Natural Hazards
https://doi.org/10.1007/s11069-023-06100-z

ORIGINAL PAPER

A new method for predicting hurricane rapid intensification based on co-occurring environmental parameters

Anushka Narayanan[1] · **Karthik Balaguru**[1] · **Wenwei Xu**[1] · **L. Ruby Leung**[1]

Received: 28 December 2022 / Accepted: 9 July 2023

Abstract
Tropical cyclones (TCs) that undergo Rapid Intensification (RI) can pose serious socio-economic threats and can potentially result in major damaging impacts along coastal areas. Considering the complexity of various physical mechanisms that play a role in RI and its relatively low probability of occurrence, predicting RI remains a major operational challenge. In this study, we propose a simple deterministic binary classification model based on the co-occurrence of environmental parameters (MCE) to predict an RI event. More specifically, the model determines the possibility of RI based on a simple count of the number of environmental predictors deemed favorable and unfavorable. We compare our model results to logistic regression (LR) and decision tree (DT) models, well-trained using the same set of environmental predictors. Results reveal that at an RI threshold of 30 kt, the MCE exhibits a critical success index score of 0.233 which is 14% higher than DT and LR model performances. When tested at multiple RI thresholds, the MCE displays relatively higher skill scores across multiple metrics. By simultaneously evaluating the favorability of predictors, the MCE is able to comparatively reduce the number of false alarms predicted when certain predictors are unfavorable toward RI. Interpreting these model results to gain a physical understanding of how co-occurring environmental parameters can affect RI, we highlight future directions for using models based on the MCE approach to understand and predict TC RI as well as other meteorological extremes.

Keywords Tropical cyclone rapid intensification · Co-occurring environmental parameters · Forecasting techniques · Model performance/evaluation

1 Introduction

The development and landfall of hurricanes or tropical cyclones (TCs) can potentially result in significant damages for coastal regions. The mechanisms behind TC intensification and weakening are complex, making TC intensification difficult to predict accurately. Even more so, TC rapid intensification (RI), defined as an instance when the storm's maximum sustained surface wind speed increases by 30 kt or more over a 24-hour period

✉ Anushka Narayanan
anushka0623@gmail.com

[1] Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, WA, USA

(Kaplan and DeMaria 2003), is substantially harder to predict given the low probability of occurrence. Historically, almost all TCs that have attained Category 4 or 5 strength underwent RI during their lifetimes (Kaplan and DeMaria 2003), further emphasizing the need to improve RI forecasts. For instance recently, Hurricane Ian underwent RI and made landfall over Florida peninsula as a strong and catastrophic Category 4 hurricane in September 2022 (Espinel et al. 2022). Similarly, in August 2021, Hurricane Ida underwent RI and struck New Orleans, Louisiana as a Category 4 storm and inflicted substantial damages (Zhu et al. 2022). In the operational RI forecasting space, a suite of dynamical models and statistical methods have been employed for RI prediction. With modest advancements in RI prediction over time, there lies considerable room for improvement (DeMaria et al. 2021) as predicting RI events remains a high priority for the National Hurricane Center (Rappaport et al. 2009), (Kaplan et al. 2010).

A multitude of challenges must be dealt with when predicting RI. First, there is uncertainty surrounding the underlying physical processes of a TC undergoing RI. Numerous studies (Kaplan et al. 2010), (Kaplan et al. 2015) have explored which select few environmental predictors favor RI. However, the relative significance of these environmental predictors for RI remains unclear. More confounding is that no one particular set of environmental conditions can guarantee an RI event (Rozoff et al. 2015). In addition, the availability of RI data is limited compared to non-RI data. Consequently, training of RI prediction models can be severely hampered by datasets. Common issues that RI prediction models face are poor probabilities of detection (POD) and high false alarm ratios (FAR) (DeMaria et al. 2021). Despite these challenges, numerous studies have addressed this problem to gain a better physical understanding of the phenomenon, and consequently, improve our ability to predict the possibility of RI.

One of the first notable statistical models developed for RI forecast was the Statistical Hurricane Intensity Prediction Scheme (SHIPS) which used multiple linear regression techniques to predict intensity changes for Atlantic TCs (DeMaria and Kaplan 1994, 1999; DeMaria et al. 2005). Kaplan and DeMaria (2003) extended this to implement an RI-specific, threshold-based probabilistic prediction scheme called SHIPS - Rapid Intensification Index (SHIPS-RII). This was later expanded upon by Kaplan et al. (2010) to include the North Pacific basin and different RI thresholds (25, 30 and 35 kt). To improve RI prediction skill, Kaplan et al. (2010) designed the SHIPS-Linear Discriminant Analysis model (SHIPS-LDA) that performed an LDA (Daniel and Wilks 2006) to obtain individual weights for each environmental predictor based on their contribution to RI. The scaled predictor values with individual weights were summed up to obtain a probabilistic prediction of RI. More recently, DeMaria et al. (2021) developed a model that has been in NHC's operational use since 2018. The Deterministic to Probabilistic Statistical model (DTOPS) converts deterministic intensity forecasts from SHIPS, statistical-dynamical, regional-dynamical and consensus models into probabilistic RI forecasts using a binary logistic regression. DTOPS demonstrates improved RI forecast skill. A more detailed report on the history and the comparative performance of the various NHC operational RI forecasting models discussed above is presented in DeMaria et al. (2021).

Machine learning (ML) techniques are increasingly being used for TC intensity prediction, including RI, with the aim of capturing the nonlinear relationships between environmental predictors and storm behavior. Initial ML efforts started with Rozoff and Kossin (2011) that introduced a logistic regression (SHIPS-LR) and a naive Bayesian model (SHIPS-NB). These models are used with the SHIPS-LDA to create a consensus SHIPS model (SHIPS-C) (Kaplan et al. 2015) that is used by the NHC in an operational setting. SHIPS-LR employs the logistic regression technique which is commonly

used for binary predictands. It assigns a regression weight to each individual predictor which is summed up with predictor values to output a probabilistic prediction of RI. The Naive Bayesian equation (Kossin and Sitkowski 2009) in SHIPS-NB uses Bayes theorem to predict the conditional probability of an RI event. There have been many studies that investigate the intensification of TCs using association rule algorithms (Yang et al. 2007, 2008, 2011), data mining techniques (Yang 2016) classification and regression trees (Wei and Yang 2021), decision trees (Zhang et al. 2013), (Kim et al. 2021), long short-term memory models (Yang et al. 2020), multi-layer perceptron models (Xu et al. 2021) and ML ensembles (Mercer and Grimes 2017; Su et al. 2020) that show skill in predicting RI.

Across the multitude of RI prediction approaches, there is little work that explores particularly how environmental parameters can concurrently contribute to a potential RI event and how that can be leveraged into an RI prediction model. Most models evaluate RI probability using methods which tend to be influenced by a limited set of strong predictors rather than looking at whether the large-scale environment is comprehensively conducive across the board. Herman and Schumacher (2018) explored the advantages and disadvantages of logistic regression (LR) models for forecasting extreme precipitation, which has a low probability of occurrence similar to RI. They suggest that a shortcoming that LR faces is that predictor regression coefficients are applied uniformly across data samples. For example in models like SHIPS-LR and SHIPS-LDA, if certain predictors independently cannot largely influence RI, but in conjunction with other predictors can lead to high chance of RI, a trained model would still assign weaker coefficients to these predictors and stronger weights to the other predictors (Herman and Schumacher 2018). This could lead to cases where a select few highly weighted predictors strongly influence the model's RI prediction, causing the model to overlook other predictors that might not favor RI, leading to a potential false alarm.

In this study, we propose a simple deterministic Model based on the co-occurrence of environmental parameters (MCE) to predict an RI event. We compare our model results to logistic regression and decision tree based approaches and interpret these models' results to explore the potential dynamics of how simultaneously co-occurring environmental parameters can affect a possible RI event. A logistic regression model was chosen for comparison, considering its use in the SHIPS-Consensus model and its contrasting method that does not explicitly evaluate predictors concurrently to predict RI. The decision tree model was also used for comparison to explore how environmental predictors are evaluated in the hierarchical structure of the model's decision rules. (Zhang et al. 2013).

In this study, we have chosen to focus on using environmental predictors as model inputs. We acknowledge the vital role played by TC internal processes in the development of RI; however, in this study, we chose 5 environmental predictors to investigate how 3 different methods can predict RI and the roles that co-occurring favorable environmental conditions can have on RI. The main objectives of the study are as follows:

1. To demonstrate the utility of the MCE for predicting RI events using co-occurring environmental parameters.
2. Interpreting the MCE results and the physical implications of co-occurring environmental parameters for TCs undergoing RI.
3. To compare the performance of three different RI prediction methods given the same input environmental parameters

2 Data

2.1 SHIPS dataset

In this study, environmental predictors from the Statistical Hurricane Intensity Prediction Scheme (SHIPS) database are used. The SHIPS database records the environmental conditions experienced by a TC in 6-hr timesteps from -12hr to 120hr relative to the current position. Data are obtained from model operational analyses and from satellite observations. For the environmental predictors chosen for this study, all fields are relative to the TC storm center determined by the NHC Best Track and only those data points where a storm is at least of Tropical Storm (maximum sustained surface winds above 34 kt) strength are considered. The version of the SHIPS database that was used for the study at the time contained data from 1982 to 2020 for the Atlantic, Eastern Pacific and Central Pacific basins. The North Indian Ocean and Western North Pacific basins contained data from 1990 to 2020, while the Southern Hemisphere contained data from 1998 to 2020.

Most RI prediction studies (for e.g., (Kaplan et al. 2010), (Mercer and Grimes 2017), (Yang 2016)) were conducted at the basin scale; however, this study's main aim is to look at how considering the co-occurrence of environmental parameters may improve RI prediction at the global scale. Further, this can lead to a more general understanding of the underlying physical mechanisms of TC RI. In addition, inclusion of all the SHIPS basins increases the amount of available training and testing data for a more robust and informed model.

2.2 Environmental predictor selection

In this study, we focus our efforts on 5 environmental predictors from the SHIPS global dataset that were used in Kaplan et al. (2015) for the revised SHIPS-RII: Potential Intensity (POT), Vertical Wind Shear (SHRD), Relative Humidity at 700 hPa (RHLO), Divergence at 200 hPa (D200) and Ocean heat Content (OHC). Kaplan et al. (2015) showed that these variables exhibited statistically significant differences at the 99.9% level using a two-sided Behrens-Fisher t-test (Dowdy and Wearden 1991) between the RI and non-RI data samples for RI thresholds of 25, 30 and 35 kt. Higher values of POT, RHLO, D200 and OHC positively correlate with higher chance of RI, whereas lower values of SHRD favor RI (Kaplan et al. 2015). We limit our predictor selection for simplicity and ease of understanding. However, the technique presented here could be extended to include different and larger selections of environmental predictors. The average RI and non-RI values of the environmental predictor data used in this study are shown in Fig. 1.

A table outlining the specific SHIPS predictors used to derive the environmental predictors and whether they are taken at t = 0 of the storm or a temporal average from $t = 0$ to $t = 24$ h is shown here in Table 1. POT is the potential intensity calculated following the method described in Kaplan and DeMaria (2003) by subtracting the intensity of the current storm (VMAX), from the average of the maximum potential intensity (VMPI) from $t = 0$ to $t = 24$. OHC is derived from the NCODA analysis (denoted as NOHC in the SHIPS database). Since at the time of the study, NOHC was not available for the Western North Pacific, Northern Indian and Southern Ocean basins, the RHCN derived from satellite altimetry data is used in place of NOHC for these basins. If RHCN data are missing, PHCN, which is the estimated ocean heat content from climatology and the current SST anomaly, is designed to fill in for RHCN as per the SHIPS documentation.

Table 1 Predictors used in this study and the corresponding SHIPS predictors used to derive them

Predictors	Definition	Unit	SHIPS Predictors
POT	Potential intensity	kt	VMPI, VMAX (time avg)
SHRD	Vertical shear of horizontal wind between 850 and 200 hPa	kt	SHDC (time avg)
RHLO	Relative humidity between 850 and 700 hPa	%	RHLO (initial time)
D200	200 hPa divergence averaged from 0 to 0 km from the storm center	$10^{-7}\,s^{-1}$	D200 (time avg)
OHC	Ocean heat content relative to 26 isotherm	$kJ\,cm^{-2}$	NOHC, RHCN, PHCN (time avg)

Time avg indicates a temporal average of $t = 0$ to $t = 24$ of the predictor was taken. Initial time indicates the predictor value was taken at $t = 0$

Any non-existent variables for the predictors are removed from the dataset. All overland TC locations are removed from the analysis. Further, those TC locations over water that had a landfall in the − 12 to 24 h time frame are also removed to ensure that our results are not contaminated by land effects.

2.3 Data pre-processing

A scaling method similar to that in Kaplan et al. (2010) is used to scale each predictor's values from 0 to 1. In this study, a scaled value of 0.0 is assigned to the dataset's minimum (maximum) value of POT, RHLO, D200, OHC (SHRD) and a scaled value of 1.0 is assigned to the dataset's maximum (minimum) value of POT, RHLO, D200, OHC (SHRD). Values in between are interpolated linearly. In the resulting dataset, the closer the predictor's values are to 1.0, the more favorable the value is for RI.

We approach RI prediction in these models as a binary classification problem. To create the predictand dataset, each data sample is categorized as an RI event (1) or non-RI event (0) when the DELV variable from the SHIPS dataset, which represents the intensity change (kt) in maximum sustained surface wind speed from t=0 to t=24 h, is above a set RI threshold of 30 kt. The resulting predictor and predictand datasets are split into training and testing datasets for model optimization and evaluation, respectively. The data samples from the last 5 years (2016–2020) from all basins are reserved so the models can be tested on unseen data. The remaining data samples from 1982 to 2015 are arranged chronologically and are used for model training purposes. It is important to note for the North Indian Ocean and Western Pacific the training data starts at 1990 and for the Southern Hemisphere it starts at 1998. In the training dataset, there were 1244 RI cases and 14,877 non-RI cases. The testing dataset had 502 RI cases and 5617 non-RI cases.

3 Methods

3.1 Model based on co-occurring environmental parameters (MCE)

The MCE method is intentionally kept simple to focus on the simultaneously co-occurring environmental parameters and gain insight into how large-scale environment's

conduciveness can affect RI. To develop our model, the training dataset (1982–2015) is used exclusively and we reserve the testing dataset (2016–2020) to test model performance. For each environmental predictor, we use a threshold similar to Kaplan and DeMaria (2003) to determine whether the value of the predictor is favorable and a separate threshold to determine whether the predictor is unfavorable toward RI.

We detail the method used in finding the optimal thresholds employed in the MCE model in our study. The threshold is defined as the standard deviation of the predictors' RI values multiplied by a particular multiplier subtracted from the average of the predictors' RI (NRI) values for the favorable (unfavorable) threshold. A grid search is employed over different multiplier values for each environmental predictor. A grid search method was chosen to ensure an exhaustive approach to the numerous combinations of available thresholds for all 5 predictors. The grid search is conducted in 2 parts over the different multiples of each predictor's standard deviation. The first grid search looks through multiplier values of 0 through 2.5 in increments of 0.5 for each predictor. A second narrower grid search is conducted in increments of 0.1 of multiplier values centered around the value that was produced from the first grid search. The multipliers for each environmental predictor for the top 10% best performing models are averaged to avoid possibly any extreme multiplier values. Once the optimal multiplier values are found, the threshold is defined as the standard deviation of the predictors' RI values multiplied by the optimal multiplier value found in the grid search subtracted from the average of the predictors' RI (NRI) values for the favorable (unfavorable) threshold. The two-part grid search is conducted to find the favorable thresholds. A second two-part grid search is conducted to find the unfavorable thresholds. The grid search ensures the unfavorable thresholds for the environmental predictors always remain lower than the favorable threshold values. The threshold values for each environmental predictor are outlined in Table 2.

The grid search is conducted using the predictor training dataset. Each model from the grid search is evaluated against the predictand training dataset and uses the critical success index (CSI) metric to determine the best performing models. CSI (Roebber 2009), also known as Threat Score, was chosen to evaluate the models due to its wide use in operational forecasting of severe weather events (Yang et al. 2020; Tam et al. 2021). Higher CSI scores indicate better performing models. Further, in Doswell et al. (1990) and Roebber (2009)'s analyses of forecasting metrics, they underline the suitability of CSI for forecasting rare events since the metric disregards skill from true negative (TN) predictions. This is when the model correctly predicts a non-event, which in our case is a non-RI event. In rare event forecasting, since non-events are more abundant, skill can be overinflated when metrics account for TNs, so CSI is chosen as for the grid search metric to avoid any such skill inflation.

For each predictor that is favorable, the net favorable predictor count increases by 1. If a predictor is deemed as unfavorable, the net favorable predictor count is reduced by 1. If a predictor is neither favorable nor unfavorable, the count is not affected. This allows the MCE to account for potentially non-conducive predictors and use only the net favorable environmental predictors when predicting RI. Finally, if the net favorable predictor count exceeds a certain count threshold, the model predicts an RI event. Details on this classification method are given in Fig. 2.

3.2 Logistic regression classifier (LR)

A logistic regression classifier (Daniel and Wilks 2006) was developed alongside the MCE to compare RI prediction performance. Logistic regression is used in the SHIPS-RII consensus model for NHC operational use. Since the logistic regression model does not explicitly depend on co-occurring environmental parameters, a comparison of the LR model with MCE will provide useful insight into the physical nature of RI. The logistic regression model is fit on the training dataset and produces a RI probability value. Depending on whether the probabilistic output exceeds a pre-determined probability threshold, the LR model predicts an RI event. To determine the best performing model, the GridSearchCV method with Leave-One-Year-Out (LOYO) cross validation evaluated on critical success index (CSI) is used for tuning the model hyperparameters and finding the optimal probability threshold. More information on CSI is given in Table 3. Details regarding the specific parameter search are included in the Supplementary Information.

3.3 Decision tree classifier (DT)

A decision tree (DT) binary classifier was developed to compare RI prediction performance with the MCE models. The tree-like structure of the decision-making process evaluates the conduciveness of the environment by hierarchically checking whether environmental predictors meet certain thresholds set by a trained decision tree model. A key difference between the MCE and DT is that in the hierarchical tree like structure, the DT uses certain predictors more often than other predictors at each node classification depending on each predictor's feature importance. In this instance, the model may be more influenced by certain predictors when predicting RI. On the other hand, the MCE considers all predictors equally during classification. The trained DT model outputs an RI probability value depending on whether this exceeds a pre-determined probability threshold. Similar to the LR, the GridSearchCV with LOYO cross validation method evaluated on CSI is used to determine the best performing model. Additional details regarding the parameter space and grid search results are given in the Supplementary Information.

3.4 Model evaluation metrics

The purpose of using multiple evaluation metrics to test the various models is to obtain a comprehensive overview of model performance and to pinpoint how each model approaches the RI prediction problem. In binary forecasts where models predict an event or nonevent for each data sample, evaluation metrics largely comprise of elements from a 2×2 contingency table that compare observations to model forecasts. The table captures

Table 2 Environmental predictors' favorable and unfavorable threshold non-scaled values used by MCE

Predictor	Unfavorable threshold	Favorable threshold
POT	37 kt	47 kt
SHRD	8 kt	5 kt
RHLO	62 %	65 %
D200	$-53\,10^{-7}\,s^{-1}$	$61\,10^{-7}\,s^{-1}$
OHC	6 kJ cm^{-2}	55 kJ cm^{-2}

the number of true positives (TP) where the model forecast RI and RI was observed, false positives (FP) where the model forecast RI and RI was not observed, false negatives (FN) where the model did not forecast RI and RI was observed and lastly, true negatives (TN) where the model did not forecast RI and RI was not observed. The common forecast evaluation metrics used in this study that are derived from these elements are described in table 3.

4 Results

4.1 Model performance analysis

We analyze the MCE's performance in predicting RI using the optimized thresholds found using the grid search methods outlined above. In, Fig. 3, we first look at a case where the MCE predicts RI using a single set of favorable thresholds, to determine if the environmental predictor is favorable toward RI. In this case, we set a count threshold of 3 indicating that if a data sample has at least 3 simultaneously favorable environmental predictors, the MCE predicts RI. We use the resulting best MCE model derived from the grid search to determine optimal set of only favorable thresholds for a count threshold of 3. The results show that the MCE exhibits a CSI score of 0.17 with a POD of around 0.56 and a high FAR around 0.81.

In comparison, for the same model, if we set a count threshold of 4 favorable predictors in order to predict RI, the MCE exhibits an increase in CSI score to around 0.21. In this case, we derived a new set of favorable thresholds using the grid search and a count threshold of 4 to produce an RI event. These favorable thresholds are found in column 1 in Table 2. By ensuring a larger portion of the overall environment is favorable, the MCE demonstrates higher CSI attributed to a marked improvement in FAR outweighing the milder decline in POD. However, when we compare to the LR and DT models, the MCE's CSI scores are only comparable. The models' FAR scores are still large which is a common issue in RI prediction models.

We introduce a second set of unfavorable thresholds into the MCE using the grid search method detailed above, specifically meant to determine if environmental predictors are unfavorable toward RI which can be found in column 2 in Table 2. By doing so, we see a further improvement of around 10% in the MCE's CSI score over LR and DT models seen in Fig. 3. Given non-RI events are abundant in our dataset, the CSI metric ignores the models' easily correct non-RI predictions which can tend to over-inflate other metrics of skill that account for these TNs (Daniel and Wilks 2006), thus giving a truer picture of model skill in predicting RI.

The number of unfavorable predictors negatively impact the net favorable predictor count, as described in the MCE methods section. In this instance, keeping the count threshold at 4 net favorable predictors ensures that there are only 2 scenarios in which the MCE predicts an RI event. This is when either 4 or 5 environmental predictors are favorable toward RI. However, an RI event is not predicted in cases where 4 environmental predictors are favorable and 1 environmental predictor is unfavorable, since the unfavorable predictor negatively impacts the net count of favorable predictors to reach the set threshold of 4. By ensuring the MCE does not predict RI when one or more unfavorable predictors are present, there is a 5% reduction in FAR. The MCE performs best when both sets of favorable and unfavorable thresholds are included to evaluate the data samples and ensuring 4 net

Table 3 Metrics used in study, formula derivations and definitions

Metric	Formula	Definition
Probability of detection (POD)	$\frac{tp}{tp+fn}$	Fraction of the observed RI events that were correctly forecast
False Alarm ratio (FAR)	$\frac{fp}{fp+tp}$	Fraction of the predicted RI events that actually did not occur
Critical success index (CSI)	$\frac{tp}{tp+fn+fp} = \frac{1}{\frac{1}{1-FAR}+\frac{1}{POD}-1}$	How well did the forecast RI events correspond to the observed RI events?
Peirce skill score (PSS)	$\frac{tp}{tp+fn} - \frac{fp}{tn+fp}$	How well did the forecast separate the RI events from the non-RI events?
F-1 score	$\frac{tp}{tp+0.5(fn+fp)}$	Harmonic mean of precision and recall
Gilbert skill score (GSS)	$\frac{tp-tp_{random}}{tp+fn+fp-tp_{random}}$, $tp_{random} = \frac{(tp+fn)(tp+fp)}{tp+fn+fp+tn}$	How well did the forecast RI events correspond to the observed non-RI events (accounting for hits due to chance)?

favorable predictors are necessary to predict RI. The MCE exhibits a CSI score of 0.233, 14% higher than LR and DT models as well as a higher POD score and lower FAR score. Given that the MCE with both sets of favorable and unfavorable thresholds with a count threshold of 4 produces the best performance, we continue to use these sets of thresholds in the following experiments in this study.

4.2 Model performance basin-wise

We compare model performance basin-wise in Fig. 4 for the Atlantic, Eastern Pacific, Western Pacific and Southern Hemisphere basins. Given the low sample testing size for the Indian and Central Pacific basins, we do not include those results here. In comparing model performance, we see that the MCE consistently outperforms LR and DT models in the Eastern Pacific basin. In the Atlantic and Southern Hemisphere basins, we find the MCE's performance is similar to that of LR and DT models. In the Western Pacific, the MCE outperforms the DT model but does not perform as well as the LR model. This is broadly consistent with the results presented in Bhatia et al. (2022) where the authors, using a similar method that incorporate critical predictor thresholds for RI probability, show that for 4 fulfilled RI thresholds, the probability of RI is highest for the Eastern Pacific basin and lowest for the Atlantic and Southern Hemisphere basins. This can potentially indicate that TCs in the Eastern Pacific could be more dependent on multiple external environmental conditions to be favorable in order for an RI event to take place. In addition, different environmental conditions can play varying roles in TC intensification for different basins as shown in Foltz et al. (2018), where the authors explored how the role of SST in different basins affect hurricane intensification. Further in-depth analysis must be conducted to explore how co-occurring environmental conditions varies across basins which we have reserved for future work.

4.3 A comparison of model 2 × 2 contingency scores

A summary of the models' 2 × 2 contingency table scores is shown in Fig. 5.

In the overall testing dataset, around 9% of the cases were RI events. When we look at the performance of various models, we find the MCE has higher TPs and has lower FNs compared to LR and DT. The MCE also shows significantly lower FPs and higher TNs than LR though not as many TNs and not as few FPs compared to DT.

4.4 Model sensitivity to multiple RI thresholds

We test the sensitivity of the MCE through added testing at multiple RI severity thresholds in addition to the 30 kt RI threshold. In this case, we use the same favorable and unfavorable thresholds outlined in Table 2. From the reserved testing dataset, the models were tested on samples that showed RI at thresholds of 25 kt, 35 kt, and 40 kt. There are 718 (338, 230) RI cases and 5401 (5781, 5889) non-RI cases in the additional testing datasets based on a 25 (35, 40) kt RI threshold.

Model performance results across RI thresholds are shown in Fig. 6. We see a common trend in model performance as we increase the RI thresholds. As we narrow down the testing data to see how the models perform for more severe cases of RI, the POD increases, FAR increases and overall CSI decreases. This could be attributed to the testing RI sample

size decreasing at higher RI thresholds. In addition, cases of more severe RI tend to have even lower probabilities of occurrence and usually require highly favorable environmental predictors. It is likely that most models predict RI when predictors are highly favorable in these severe RI circumstances. This can explain the increase in POD and FAR across the three models. Interestingly though, across the increasingly severe RI cases, the MCE consistently shows higher CSI scores driven by a lower false alarm ratio compared to LR and DT models at the same RI thresholds. This indicates that for higher RI cases, in comparison to models like the LR and DT that are not explicitly dependent on co-occurring parameters, the MCE shows comparatively improved skill in RI prediction.

Results from further analysis into model performance using the aforementioned evaluation metrics are shown in Fig. 7. The MCE consistently shows higher PSS, F1 and GSS scores and lower FAR scores compared to LR and DT models. This indicates that in spite of the MCE's simple threshold-based decision-making, the model consistently outperforms across multiple metrics of skill.

In Fig. 7a, the MCE exhibits higher POD scores than the DT model across RI thresholds but in comparison to LR, the MCE has POD scores only on par. Interestingly, despite similar POD scores between MCE and LR models, the MCE exhibited higher PSS scores as shown in Fig. 7c. Since PSS can also be defined as the difference between the POD and probability of false detection (POFD) (Daniel and Wilks 2006), this indicates the MCE had lower rates of POFD compared to LR. In other words, in comparison to LR across RI thresholds, the MCE had a lower rate of observed non-RI events being incorrectly forecast as RI events. PSS differs from the FAR metric which evaluates the fraction of predicted RI events that were actually observed non-RI events.

4.5 Feature analysis

An analysis of the environmental features for each model, and their relative significance, is discussed below. The feature importance scores determined by the DT model and the linearly scaled feature weight scores determined by the LR model are in Table 4. They show that certain predictors play a relatively more important role in RI prediction. For example, the DT model places high importance on POT and SHRD when evaluating a data sample, indicating POT and SHRD are considered more often in the decision rules set by the model. Because a larger percentage of the DT decisions are determined by the POT and SHRD values, they hold larger influence when the model predicts RI.

Similarly, in the LR model, SHRD has a considerably larger weight compared to the other predictors. This indicates that with lower values of SHRD, the odds ratio for RI is increased by a larger magnitude than the other environmental predictors. Hence an increased SHRD predictor value would affect the resulting RI probability value returned by the LR model by a larger percentage. The LR model may attribute a higher RI probability to a non-RI data sample despite the data sample having unfavorable values of the other predictors.

Comparatively, the principle of the MCE is to consider each predictor simultaneously at equal proportions to evaluate the environment's favorability toward RI. We can understand the relative significance of each of the predictors in the MCE's predictions in Fig. 8. In Fig. 8 panel (a), we see of the MCE's true positive (TP) predictions, the percentage of the time that each predictor was favorable, neutral or unfavorable. Since, the MCE does not predict RI in cases where one or more predictor is unfavorable, we can see that no predictors are unfavorable for the TP event. In Panel (b), we show a

Table 4 Decision tree feature importance scores and linearly scaled logistic regression feature regression coefficient weights determined by model training

Predictor	Decision tree feature score	Logistic regression feature weight
POT	0.354	0.180
SHRD	0.335	0.578
D200	0.093	0.083
RHLO	0.022	0.040
OHC	0.197	0.119

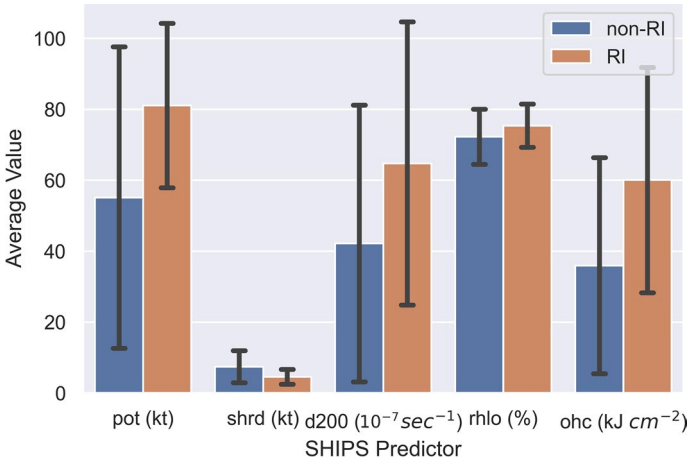

Fig. 1 Bar chart of average values of environmental predictors RI samples (orange) and non-RI samples (blue) with error bars depicting the standard deviation

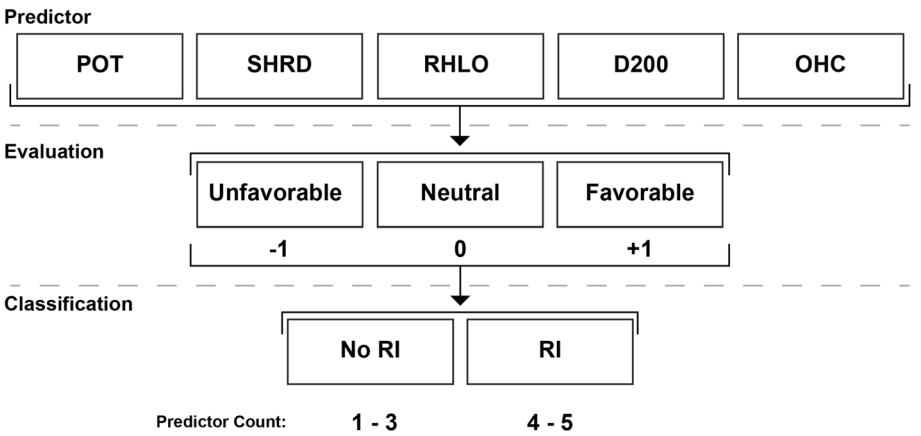

Fig. 2 Flowchart depicting the workflow of the MCE to determine RI event. For each predictor, dependent on where its value falls relative to the specified thresholds, the predictor is classified as favorable, neutral or unfavorable. In each of these cases, they can add to, not affect or subtract from the net favorable predictor count. The net favorable predictor count ultimately decides how the model predicts RI dependent on whether the count exceeds a set number of co-occurring environmental parameters

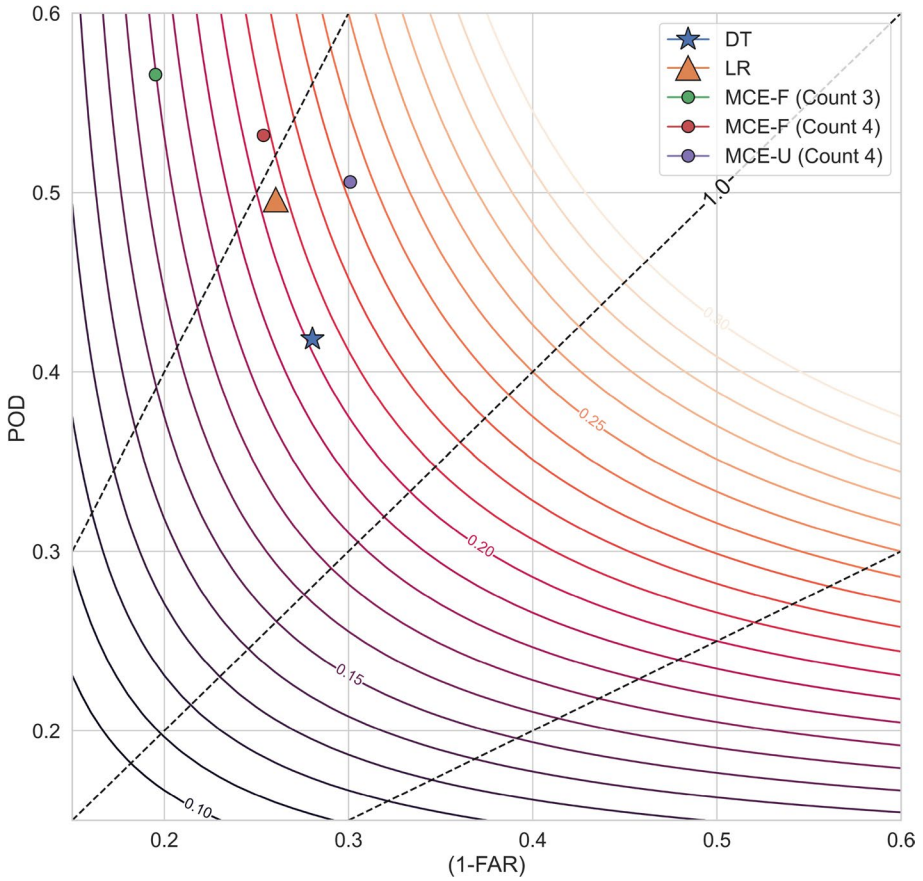

Fig. 3 Performance diagram summarizing multiple performance metrics of decision tree (star), logistic regression (triangle), and MCE (circle). Diagram compares the MCE-F using a count threshold of 3 favorable predictors (green), MCE-F using a count threshold of 4 favorable predictors (red) and MCE-U that utilized additional unfavorable thresholds (purple). Models were tested using an RI threshold of 30 *kt*. *x* axis shows Success Ratio = (1-FAR). POD on *y* axis. Contour lines show CSI scores. Dotted diagonal line represents bias scores

similar plot of the percentage of time each predictor was favorable, neutral or unfavorable for the MCE's true negative (TN) predictions. We can therefore use how often a predictor was favorable for a TP prediction (green bar in Panel (a)) and how often a predictor was unfavorable for a TN prediction (red bar in Panel (b)) to gage predictor significance for the MCE. By adding the two variables above, each divided by the total number of true positives or true negatives respectively, we can gage feature importance for the MCE, as shown in panel (c). We find that the MCE is broadly consistent with LR and DT models with SHEAR and POT being most significant. However, we see that given the MCE is trained with equal weights for each predictor, the significance for all 5 predictors does not vary as much as compared to LR and DT models in panel (c). This can indicate that a select few predictors can affect the LR and DT's

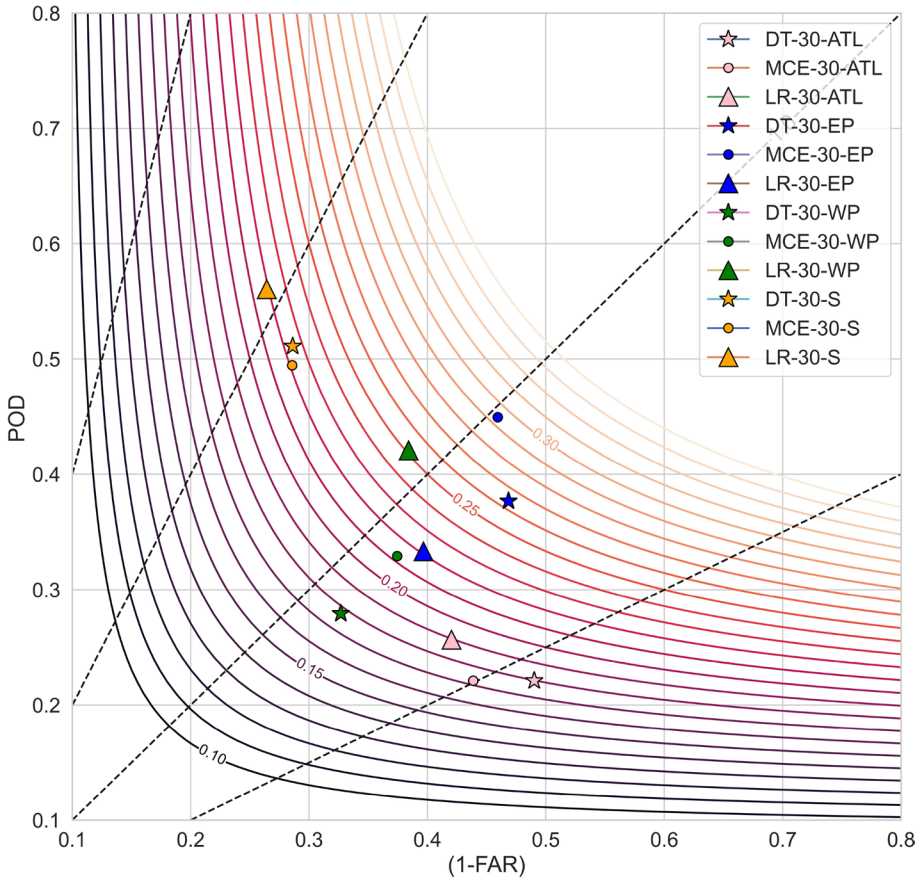

Fig. 4 Performance diagram summarizing multiple performance metrics at RI 30 kt for decision tree model (star), logistic regression (triangle) and MCE (circle) tested in multiple basins; Atlantic Ocean (pink), Eastern Pacific (blue), Western Pacific (green) and Southern Hemisphere (orange). *X* axis shows Success Ratio = (1-FAR). POD on *y* axis. Contour lines show CSI scores. Dotted diagonal line represents bias scores

RI prediction. For example, in the framework of LR and DT, if a few environmental parameters are unfavorable for RI, it could still lead to a RI prediction if the other parameters are highly favorable. This could explain the increase in FPs for the LR and DT prediction models. On the other hand for the MCE, the relative significance of each predictor is not as highly variable, indicating that multiple co-occurring predictors can play an important role in predicting RI.

5 Conclusions and discussion

The aim of the study was to explore how accounting for co-occurring environmental parameters can improve RI prediction. We create a simple binary RI prediction model, one that solely depends on ensuring there are multiple favorable environmental predictors while accounting for those predictors that can be unfavorable. The MCE model predicts

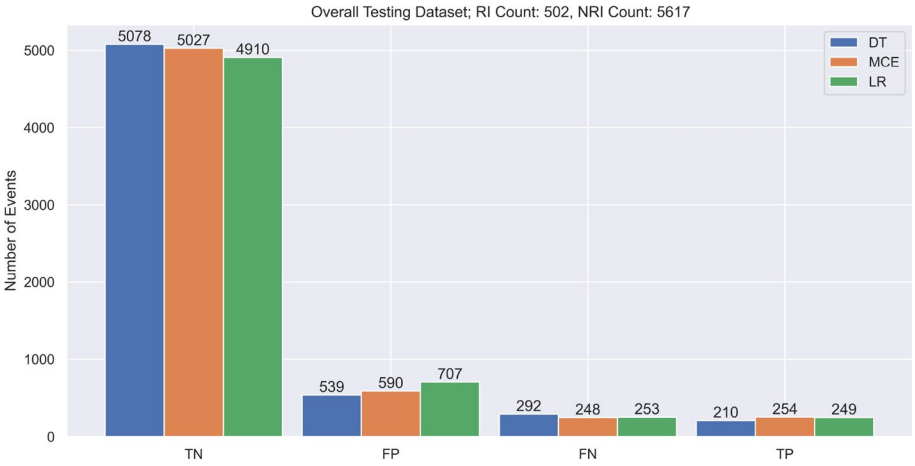

Fig. 5 True Negative, False Positive, False Negative, True Positive Scores of Decision Tree, MCE and Logistic Regression at RI threshold of 30 kt for the overall testing dataset

RI if there are at least 4 net favorable predictors out of the 5 listed environmental predictors used in this study. Overall, results show that the MCE outperforms a well-trained LR and DT model across multiple performance metrics. When evaluated at an RI threshold of 30kt, the MCE had a CSI score of 0.23 which is around 14% higher than LR and DT models. From the model 2×2 contingency scores, we can see that the MCE shows improved skill over the LR and DT models, primarily with more accurate RI predictions in the overall testing dataset. When evaluated at higher RI thresholds, the MCE consistently exhibits lower FAR scores showing that concurring favorable environmental parameters can be particularly important for predicting cases of higher rates of intensification.

We can see from our feature analysis, the LR model assigns higher regression weights to certain predictors, such as wind shear, when predicting RI. The DT model accounts for environmental predictors to hierarchically meet certain thresholds in the model rules. However, the model favors incorporating certain predictors like potential intensity and wind shear at a much higher percentage in the model's decisions over other predictors. On the other hand, the MCE evaluates the predictors simultaneously without varying assigned predictor weights and considers the overall environment's favorability for RI. The importance of co-occurring predictors can be seen in the MCE's derived feature scores which overall exhibit similar relative significance for each predictor.

Beyond improving prediction, these results can help improve our physical understanding of RI. They suggest that RI is more likely to happen when several environmental parameters align together rather than in situations where only one or two parameters are highly favorable. Further, the MCE suggests that the occurrence of both favorable and unfavorable environmental parameters plays an important role in distinguishing RI and non-RI events.

Given the limited number of predictors used in this study, future work can involve broadening the scope of environmental predictors and adding relevant predictors related to TC internal processes for use in the MCE. Further exploration can also be conducted to investigate how the role of co-occurring environmental predictors varies in different basins. The model can also be expanded to include RI predictions at lead times longer than 24 h

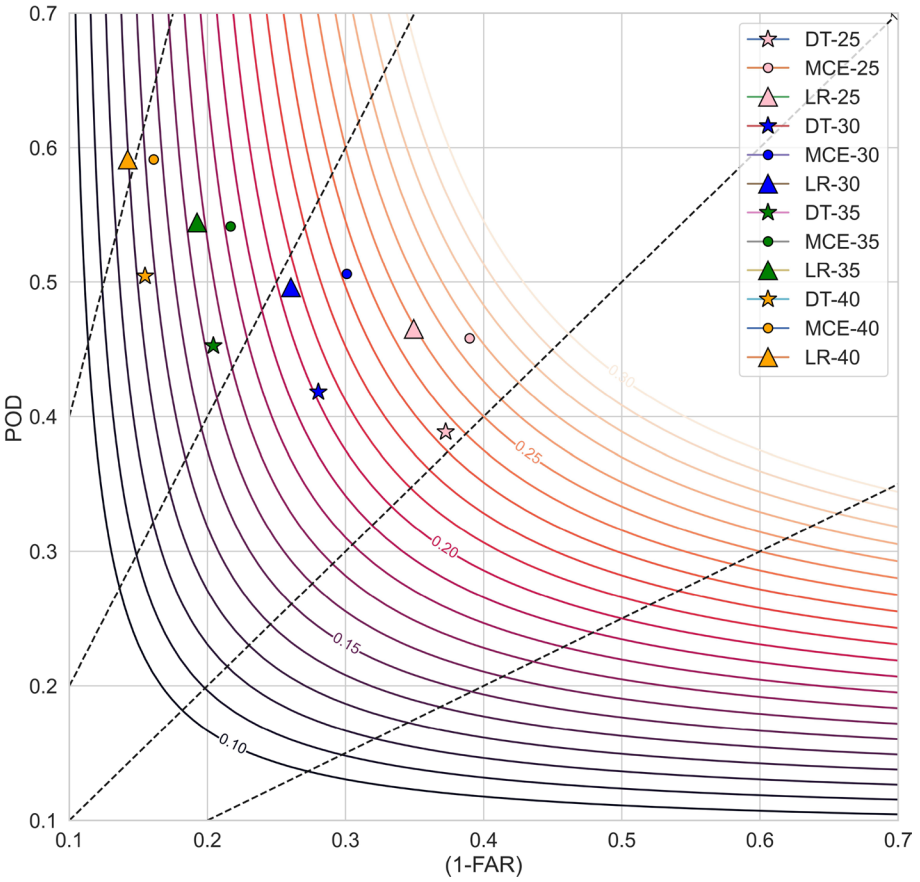

Fig. 6 Performance diagram summarizing multiple performance metrics of decision tree model (star), logistic regression (triangle) and MCE (circle) tested at multiple RI thresholds of 25 kt (pink), 30 kt (blue), 35 kt (green) and 40 kt (orange). *X* axis shows Success Ratio = (1-FAR). POD on *y* axis. Contour lines show CSI scores. Dotted diagonal line represent bias scores

and results can be analyzed at the individual basin level. Also, there is an opportunity to expand the MCE framework to account for the degree of favorability of an environmental predictor in the current threshold levels. Incorporating the MCE's simple method of evaluating large-scale environmental conduciveness to RI within other skillful RI models could potentially improve RI prediction performance. Additionally, the concept of the MCE can be applied to understand the physical underpinnings of other low-probability weather and climate extremes that tend to have substantial societal impacts.

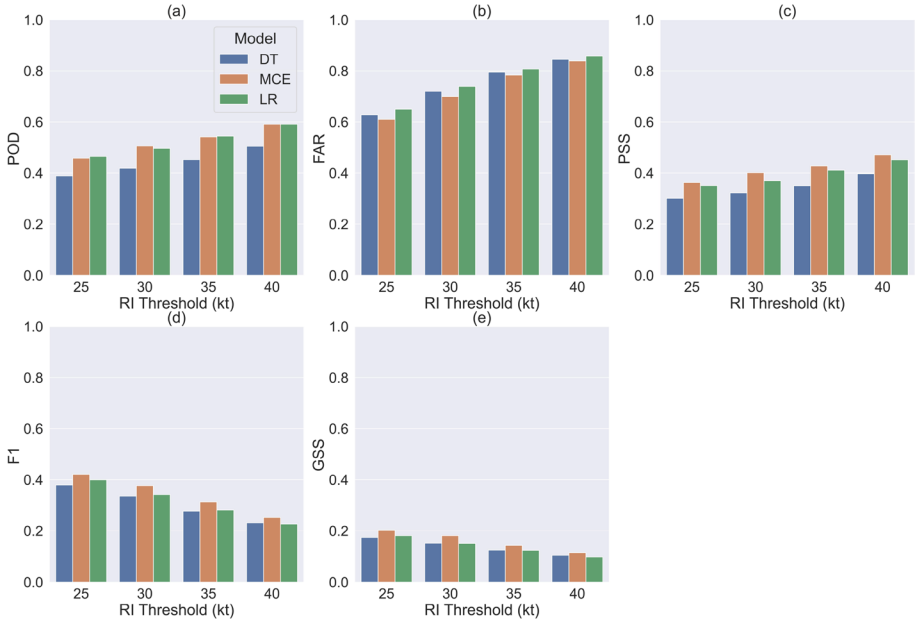

Fig. 7 POD (5a), FAR (5b), PSS (5c), F1 Score (5d), GSS (5e) scores of Decision Tree (blue), Logistic Regression (green) and MCE (orange) at multiple RI thresholds

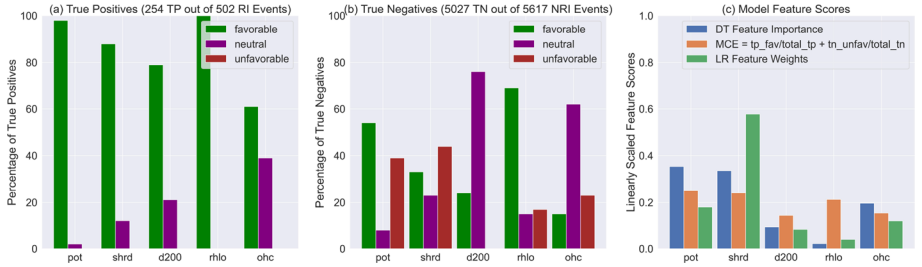

Fig. 8 In Panel (a) we can see for the MCE's true positive predictions, the percentage of time each predictor was favorable, neutral or unfavorable. In Panel (b), we can see for the MCE's true negative predictions, the percentage of time each predictor was favorable, neutral or unfavorable. To gage predictor significance in the MCE's predictions, we take the number of times a predictor was favorable for true positive predictions (green bar in Panel a) out of the total true positives and add it to the number of times a predictor was unfavorable (red bar in Panel b) for a true negative prediction out of the total true negative events which is then linearly scaled (orange bar in panel (c)). In Panel (c), the blue bar represents DT feature scores and the green bar represents LR's linearly scaled feature weights

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11069-023-06100-z.

Funding A.N., K.B., W.X. and L.R.L. are supported by the Office of Science (BER) of the U.S. Department of Energy as part of the Regional and Global Model Analysis (RGMA) program area. A.N. and K.B. also acknowledge support from NOAA's Climate Program Office, Climate Monitoring Program (Award Number: NA17OAR4310155). The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. The authors would also like to acknowledge Sujith Krishnakumar's efforts during the preliminary stages of this study.

Data availability The data that support the findings of this study are publicly available online at https://rammb2.cira.colostate.edu/research/tropical-cyclones/ships/.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Bhatia K, Baker A, Yang W et al (2022) A potential explanation for the global increase in tropical cyclone rapid intensification. Nat Commun 13(1):6626

Daniel S, Wilks W (2006) Statistical methods in the atmospheric schiences. Academic Press

DeMaria M, Kaplan J (1994) A statistical hurricane intensity prediction scheme (ships) for the atlantic basin. Weather Forecast 9(2):209–220

DeMaria M, Kaplan J (1999) An updated statistical hurricane intensity prediction scheme (ships) for the atlantic and eastern north pacific basins. Weather Forecast 14(3):326–337

DeMaria M, Mainelli M, Shay LK et al (2005) Further improvements to the statistical hurricane intensity prediction scheme (ships). Weather Forecast 20(4):531–543

DeMaria M, Franklin JL, Onderlinde MJ et al (2021) Operational forecasting of tropical cyclone rapid intensification at the national hurricane center. Atmosphere 12(6):683

Doswell C, Davies-Jones R, Keller DL (1990) On summary measures of skill in rare event forecasting based on contingency tables. Weather Forecast 5(4):576–585

Dowdy S, Wearden S (1991) The analysis of variance model. Statistics for research, pp 339–367

Espinel Z, Nogueira LM, Gay HA et al (2022) Climate-driven atlantic hurricanes create complex challenges for cancer care. Lancet Oncol

Foltz GR, Balaguru K, Hagos S (2018) Interbasin differences in the relationship between sst and tropical cyclone intensification. Mon Weather Rev 146(3):853–870

Herman GR, Schumacher RS (2018) "Dendrology" in numerical weather prediction: what random forests and logistic regression tell us about forecasting extreme precipitation. Mon Weather Rev 146(6):1785–1812

Kaplan J, DeMaria M (2003) Large-scale characteristics of rapidly intensifying tropical cyclones in the north atlantic basin. Weather Forecast 18(6):1093–1108

Kaplan J, DeMaria M, Knaff JA (2010) A revised tropical cyclone rapid intensification index for the atlantic and eastern north pacific basins. Weather Forecast 25(1):220–241

Kaplan J, Rozoff CM, DeMaria M et al (2015) Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. Weather Forecast 30(5):1374–1396

Kim SH, Moon IJ, Won SH et al (2021) Decision-tree-based classification of lifetime maximum intensity of tropical cyclones in the tropical western north pacific. Atmosphere 12(7):802

Kossin JP, Sitkowski M (2009) An objective model for identifying secondary eyewall formation in hurricanes. Mon Weather Rev 137(3):876–892

Mercer A, Grimes A (2017) Atlantic tropical cyclone rapid intensification probabilistic forecasts from an ensemble of machine learning methods. Procedia Comput Sci 114:333–340

Rappaport EN, Franklin JL, Avila LA et al (2009) Advances and challenges at the national hurricane center. Weather Forecast 24(2):395–419

Roebber PJ (2009) Visualizing multiple measures of forecast quality. Weather Forecast 24(2):601–608

Rozoff CM, Kossin JP (2011) New probabilistic forecast models for the prediction of tropical cyclone rapid intensification. Weather Forecast 26(5):677–689

Rozoff CM, Velden CS, Kaplan J et al (2015) Improvements in the probabilistic prediction of tropical cyclone rapid intensification with passive microwave observations. Weather Forecast 30(4):1016–1038
Su H, Wu L, Jiang JH et al (2020) Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. Geophys Res Lett 47(17):e2020GL089,102
Tam HF, Choy CW, Wong WK (2021) Development of objective forecast guidance on tropical cyclone rapid intensity change. Meteorol Appl 28(2):e1981
Wei Y, Yang R (2021) An advanced artificial intelligence system for investigating tropical cyclone rapid intensification with the ships database. Atmosphere 12(4):484
Xu W, Balaguru K, August A et al (2021) Deep learning experiments for tropical cyclone intensity forecasts. Weather Forecast 36(4):1453–1470
Yang R, Tang J, Kafatos M (2007) Improved associated conditions in rapid intensifications of tropical cyclones. Geophys Res Lett 34(20)
Yang R, Sun D, Tang J (2008) A "sufficient" condition combination for rapid intensifications of tropical cyclones. Geophys Res Lett 35(20)
Yang R, Tang J, Sun D (2011) Association rule data mining applications for atlantic tropical cyclone intensity changes. Weather Forecast 26(3):337–353
Yang R (2016) A systematic classification investigation of rapid intensification of atlantic tropical cyclones with the ships database. Weather Forecast 31(2):495–513
Yang Q, Lee CY, Tippett MK (2020) A long short-term memory model for global rapid intensification prediction. Weather Forecast 35(4):1203–1220
Zhang W, Gao S, Chen B et al (2013) The application of decision tree to intensity change classification of tropical cyclones in western north pacific. Geophys Res Lett 40(9):1883–1887
Zhu YJ, Collins JM, Klotzbach PJ et al (2022) Hurricane ida (2021): rapid intensification followed by slow inland decay. Bull Am Meteorol Soc 103(10):E2354–E2369