

Statistical Prediction of Tropical Cyclone Rapid Intensification with Explainable AI

TAKESHI HORINOUCHI^{a,b}, TAKASHI YANASE^c, YUIKO OHTA^{b,c}, DAISUKE MATSUOKA^{b,d}, ASANOBU KITAMOTO^{b,e},
UDAI SHIMADA^{b,f}, RYUJI YOSHIDA^b, AND HIRONORI FUDEYASU^b

^a Faculty of Environmental Earth Science, Hokkaido University, Sapporo, Hokkaido, Japan

^b Typhoon Science and Technology Research Center, Yokohama National University, Yokohama, Kanagawa, Japan

^c Artificial Intelligence Laboratory, Fujitsu Research, Fujitsu Limited, Kawasaki, Kanagawa, Japan

^d Center for Earth Information Science and Technology, JAMSTEC, Yokohama, Japan

^e National Institute of Informatics, Tokyo, Japan

^f Meteorological Research Institute, Tsukuba, Ibaraki, Japan

(Manuscript received 5 December 2024, in final form 17 June 2025, accepted 1 July 2025)


ABSTRACT: Imperfectness of the state-of-the-art intensity forecasting of tropical cyclones (TCs) necessitates independent rapid intensification (RI) prediction schemes. Here, we report one derived with an explainable artificial intelligence Wide Learning (WL). The scheme, named Wide Learning-based TC rapid intensification prediction scheme (WRPS), version 1 (WRPS1), predicts RI in the western North Pacific by using twelve predictor variables representing environmental conditions and the state of TCs. Its prediction is based on a score that is a linear combination of whether or not (1 or 0) joint conditions on ranges of multiple variables are met, which is reproducible without WL. Relying on joint conditions allows WRPS to handle nonlinearity and interdependence among predictors, and the simpleness of the conditions provides explainability. A method to map an RI-prediction score to its probability is proposed and is used in WRPS. It is suggested that handling predictors favorable to RI when having moderate values, such as the current intensity, is a key for good RI prediction. It is demonstrated that quantifying the contribution of each predictor to the WRPS score helps one elucidate how the predictors jointly facilitated or hindered RI for each prediction case. The performance of WRPS1 is compared with RI predictions using the linear discriminant analysis, and WRPS1 is shown to perform well without using track predictions. The multiple linear regression analysis, which is customarily used for intensity prediction but not for RI prediction, is shown to perform well if the fraction of RI cases is increased when conducting regression.


SIGNIFICANCE STATEMENT: We developed a new scheme to predict rapid intensification of tropical cyclones, WRPS version 1, by using an explainable artificial intelligence. Its prediction is based only on 12 parameters, but it performs well. The WRPS prediction formula is simple and reproducible by using information available in the supplemental material of this paper. It is suggested that handling predictors favorable to RI when having moderate values, such as the current intensity, is a key for good RI prediction. It is demonstrated that quantifying the contribution of each predictor to the WRPS score helps one elucidate how the predictors jointly facilitated or hindered RI for each prediction case.

KEYWORDS: Tropical cyclones; Probability forecasts/models/distribution; Statistical forecasting

1. Introduction

Accurate predictions of tropical cyclones (TCs) have enormous socioeconomic benefits. Although their track forecasts have been steadily improving, their intensity forecasts suffer from slow progress (e.g., DeMaria et al. 2014; Zhang et al. 2023; Wang et al. 2023). Because of the limited performance of physically based numerical atmospheric forecasts, empirical statistical methods are widely used supplementarily in operational forecasts.

 Denotes content that is immediately available upon publication as open access.

 Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-24-0228.s1>.

Corresponding author: Takeshi Horinouchi, horinout@ees.hokudai.ac.jp

The most widely used empirical intensity forecast method is the Statistical Hurricane Intensity Prediction Scheme (SHIPS) (DeMaria and Kaplan 1994, 1999; DeMaria et al. 2005). SHIPS relies on the multiple linear regression (MLR) of intensity changes onto some predictors, which are numerical quantities regarding along-track environmental conditions (e.g., ocean heat content, SST, and the horizontal wind shear between the upper and lower troposphere), the current state of the TC (e.g., maximum sustained wind, which shall be called VMAX in what follows), and their combinations [e.g., maximum potential intensity (MPI) minus VMAX, which is customarily called as potential intensification (POT)]. In addition to its practical merit by its performance, SHIPS provides knowledge on intensification mechanisms through regression coefficients of the predictors. Also, SHIPS provides explanation on each prediction by elucidating the contribution of each predictor.

Each SHIPS model is derived for each ocean basin. DeMaria et al. (2005) derived it for the Atlantic and the eastern North Pacific. Knaff et al. (2005) and Yamaguchi et al. 2018 (Y18 hereinafter) derived ones for the western North Pacific. SHIPS is based on the “perfect prog” approach, in which the predictors

DOI: 10.1175/WAF-D-24-0228.1

© 2025 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Unauthenticated | Downloaded 09/13/25 02:26 AM UTC

for training are derived from observations and/or objective (reanalysis) products. For environmental variables, it is customary to use temporal averages along observed tracks over the forecast period (say, $t = 0$ –24 h, where t denotes the forecast time). This approach is adequate when the actual forecast of the TC track and the environmental variables are good enough. More recently, many studies were conducted to forecast TC intensity using machine learning (ML; e.g., Sharma et al. 2013; Chaudhuri et al. 2013; Wimmers et al. 2019; Cloud et al. 2019; Xu et al. 2021; Meng et al. 2023; Chen et al. 2023; Shimada 2024).

The rapid intensification (RI) of TCs does not have a universal definition but is customarily defined as a 1- or 10-min VMAX increase of around 30 kt (1 kt $\approx 0.51 \text{ m s}^{-1}$) over 24 h (e.g., Kaplan and DeMaria 2003). It is sometimes defined with respect to the central pressure drop (Holliday and Thompson 1979). Predicting RI is beneficial for society but is challenging as it is to predict relatively infrequent intensity changes (Fudeyasu et al. 2018). Fudeyasu et al. (2018) statistically investigated the characteristics of TCs undergoing RI in the western North Pacific over 37 years from 1979 to 2015. Out of the 900 TCs, 201 TCs underwent RI defined by a 10-min VMAX increase of 30 kt or more over 24 h.

The imperfectness of intensity prediction necessitates independent RI prediction. Kaplan et al. (2010; K10 hereinafter) developed a deterministic and probabilistic RI prediction scheme based on the linear discriminant analysis (LDA) using SHIPS predictors, and they applied it to the Atlantic and the eastern North Pacific. Rozoff and Kossin (2011; R11 hereinafter) also used SHIPS predictors for RI prediction and tested two additional methods, logistic regression (LR) and a Naïve Bayesian probabilistic model. They found that LR yielded better skill scores than LDA and the Bayesian model. They also showed that a consensus prediction in which the probabilities from the three methods are averaged tends to provide better prediction than any of the three, indicating the importance of using multiple prediction methods. Kaplan et al. (2015) also tested the three models and obtained similar results. Knaff et al. (2020) describe an operational RI prediction system for the Joint Typhoon Warning Center TC forecast area of responsibility. Sampson et al. (2023) present several RI predictions schemes and the formulation of a consensus RI methodology for use in the western North Pacific, Southern Hemisphere, and north Indian Ocean.

Recent development of RI predictions naturally includes the use of ML techniques. Gao et al. (2016) introduced a decision tree approach to statistical RI forecast. Shaiba and Hahsler (2016) tried several methods including the support vector machines (SVMs). Mercer and Grimes (2017) and Su et al. (2020) tried to optimize the combinations of several RI prediction methods including the LR, random forest, decision tree, and SVM. Griffin et al. (2022) fed geostationary satellite infrared images in addition to SHIPS predictors to a convolutional neural network. Chen et al. (2023) employed a deep neural network by using satellite infrared images and the “mimic” passive microwave images (PMWs) generated from the infrared images. Kim et al. (2024) tested the use of the net energy gain rate index, which is qualitatively similar to POT.

In this study, we develop a deterministic as well as probabilistic RI prediction scheme by using an ML algorithm called Wide Learning (WL) developed by Fujitsu Limited (Iwashita et al. 2020). WL is an explainable artificial intelligence (XAI). WL scores each prediction based on a linear combination of compound predictors (see section 2). This provides explanations like SHIPS while allowing nonlinear dependence on raw predictors through compounding. As the first step of RI prediction with WL, we only use SHIPS predictors in this study, which helps comparisons with previous studies. We plan to use additional data in the future. For convenience of referencing, we name our method as Wide Learning-based TC rapid intensification prediction scheme (WRPS).

For comparison with WRPS, we mainly use the LDA method, which can serve as a benchmark because it has been widely used since K10. We also use MLR as in SHIPS. Operational centers do not use SHIPS for RI forecasts presumably because of its poor performance for RI predictions (DeMaria et al. 2021; Knaff et al. 2023). However, we show that it can have a comparable performance to LDA if the ratio of the RI cases is increased by undersampling non-RI cases in the training dataset. Such a technique is frequently used in ML practices, and it is actually used for WL in this study. A similar approach can be seen in the study of Shimada (2024) who developed a SHIPS intensity forecast in which regression models are prepared separately for intensifying, nearly steady, and weakening TCs; the choice from the three is made based on a random forest classification.

The rest of the paper is organized as follows. Data and methods are described in section 2. The performance of the WL predictions is investigated and compared with other methods in section 3. The overall features of the WL predictions are investigated in section 4, and case studies are conducted in section 5. Conclusions are drawn in section 6.

2. Data and methods

a. Data

We use SHIPS developmental data for the western North Pacific published by the Cooperative Institute for Research in the Atmosphere (https://rammb-data.cira.colostate.edu/ships/data/ships_predictor_file.pdf). This dataset is designed for developing SHIPS-like empirical–statistical intensity predictions, and it consists of 6-hourly data along TC tracks derived from the National Hurricane Center and JTWC best track data, the NCEP global model analyses, and geosynchronous meteorological satellites. Slocum et al. (2022) conducted a comparison among the SHIPS developmental data, fifth generation European Centre for Medium-Range Weather Forecasts atmospheric reanalysis (ERA5) data, and dropsonde data.

The period of the data used is 2002–21. The year 2021 is the last year available at the time when the study is conducted. We limit our analysis to TCs over the ocean at the prediction initial time $t = 0$, which is done by excluding the cases in which the “INCV” field, the intensity change from the previous time, is set to missing. Therefore, our analysis includes a small number of cases in which the TCs are over land at the

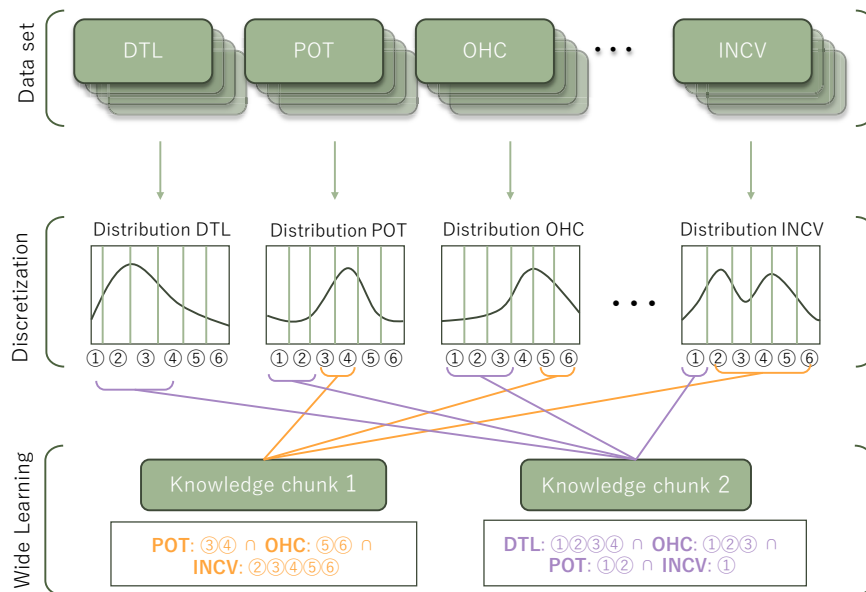


FIG. 1. Schematic illustration of data processing in WRPS1. Each variable is discretized unevenly into n ($=6$) ranges maximizing information entropy. Then, WL is applied to extract knowledge chunks, each of which consists of consecutive ranges of L ($=4$) or fewer variables. The knowledge chunks are conditions favorable or against RI, and these are used to estimate RI probability. In this figure, “POT: ③④,” for example, means that the value of POT falls in its third or fourth range.

forecast times. We exclude the first record time for each TC because INCV is not available, and we also exclude the last recorded 18 h for each TC because the intensity change to $t = 24$ h is not available.

For this study, we use the best track to estimate RI occurrence and accept the uncertainties associated with the intensity estimates. The total number of the 6-hourly data used is 9648. We define RI as the 1-min VMAX increase of 30 kt or greater in 24 h. Then, RI occurred in 1018 cases, so the RI ratio is 10.6%.

We employ two types of predictor treatment regarding the times at which the training data are taken. The one mainly used is to use predictor values only at the forecast initial time ($t = 0$) or changes from $t = -6$ h to $t = 0$. We shall call this as the “no forecast” (NF) type. The other is to use some of the environmental values averaged from $t = 0$ to the forecast time, which is $t = 24$ h in this study, while the other predictors as in the NF type. This is called the “perfect environmental forecast” (PEF) type. The PEF approach is conventionally used to derive the SHIPS models, and it is suitable for operations. However, it might be too advantageous in evaluating performance, since we do not use operational forecasts in our tests. Therefore, we mainly use NF cases. Note that both NF and PEF are within the perfect prog framework as defined in the glossary of meteorology of American Meteorological Society (https://glossarytest.ametsoc.net/wiki/Perfect_prognostic).

b. Wide Learning

WL is an XAI whose core algorithm is described by Iwashita et al. (2020). WL targets classification or discrimination problems and learns well using limited training data. It exhaustively

searches training data in an efficient way for all combinations of input variables, and it extracts combinations suitable for the target discrimination problem as important hypotheses, which are called “knowledge chunks.” WL version V2210 is used in this study.

Figure 1 illustrates our data process to use WL. The input variables of WL are allowed to have only two values: true or false. Therefore, a continuous predictor variable is discretized into a fixed number of ranges so that its value can be classified according to the range it resides. The knowledge chunks are the combinations of variable ranges. For example, consider predictor variables a , b , and c , which are continuous, and their values are classified into n ranges relative to the threshold values a_i , b_i , and c_i for $i = 1, 2, \dots, n-1$, which, for example, gives the n ranges of $a < a_1$, $a_1 \leq a < a_2$, \dots , $a_{n-1} \leq a$ for the variable a . A knowledge chunk expresses a condition regarding a subset of the predictors, say, a and c , as $(a_{\{i\}} \text{ AND } c_{\{k\}})$, where $a_{\{i\}}$ designates whether the value of a falls in some consecutive ranges of a . Here, $\{i\}$ refers to a group of consecutive i values between 1 and n . For example, when $\{a_i\} = \{1, 2, 5, 8, 10\}$ and $\{i\} = \{2, 3\}$ (i.e., the value of a is located in the second or third bin), $a_{\{i\}}$ is equivalent to whether or not $a_1 \leq a < a_3$, namely, $1 \leq a < 5$. Also, $c_{\{k\}}$ is similarly defined. The number of the variables used in a chunk is equal to or less than a threshold value of L . After testing, this study uses $L = 4$.

The discretization threshold values of a predictor are adaptively determined unevenly to maximize information entropy, the average amount of information, so that the correlation between the discretized predictor and the prediction target is largely maintained (Fayyad and Irani 1993). It might appear

TABLE 1. The SHIPS predictor variables used in the 12-predictor version of the WL predictions and in our LDA. The predictor types are classified into C (climatological or geographical values dependent solely on the location and the day of the year), T (measures on the state of the TC), and E (environmental conditions). POT is classified as E&T because MPI, which is the maximum potential intensity from Kerry Emanuel's equation (kt), is E but VMAX is T; Z850 is E&T because of the wide areal averaging; and R000 is predominantly E, but it may partly be T for large TCs. The numbers in the last column are the orders in which the predictors are selected in the stepwise regression for VMAX prediction at $t = 24$ h in the SHIPS model of Y18 for the western North Pacific. The symbol “(*)” indicates that their definitions are not the same as in this table (e.g., in Y18, OHC is not climatological, and VMAX is for the 10-min sustained wind).

Acronym	Description (units) (radial range if areal average)	Time mean if PEF	Type	Importance in Y18
COHC	Climatological ocean heat content (kJ cm^{-2})	y	C	7(*)
DTL	Distance to nearest major landmass (km)	y	C	26
VMAX	Maximum sustained wind (kt)		T	6(*)
POT	MPI minus VMAX (kt)	y for MPI	E&T	1
INCV	Intensity change from $t = -6$ to 0 h (kt)		T	2(*)
SDIR	Standard deviation of IR brightness temperature (10^{-1}°C) (0–200 km)		T	8
SHRD	850–200-hPa shear magnitude (10^{-1} kt) (200–800 km)	y	E	3(*)
T150	150-hPa temperature (10^{-1}°C) (200–800 km)		E	23
T200	200-hPa temperature (10^{-1}°C) (200–800 km)		E	24
T250	250-hPa temperature (10^{-1}°C) (200–800 km)		E	19
Z850	850-hPa vorticity (10^{-7} s^{-1}) (0–1000 km)		E&T	11
R000	1000-hPa relative humidity (%) (200–800 km)		E&T	9

that the number of discretization bins n is better to have a sufficiently large value, but using too many divisions can deteriorate predictions. After trials, n is fixed to 6 in this study. The discretization based on information entropy is not a part of WL, but it is used frequently in its practices.

The climatological fraction of the RI cases is much lower than non-RI cases. Many ML classification methods do not perform well if trained with highly unbalanced data, which is the case for WL too. Therefore, we reduce the fraction of non-RI cases in the training dataset by undersampling. The undersampling is made randomly in terms of the six hourly data, so most of the TCs are used. After some trials, we set the fraction of RI cases to 30%. Undersampling is not used for testing as a matter of course.

WL conducts logistic regression for knowledge chunks, so the RI probability at the forecast time in the undersampled training dataset, p_u , is expressed as

$$\log \frac{p_u}{1 - p_u} = \sum_{j=1}^J w_j x_j + c. \quad (1)$$

Here, J is the number of chunks, x_j is 1 (0) if the chunk j is true (false), w_j is the weight for the chunk, and c is a constant offset; w_j and c are obtained by training. The deterministic prediction by WL is RI (non-RI) if p_u is greater than or equal to (less than) 0.5. Thresholding at $p_u = 0.5$ may appear ad hoc, since we conduct undersampling. This threshold is rather an empirical choice that provides good performance (see the discussion based on the p - r diagram in section 3b). For probabilistic predictions, p_u is adjusted to derive the predicted RI predictability p by the method described in section 2e.

Our method is similar to R11's in the sense that it conducts LR by using SHIPS predictors. However, the variable x_j is created from a combination of SHIPS predictors. Therefore, p_u can express nonlinear dependence on them.

Equation (1) indicates that the degree of freedom of WL predictions is J . As shown later, J becomes only a few tens, so

WL is good at avoiding overfitting as in the simple LR and LDA. Another merit of WL is that it can secure full reproducibility by simply providing the conditions, the coefficients $\{w_j\}$, and c , as is done in this study.

As described in section 3a, we start by using most of the variables available in the SHIPS developmental data and subsequently reduce the number of variables by checking performance. Table 1 shows our final 12 predictors. The variables have little data missing (less than 0.03%) except for SDIR whose data missing fraction is 14%.

c. Linear discriminant analysis

To compare with the WL results, LDA is conducted as in K10. The predictors used in this study are set to the same as those used in WL, so one can directly evaluate the difference among the methods. We adopt K10's thresholding for deterministic predictions. For probabilistic predictions, we derive probabilities with the method described in section 2e, which is shown to perform better than the method of K10 (section 3c).

d. Multiple linear regression

The MLR method used in SHIPS is not customarily used for RI predictions. However, in the course of the present study, we found that its performance becomes comparable to LDA if the training data are undersampled as described in section 2b. For MLR, we use the 12 predictors introduced above and, in addition, nine variables including nonlinear combinations of predictors as in SHIPS as in Table 2. With MLR, we predict RI when the predicted 24-h intensity change is greater than or equal to 30 kt.

e. Probabilistic prediction

The RI probability p_u for the undersampled datasets (section 2b) can be converted into the probability without undersampling as

TABLE 2. Predictors used in MLR in addition to the 12 predictors in Table 1. In the last column, the symbol (*) indicates the existence of differences between Y18 and this study, which is largely due to the differences described in the caption of Table 1; see Y18 for other differences.

Acronym	Definition	Importance in Y18
VMA2	(VMAX) ²	6(*)
VMPE	VMAX × INCV	14(*)
OHC2	(COHC) ²	5(*)
SHSH	(SHRD) ²	Unused
VMSH	VMAX × SHRD	22(*)
SHVM	VMAX/SHRD	16(*)
POT2	(POT) ²	10
PMPE	(MSLP − 880) × INCV	12(*)
PMSH	(MSLP − 880) × SHRD	18(*)

$$p_l \equiv \frac{\alpha p_u}{1 - (1 - \alpha)p_u}. \quad (2)$$

Here, α is the rate at which the non-RI cases are sampled so that only 100 α % of non-RI cases are used. Equation (2) is derived from the relation $p_u = p_l/[p_l + \alpha(1 - p_l)]$. It is readily shown that

$$\alpha = \frac{p_c}{1 - p_c} \frac{1 - X}{X}, \quad (3)$$

where p_c , which is 0.1055, is the RI ratio in the original dataset and X is the ratio of RI in the undersampled training data, which is set to 0.3 in this study. Then, $\alpha = 0.275$.

If the ML method used is adequate, the probability p_l should be somewhat close to the actual RI ratio in the training data, but it can be biased. Therefore, in the probabilistic prediction, a better performance is expected if the RI probability P is derived by correcting the bias in p_l . More generally, p_l (or p_u) can be treated as a score indicating RI probability. Let s be such a score; for WL, $s \equiv p_l$. We estimate the RI probability P as a function of s as

$$P(s) = \sum_{k=1}^K c_k \rho_k(s), \quad (4)$$

where, for each k , $\rho_k(s)$ is a known analytic function of s and c_k is a constant to be derived; K is the number of polynomials used. We derive c_k by the least square fitting to minimize the Brier score (BS):

$$\text{BS} \equiv \frac{1}{N} \sum_{n=1}^N \{P(s_n) - a_n\}^2, \quad (5)$$

where N is the number of the cases used for training, s_n is the score for the n th case, and

$$a_n \equiv \begin{cases} 1, & \text{if RI occurred actually,} \\ 0, & \text{if RI did not occur.} \end{cases} \quad (6)$$

Note that the resultant $P(s)$ should be a monotonic function of s if the score s is adequate as a measure of RI probability.

TABLE 3. “Confusion matrix” introducing the symbols for case counts a – d in deterministic predictions.

Actual	Prediction	
	RI	Non-RI
RI	a (true positive)	b (false negative)
Non-RI	c (false positive)	d (true negative)

The choice of the functions $\rho_k(s)$ and their number K should be made under this principle. In this study, we simply set as $\rho_k(s) = s^k$; a nonzero intercept is not introduced from the examinations of the results.

K10 related the LDA score to the RI probability through a linear interpolation/extrapolation by coarsely binning RI ratio along s . The relationship between K10’s and our methods is shown in the appendix. Like ours, K10’s method is designed to reduce biases by using classification results, and it is close to minimizing BS.

f. Evaluation metrics

For the deterministic forecast, we use the following measures (see Table 3 for symbol definitions):

- Precision: $p = a/(a + c)$ (the ratio of true predictions among all positive predictions).
- Recall (the probability of detection): $r = a/(a + b)$ (the ratio of true predictions among all RI cases).
- F1 measure: $F = 2pr/(p + r) = a/[a + (b + c)/2]$ (balanced measure of precision and recall).
- False alarm rate (FAR): $f = c/(c + d)$ (the ratio of false predictions among all non-RI cases).
- Peirce skill score (PSS) = $r - f$ (a balanced measure of true and false detections).

As in many ML-related studies, we treat F1 as the most important measure, since it balances precision and recall that tend to trade off each other. Note that F1 is qualitatively similar to but is greater than the threat score, $a/(a + b + c) = pr/(p + r - pr)$.

Probabilistic predictions are evaluated with the Brier skill score (BSS):

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_c}. \quad (7)$$

Here, BS is defined by Eq. (5), and BS_c is the BS in which the predicted probability is replaced by the constant climatological one, p_c .

g. Training and testing sets and cross validation

For evaluation, we use four-fifths of the 20 years for training (16 years) and the remaining one-fifth (4 years) for testing. Cross validation is conducted by averaging the performance from the five cases in which test is conducted by using the years 1) 2002, 2007, 2012, and 2017; 2) 2003, 2008, 2013, and 2018; 3) 2004, 2009, 2014, and 2019; 4) 2005, 2010, 2015, and 2020; and 5) 2006, 2011, 2016, and 2021. The performance metrics for cross validations are the averages of those for the five sets.

TABLE 4. The cross-validation performance metrics for deterministic predictions to compare the 71-, 47-, and 12-predictor versions of the NF-type WL prediction for the RI from $t = 0$ to 24 h. The values are shown as the mean \pm the standard error over the five test sets, which is the sample standard deviation divided by $\sqrt{5}$. The F1 and BSS measures are shown with bold fonts, since they are treated as the most important metrics.

No.	Precision	Recall	F1	FAR	PSS	BSS
71	0.420 \pm 0.031	0.518 \pm 0.013	0.463 \pm 0.024	0.086 \pm 0.007	0.432 \pm 0.018	0.199 \pm 0.035
47	0.420 \pm 0.025	0.487 \pm 0.028	0.451 \pm 0.026	0.079 \pm 0.004	0.408 \pm 0.027	0.183 \pm 0.030
12 (WRPS1)	0.429 \pm 0.028	0.553 \pm 0.020	0.482 \pm 0.024	0.088 \pm 0.006	0.465 \pm 0.023	0.225 \pm 0.035

The final WL model of this study is created by using the entire 20 years for training. Its metrics are derived by using the entire data for the same 20-yr period without conducting undersampling. We shall call this the training data test.

3. Parameter selection and prediction performance

a. Reduction of predictors

The SHIPS developmental data for the western North Pacific has ~ 100 variables, and high correlations exist among some of them. Using so many variables makes interpretation a difficult task, even though it may not hinder achieving good performance. We began by using 71 predictors, which are listed in Table S1 in the online supplemental material. Then, the number was first reduced to 47 by excluding predictors that are redundant, have high correlation with others, or are produced secondarily from other variables (see the caption of Table S1 for further explanation). After some further trials and referencing Y18 (see the caption of Table S1), we finally reduced the set to the 12 predictors in Table 1. The 12 predictors are not a simple subset of the 47 or 71 predictors, since MPI is used instead of POT in the 47- and 71-predictor versions. Also, use of SDIR is introduced for the first time in the 12 predictors. We call the 12-predictor version of the RI prediction with WL as WRPS version 1 (WRPS1).

We provide here a few remarks on some of the 12 predictors that are not frequently used for SHIPS or LDA. In many studies (e.g., DeMaria et al. 2005, K10, Kaplan et al. 2015, Y18), the intensity changes over the past 12 h are used, but we use the 6-h changes because it provided better performance in our test. Customarily used ocean heat content (OHC) is time dependent, changing year to year. Instead, we use the climatological OHC (COHC) just because time-dependent OHC is

not available in the SHIPS developmental data for the western North Pacific. As for low-level humidity, averages over 850–750 hPa are used more frequently. Our choice to use R000 is not a result of the performance test, but it comes from the fact that it was found useful by Y18 for 24-h VMAX prediction for the western Pacific, which was systematically selected in their study. It will be shown below that R000 is infrequently used in our result, so it might be desirable to change it or simply not to use it in future. Use of three pressure levels (150, 200, and 250 hPa) for environmental temperature in such a limited predictor set may be redundant. Again, this choice is not elaborated by performance, so further investigation would be desirable if further simplification is wanted.

The performance metrics obtained by cross validation for the 71-, 47-, and 12-predictor versions are shown in Table 4. The overall performances are similar among the three. The mean F1 and BSS values of the 12-predictor version (WRPS1) are rather slightly better than the other two. The improvement might be due to the use of POT instead of MPI, or it might be due to the use of SDIR. However, the improvement is rather small in light of the standard errors.

b. Performance of the NF-type 24-h predictions

In the rest of this paper, we solely use the 12 predictors for WL (WRPS1). Table 5 compares performance metrics for WL and the other methods obtained by the cross validation of the NF-type 24-h predictions. It also shows the training data test results for our final WL model created by using the 20-yr data for training.

The cross validation shows that WRPS1 performs best among the four methods. The F1 and PSS values, 0.482 and 0.465, respectively, are quite high. These values are comparable to those reported by Kim et al. (2024), who conducted the NF-type

TABLE 5. The performance metrics for the NF-type predictions for the RI from $t = 0$ to 24 h: from precision to PSS for deterministic predictions and BSS for probabilistic predictions. (second row–fourth row) The cross validation for WL (WRPS1), LDA, and MLR with and without undersampling (United States). (bottom row) The results of the training data test, in which the entire (not undersampled) data for the 20-yr training period are used for evaluation. Values from the cross validation are shown as in Table 4.

Test type	Method	Precision	Recall	F1	FAR	PSS	BSS
Cross validation	WRPS1	0.429 \pm 0.028	0.553 \pm 0.020	0.482 \pm 0.024	0.088 \pm 0.006	0.465 \pm 0.023	0.225 \pm 0.035
	LDA	0.354 \pm 0.030	0.471 \pm 0.031	0.404 \pm 0.031	0.102 \pm 0.004	0.370 \pm 0.034	0.140 \pm 0.023
	MLR w/ United States	0.444 \pm 0.048	0.389 \pm 0.036	0.414 \pm 0.041	0.058 \pm 0.005	0.331 \pm 0.040	
	MLR w/o United States	0.518 \pm 0.061	0.074 \pm 0.011	0.128 \pm 0.018	0.008 \pm 0.001	0.066 \pm 0.011	
Training data test	WRPS1	0.453	0.601	0.517	0.086	0.515	0.260

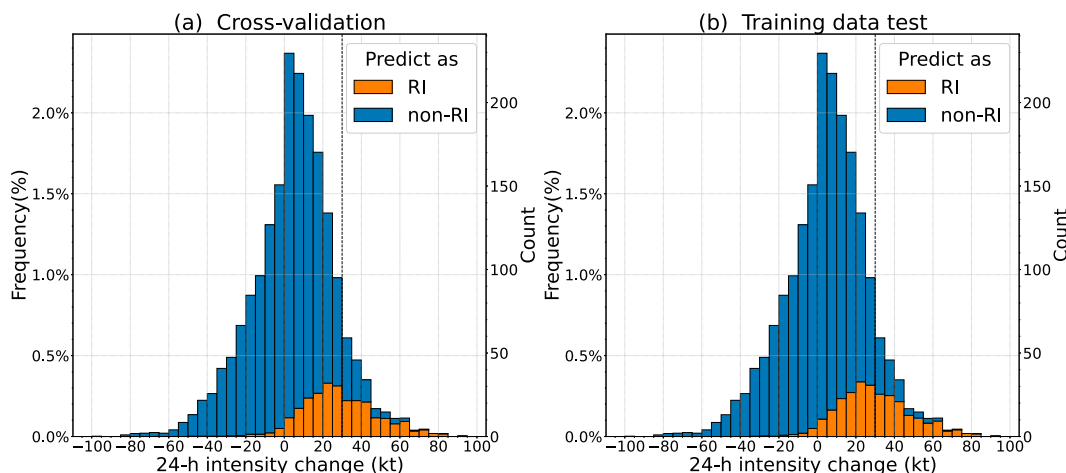


FIG. 2. Histogram showing the actual 24-h intensity (1-min VMAX) changes in the data we used. The orange (blue) portions indicate those predicted to RI (not to RI) by WRPS1: (a) cross validation and (b) the training data test of the final WL model obtained by using the 20-yr training data.

prediction too, even though we did not introduce their thresholding to limit to the cases where VMAX is 34 kt or higher. According to their Table 1, their RI ratio is 15.5%, which is higher than ours (10.6%). It should be noted that classification problems generally become easier if events are not as rare. Therefore, a direct comparison of our results with their results is not possible. The F1 value of the RI prediction reported by Chen et al. (2023) is 0.36. This is lower than our result, but they used 24-h 30-kt increase in the 10-min maximum wind to define RI, which is less frequent than that for the 1-min maximum wind (7.9% in their study). Therefore, the direct comparison is again unavailable.

As shown in Table 5, the MLR performs poorly if the entire training data are used, but it performs comparably to LDA if the training data are undersampled. This is understood because MLR is not good at handling outliers. Our sampling strategy is somewhat in liaison with Shimada (2024) who prepared multiple MLR models for different VMAX-change ranges (see section 1).

In comparison with other ML results (e.g., Kim et al. 2024), a remarkable feature of the WRPS1 results is that the metrics values are not very different between the training data test and the cross validation (see Table 5). This is likely because the degree of freedom is limited, as stated in section 2b. In this respect, WRPS behaves similarly to LDA and simple LR.

Figure 2 shows the histogram of the actual (recorded) 24-h intensity changes along with classification results by WRPS1 shown with colors. In both cross validation and training data test, the prediction to RI is mostly limited to intensifying cases. Also, most of the cases in which the actual intensification was greater than 40 kt are correctly predicted as RI. The ratio at which RI is predicted is around 14% in both cases (Table 6).

Figure 3 shows the histograms in terms of the predicted RI probability p_l which is the score from WL. From this figure, one can see how deterministic prediction changes according to the threshold in p_u ; the value used, $p_u = 0.5$, corresponds to $p_l = 0.215$ when $p_c = 0.1055$. As shown in Fig. 4, precision and recall trade off each other, and the figure indicates that the present threshold is adequate in terms of the F1 measure.

c. Probabilistic prediction

As described in section 2e, we estimate RI probability as a function of p_l by minimizing BS. This optimization is done solely based on the training data test using the entire 20-yr data. We set $p_k(p_l) = p_l^k$ and tested the polynomial orders K from 1 to 5 and additionally 10. The BSS is 0.222 for $K = 1$ and 0.225–0.226 for all the rest ($K = 2, 3, 4, 5, 10$). Therefore, we use the simplest among the latter, $K = 2$, which resulted in the following for the NF-type $t = 0$ –24-h RI prediction:

$$P(p_l) = 1.134p_l - 0.296p_l^2. \quad (8)$$

The black solid curves in Fig. 5 visualize Eq. (8). As expected, $P(p_l)$ matches the actual RI ratio (orange bars) in the training data test (Fig. 5b). Furthermore, it also matches the actual RI ratio in the cross validation (Fig. 5a). This result indicates that the WL-based probability prediction is largely unbiased. This nice feature is presumably because the limited degree of freedom in WL predictions suppresses overfitting.

To be unbiased, however, is not sufficient to be a good prediction. A perfect prediction with BSS equal to 1 is only achievable when the predicted probability is only 0 or 1, and furthermore, the prediction is 100% correct. We find from Fig. 3 that a considerable portion of the probability prediction falls around the climatological RI probability of 11%. For better prediction, we

TABLE 6. Case counts and percentage from the NF-type deterministic WRPS1 predictions for RI from $t = 0$ to 24 h.

		Prediction	
		RI	Non-RI
Cross validation			
Actual	RI	566 (5.9%)	452 (4.9%)
	Non-RI	752 (7.8%)	7878 (81.7%)
Training data test			
Actual	RI	612 (6.3%)	406 (4.2%)
	Non-RI	739 (7.7%)	7891 (81.8%)

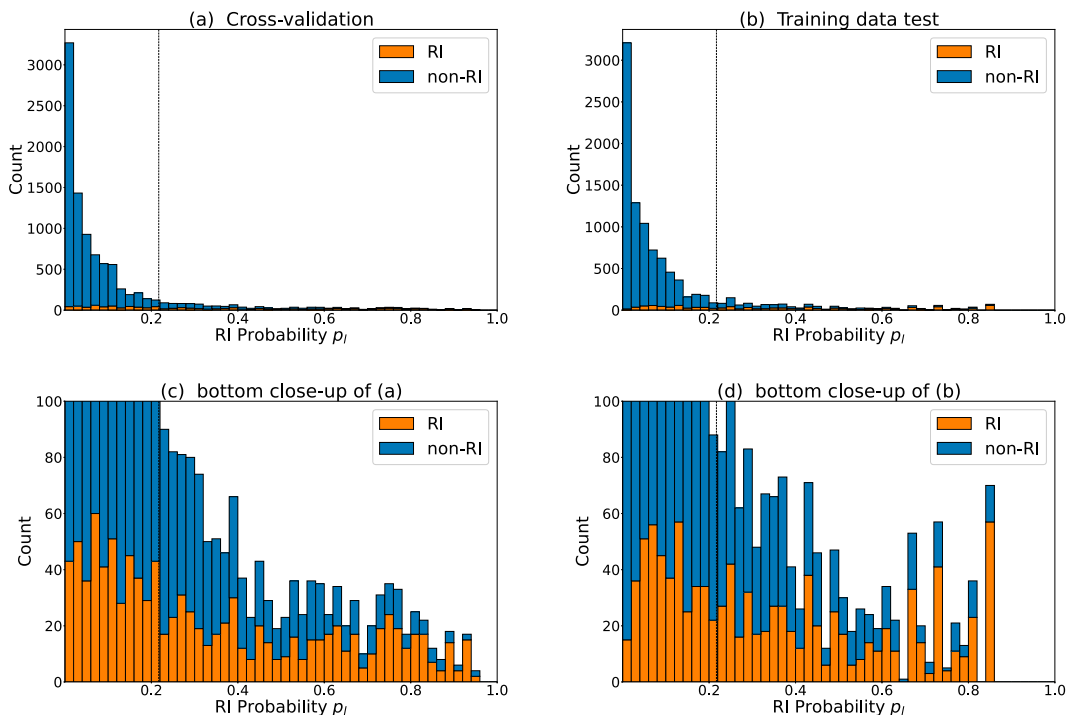


FIG. 3. Histograms with respect to p_i , the WL-based probability defined by Eq. (2) for (a) cross validation and (b) training data test. (c),(d) Close-ups of (a) and (b), respectively, for low values. The coloring is as in Fig. 2. The vertical dotted lines show the threshold p_i corresponding to $p_u = 0.5$.

need not only unbiased but also a bimodal distribution along $P(p_i)$, which poses further challenges.

It is not straightforward to compare the present BSSs with previous studies. For reference, R11, who employed PEF-type predictions and conducted the leave-one-out cross

validation, reported BSSs for the simple LR and LDA predictions of 30-kt RI around 0.15 and 0.11, respectively, for the Atlantic; the values are around 0.28 and 0.22 for the eastern Pacific (see their Fig. 1). Therefore, there is a large basin dependence in the difficulty of predicting RI.

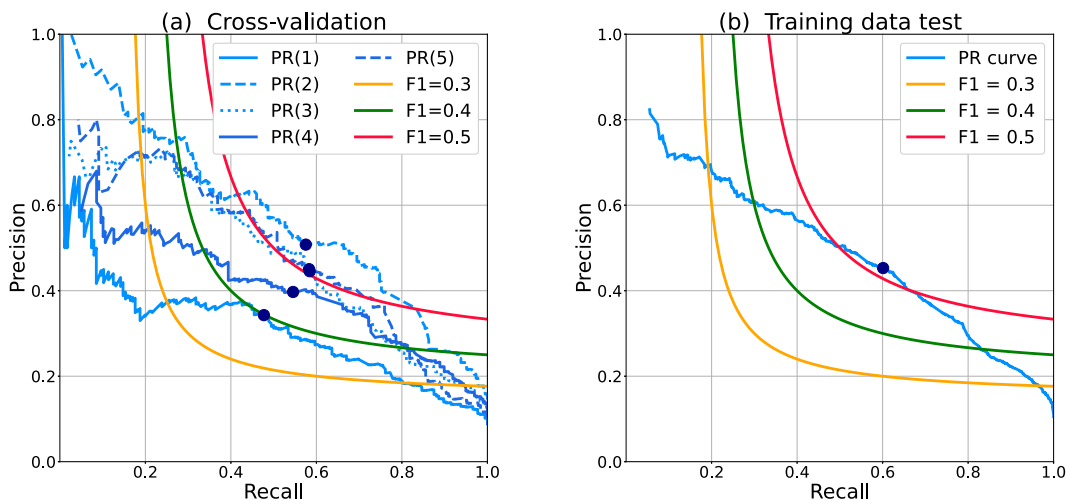


FIG. 4. The precision–recall diagrams that visualize the dependence on changing the RI threshold: (a) from the cross validation, for which the results for the five validation sets are shown by blue curves with different line types [PR(i) = tested with the years $2001 + i + 5l$, $l = 0, 1, 2, 3$], and (b) from the training data test. The p – r values for the deterministic RI predictions ($p_u = 0.5$) are marked by the bullets. F1 values corresponding to the combinations of p and r are indicated by the colored contours.

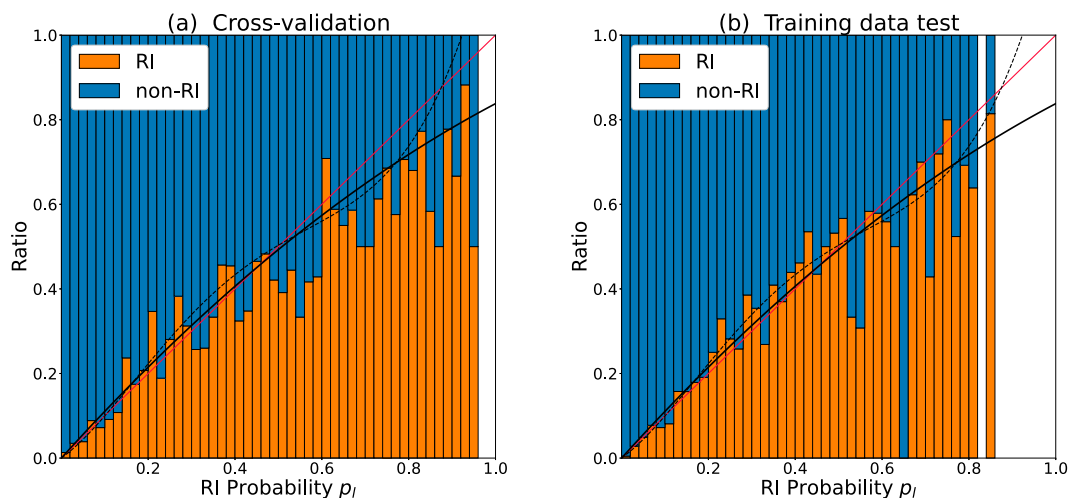


FIG. 5. The ratio of actual RI depending on p_i (orange bars) for (a) the cross validation and (b) training data test for the NF-type 24-h WRPS1 predictions. The black solid curves show the optimized RI probability $P(p_i)$ derived from the 20-yr training result, Eq. (8), where $K = 2$. For reference, the dotted curves show $P(p_i)$ when $K = 5$, which is not used in our probabilistic prediction.

Contrary to the WRPS1 results, $P(p_i)$ for the LDA prediction is nonmonotonic (Fig. 6; p_i for LDA is derived from its assumption of the normal distribution). The BSS for LDA in Table 5, which is 0.140, is for $K = 2$. We speculate that the nonmonotonicity arose from the presence of the “range”-type crucial predictors, VMAX and POT.

As mentioned in section 2f, the probability assignment method by K10 is designed to reduce biases, as it is the case in the present method. Therefore, we also tried their method, but the resultant BSS for WRPS1 was only 0.129, much lower than that obtained by the present method. This result indicates the merit of our probability assignment.

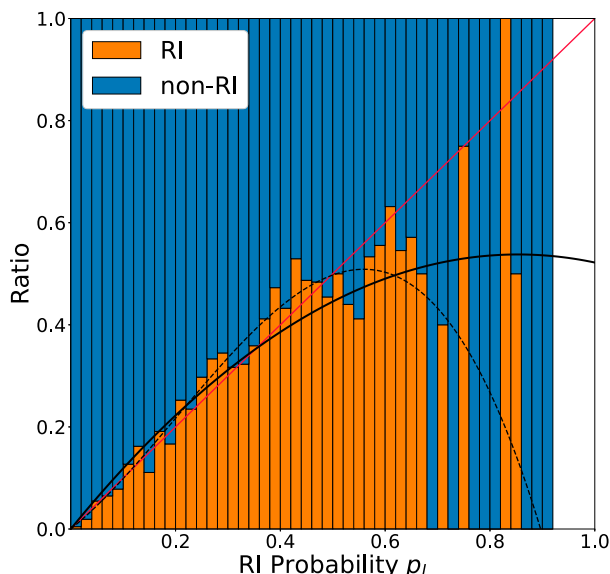


FIG. 6. As in Fig. 5b, but for LDA. The black solid curve shows the optimized RI probability $P(p_i)$ for $K = 2$, while the black dotted curve shows that for $K = 3$, for reference.

d. Performance of the PEF-type 24-h predictions

Here, we examine the performance of the PEF-type RI predictions from $t = 0$ –24 h. The difference from the NF-type predictions is that some environmental predictor variables are the averages over $t = 0$ –24 h. This is close to the practice of operational forecast, but our evaluation uses SHIPS developmental data, which is based on observations and reanalysis. To avoid excessively good performance by using such data, we take a conservative approach to apply time averaging only to four variables, COHC, DTL, MPI (for POT), and SHRD, as indicated in Table 1. These parameters depend on TC tracks and the environmental shear.

The resultant performance is summarized in Table 7. The F1 value for WRPS1 only slightly increased from 0.482 to 0.488, well within the reach of the standard errors. This result could be interpreted to mean that, when WL is used, it is not important to use predicted TC track and environmental shear. However, this feature might be limited to WRPS1, as it may change, for example, by using predictors unused in WRPS1.

The PEF and NF differences are slightly greater for the LDA and MLR. The skill scores for probabilistic predictions, BSS, of WRPS1 and LDA are also slightly higher for PER than NF.

e. Performance of the NF-type $t = 24$ –48-h predictions

It is of interest to know how long the $t = 0$ predictors possess information on the RI at later times. A possible way to examine it is to define RI for $t = 0$ –36 or 48 h (e.g., Kaplan et al. 2015). In this study, we rather take a simple approach to examine the predictions of RI that occur from $t = 24$ to 48 h. This is an ad hoc choice, but we can use the same threshold, 30 kt $(24 \text{ h})^{-1}$, and make a direct comparison with the prediction of the $t = 0$ –24 h RI. Here, we only conduct the NF-type predictions.

TABLE 7. As in Table 5, but for the PEF-type RI predictions for $t = 0\text{--}24$ h.

Test type	Method	Precision	Recall	F1	FAR	PSS	BSS
Cross validation	WRPS1	0.437 ± 0.022	0.555 ± 0.033	0.488 ± 0.024	0.085 ± 0.005	0.471 ± 0.032	0.239 ± 0.035
	LDA	0.367 ± 0.025	0.496 ± 0.024	0.422 ± 0.025	0.101 ± 0.003	0.395 ± 0.026	0.154 ± 0.025
	MLR w/ United States	0.472 ± 0.045	0.417 ± 0.030	0.442 ± 0.036	0.055 ± 0.004	0.361 ± 0.034	
	MLR w/o United States	0.560 ± 0.056	0.082 ± 0.010	0.143 ± 0.017	0.007 ± 0.001	0.075 ± 0.010	
Training data test	WRPS1	0.465	0.607	0.527	0.082	0.525	0.288

Table 8 summarizes the results. As expected, the performance deteriorates from the 0- to 24-h prediction. However, the F1 value still amounts to 0.397 (Table 8) as opposed to 0.482 in Table 5. The BSS for the $t = 24\text{--}48\text{-h}$ WRPS1 prediction, 0.124, is also lower, but it is not meaningless. These results indicate that predictability exists in the SHIPS predictors on RI over 2 days.

4. Overall features of the WRPS1 predictions

Here, we examine the results of the $t = 0\text{--}24\text{-h}$ NF-type WRPS1 trained by the 20-yr data, our final scheme for the 24-h RI prediction. The constant in Eq. (1) is $c = -0.3536$. The knowledge chunks, which are the conditions to evaluate x_j , and their coefficients w_j are fully provided in Tables S2 and S3 for positive and negative chunks, respectively; here, positive (negative) chunks denote the chunks with positive (negative) w_j values, which are favorable (detrimental) to the occurrence of RI. The tables also provide some statistical properties of the chunks such as the fraction of the cases satisfying the condition (support) and the fraction of the RI occurrence in the following 24 h among them (confidence).

The ways in which the predictors appear in the chunks are summarized in Table 9. It shows the number of chunks in which each predictor is used. The counts are classified into six types: P/pos , R/pos , N/pos , P/neg , R/neg , and N/neg . This is based on whether it appears in the positive (pos) or negative (neg) chunks and whether it is favorable to RI when the value is large (P), small (N), or in a range (R). The table also shows the summation of the weights associated with each variable.

Not surprisingly, COHC is the most frequently used predictor, as latent heat supply is crucial for TC intensification. On the other hand, R000 is used only in a chunk that has a small weight (Table S2), so removing it from the predictor set would have little impact on the prediction. The linear RI models (LDA, MLD, and simple R11-type LR) treat their predictors equally for the positive and negative predictions. In WL, positive and negative impacts are treated separately, and there is asymmetry between them. For example, VMAX appears only in the positive chunks, and T200 appears mainly

in the negative chunks (Table 9). T250 appears both in the positive and negative chunks, but it is always as $T250 > -39.6^\circ\text{C}$, which is meaningless if it were not combined with other parameters differently between the positive and negative chunks. Since the offset c is negative, RI is never predicted if no positively weighted condition is met. Therefore, when COHC is lower than 63 kJ cm^{-2} , RI is always missed (Table S2).

The positive chunks are fewer than the negative chunks, while the negative chunks tend to consist of fewer predictors than the positive chunks (Tables S2 and S3). One of the reasons for this asymmetry should be the existence of predictors that are favorable to RI at moderate values. If, for example, RI tends to occur when the predictors a and b satisfy $a_1 < a < a_2$ AND $b > b_1$, the unfavorable conditions can be expressed as $a < a_1$ OR $a > a_2$ OR $b < b_1$. Such a tendency is broadly seen in joint distributions of the predictors as partly shown in Fig. 8, where RI cases tend to be distributed more narrowly. It is also noteworthy that positive (negative) chunks tend to have relatively high (low) support and low (high) confidence (Tables S2 and S3).

In the positive chunks, POT and VMAX are used somewhat interchangeably (Table S2). This feature is understood from the POT–VMAX joint distributions (Fig. 8a), where upper bounds of POT (=MPI – VMAX) tend to depend on VMAX because MPI tends to be limited. Contrarily, only POT is used in the negative chunks (Table 9).

Based on the counts shown in Table 9, the overall evaluation of each predictor is summarized in the last column of the table. The summary is about whether the predictor is favorable to RI when it is large (pos), small (neg), having moderate values (range), or can be positive or negative (mixed). Most of the evaluations are consistent with the differences between the RI- and non-RI-case distributions shown in Fig. 7 except for the mixed-type ones. For example, VMAX and POT are favorable to RI when having moderate values, which is consistent with being range-type variables. Note that range-type variables are not quite suitable for LDA or simple LR. Two frequently used variables, VMAX and POT, are classified as range-type variables. All conditions regarding VMAX are ranges. POT has R/pos and P/neg counts, so it might be better

TABLE 8. As in Table 5, but for the NF-type RI predictions for $t = 24\text{--}48$ h and for the omission of the training data test.

Test type	Method	Precision	Recall	F1	FAR	PSS	BSS
Cross validation	WRPS1	0.321 ± 0.027	0.525 ± 0.039	0.397 ± 0.029	0.127 ± 0.011	0.397 ± 0.034	0.124 ± 0.024
	LDA	0.272 ± 0.031	0.398 ± 0.035	0.322 ± 0.034	0.123 ± 0.006	0.275 ± 0.038	0.090 ± 0.024
	MLR w/ United States	0.300 ± 0.036	0.259 ± 0.021	0.274 ± 0.025	0.072 ± 0.009	0.187 ± 0.022	
	MLR w/o United States	0.024 ± 0.024	0.006 ± 0.006	0.010 ± 0.009	0.005 ± 0.004	0.001 ± 0.002	

TABLE 9. The numbers of times at which each predictor appears in the knowledge chunks of the NF-type 24-h RI prediction of WRPS1, which are described in Tables S2 and S3; also shown are the total weight and the overall evaluation of each predictor regarding RI prediction. The chunks are classified into the positive and negative ones: the positive (negative) chunks are the ones having positive (negative) weights in Eq. (1), whose total number is 18 (55) as in Tables S2 and S3. The counts are classified into six categories: *P/pos*: The variable is favorable to RI when large by having a lower threshold in positive chunks; *N/pos*: The variable is favorable to RI when small by having an upper threshold in positive chunks; *R/pos*: The variable has a favorable range for RI in positive chunks; *P/neg*: The variable is favorable to RI when large by having an upper threshold in negative chunks; *N/neg*: The variable is favorable to RI when small by having a lower threshold in negative chunks; and *R/neg*: The variable has a favorable range for RI in negative chunks (note that *R/neg* is 0 for all the predictors). Asterisks (*) are added to the counts greater than one-fourth of the total chunk counts, 18 for the positive and 55 for the negative chunks. The column “Sum (pos)” shows the summation of the weights of the positive chunks that include the variable. The column “Sum (neg)” is the same but for the negative chunks. In the last column, the overall predictor-value evaluations are classified into one of “pos” if favorable to RI when large, “neg” if favorable to RI when small, “range” if favorable to RI when having a moderate value, and “mixed” if the effects are positive or negative depending on chunks.

Acronym	<i>P/pos</i>	<i>R/pos</i>	<i>N/pos</i>	Sum (pos)	<i>P/neg</i>	<i>R/neg</i>	<i>N/neg</i>	Sum (neg)	Evaluation
COHC	7*/18	11*/18	0/18	3.388	31*/55	0/55	0/55	−4.833	Pos
DTL	1/18	0/18	0/18	0.137	19*/55	0/55	0/55	−1.547	Pos
VMAX	0/18	11*/18	0/18	2.360	0/55	0/55	0/55	0	Range
POT	0/18	7*/18	0/18	1.028	30*/55	0/55	0/55	−2.265	Range
INCV	5*/8	0/18	0/18	0.840	30*/55	0/55	0/55	−4.095	Pos
SDIR	0/18	0/18	11*/18	1.733	3/55	0/55	8/55	−1.348	Neg
SHRD	0/18	0/18	7*/18	1.942	0/55	0/55	6/55	−0.958	Neg
T150	0/18	0/18	2/18	0.296	0/55	0/55	8/55	−1.253	Neg
T200	0/18	1/18	0/18	0.731	0/55	0/55	16*/55	−1.022	Neg
T250	4/18	1/18	0/18	0.532	0/55	0/55	6/55	−0.605	Mixed
Z850	0/18	0/18	3/18	0.517	8/55	0/55	0/55	−1.163	Mixed
R000	0/18	0/18	1/18	0.049	0/55	0/55	0/55	0	Pos

classified as both the pos and range types. COHC is classified into the pos type despite a large *R/pos* count, since the upper limits there are large, so the conditions are not very different from being *P/pos*.

5. Case studies

In this section, we conduct case studies and show how one can utilize WRPS1 results for case diagnosis. We examine the results for 2021 based on the NF-type WRPS1 predictions trained by the 20-yr data. Figures 9–11 show selected cases, and the results for all the remaining named tropical storms and typhoons are shown in the supplemental material as Figs. S1–S16.

Figure 9 shows the result for Typhoon Surigae (2021) whose lifetime high VMAX is 170 kt, well above the category-5 threshold of 137 kt. It experienced consecutive 24-h RIs from 15 to 17 April 2021 as indicated by the red bullets underneath the intensity. The predicted RI start times (orange bullets) largely reproduce the actual RI start times, indicating the usefulness of WRPS1. The predicted RI starts with two false alarms at 0000 UTC 15 April and 0600 UTC 15 April, but all the succeeding actual RIs are predicted without further false alarms. The figure shows the contributions of individual predictors as a part of rhs of Eq. (1):

$$S_{\text{Vars}} \equiv \sum_{j \in \{\text{chunks including Vars}\}} \frac{w_j x_j}{m_j}, \quad (9)$$

where Vars can be a single variable such as COHC or can be a list of variables such as VMAX and POT and m_j is the

number of variables used in the chunk j ($1 \leq j \leq 4$). For example, the condition (chunk) having the highest weight (at the top of Table S2) consists of ranges of four variables, COHC, SHRD, T200, and VMAX, so the weight 0.731 divided by 4, 0.0183, is added to S_{COHC} , S_{T200} , S_{SHRD} , and $S_{\text{VMAX,POT}}$ (here, $S_{\text{VMAX,POT}} \equiv S_{\text{VMAX}} + S_{\text{POT}}$). As such, plotting S_{Vars} visualizes how the predictors jointly contribute to the prediction. The summation of Eq. (9) for all the predictors is equal to the summation in the rhs of Eq. (1).

Figure 9 suggests that, in the prolonged extreme RI of Surigae, all of locations (COHC), intensity or intensification potential (VMAX or POT), environmental wind shear (SHRD), upper-tropospheric temperature (T150, T200, or T250), and upper-level cloud distributions (SDIR) were favorable to RI. The previous 6-h intensity change (INCV) was slightly against RI in the early phase of the prolonged RI, and it became weakly favorable later. The continuation of many favorable factors over several days appears the source of the extreme intensification of Surigae. One may wonder why the impact of INCV can be negative even when the TC is intensifying. This is because there exist a number of negatively weighted chunks that require INCV to be not too large (Table S3), which indicates that some of the negative impacts associated with other predictors can be ineffective when the current intensification is at a high pace.

Figure 10 (Typhoon Conson) and Fig. 11 (Typhoon Rai) show similar cases in terms of tracks and environmental conditions. While RI was predicted for both typhoons, actual RI occurred only in the latter. RI was predicted four times for the former (Conson) not only because COHC and “VMAX

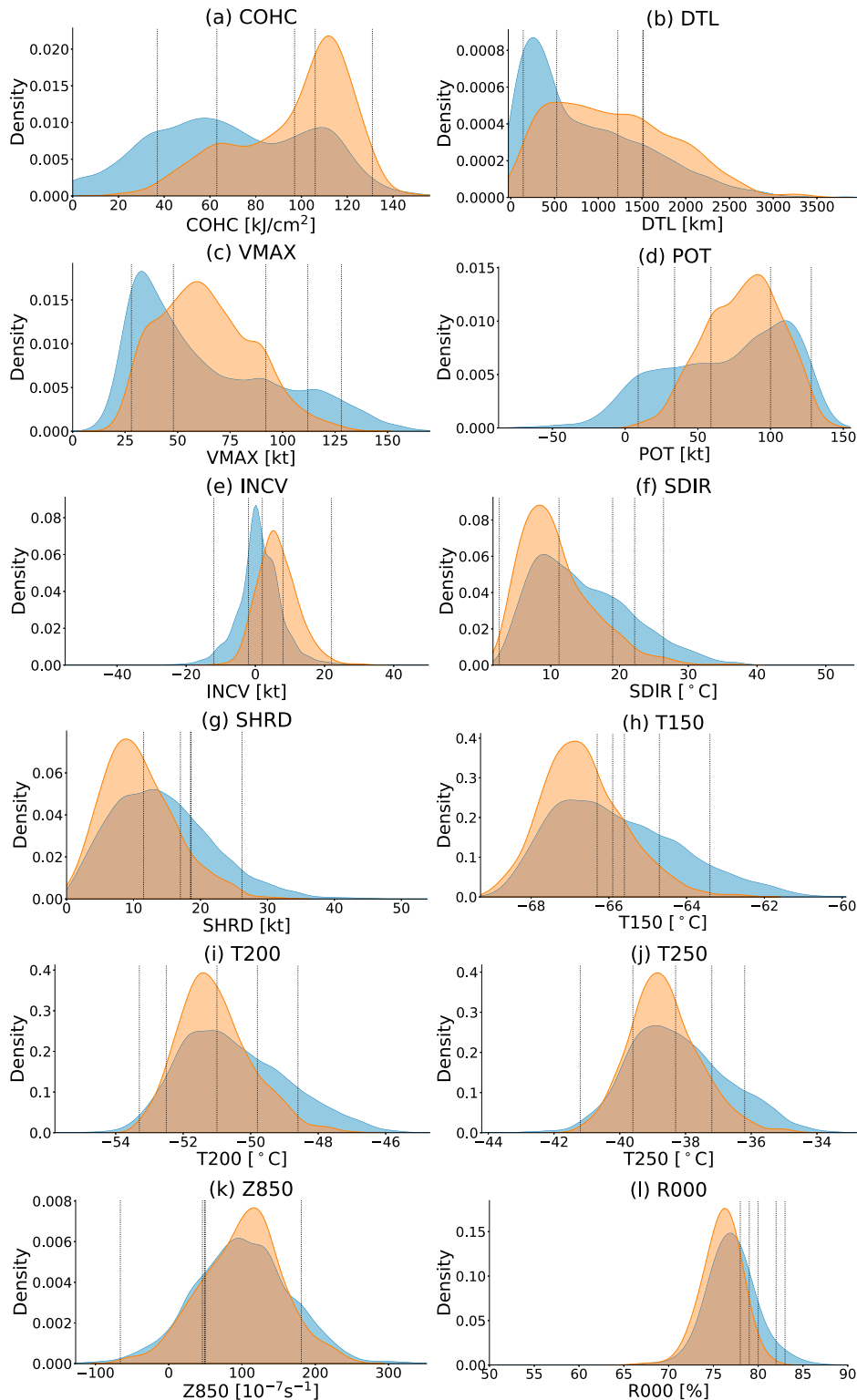


FIG. 7. Frequency distributions of the 12 predictors for each of RI (orange) and non-RI (blue) cases, which are smoothed by using the kernel density estimation (optimized except for INCV. See statistics textbooks for the standard way of kernel density estimation. Since INCV is coarsely discretized by 5 kt, a wider smoothing was used to avoid wiggles.) Vertical dotted lines are the discretization boundaries obtained for the NF-type 24-h prediction.

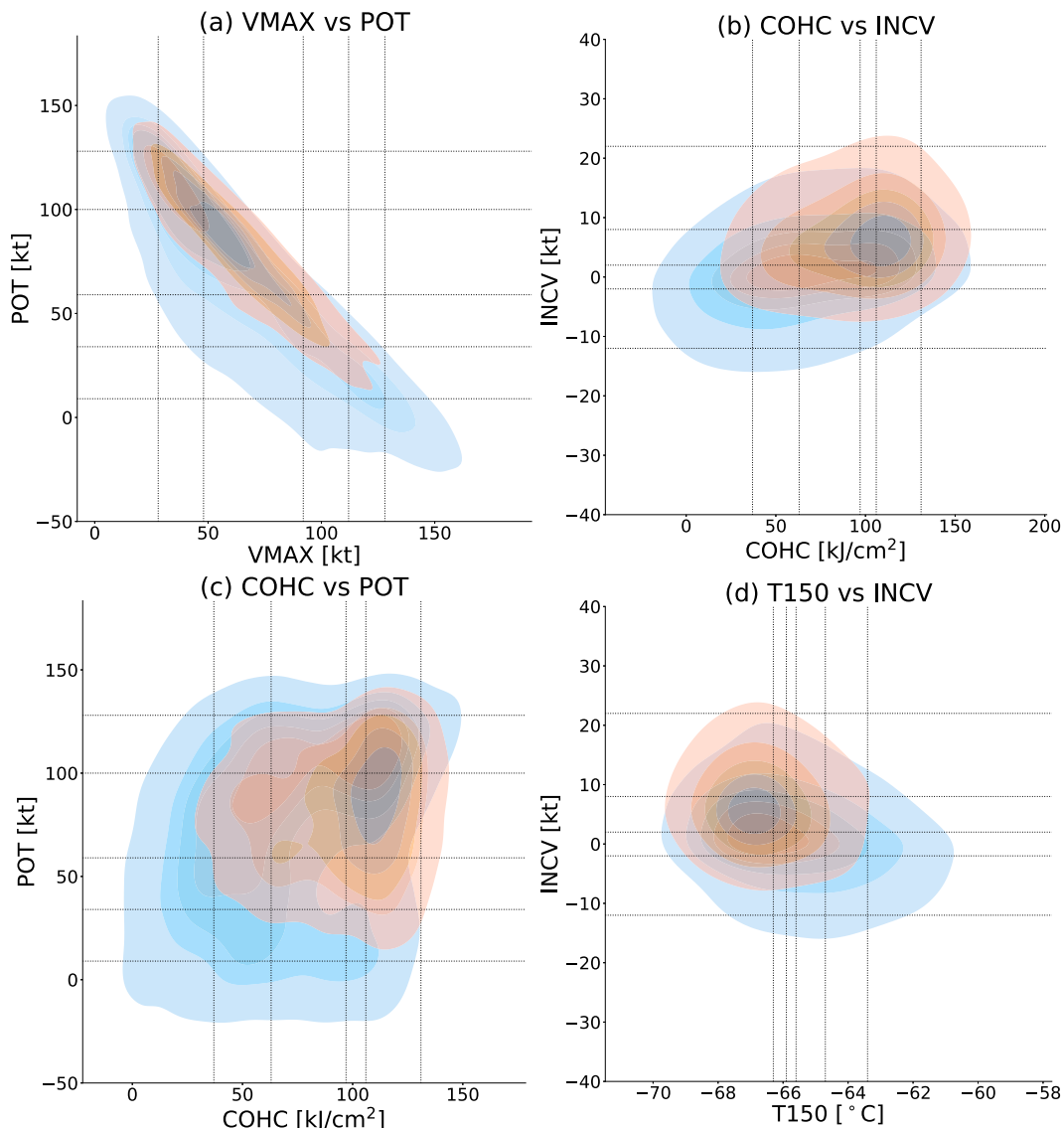


FIG. 8. Selected two-predictor joint distributions for each of RI (orange) and non-RI (blue) cases: (a) VMAX vs POT, (b) COHC vs INCV, (c) COHC vs POT, and (d) T150 vs INCV. Dotted lines are discretization thresholds for the NF-type $t = 0-24$ -h RI predictions. Smoothing is applied as in Fig. 7. Color shading is made to equally divide from 0 to the maximum into six ranges (every 16.6%), where the lowest range is not colored.

or POT” were favorable but also because shear was weak and the upper-tropospheric temperature was low (Fig. 10). The RI was predicted when the TC was close to Philippines, so the false alarm might be due to the detrimental effect of DTL is not sufficiently diagnosed. However, Fig. 11 shows that the second RI of Rai also occurred near Philippines, so finding appropriate conditions regarding DTL might not be an easy task. Typhoon Rai was born in December, so even though its track was slightly south of Conson, the climatological ocean conditions were more favorable to the RI of Conson (the predicted RIs of Conson occurred when COHC was in the third and fourth highest range, while the predicted RIs of Rai occurred solely when COHC was in the fourth highest range). We checked SST data for Conson and found that it was close

to the climatological value. Therefore, the false alarm to predict RI of Conson appears to suggest the necessity to consider factors that affect intensity changes other than the present predictors.

It is interesting that the two RIs of Rai were predicted though not perfectly (Fig. 11). The COHC-related score summation S_{COHC} in Fig. 11 drops sharply at the end of the first consecutive RI start times on 15 December. This drop is not by the fall of COHC, which remained in the fourth highest range (not shown). The fall rather occurred because VMAX increased to 120 kt, which is in the second highest VMAX range, and POT decreased to 30 kt, which is in the second lowest POT range. Thus, positive conditions, all of which include COHC (Table S2), are not met anymore. The prediction of

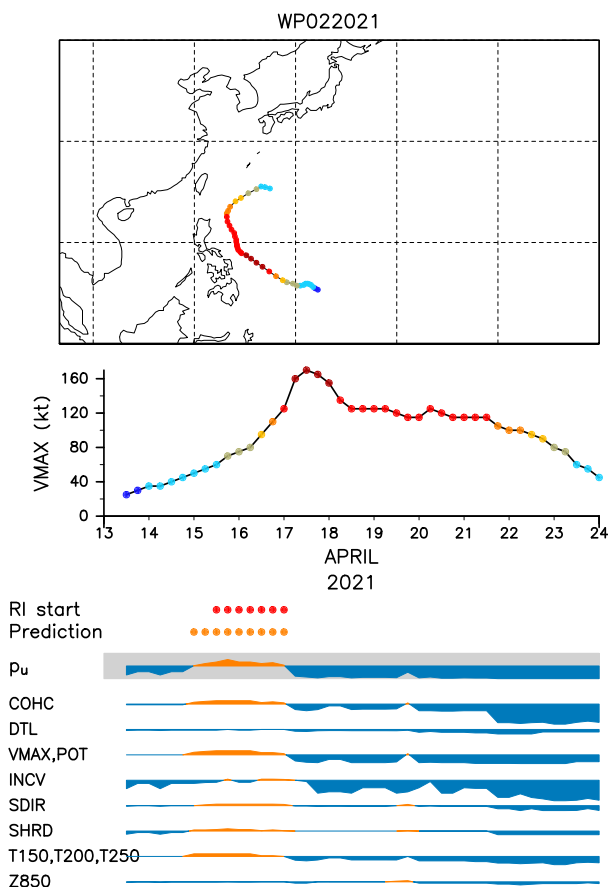


FIG. 9. The track, intensity, and RI predictions for Typhoon Surigae (2021). Intensities are colored with respect to the Saffir-Simpson scale. The start times of the actual RI over the following 24 h are shown by red bullets underneath the intensity plot, and the RI start times predicted by the NF-type WRPS1 are shown by orange bullets. The RI predictions are also shown by p_u in Eq. (1); its background gray shading shows the range from 0 to 1. The values greater than (lower than or equal to) 0.5 are indicated by the orange (blue) color. Also shown are contributions of the predictors: S_{COHC} , S_{DTL} , $S_{VMAX,POT}$, S_{INCV} , S_{SDIR} , S_{SHRD} , $S_{T150,T200,T250}$, and S_{Z850} . Here, S_{R000} is not shown because the chunk including R000 was not used at any time for this typhoon. Orange (blue) color shows positive (negative) values. The interval between the plots is 2.5, so, for example, both S_{COHC} and S_{INCV} were around -2 from 22 to 24 Apr.

the second RI was made possible by the drop of VMAX (and thereby the increase of POT) after the first intensity peak. In detail, the beginning of the second actual RI at 0600 UTC 17 December was not predicted because the intensity was decreasing, so INCV was negative. Note that the actual 24-h intensity change from 0600 UTC 17 December marginally satisfies the RI condition, and the 24-h change from 1200 UTC 17, to which the RI prediction was successful, is much greater.

Figures S1–S18 show the remaining cases in 2021. These figures indicate the overall adequacy of the WRPS1 predictions. It also provides further insight into the conditions of RI occurrence and how they can be visualized.

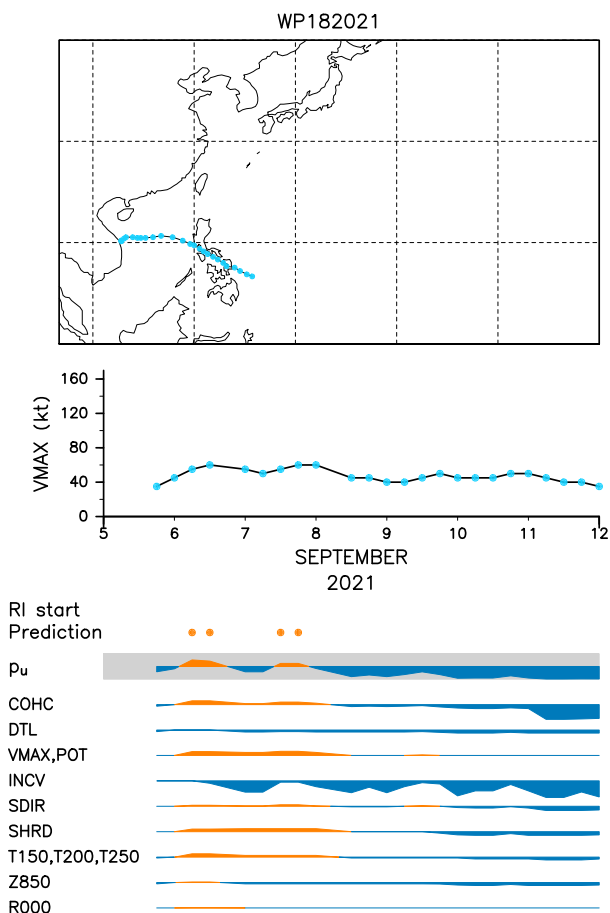


FIG. 10. As in Fig. 9, but for Typhoon Conson (2021).

6. Conclusions

We have developed a deterministic as well as probabilistic RI prediction scheme WRPS, version 1 (WRPS1), based on WL, an XAI. It uses only 12 predictor (input) variables that are available in the SHIPS developmental data for the western North Pacific. The 12 predictors consist of parameters on the climatological or environmental conditions such as environmental shear and the state of TCs such as intensity. The predictors are basically evaluated at the initial time of the prediction, $t = 0$ (NF-type prediction), but we also tested some of the environmental variables averaged from $t = 0$ to 24 h along TC tracks, as is done in many RI prediction schemes (PEF-type prediction).

Each of the WRPS1 predictors is discretized into six ranges to maximize information entropy in terms of distinguishing RI and non-RI cases. Then, WL is used to extract joint conditions called knowledge chunks, each of which consist of ranges of four or less predictors; the chunks are either favorable or detrimental to RI occurrence. Finally, LR is conducted to obtain a formula to evaluate RI probability. While this probability is directly used in the deterministic RI prediction, we have devised a method to adjust it by using a least square fitting to minimize the Brier score. This adjustment

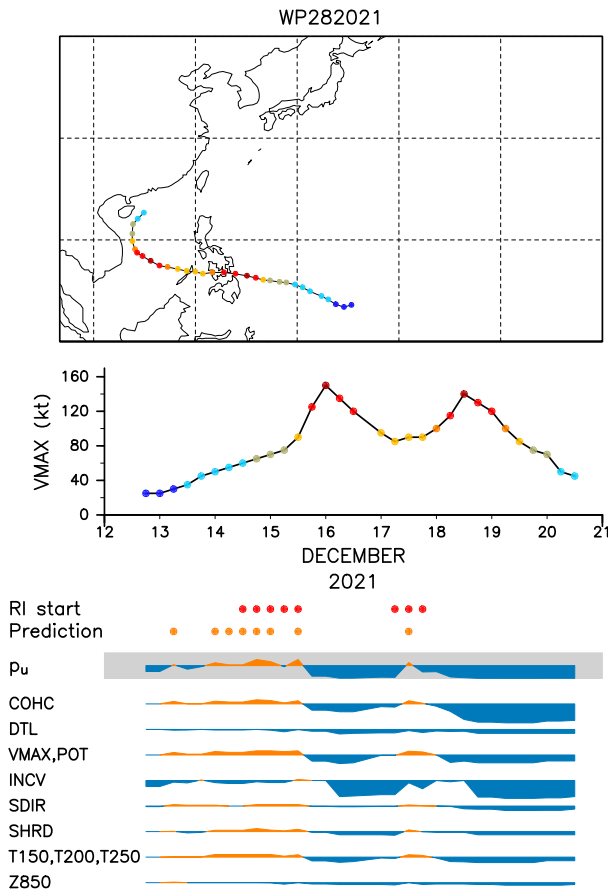


FIG. 11. As in Fig. 9, but for Typhoon Rai (2021).

method is general, so it is useful to improve other methods such as LDA and the simple LRs using raw predictors. The chunks of WRPS1 and the coefficients in Eq. (1) are fully provided in the supplemental material, so one can precisely reproduce WRPS1 predictions without using WL. Since the chunks are simple combinations of predictor ranges, one can grasp conditions favorable or detrimental to RI, which is how WL is explainable. These are the unique features of WRPS among ML-based schemes, which tend to be in black boxes. It was found that WRPS1 does not suffer much from overfitting.

The fact that the chunks consist of ranges of multiple variables enables WRPS1 to handle interdependence on predictors, and it also facilitates handling predictors favorable to RI at moderate values (range-type predictors). These features are not covered by MLR or the simple LRs. It was shown that the range-type predictors, VMAX and POT, are indeed important and that the separation into the positive and negative chunks facilitates their proper handling. Note that SHIPS deals with range-type predictors by including predictors that are the squares of some predictors. It was shown that SHIPS exhibits good performance in deterministic RI prediction if the RI ratio in the input data is increased.

In the case studies, we demonstrated how WRPS predictions can be analyzed to provide insights on RI occurrence. This is done by quantifying the impacts of individual predictors on joint

conditions favorable or detrimental to RI. Such a diagnosis can supplement implications from LDA or SHIPS. The case studies also demonstrated the overall adequacy of WRPS1.

Still, both F1 and BSS (deterministic and probabilistic prediction scores, respectively) of WRPS1 are much smaller than the perfect value of 1, so further improvement is wanted. That would likely require one to make use not only of further environmental conditions but also of internal conditions of TCs. The ability of WL to handle nonlinear dependence among predictors leaves room for such improvements, and its explainability could then contribute to enhance scientific understanding on the conditions to RI, so we envision further development.

The present study is based on the SHIPS developmental data, which includes postprocessed data such as the best track. However, from the overall robustness shown in this study, we expect that WRPS1 performs similarly when applied to real-time data.

Acknowledgments. This study was conducted under the cooperation of Yokohama National University (YNU) and Fujitsu Ltd. through the YNU-Fujitsu Small Research Laboratory. The authors declare no competing interests. We thank the comments by Dr. Mark DeMaria, Dr. John Knaff, and an anonymous reviewer that helped us improve the manuscript.

Data availability statement. The SHIPS developmental data are available from <https://rammb2.cira.colostate.edu/research/tropical-cyclones/ships/>; the dataset we used is the 5-day SHIPS predictor file for the western North Pacific available at https://rammb-data.cira.colostate.edu/ships/data/WP/lstdiagw_1990_2021_5day.txt. The data sufficient to reproduce WRPS1 predictions (and apply WRPS1 to future typhoons) are available in the paper's supplemental material.

APPENDIX

Relationship between BS-Minimizing and Histogram-Based Probability Mappings

We show here that the present probability mapping to minimize BS by using the least square fitting is closely related to K10's mapping method to derive RI probability. Equation (5) is easily rewritten as follows:

$$BS = \int_0^1 f(s) \{P(s)(s-1)^2 + [1-P(s)]s^2\} ds, \quad (A1)$$

where s is the score of the incidence to predict, which is to RI here; $f(s)$ is the number distribution of the cases whose integration is normalized to 1, i.e., the sample-based probability distribution function; and $P(s)$ is the predicted probability of the incidence as a function of the score.

Since $f(s)$ is based on samples, it is not continuous or smooth, so making a histogram along s is done customarily, as in Fig. 3. Binning Eq. (A1) yields the following approximation:

$$\begin{aligned} \text{BS} &\cong \sum_{j=1}^C f_j \{p_j(P_j - 1)^2 + (1 - p_j)P_j^2\}, \\ &= \sum_{j=1}^C f_j \{(P_j - p_j)^2 + p_j(1 - p_j)\}, \end{aligned} \quad (\text{A2})$$

where f_j is the total number of the cases in the bin $j = 1, 2, \dots, C$; P_j is the predicted probability of the incidence; and p_j is the actual fraction of the cases in which the incidence occurred. From Eq. (A2), BS is approximately minimized by setting

$$P_j = p_j. \quad (\text{A3})$$

K10's approach is to make four bins along s . Each of them includes the same number of RI cases. By using p_j , they defined $P(s)$ that is piecewise linear between the s boundaries, so it minimizes BS to the extent allowed by the coarse binning.

REFERENCES

- Chen, B.-F., Y.-T. Kuo, and T.-S. Huang, 2023: A deep learning ensemble approach for predicting tropical cyclone rapid intensification. *Atmos. Sci. Lett.*, **24**, e1151, <https://doi.org/10.1002/asl.1151>.
- Chaudhuri, S., D. Dutta, S. Goswami, and A. Middey, 2013: Intensity forecast of tropical cyclones over North Indian Ocean using multilayer perceptron model: Skill and performance verification. *Nat. Hazards*, **65**, 97–113, <https://doi.org/10.1007/s11069-012-0346-7>.
- Cloud, K. A., B. J. Reich, C. M. Rozoff, S. Alessandrini, W. E. Lewis, and L. Delle Monache, 2019: A feed forward neural network based on model output statistics for short-term hurricane intensity prediction. *Wea. Forecasting*, **34**, 985–997, <https://doi.org/10.1175/WAF-D-18-0173.1>.
- DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, [https://doi.org/10.1175/1520-0434\(1994\)009<0209:ASHIPS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2).
- , and —, 1999: An updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **14**, 326–337, [https://doi.org/10.1175/1520-0434\(1999\)014<0326:AUSHIP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0326:AUSHIP>2.0.CO;2).
- , M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543, <https://doi.org/10.1175/WAF862.1>.
- , 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>.
- Fayyad, U. M., and K. B. Irani, 1993: Multi-interval discretization of continuous-valued attributes for classification learning. *Proc. 13th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Chambéry, France, IJCAI, 1022–1029.
- Fudeyasu, H., K. Ito, and Y. Miyamoto, 2018: Characteristics of tropical cyclone rapid intensification over the western North Pacific. *J. Climate*, **31**, 8917–8930, <https://doi.org/10.1175/JCLI-D-17-0653.1>.
- Gao, S., W. Zhang, J. Liu, I.-I. Lin, L. S. Chiu, and K. Cao, 2016: Improvements in typhoon intensity change classification by incorporating an ocean coupling potential intensity index into decision trees. *Wea. Forecasting*, **31**, 95–106, <https://doi.org/10.1175/WAF-D-15-0062.1>.
- 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>.
- Holliday, C. R., and A. H. Thompson, 1979: Climatological characteristics of rapidly intensifying typhoons. *Mon. Wea. Rev.*, **107**, 1022–1034, [https://doi.org/10.1175/1520-0493\(1979\)107<1022:CCORIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1979)107<1022:CCORIT>2.0.CO;2).
- Iwashita, H., T. Takagi, H. Suzuki, K. Goto, K. Ohori, and H. Arimura, 2020: Efficient constrained pattern mining using dynamic item ordering for explainable classification. arXiv, 2004.08015v1, <https://doi.org/10.48550/arXiv.2004.08015>.
- 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>.
- Kim, S.-H., W. Lee, H.-W. Kang, and S. K. Kang, 2024: Predicting rapid intensification of tropical cyclones in the western North Pacific: A machine learning and net energy gain rate approach. *Front. Mar. Sci.*, **10**, 1296274, <https://doi.org/10.3389/fmars.2023.1296274>.
- Knaff, J. A., C. R. Sampson, and M. DeMaria, 2005: An operational statistical typhoon intensity prediction scheme for the western North Pacific. *Wea. Forecasting*, **20**, 688–699, <https://doi.org/10.1175/WAF863.1>.
- , —, 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>.
- , —, A. Brammer, and C. J. Slocum, 2023: A Rapid Intensification Deterministic Ensemble (RIDE) for the joint typhoon warning center's area of responsibility. *Wea. Forecasting*, **38**, 1229–1238, <https://doi.org/10.1175/WAF-D-23-0012.1>.
- Meng, F., Y. Yao, Z. Wang, S. Peng, D. Xu, and T. Song, 2023: Probabilistic forecasting of tropical cyclones intensity using machine learning model. *Environ. Res. Lett.*, **18**, 044042, <https://doi.org/10.1088/1748-9326/acc8eb>.
- Mercer, A., and A. Grimes, 2017: Atlantic tropical cyclone rapid intensification probabilistic forecasts from an ensemble of machine learning methods. *Proc. Comput. Sci.*, **114**, 333–340, <https://doi.org/10.1016/j.procs.2017.09.036>.
- Rozoff, C. M., and J. P. Kossin, 2011: New probabilistic forecast models for the prediction of tropical cyclone rapid intensification. *Wea. Forecasting*, **26**, 677–689, <https://doi.org/10.1175/WAF-D-10-05059.1>.
- Sampson, C. R., 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>.
- Shaiba, H., and M. Hahsler, 2016: Applying machine learning methods for predicting tropical cyclone rapid intensification events. *Res. J. Appl. Sci. Eng. Technol.*, **13**, 638–651, <https://doi.org/10.19026/rjaset.13.3050>.
- Sharma, N., M. M. Ali, J. A. Knaff, and P. Chand, 2013: A soft-computing cyclone intensity prediction scheme for the western North Pacific Ocean. *Atmos. Sci. Lett.*, **14**, 187–192, <https://doi.org/10.1002/asl2.438>.
- Shimada, U., 2024: Tropical cyclone intensity forecasting with three multiple linear regression models and random forest classification. *J. Meteor. Soc. Japan*, **102**, 555–573, <https://doi.org/10.2151/jmsj.2024-030>.
- Slocum, C. J., M. N. Razin, J. A. Knaff, and J. P. Stow, 2022: Does ERA5 mark a new era for resolving the tropical cyclone environment? *J. Climate*, **35**, 7147–7164, <https://doi.org/10.1175/JCLI-D-22-0127.1>.
- 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>.
- Wang, W., and Coauthors, 2023: A review of recent advances (2018–2021) on tropical cyclone intensity change from operational perspectives, part 2: Forecasts by operational centers. *Trop. Cyclone Res. Rev.*, **12**, 50–63, <https://doi.org/10.1016/j.tcr.2023.05.003>.
- Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Mon. Wea. Rev.*, **147**, 2261–2282, <https://doi.org/10.1175/MWR-D-18-0391.1>.
- 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>.
- Yamaguchi, M., H. Owada, U. Shimada, M. Sawada, T. Iriguchi, K. D. Musgrave, and M. DeMaria, 2018: Tropical cyclone intensity prediction in the western North Pacific basin using SHIPS and JMA/GSM. *SOLA*, **14**, 138–143, <https://doi.org/10.2151/sola.2018-024>.
- Zhang, Z., and Coauthors, 2023: A review of recent advances (2018–2021) on tropical cyclone intensity change from operational perspectives, part 1: Dynamical model guidance. *Trop. Cyclone Res. Rev.*, **12**, 30–49, <https://doi.org/10.1016/j.tcr.2023.05.004>.