test Bed: its first decade of tropical cyclone research-to-operations activities reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371–380, <https://doi.org/10.1175/BAMS-D-11-00037.1>.
- Reasor, P. D., M. D. Eastin, and J. F. Gamache, 2009: Rapidly intensifying Hurricane Guillermo (1997). Part I: Low-wavenumber structure and evolution. *Mon. Wea. Rev.*, **137**, 603–631, <https://doi.org/10.1175/2008MWR2487.1>.

- Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, <https://doi.org/10.1175/2008WAF2222159.1>.
- Rogers, R., 2010: Convective-scale structure and evolution during a high-resolution simulation of tropical cyclone rapid intensification. *J. Atmos. Sci.*, **67**, 44–70, <https://doi.org/10.1175/2009JAS3122.1>.
- Sampson, C. R., and A. J. Schrader, 2000: The Automated Tropical Cyclone Forecasting System (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, [https://doi.org/10.1175/1520-0477\(2000\)081<1231:TATCFS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2).
- , J. Cummings, J. A. Knaff, M. DeMaria, and E. A. Serra, 2022: An upper ocean thermal field metrics dataset. *Meteorology*, **1**, 327–340, <https://doi.org/10.3390/meteorology1030021>.
- , J. A. Knaff, C. J. Slocum, M. J. Onderlinde, A. Brammer, M. Frost, and B. Strahl, 2023: Deterministic rapid intensity forecast guidance for the Joint Typhoon Warning Center's area of responsibility. *Wea. Forecasting*, **38**, 2631–2640, <https://doi.org/10.1175/WAF-D-23-0084.1>.
- Shanker, M., M. Y. Hu, and M. S. Hung, 1996: Effect of data standardization on neural network training. *Omega*, **24**, 385–397, [https://doi.org/10.1016/0305-0483\(96\)00010-2](https://doi.org/10.1016/0305-0483(96)00010-2).
- Shapiro, L. J., and H. E. Willoughby, 1982: The response of balanced hurricanes to local sources of heat and momentum. *J. Atmos. Sci.*, **39**, 378–394, [https://doi.org/10.1175/1520-0469\(1982\)039<0378:TROBHT>2.0.CO;2](https://doi.org/10.1175/1520-0469(1982)039<0378:TROBHT>2.0.CO;2).
- Shapley, L., 1953: A value for n-person games. *Contributions to the Theory of Games*, Vol. 2, H. Kuhn and A. Tucker, Eds., Princeton University Press, 307–317.
- Singh Bansal, A., Y. Lee, K. Hilburn, and I. Ebert-Uphoff, 2022: Tools for extracting spatio-temporal patterns in meteorological image sequences: From feature engineering to attention-based neural networks. arXiv, 210.12310v2, <https://arxiv.org/abs/2210.12310>.
- Slocum, C. J., J. A. Knaff, and S. N. Stevenson, 2023: Lightning-based tropical cyclone rapid intensification guidance. *Wea. Forecasting*, **38**, 1209–1227, <https://doi.org/10.1175/WAF-D-22-0157.1>.
- Su, H., L. Wu, J. H. Jiang, R. Pai, A. Liu, A. J. Zhai, P. Tavallali, and M. DeMaria, 2020: Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophys. Res. Lett.*, **47**, e2020GL089102, <https://doi.org/10.1029/2020GL089102>.
- Tang, X., F. Ping, S. Yang, M. Li, and J. Peng, 2018: Relationship between convective bursts and the rapid intensification of Typhoon Mujigae (2015). *Atmos. Sci. Lett.*, **19**, e811, <https://doi.org/10.1002/asl.811>.
- Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, Vol. 100, Academic Press, 627 pp.
- 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>.
- Zehr, R. M., and J. A. Knaff, 2007: Atlantic major hurricanes, 1995–2005—Characteristics based on best-track, aircraft, and IR images. *J. Climate*, **20**, 5865–5888, <https://doi.org/10.1175/2007JCLI1652.1>.