ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED PUBLISHING™ Part of the ENVIRONMENTAL RESEARCH SERIES

LETTER • OPEN ACCESS

Improving tropical cyclone rapid intensification forecasts with satellite measurements of sea surface salinity and calibrated machine learning

To cite this article: Ryan Eusebi *et al* 2025 *Environ. Res. Lett.* **20** 034010

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
8 October 2024

REVISED
18 January 2025

ACCEPTED FOR PUBLICATION
21 January 2025

PUBLISHED
11 February 2025

LETTER

Improving tropical cyclone rapid intensification forecasts with satellite measurements of sea surface salinity and calibrated machine learning

Ryan Eusebi[1,*], Hui Su[2,3,*], Longtao Wu[4], Pingping Rong[2], Karthik Balaguru[5], Ruby Leung[5], Yong-Sang Choi[6], Pak Wai Chan[7], Jianping Gan[2], Mark DeMaria[8] and Galina Chirokova[8,*]

[1] California Institute of Technology, Pasadena, CA, United States of America
[2] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong Special Administrative Region of China, People's Republic of China
[3] Center for Ocean Research in Hong Kong and Macau (CORE), The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong Special Administrative Region of China, People's Republic of China
[4] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, United States of America
[5] Pacific Northwest National Laboratory, Richland, WA, United States of America
[6] Ewha Womans University, Seoul, Republic of Korea
[7] Hong Kong Observatory, Hong Kong Special Administrative Region of China, People's Republic of China
[8] Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, United States of America
[*] Authors to whom any correspondence should be addressed.

E-mail: reusebi@caltech.edu, cehsu@ust.hk and Galina.Chirokova@colostate.edu

Keywords: tropical cyclone, satellite, hurricane, rapid intensification, salinity, machine learning

Supplementary material for this article is available online

Abstract

Forecasting rapid intensification (RI) of tropical cyclones (TC) is a mission known for large errors. One under-researched factor that affects TC intensification is salinity, which is important for density stratification in certain ocean regions and can affect the surface enthalpy flux under a strengthening hurricane. To investigate the impact and efficacy of using salinity information in state-of-the-art forecasting, we use a statistical model consisting of a variety of machine learning (ML) methods. For salinity data, we use satellite measurements of pre-storm sea surface salinity (SSS) as a proxy for the salinity stratification. We train and test the model on various ocean basins, including the Atlantic, eastern North Pacific and western North Pacific. A calibrator is trained on top of the ML models to correct and enhance probability forecasts. The calibrator significantly improves probability forecasts relative to recent works. The ML model performance is improved with the addition of SSS in the Eastern North Pacific, western North Pacific, and the Caribbean subregion of the North Atlantic, and the overall model performance is better than previous studies. SSS decreases model skill for a model trained on the full Atlantic basin. In the Indian Ocean, SSS is also notably correlated with RI occurrence, but the TC samples are not sufficient to train ML models.

1. Introduction

Adequate warning and preparedness can help mitigate the severe socio-economic impacts of landfalling tropical cyclones (TC), thus necessitating accurate forecasting of TC track and intensity. While the NOAA National Hurricane Center (NHC) TC track forecast errors have steadily decreased over recent years, the improvements have been slower for intensity forecasts, with stagnated forecast errors for long periods of time (Rappaport *et al* 2009, DeMaria *et al* 2014, 2021). Rapid intensification (RI), defined as a minimum increase of 25, 30, or 35 knots (kt hereinafter, with 1 kt = 0.514 m s^{-1}) in the maximum sustained surface winds of a TC within a 24-hour period, is particularly difficult to predict and has been a high priority in operational TC predictions (Rappaport *et al* 2012, Gall *et al* 2013). Given

the fact that the most destructive hurricanes, i.e. those of category 4 strength or greater, all undergo RI at a certain point in their lifetime (Kaplan and DeMaria 2003), and that the frequency of unusual events such as TCs undergoing RI just before landfall may increase as a result of global warming (Emanuel 2017), more in-depth studies are needed to better understand the contributing factors and how they shall be used to best predict RI.

The NHC's Statistical Hurricane Intensity Prediction Scheme (SHIPS) Rapid Intensification Index (RII) adopted environmental parameters such as sea surface temperature (SST), lower and middle troposphere moisture, vertical wind shear, and upper tropospheric divergence as the primary predictors for RI (Kaplan and DeMaria 2003, Kaplan *et al* 2010, Kaplan *et al* 2015, hereafter K15). Additionally, given favorable environmental conditions, internal storm dynamics are also an important factor in RI processes (Shapiro and Willoughby 1982, Willoughby 1990, Nolan and Grasso 2003, Nolan *et al* 2007, Hendricks *et al* 2010, Molinari and Vollaro 2010, Jiang 2012, Chen and Zhang 2013, Stevenson *et al* 2014, Rogers *et al* 2015, Hazelton *et al* 2017, Chen *et al* 2018). The SHIPS RII model incorporates information about storm internal processes through metrics of geostationary infrared brightness temperature (BT), including mean and standard deviation, within a 50–200 km radius of the storm center (K15). An in-depth description of the predictors used in the SHIPS RII model can be found in K15.

In this paper, we aim to understand the role of the upper ocean salinity environment, represented by sea surface salinity (SSS), in modulating TC RI using statistical models for forecasting RI. The effects of upper ocean salinity on TC RI were explored by (Balaguru *et al* 2020, hereafter B20). It is known that TC intensification relies on the heat supply from the ocean, and therefore parameters such as SST are important factors in RI prediction models (Emanuel 1999, Cione and Uhlhorn 2003, B20). However, along TC tracks, wind-induced oceanic vertical mixing and subsequent sea surface cooling can contribute to TC weakening. Thus, the upper ocean density stratification plays an important role in TC intensification, as it controls to what extent the wind-induced vertical mixing can entrain colder subsurface water to the surface (Price 1981, Bender and Ginis 2000, Cione and Uhlhorn 2003, B20).

B20 found that, in the North Atlantic, density stratification can be dominated either by temperature stratification or salinity stratification, depending on the region. In the Gulf of Mexico, temperature tends to dominate density stratification. Thus, SST and oceanic heat content (OHC) are sufficient within the SHIPS RII to provide information on the available heat supply. However, John *et al* (2023) have found evidence of freshwater fluxes on density stratification and TC intensification in the Gulf of Mexico. In the Eastern Caribbean and western tropical Atlantic, upper-ocean density stratification is largely controlled by salinity due to the surface-level flux of freshwater from the Amazon-Orinoco River system (B20). Previous research has shown that this freshwater flux, which would decrease surface salinity, can increase the near-surface density stratification and consequently weaken the TC-induced ocean mixing. Since mixing brings up colder sub-surface water to the surface which can contribute to weakening cyclones, low surface salinities, through weakening this mixing, can contribute to storm strengthening (Balaguru *et al* 2012, Grodsky *et al* 2012, B20). In this region, SST and OHC do not account for the effects of salinity on the vertical density structure (Balaguru *et al* 2015), so most RI prediction models lack a predictor to account for density stratification where it is dominated by salinity. Since measurements of near-surface vertical structure of salinity are not readily available and are difficult since the freshwater flux forcing over the ocean tends to be more localized, SSS can be used as a proxy for salinity stratification. The area affected by the influx of the Amazon-Orinoco River system is such an example.

B20 used logistic regression to verify the significance of SSS (derived from reanalysis products) for RI prediction in this region. B20 stresses that in order to demonstrate feasibility for implementation in real-time forecasting systems, satellite-derived estimates of SSS should be used in the model. In this work, we experiment with using satellite-derived pre-storm SSS as a predictor for RI. New from B20, we also use a more sophisticated statistical model (an adapted version of the machine learning model used in Su *et al* 2020, hereafter S20). The model used is capable of uncovering complex and non-linear relationships. Demonstrating success and significance of SSS as a predictor in a complex model is essential to demonstrate its utility and impact as a predictor in modern forecasting systems. Finally, while B20 focused on a sub-region of the Atlantic Ocean, we test its effectiveness as a predictor in a variety of ocean basins, including the whole Atlantic, the northern East Pacific, the northern West Pacific, and the Indian Ocean, and find it to be especially useful in the northern East and WestPacific basins.

Machine learning (ML) applications to TC forecasting are becoming more frequent and powerful, and have found recent success in forecasting TC track (Giffard-Roisin *et al* 2020, Boussioux *et al* 2022) and intensity (Boussioux *et al* 2022, Meng *et al* 2023, Chen *et al* 2018, Xu *et al* 2021, Narayanan *et al* 2023, Ko *et al* 2023, S20). Some physics-informed ML applications to TCs exist, including reconstructions of TC wind fields to aid in dynamical forecast initialization

(Eusebi *et al* 2024). In this study, we adapt the statistical ML model for RI prediction built by S20 that tested the role of satellite-measured precipitation in RI in combination with conventional SHIPS predictors. This paper builds on the work of B20 and S20 to elucidate the role of salinity in TC RI and use ML models to enhance RI forecasting.

2. Methods

2.1. Data

We consider storms from the North Atlantic (ATL), northern East Pacific (EPAC), the northern West Pacific (WPAC), and the Indian Ocean (IO). Additionally, we consider in isolation the subregion of the Atlantic basin in which the density stratification was found to be dominated by salinity stratification (B20). We refer to this sub-basin as WATL. This region contains TC data points within 80° W–40° W and 0°–30° N, which roughly overlaps with but is slightly wider than the region noted by B20 to obtain a greater number of data points available for statistical analysis.

The TC data points we used in this study are from the NHC Joint Typhoon Warning Center's Best Track Archive dataset between the years 2010 and 2019 (only 2010–2017 for the IO and WPAC). For a given storm, the TC location (latitude and longitude) and the maximum sustained surface wind speed were reported every 6 h with a 5 kt resolution for the maximum winds.

For SSS measurements, we use the version v2.31 European Space Agency Climate Change Initiative SSS dataset (Boutin *et al* 2020). This data set is a combination of the measurements from the Soil Moisture Active Passive, Soil Moisture and Ocean Salinity, and Aquarius SSS datasets. The SSS dataset has a spatial resolution of 50 km and a time resolution of 1 week. It is resampled on a 25 km Equal Area Scalable Earth Grid and 1 day time interval. Other TC predictors implemented in the statistical model aside from SSS are obtained from the SHIPS developmental dataset. The pre-storm SSS measurements were extracted based on the storm time and location in the Best Track dataset. All storms throughout years 2010–2019 are used. For each data point, SSS values are averaged within the radii of 200 km from the storm center (other radii between 100 km and 400 km yielded similar results). The SSS values are temporally averaged over days 4–10 prior to the storm passing over that location (to rule out any effects caused by sea surface mixing induced by storm winds as it passes over that location). Since the SSS measurements have a time resolution of 7 d, using data from the previous 4–10 d period also mimics what would be available in real-time for operational forecasting.

2.2. The machine learning model

The statistical ML model in S20 is used with modifications. The models are trained using the SHIPS RII predictors and SSS. The predictors used by the model (the SHIPS RII predictors plus SSS) are described in table S1. We use 9 of the 10 SHIPS RII predictors that were utilized in K15, excluding the inner-core dry-air predictor. The predictors are described in table S1. We aim to not only compare the results of the ML model against the NHC's operational RI consensus model (CON-RII), but also to compare between the results of the ML model with and without SSS as a predictor to understand the significance of SSS for RI prediction. The CON-RII model results are available only in the ATL and EPAC basins, and therefore our ML model analysis in the WPAC solely focuses on the impact of SSS on ML model results. The ML models are trained for each basin separately (including the sub-basin WATL) using the SHIPS developmental database, which are based on Climate Forecast System Reanalysis and GOES IR data. Note that the OHC predictor is not available for the IO and WPAC, so the ML model is constructed without it for those basins. The CON-RII model is the average of the logistic regression, Bayesian, and the discriminant analysis-based SHIPS-RII probabilistic forecast results and is the most skillful model for RI at the NHC according to K15 for 2008–2013 TCs. Following Kaplan *et al* 2010, K15, and S20, three RI thresholds are considered: the 24-hour intensity increase (DV24) $\geqslant$ 25, 30 and 35 kt. Evaluating multiple RI thresholds is important to understand how the model performs with rarer event scenarios, such as with 35 kt RI vs. 25 kt RI prediction. Deterministic and probabilistic forecast skills are assessed. Note that the SHIPS developmental dataset was used, so this is data that has been corrected post-storm, and is not necessarily what was available at forecast time, so the comparison between the ML model and NHC model will not be fully accurate. The number of RI and non-RI cases used by the ML model for different basins and RI thresholds is shown in table S2.

We evaluate the predictive skill of the forecast models using a variety of metrics: probability of detection (POD), false alarm ratio (FAR), the Threat Score (TS), and the Peirce Skill Score (PSS). Note that TS is also referred to as the Critical Success Index. These metrics are defined as $\mathrm{POD} = \frac{a}{a+c}$, $\mathrm{FAR} = \frac{b}{a+b}$, $\mathrm{TS} = \frac{a}{a+b+c}$, and, $\mathrm{PSS} = \frac{ad-bc}{(a+c)(b+d)}$, where a, b, c, and d refer to the numbers of true positive, false positive, false negative, and true negative, with the best scores being POD = 1, FAR = 0, TS = 1, and PSS = 1, respectively. Contrary to S20, in this paper we focus on TS instead of PSS due to its popularity as an evaluation metric in recent years (e.g. Narayanan *et al* 2023) and its utility for rare-event problems (Doswell *et al* 1990, Roebber

2009) since it ignores skill from true negatives. Our results indicate that PSS tends to be maximized at high FAR, which is not always desirable for an RI model (Demaria *et al* 2021). Note that for models which make predictions from probabilities, deterministic metrics like TS and PSS are dependent on the choice of a cutoff probability. If the model predicts a probability above (below) this cutoff value, the event is classified as RI (not RI). Comparing TS and PSS results from different models without the context of how these values vary with cutoff threshold can be misleading, so in section 3.2 we show the best TS and PSS the model can achieve across all threshold probabilities (we choose the threshold that maximizes TS or PSS). In the supplementary figures, we show how these metrics vary with the chosen cutoff probability for a fairer comparison between models.

For the probability forecasts, the Brier score is used, defined as $\mathrm{BSM} = \frac{1}{N}\sum_{i=1}^{N}(f_i - O_i)^2$, where f_i is the forecast probability from the model and O_i is 1 or 0 if the event is observed or not, respectively; i and N represent the event index and the total number of events, respectively. A perfect Brier Score is 0. The Brier Skill Score (BSS), defined as $\mathrm{BSS} = \left(1 - \frac{\mathrm{BSM}}{\mathrm{BSC}}\right)$, further normalizes the BSM relative to the climatological brier score, BSC (e.g. see Kaplan *et al* 2010, K15, S20). A BSS of 1 corresponds to a perfect model since low BSM is expected for a skillful prediction. In this study, the climatological probability of RI in the ATL (including WATL) and EPAC is taken from K15, and for the IO and WPAC we find the climatological probability based on the values listed on the SHIPS RII developmental dataset for 2010–2017. These probabilities are shown in table S3.

The Machine Learning Hurricane Intensity Forecast Scheme used in this study (ML.HIFS hereinafter) is an ensemble classifier model consisting of 4 ML models of increasing complexity available from Python's standard Scikit-learn library (Pedregosa *et al* 2011) as in S20: logistic regression (LR), decision tree (DT), random forest (RF), and extra trees classifier (ET). A description of these models can be found in the Supplementary Information. New from S20, we also train a calibrator on top of each individual model and the final ensemble model. We use an isotonic calibrator (Niculescu-Mizil & Caruana 2005) available through the python package scikit-learn. The objective of the calibrator is to learn an additional regression on top of the existing model to transform the model's probability outputs to a distribution of properly 'calibrated' probabilities. In this context, a properly calibrated model means that events that are predicted to undergo RI with a probability of X% should experience RI X% of the time. Calibrators are especially necessary for realistic probability forecasts from tree-based models whose output probabilities often do not correspond to true probabilities, as they are intended more for deterministic classifications (Niculescu-Mizil & Caruana 2005).

Also new from S20, data augmentation (e.g. Ko *et al* 2023) is used to create a larger dataset for each basin with equal numbers of RI and non-RI cases. A form of synthetic minority oversampling technique (SMOTE, Chawla *et al* 2002) called Borderline-SMOTE (Han *et al* 2005) is used to generate new samples of RI cases in the training set within each given year to create a balanced training dataset. The resampling is done for each year separately so data points from different years are not generated that are correlated with each other (and would then get placed separately in training and test sets).

Leave-one-year-out nested cross validation (Vabalas *et al* 2019) was used to maximize the amount of data that could be used to train and validate the model, while still yielding predictions and evaluation metrics on unseen data for the whole dataset. Nested cross validation consists of two layers of cross-validation. In the outer layer, all data points from one year of data are separated into a test set, and data from the other years is separated into a training set. On this training set, leave-one-year-out cross validation is performed again to find the best performing model and hyperparameters. This model is then trained on the full outer training set and evaluated on the held-out test set to obtain unbiased predictions and evaluation metrics. This process is repeated using each year of data as a held-out test set to obtain unbiased predictions by the trained model across the full dataset. This full nested cross-validation process is repeated 36 times, each time setting different random state parameters for the data resampler and the ML models, to calculate uncertainty bounds on evaluation metrics. All results shown will be the mean across the 36 iterations, with error bars indicating 95% confidence intervals across the 36 iterations.

During the training process, each model is first trained individually on the augmented dataset with random combinations of hyperparameters. Using the validation dataset from the inner cross validation on the augmented data, a calibrator is trained on top of each model to correct model probability predictions. The validation data is also used to select the best performing model and hyperparameters by selecting the model with the highest BSS. An ensemble model is built using all four models, and the optimal weighting that is selected is that which reduces BSS on the validation set. Finally, the ensemble model is built with the determined optimal weighting after the inner cross validation, and another isotonic calibrator is trained on top of the ensemble on the original (not augmented) full training set before making predictions on the test set (which is also not augmented). The original data is used because we want the model's probability predictions to be calibrated to the realistic distribution of RI data. Instead of a calibrator,

the uncalibrated model still has its probabilities transformed to be representative of a distribution with the true average RI rate, not a 50% rate, as described in Saerens *et al* (2002). BSS was used instead of TS for model selection because optimizing the probability forecasts should lend itself to better deterministic forecasts since the model will learn to discern the chances of RI between events.

We apply the ML.HIFS model to the 4 ocean basins listed above (including the sub-basin WATL) and evaluate the results in section 3.2. We compare BSS and other metrics, and further examine the significance of the different predictors and the calibrator. Where stated, statistical significance refers to p-values less than 0.05.

3. Results

3.1. Justification for satellite-measured SSS as an RI predictor

Figure 1 shows SSS distributions for all storm data points in the NHC best track for which we have SSS data. Each panel shows the distributions for RI and non-RI data points in the given basin, with the RI threshold as 25 knots in a 24 h period. Panels 1a and 1b show the ATL and WATL distributions. In both regions, at lower salinity levels (especially below 36 pss), RI cases tend to be more common than non-RI cases. In the WATL, the difference in distributions is clearer with a stronger bimodality associated with a secondary peak in RI cases between 34 and 35.5 pss. This supports the findings of B20 that in the WATL region below 36 pss, freshwater fluxes can contribute to RI. For comparison, the subset of ATL with all points outside of WATL is shown in panel 1d, where the RI distribution is stronger at some lower salinity levels, but confined to the 35.5–36.5 pss range. Nonetheless, across the ATL RI distributions tend to be stronger in lower salinity regimes relative to non-RI distributions.

Figures 1(c) and (e) indicate that there is a very distinct difference in the RI and non-RI distributions in both the EPAC and WPAC basins characterized by offsets in the mean of 0.6 pss and 0.5 pss, respectively. These offsets are considered significant since the standard deviations of the distributions are about 0.7 pss. No previous studies have addressed the salinity variability in the Pacific basin and the associated non-RI and RI occurrences, as is done for the ATL basin in B20.

IO does not have enough data points coincident with satellite-based SSS measurements to build a statistical model, but we still examined the relationship between SSS and RI in this basin, shown in figure 1(e), which exhibits a distinct bi-modal behavior. Nearly all the storms with SSS below 34 belong to the Bay of Bengal while the storms with SSS above 35 mostly belong to the saltier Arabian Sea. In the high-salinity lobe, the distributions of RI and non-RI cases have similar means and shapes, suggesting the insensitivity of RI to salinity in this region. In the low-salinity lobe, however, given the lower number of total cases, it clearly shows that peaks separate between the RI and non-RI cases.

3.2. Machine learning model results

Figure 2 shows model results in the ATL, WATL, EPAC, and WPAC basins for the NHC model, the uncalibrated ML.HIFS, and the calibrated ML.HIFS with and without SSS. Uncertainty bounds show the 95% confidence interval of the metrics calculated from the 36 separate training iterations. Note that the resulting evaluation scores for the CON-RII model have been verified against the results of K15 to make sure they are consistent with past model performance. While the training methodologies are slightly different, the uncalibrated ML.HIFS model (green in the figure 1) can be considered structurally identical to the model of S20. In terms of TS, the ML models outperform the NHC model in all basins and for all RI thresholds. In terms of BSS, the calibrated ML models outperform the NHC model in all basins and for all RI thresholds. Figure S1 shows the PSS for all the models, and similarly the ML models outperform PSS in every basin and RI threshold. In addition to the already mentioned metrics, figure S2 shows receiver operating curves (ROC) for each model in each basin for 25 kt RI. The ML.HIFS AUC is substantially and significantly better than the NHC-CON RII model with AUC (area under the curve) ranging from 0.90 to 0.93.

In all basins, the impact of the calibrator is clear in the BSS. The calibrator generally increases BSS in all basins for all RI thresholds by 0.05–0.1. Figure S3 shows reliability diagrams for all the models, which many studies cite as an important marker of a probabilistic model's prediction capabilities (Demaria *et al* 2021, Griffin *et al* 2022, Meng *et al* 2023). The NHC models and the calibrated ML.HIFS are all very well calibrated, especially for the EPAC and WPAC basins which have more data points. The uncalibrated ML.HIFS, however, has a tendency to output too-low probabilities, which negatively impacts its BSS. Note that a calibrator can only change probabilistic metrics, not deterministic metrics for a model. But in the results, the deterministic metrics change between the uncalibrated and calibrated models. This is because the calibration affects the selected weightings of the ML models in the final ensemble. Generally, the calibrated model has TS that are significantly better or not statistically different, because the calibration might allow the ensemble to select models with more predictive power. In some cases, though, the calibration yields significantly worse (p-value < 0.05) results. This happens, for instance, at the 35 kt RI threshold for ATL and EPAC basins, and could indicate that

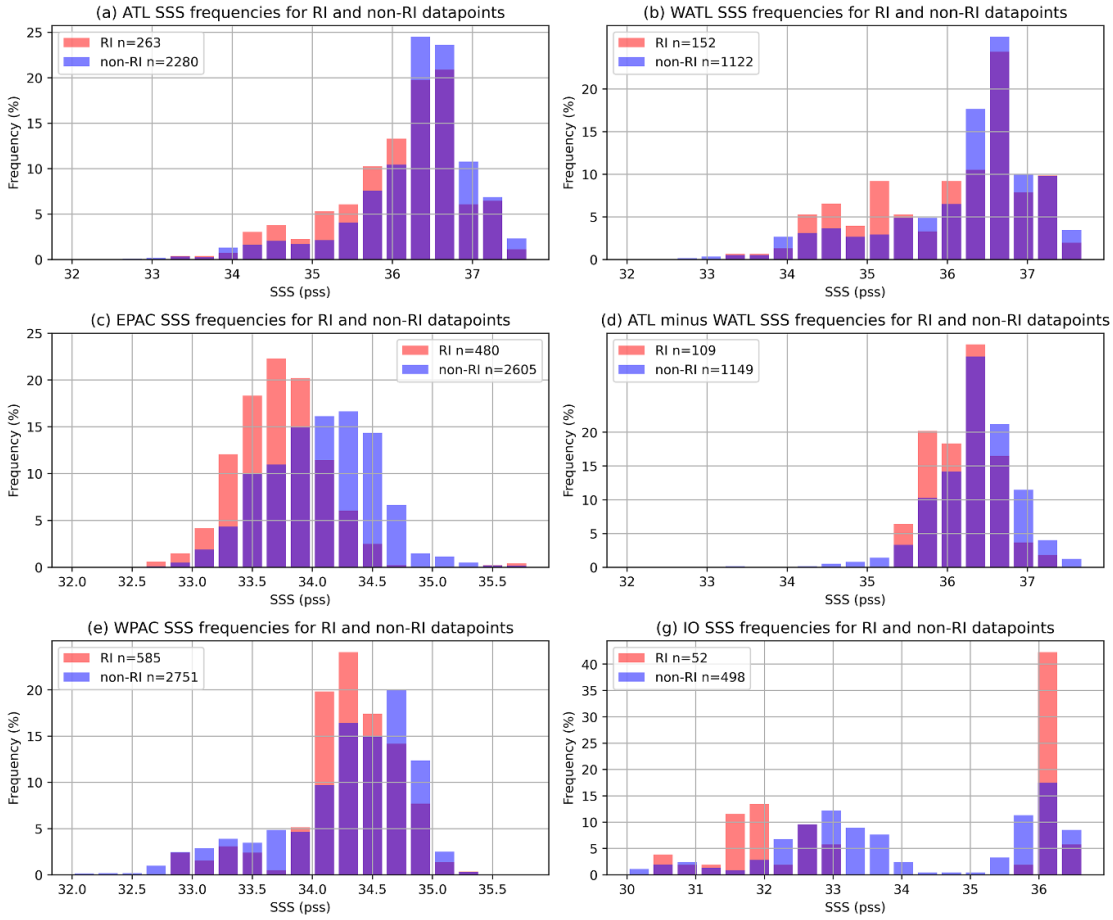

Figure 1. SSS distributions for all storm instances in the NHC best track dataset for the (a) north Atlantic, (b) western north Atlantic, (c) east Pacific, (d) The subset of ATL excluding all points in WATL, (e) west Pacific, and (f) Indian Ocean basins. RI and non-RI distributions are plotted in each plot, with RI distributions containing all storm instances which underwent 25 knots or greater of intensification in the subsequent 24 h. Non-RI distributions contained all other storm instances. The legend indicates the number of data points used for each case. Note that the number of data points used for each basin in the ML model is lower than these numbers, due to some points not having data for other predictors or other reasons.

the calibrator is helping overfit in scenarios where the number of RI cases is small. Figure S4 shows the average learned optimal weights of the individual models within the ensemble model across all 36 training iterations. For the 25 kt RI threshold across all basins, the calibrated model tends to prefer the ET more than in the uncalibrated models. ETs are inherently very uncalibrated due to the random nature of their construction, but very powerful, so calibration should allow them to become more prevalent in the ensemble, as is shown. However, at the 35 kt threshold, this is not always the case, as overfitting from the calibration sometimes leads to a preference for the simpler LR over ET. Generally, all ensembles tend to lend more weight to the RF model.

In the ATL basin, the addition of SSS as a predictor has no statistically significant effect at the 25 kt threshold, but the model with SSS has significantly worse BSS at 30 and 35 kt threshold, and significantly worse TS at 30 kt threshold. As explained in the introduction, SSS is expected to be significant mostly in the WATL region, so using SSS in a model for the whole Atlantic likely confuses the model and hurts performance. SSS shows promise in the WATL subbasin at the 25 kt threshold with significantly better performance than the model without SSS. However, there is no significant difference at the 30 kt and 35 kt threshold, where the model might be suffering from too few RI cases. Interestingly, SSS shows great promise as a predictor in the EPAC and WPAC basin. The ML.HIFS + SSS has significantly higher BSS and TS for all RI thresholds in the EPAC and the 25 and 30 kt thresholds in the WPAC compared to the model without SSS.

The metrics (including BSS, TS, PSS) the ML.HIFS + SSS the model achieves in this work are significantly improved from both NHC CONRII model and the ML model of S20. They are also significantly better than the results obtained by other NHC models, including the newest deterministic to probabilistic statistical model (DTOPS, Demaria *et al* 2021), with the caveat that our ML model was trained using SHIPs predictors from the developmental database, which corrects some of the

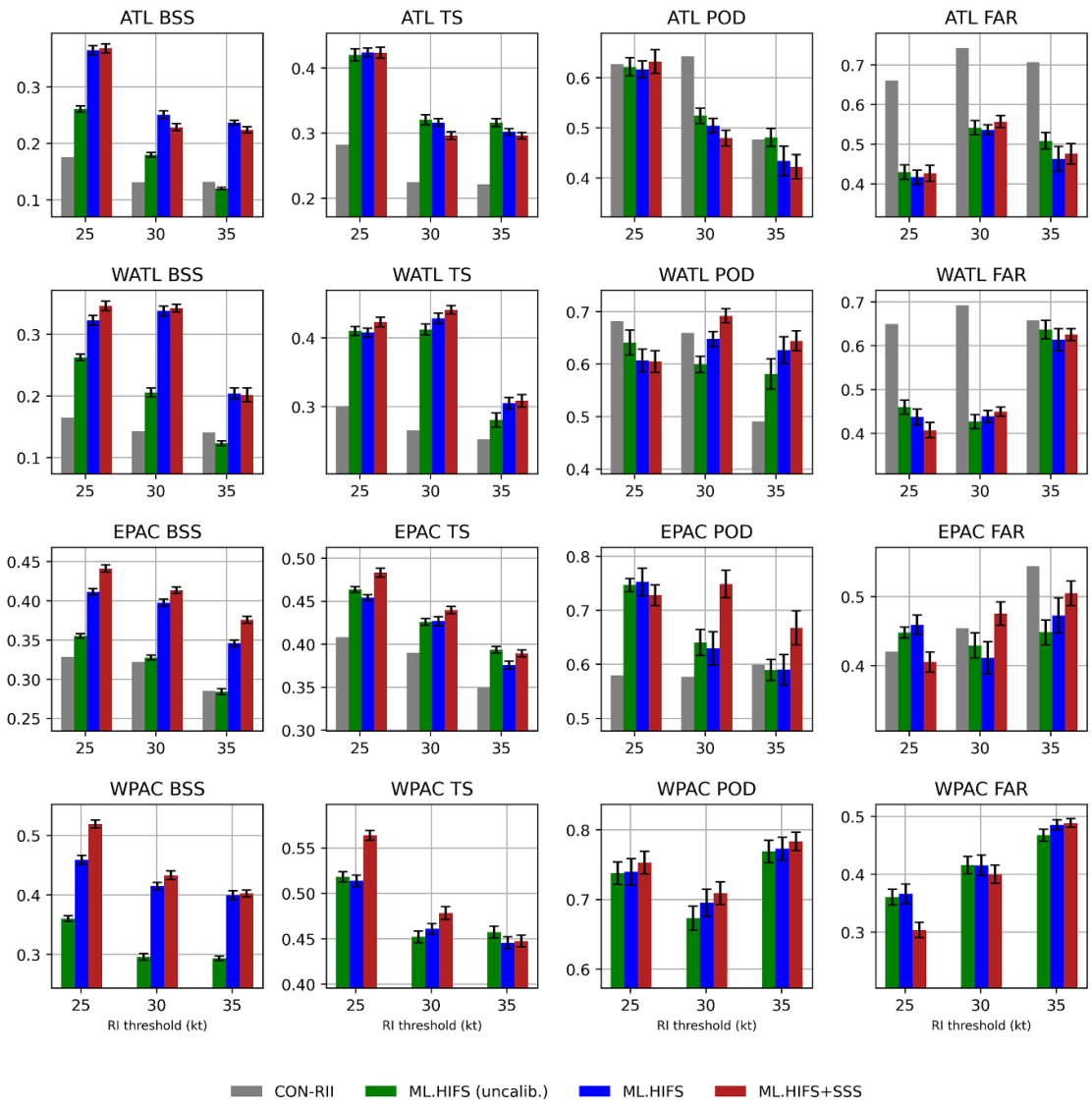

Figure 2. Predictive skill of the ML.HIFS model on unseen data for RI occurrence in the northern Atlantic (top), the WATL subset of the northern Atlantic between 80° W and 40° W, and 10° N and 30° N (second row), the northern East Pacific (third row), and northern West Pacific (bottom) using leave-one-year-out nested cross validation. From left to right are the Brier Skill Score, the Threat Score, the Probability of Detection, and the False Alarm Ratio. The TS is the maximum TS the model can obtain across the range of possible cutoff probabilities. POD and FAR for each model are those obtained when using the cutoff probability that maximizes TS. Results are shown for the NHC's RII-Consensus model (gray), the uncalibrated ML.HIFS model without SSS, the calibrated ML.HIFS without SSS (blue), and the calibrated ML.HIFS model constructed using SSS azimuthally averaged over a 200 km radius (red). The uncalibrated ML.HIFS model without SSS is structurally identical to the ML model of Su *et al* 2020. Uncertainty bars indicate 95% confidence intervals from the 36 training iterations.

variables using reanalysis data not available at forecast time. Our model also seems to perform similar or better compared to other ML-based models from recent literature (e.g. Griffin *et al* 2022, Ko *et al* 2023, Yang *et al* 2024). These comparisons are not exactly direct, however, due to differences in training and testing data.

We stress that the TS, POD, and FAR metrics shown in figure 2 (and S1) are dependent on the cutoff probability for RI. For fair comparison to the CON-RII model, we show these metrics for each model using the cutoff probability that maximizes the respective model's TS. Figures S5 and S6 show how the TS and PSS for the ML.HIFS and CON-RII model vary with the cutoff probability. The PSS for the CON-RII model in figure 2, for instance, are much higher than those in K15 and S20 because we choose a different cutoff probability that results in a much higher PSS for a fairer comparison to the ML model. Looking at the cutoff probability used in K15 and S20 in figure S6, one can obtain a similar PSS to that obtained in K15 and S20. Figures S7 and S8 shows how the POD and FAR similarly change with varying the cutoff probability.

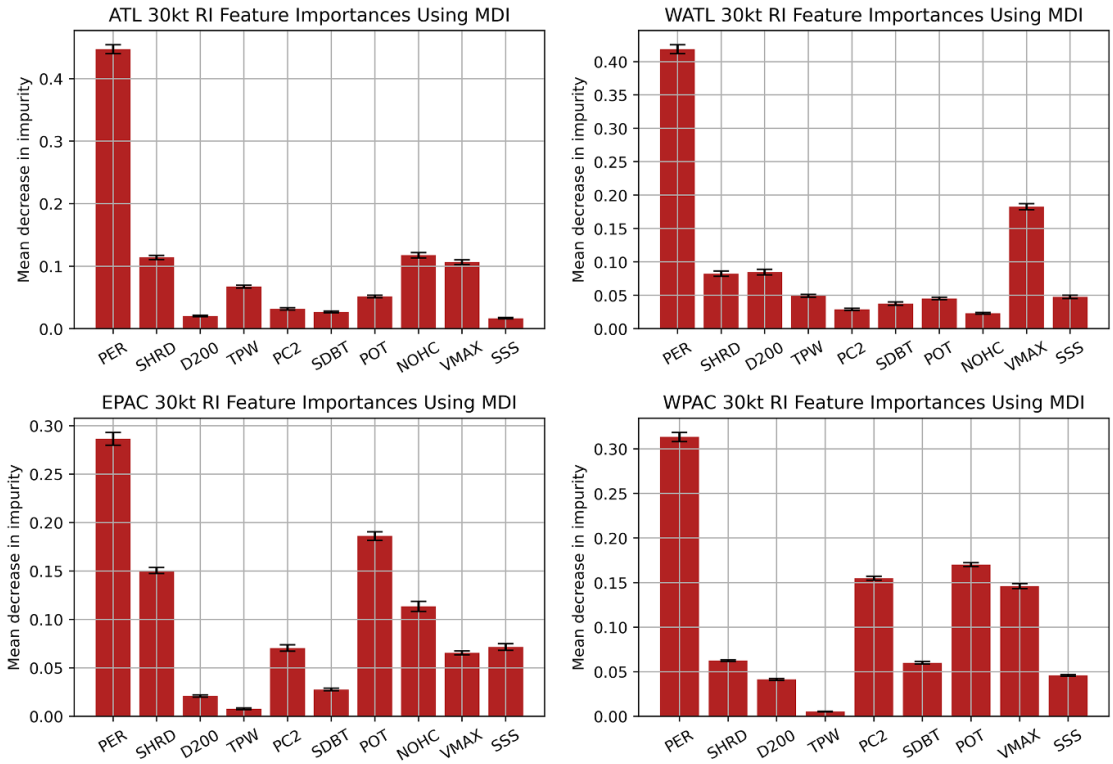

Figure 3. Feature importances in the random forest classification model, calculated using mean decrease in impurity. All importances are assessed relative to each other, and their absolute values have no meaning individually. These are calculated for the models trained for the (a) ATL, (b) WATL, (c) EPAC, and (d) WPAC basins for the 30 kt RI threshold. The values indicated are the average across the 36 independent training iterations, and error bars indicate 95% confidence. Importances have been normalized so that the sum across all features is 1. The features from left to right (in the ATL basin) stand for previous intensity change (PER), vertical wind shear (SHRD), 200 hPa divergence (D200), precipitable water (TPW), 2nd principal component of IR satellite image (PC2), standard deviation of satellite brightness temperature (SDBT), potential intensity (POT), oceanic heat content (OHC), maximum sustained wind (VMAX), and sea surface salinity (SSS). Predictors are described in table S1.

Finally, we examine the relative feature importances. We focus on the RF model for feature importances because this is the model that the ensemble models tended to prefer in their learned optimal weighting (figure S4). Each feature importance represents to what extent a given feature in the RF contributes towards minimizing the cost function that optimizes the correct number of classifications as RI or non-RI. It is calculated for each individual DT in the RF, and the final importance is the average across all trees.

The feature importances for each predictor in each basin for the 30kt RI RF model are shown in figure 3. In the ATL SSS does not appear too significant, but it is comparable to some of the other less important features. SSS appears more relatively important in WATL than ATL, as expected, and is more important than several other SHIPS predictors. Interestingly, the relative importance of OHC to SSS decreases from the WATL to ATL, corroborating the idea that in the WATL region, salinity might be a more important marker of density stratification than temperature. In the EPAC and WPAC, SSS is a much more important predictor relative to the ATL, which supports its significant impact on the model's performance in these basins. Feature importances for the RF model at the 25 kt and 35 kt thresholds are shown in figures S9 and S10, and show a similar story.

4. Conclusions and discussion

Aiming for imminent application of satellite observations in operational forecasts, we present an ML framework that combines satellite products and conventional predictors employed at the NOAA NHC for statistical RI forecasts. An ensemble average of linear and nonlinear ML models is constructed. The ML model outperforms the NHC's CON-RII model in terms of BSS and TS in the three RI threshold categories we explored in this work. Especially, the ML model performs far better in the 35kt RI threshold in each basin; for such intense RI episodes, the NHC model performance has been very poor in the Atlantic (PSS < 0.2) (Kaplan *et al* 2010, K15, S20). The ML model's performance comes with the caveat that it was tested using reanalysis data as opposed to forecast data.

In the WATL basin, as expected from B20, the inclusion of SSS significantly improves model performance in terms of our chosen evaluation metrics. Contrary to what was expected, its inclusion does not significantly change the skill at the 30 and

35 kt threshold, possibly because the model suffers from too few datapoints. The results of this study indicate that SSS might actually decrease model skill if data from the whole ATL basin is considered, possibly because the model is not able to reconcile the varying roles of salinity across the ATL basin without any location-specific information about the data. Including SSS significantly improves model performance in the EPAC and WPAC basin across nearly every RI threshold, indicating it can be a valuable predictor in RI models in these basins. However, the exact mechanism by which SSS is influencing RI in these regions is unclear and merits future research. There is insufficient data in the IO to draw any firm conclusions yet from a statistical model. However, it is intriguing that the limited data appear to support a strong relationship between SSS and TC RI in the Bay of Bengal. Given storms that undergo RI in the Bay of Bengal could affect large populations in India and Sri Lanka within relatively short time, it is prudent to further investigate what drives the SSS spatiotemporal variability and the implications for RI in this basin.

New from S20, we train a calibrator on top of our ensemble classifier which has significantly improved our probabilistic forecasts (and thus substantially improved BSS scores of the model) relative to the model of S20 (figure 2). We thus encourage all future researchers using ML models, and especially deterministic ML models (such as tree-based methods like RF and ET), to utilize calibrators to improve their probability forecasts. The BSS obtained in this study are higher than those obtained in other recent works, including (K15 and S20).

To take full advantage of the sophistication of the RF and ET models, and the usefulness of the data augmentation technique, more years of data are essential for improving ML model performance. As more years of satellite SSS measurements become available, a more reliable ML model can be built through which the role of SSS as an RI predictor can be better understood. Further work should also aim to better understand the mechanisms through which SSS might be a proxy of ocean density stratification in various regions, especially in the EPAC and WPAC. Like the ATL basin, there may be subregions in the WPAC, analogous to the WATL region, where ocean density stratification depends heavily on salinity stratification, and in those regions SSS may become a more important RI predictor.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files). The SSS dataset is available at https://dx.doi.org/10.5285/9ef0ebf847564c2eabe62cac4899ec41. The Best Track data is available at www.nhc.noaa.gov/data/#hurdat. The SHIPS developmental dataset is available at http://rammb.cira.colostate.edu/research/tropical_cyclones/ships/developmental_data.asp website. Code for training models is available upon request.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2139433. Part of the research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). HS thanks the funding support from the Hong Kong Jockey Club Charities Trust (FA123), the Innovation and Technology Commission(P0413), and the Center for Ocean Research in Hong Kong and Macau (CORE). KB and LRL are supported by the Office of Science, U.S. Department of Energy (DOE) Biological and Environmental Research through the Water Cycle and Climate Extremes (WACCEM) Scientific Focus Area funded by the Regional and Global Model Analysis program area. Pacific Northwest National Laboratory (PNNL) is operated for DOE by Battelle Memorial Institute under Contract DE-AC05-76RL01830. The authors thank three anonymous reviewers for very helpful comments and feedback which helped shape the methodology of the paper.

ORCID iDs

Ryan Eusebi https://orcid.org/0009-0008-0396-7731
Pingping Rong https://orcid.org/0000-0002-9923-0652
Karthik Balaguru https://orcid.org/0000-0003-0181-2687
Ruby Leung https://orcid.org/0000-0002-3221-9467
Mark DeMaria https://orcid.org/0000-0003-4746-4462

References

Balaguru K, Chang P, Saravanan R, Leung L R, Xu Z, Li M and Hsieh J-S 2012 Ocean barrier layers' effect on tropical cyclone intensification *Proc. Natl Acad. Sci. USA* **109** 14343–7

Balaguru K, Foltz G R, Leung L R, Asaro E D, Emanuel K A, Liu H and Zedler S E 2015 Dynamic potential intensity: an improved representation of the ocean's impact on tropical cyclones *Geophys. Res. Lett.* **42** 6739–46

Balaguru K, Foltz G R, Leung L R, Kaplan J, Xu W, Reul N and Chapron B 2020 Pronounced impact of salinity on rapidly intensifying tropical cyclones *Bull. Am. Meteorol. Soc.* **101** E1497–511

Bender M A and Ginis I 2000 Real-case simulations of hurricane–ocean interaction using a high-resolution

coupled model: effects on hurricane intensity *Mon. Weather Rev.* **128** 917–46
Boussioux L, Zeng C, Guénais T and Bertsimas D 2022 Hurricane forecasting: a novel multimodal machine learning framework *Weather Forecast.* **37** 817–31
Boutin J *et al* 2020 ESA sea surface salinity climate change initiative (sea_surface_salinity_cci): weekly sea surface salinity product, v2.31, for 2010–2019 (Centre for Environmental Data Analysis) (available at: https://catalogue.ceda.ac.uk/uuid/eacb7580e1b54afeaabb0fd2b0a53828)
Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57
Chen H and Zhang D L 2013 On the rapid intensification of Hurricane Wilma (2005). Part II: convective bursts and the upper-level warm core *J. Atmos. Sci.* **70** 146–62
Chen X, Wang Y, Fang J and Xue M 2018 A numerical study on rapid intensification of Typhoon Vicente (2012) in the South China Sea. Part II: roles of inner-core processes *J. Atmos. Sci.* **75** 235–55
Cione J J and Uhlhorn E W 2003 Sea surface temperature variability in hurricanes: implications with respect to intensity change *Mon. Weather Rev.* **131** 1783–96
DeMaria M, Franklin J L, Onderlinde M J and Kaplan J 2021 Operational forecasting of tropical cyclone rapid intensification at the National Hurricane Center *Atmosphere* **12** 683
DeMaria M, Sampson C R, Knaff J A and Musgrave K D 2014 Is tropical cyclone intensity guidance improving? *Bull. Am. Meteorol. Soc.* **95** 387–98
Doswell C, Davies-Jones R and Keller D L 1990 On summary measures of skill in rare event forecasting based on contingency tables *Weather Forecast.* **5** 576–85
Emanuel K A 1999 Thermodynamic control of hurricane intensity *Nature* **401** 665–9
Emanuel K A 2017 Will global warming make hurricane forecasting more difficult? *Bull. Am. Meteorol. Soc.* **98** 495–501
Eusebi R, Vecchi G A, Lai C Y and Tong M 2024 Realistic tropical cyclone wind and pressure fields can be reconstructed from sparse data using deep learning *Commun. Earth Environ.* **5** 8
Gall R, Franklin J, Marks F, Rappaport E N and Toepfer F 2013 The hurricane forecast improvement project *Bull. Am. Meteorol. Soc.* **94** 329–43
Giffard-Roisin S, Yang M, Charpiat G, Kumler Bonfanti C, Kégl B and Monteleoni C 2020 Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data *Front. Big Data* **3** 1
Griffin S M, Wimmers A and Velden C S 2022 Predicting rapid intensification in North Atlantic and Eastern North Pacific tropical cyclones using a convolutional neural network *Weather Forecast.* **37** 1333–55
Grodsky S A, Reul N, Lagerloef G, Reverdin G, Carton J A, Chapron B, Quilfen Y, Kudryavtsev V N and Kao H-Y 2012 Haline hurricane wake in the Amazon/Orinoco plume: AQUARIUS/SACD and SMOS observations *Geophys. Res. Lett.* **39** L20603
Han H, Wang W-Y and Mao B-H 2005 *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning* (Springer) (*Lecture Notes in Computer Science*) pp 878–87
Hazelton A T, Hart R E and Rogers R F 2017 Analyzing simulated convective bursts in two Atlantic hurricanes. Part II: intensity change due to bursts *Mon. Weather Rev.* **145** 3095–117
Hendricks E A, Peng M S, Fu B and Li T 2010 Quantifying environmental control on tropical cyclone intensity change *Mon. Weather Rev.* **138** 3243–71
Jiang H 2012 The relationship between tropical cyclone intensity change and the strength of inner-core convection *Mon. Weather Rev.* **140** 1164–76
John E B, Balaguru K, Leung L R, Foltz G R, Hetland R D and Hagos S M 2023 Intensification of hurricane sally (2020) over the Mississippi river plume *Wea. Forecast.* **38** 1391–404
Kaplan J *et al* 2015 Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models *Wea. Forecast.* **30** 1374–96
Kaplan J and DeMaria M 2003 Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin *Wea. Forecast.* **18** 1093–108
Kaplan J, DeMaria M and Knaff J A 2010 A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins *Wea. Forecast.* **25** 220–41
Ko M-C, Chen X, Kubat M and Gopalakrishnan S 2023 The development of a consensus machine learning model for hurricane rapid intensification forecasts with hurricane weather research and forecasting (HWRF) data *Weather Forecast.* **38** 1253–70
Meng F, Yao Y, Wang Z, Peng S, Xu D and Song T 2023 Probabilistic forecasting of tropical cyclones intensity using machine learning model *Environ. Res. Lett.* **18** 044042
Molinari J and Vollaro D 2010 Rapid intensification of a sheared tropical storm *Mon. Weather Rev.* **138** 3869–85
Narayanan A, Balaguru K, Xu W and Leung L R 2023 A new method for predicting hurricane rapid intensification based on co-occurring environmental parameters *Nat. Hazards* **120** 881–99
Niculescu-Mizil A and Caruana R 2005 Predicting good probabilities with supervised learning *Proc. 22nd Int. Conf. on Machine Learning* (*Bonn, Germany*) vol 5 pp 625–32
Nolan D S and Grasso L D 2003 Nonhydrostatic, three-dimensional perturbations to balanced, hurricane-like vortices. Part II: symmetric response and nonlinear simulations *J. Atmos. Sci.* **60** 2717–45
Nolan D S, Moon Y and Stern D P 2007 Tropical cyclone intensification from asymmetric convection: energetics and efficiency *J. Atmos. Sci.* **64** 3377–405
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O and Duchesnay É 2011 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825–30 (available at: http://jmlr.org/papers/v12/pedregosa11a.html)
Price J F 1981 Upper ocean response to a hurricane *J. Phys. Oceanogr.* **11** 153–75
Rappaport E N *et al* 2009 Advances and challenges at the National Hurricane Center *Weather Forecast.* **24** 395–419
Rappaport E N, Jiing J-G, Landsea C W, Murillo S T and Franklin J L 2012 The joint hurricane test bed: its first decade of tropical cyclone research-to-operations activities reviewed *Bull. Am. Meteorol. Soc.* **93** 371–80
Roebber P J 2009 Visualizing multiple measures of forecast quality *Wea. Forecast.* **24** 601–8
Rogers R F, Reasor P D and Zhang J 2015 Multiscale structure and evolution of Hurricane Earl (2010) during rapid intensification *Mon. Weather Rev.* **143** 536–62
Saerens M, Latinne P and Decaestecker C 2002 Adjusting the outputs of a classifier to new *a priori* probabilities: a simple procedure *Neural Comput.* **14** 21–41
Shapiro L J and Willoughby H E 1982 The response of balanced hurricanes to local sources of heat and momentum *J. Atmos. Sci.* **39** 378–94
Stevenson S N, Corbosiero K L and Molinari J 2014 The convective evolution and rapid intensification of Hurricane Earl (2010) *Mon. Weather Rev.* **142** 4364–80
Su H, Wu L, Jiang J H, Pai R, Liu A, Zhai A J, Tavallali P and DeMaria M 2020 Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast

with machine learning *Geophys. Res. Lett.* **47** e2020GL089102
Vabalas A, Gowen E, Poliakoff E and Casson A J 2019 Machine learning algorithm validation with a limited sample size *PLoS One* **14** e0224365
Willoughby H E 1990 Temporal changes of the primary circulation in tropical cyclones *J. Atmos. Sci.* **47** 242–64
Xu W, Balaguru K, August A, Lalo N, Hodas N, DeMaria M and Judi D 2021 Deep learning experiments for tropical cyclone intensity forecasts *Wea. Forecast.* **36** 1453–70
Yang W, Huang X, Fei J, Ding J and Cheng X 2024 Applying weighted salinity stratification to rapid intensification prediction of tropical cyclone with machine learning *Earth Space Sci.* **11** e2023EA002932