Development and Evaluation of an Evolutionary Programming-Based Tropical Cyclone Intensity Model

JESSE D. SCHAFFER, PAUL J. ROEBBER, AND CLARK EVANS

Atmospheric Science Program, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin

(Manuscript received 16 October 2019, in final form 27 February 2020)

ABSTRACT

A statistical–dynamical tropical cyclone (TC) intensity model is developed from a large ensemble of algorithms through evolutionary programming (EP). EP mimics the evolutionary principles of genetic information, reproduction, and mutation to develop a population of algorithms with skillful predictor combinations. From this evolutionary process the 100 most skillful algorithms as determined by root-mean square error on validation data are kept and bias corrected. Bayesian model combination is used to assign weights to a subset of 10 skillful yet diverse algorithms from this list. The resulting algorithm combination produces a forecast superior in skill to that from any individual algorithm. Using these methods, two models are developed to give deterministic and probabilistic forecasts for TC intensity every 12 h out to 120 h: one each for the North Atlantic and eastern and central North Pacific basins. Deterministic performance, as defined by MAE, exceeds that of a ''no skill'' forecast in the North Atlantic to 96 h and is competitive with the operational Statistical Hurricane Intensity Prediction Scheme and Logistic Growth Equation Model at these times. In the eastern and central North Pacific, deterministic skill is comparable to the blended 5-day climatology and persistence (CLP5) track and decay-SHIFOR (DSHF) intensity forecast (OCD5) only to 24 h, after which time it is generally less skillful than OCD5 and all operational guidance. Probabilistic rapid intensification forecasts at the 25–30 kt $(24\,\text{h})^{-1}$ thresholds, particularly in the Atlantic, are skillful relative to climatology and competitive with operational guidance when subjectively calibrated; however, probabilistic rapid weakening forecasts are not skillful relative to climatology at any threshold in either basin. Case studies are analyzed to give more insight into model behavior and performance.

1. Introduction

Tropical cyclone (TC) intensity forecasting is recognized as being particularly challenging with only slow improvements over recent years, especially at shorter lead times (Fig. 1). This lack of improvement is even more dramatic when the time series is placed alongside track errors, which are improving at 3 times the rate of intensity errors over the 24–72 h range (DeMaria et al. 2014). At shorter lead times, intensity errors are dominated by the mischaracterization of the TC's initial intensity, as well as by inner-core and eyewall processes due to our limited understanding of and ability to resolve such processes (Emanuel and Zhang 2016, 2017; Kieu and Moon 2016). Furthermore, the challenge of forecasting the magnitude and timing of rapid intensification (RI) and rapid weakening (RW) significantly contributes to large absolute forecast errors and overall forecast difficulty at shorter lead times (Rappaport et al. 2012; Kaplan et al. 2010).

Despite these challenges, some improvement in TC intensity forecasts have occurred. While improvement rates of 1%–2% yr^{-1} (as has occurred over the 24–72 h lead time from 1989 to 2012) may seem negligible, they are nonetheless statistically significant (DeMaria et al. 2014). Furthermore, at lead times longer than 72 h, more substantial improvements have occurred, with rates averaging 2%–4% yr^{-1} (DeMaria et al. 2014). However, this 2%–4% increase is largely attributed to better track forecasts, which have had similar improvement rates over the same time period, as better track predictions lead to more accurate forecasts of the environmental conditions within which a TC is embedded (Emanuel et al. 2004; DeMaria et al. 2014; Emanuel and Zhang 2016). Yet, the view that TC intensity forecasts have not improved quickly enough (Gopalakrishnan et al. 2011; Rappaport et al. 2012; DeMaria et al. 2014; Emanuel and Zhang 2016) is still indicative that these increases

Corresponding author: Dr. Clark Evans, evans36@uwm.edu.

DOI: 10.1175/MWR-D-19-0346.1

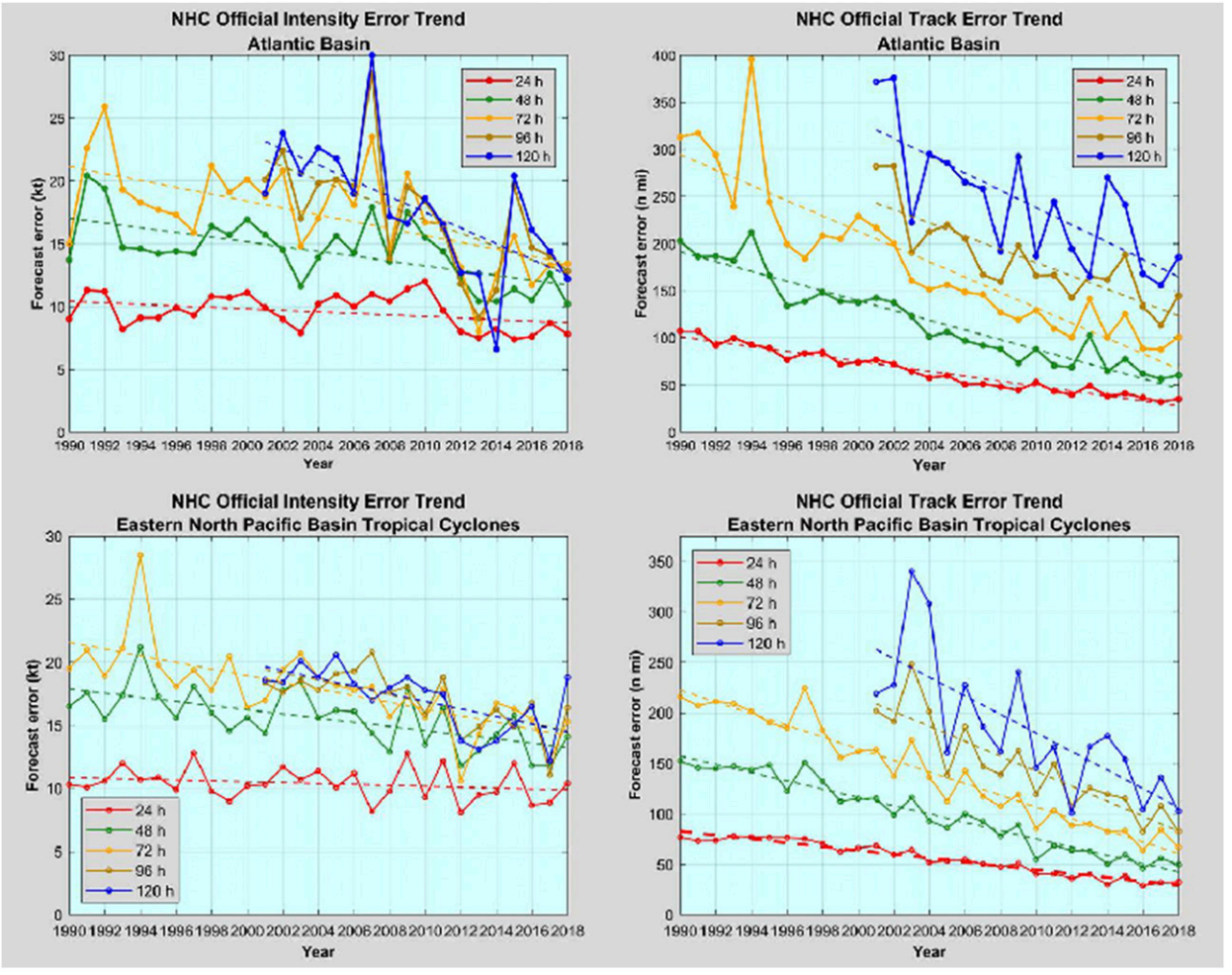

FIG. 1. Annual average of official NHC (top left) Atlantic-basin intensity errors, (top right) Atlantic-basin track errors, (bottom left) eastern North Pacific-basin intensity errors, and (bottom right) eastern North Pacific-basin track errors for the period 1990–2017, each as a function of forecast lead time (colored lines) with least squares lines (dashed) for each lead time superimposed (Cangialosi 2019).

may be too small to produce significant *practical* advantages for emergency management preparation, planning, and decision-making.

Deterministic TC intensity forecasts are typically divided into three different types of models: dynamical, statistical–dynamical, and consensus. Whereas dynamical (or numerical weather prediction) models predict TC intensity by solving the governing equations and appropriately parameterizing other processes (e.g., cloud microphysics, radiative transfer, turbulence, surface energy fluxes, etc.), statistical–dynamical models use statistical methods to assign appropriate weighting to empirical relationships derived from environmental and TC structure characteristics obtained from dynamical models and/or observations. Last, consensus models combine intensity forecasts from multiple models, whether dynamical and/or statistical–dynamical, and use a variety of methods to derive the weights (e.g., equal vs variable) for the selected models. To date, consensus models have outperformed the other intensity model types in the Atlantic and eastern and central North Pacific basins, but they are followed closely by statistical–dynamical models and recently by the best-performing dynamical models (Stewart 2014, 2016; Pasch 2015). Probabilistic TC intensity forecasts are almost exclusively derived from statistical–dynamical models (e.g., Kaplan and DeMaria 2003; Kaplan et al. 2010, 2015; Rozoff and Kossin 2011; Cloud et al. 2019), although recent attempts to use dynamical-model ensembles to predict RI shown promise (e.g., Alessandrini et al. 2018).

Here, we develop two statistical–dynamical TC intensity models, with one forecasting TC intensity, RI, and RW for the North Atlantic basin and the other doing so for the eastern and central North Pacific basins. The process for developing each model is identical with the only distinction being that the Atlantic model is

trained on data from the North Atlantic basin, while the Pacific model is trained on data from the eastern and central North Pacific basins.

The models are developed through a statistical–dynamical approach in which each model is derived from a large ensemble of algorithms, which are themselves generated through the process of evolutionary programming (EP; Fogel 1999). EP utilizes the evolutionary principles of reproduction and mutation to develop, through selective pressure, predictor combinations that maximize forecast skill. EP-generated predictor combinations have shown superior performance over dynamical models in 500-hPa height forecasts (Roebber 2013) and statistical–dynamical models like model output statistics in minimum 2-m temperature forecasts (Roebber 2010, 2015a,b). Furthermore, EP-generated algorithms provide forecast probability density functions (PDFs) superior in probabilistic and deterministic skill than many traditional models in 500-hPa height forecasts, particularly at the tail ends of the distribution (Roebber 2013). Recently, Roebber and Crockett (2019) developed a new approach to EP using a coevolution predator–prey ecosystem and applied it to both 72-h 2-m temperature forecast as well as 60-min nowcasts of convective occurrence. This new formulation incorporates competition between algorithms in a simulated ecosystem, wherein algorithms behave as members of a particular species, and their ultimate evolutionary success is tied to their ability to provide skillful forecasts. This new formulation shows improvements over not only numerical weather prediction forecasts, but also earlier EP approaches applied to the same data. The performance of the EP applications described above is due to EP being developed specifically to produce large-ensemble forecasts with a high degree of heterogeneity amongst the algorithms (Fogel 1999).

What does the EP approach bring to the TC intensity-change prediction problem relative to existing methods, particularly deterministic statistical–dynamical models such as the Statistical Hurricane Intensity Prediction Scheme (SHIPS; DeMaria and Kaplan 1994, 1999; DeMaria et al. 2005) and Logistic Growth Equation Model (LGEM; DeMaria 2009) or probabilistic RI models such as the SHIPS-Rapid Intensification Index (SHIPS-RII; Kaplan and DeMaria 2003; Kaplan et al. 2010) and logistic and Bayesian models (Rozoff and Kossin 2011; Kaplan et al. 2015)? Although the EP method as formulated here relies on large-scale cyclone and environment characteristics as does SHIPS and LGEM and is composed of primarily linear predictor combinations as is SHIPS; it has a more flexible algorithm formulation that allows particular terms to execute only if a certain criterion is met and thus is more responsive to specific cyclone and environmental attributes. Further, it is straightforward to diagnose the contributions from each predictor to the EP intensity-change forecast at each lead time, an advantage that it shares with SHIPS and that is increasingly desirable in meteorological applications of statistical and machine-learning approaches (e.g., McGovern et al. 2019). Unlike existing deterministic and probabilistic models, which are independent of each other, the approach used here results in the development of internally consistent deterministic *and* probabilistic forecast models, including the first RW model to our knowledge, although we note that this is an attribute of the overall methodology rather than EP itself. Despite not being applied here, the EP method can continually and independently (i.e., without human intervention) adapt to new and/or improved information without redeveloping the predictive model (Roebber 2015a), an attractive attribute for operational applications. Perhaps most importantly, the EP method can provide skillful and independent deterministic and probabilistic forecasts that, in turn, may contribute to improved skill for the consensus methods that are currently associated with the highest forecast skill.

The rest of this paper is structured as follows. Section 2 is broken into five parts describing the data used to train the model, EP and the training process itself, the postprocessing techniques used to generate the models' final structures, illustrations of the final models' interpretability, and the operational implementation and associated model verification methods. The deterministic and probabilistic performance of the model for each basin along with illustrative case studies are presented in section 3. The paper closes in section 4 with a summary and conclusion.

2. Data and methods

a. Data

TC intensity and predictor data for training both models are sourced from the SHIPS developmental database, which contains 0-h analysis data in 6-h intervals for all classified TCs (here including both tropical and subtropical cyclones, the latter of which make up <5% of the dataset, at all classified intensities). Only data from TCs for 2000 to 2016 are used, since 2000 coincides with the start of the period when variables are derived from the Global Forecast System (GFS; NCEP 2016) rather than the Climate Forecast System Reanalysis (CFSR; Saha et al. 2010). Forecast predictor values for TCs from the 2017 and 2018 seasons are used for independent testing, as described in more detail at the end of this subsection. Atlantic and eastern and

TABLE 1. List of chosen predictor variables used in the EP model.

DELV	Change in TC intensity over the prior 12 h
CD26	Climatological depth of 26°C isotherm beneath the TC from 2005–10 NCODA analysis
U20C	200 hPa zonal wind (over a 0–500 km radius from the center of the TC)
D200	200 hPa divergence (over a 0–1000 km radius from the center of the TC)
TWAC	Azimuthally averaged 850 hPa tangential wind (over a 0–600 km radius from the center of the TC) from the NCEP analysis
SHDC	850–200 hPa shear magnitude, computed with the TC vortex removed and averaged over a 0–500 km radius relative to 850 hPa vortex center
VMPI	Maximum potential intensity at the TC location, as determined from Kerry Emanuel's MPI equation
CFLX	Dry-air predictor based on the difference in surface moisture flux between air with the observed (GFS) RH value, and with RH of air mixed from 500 hPa to the surface, at the TC location
CONS	Constant value of 10

central Pacific TCs during the 2017–18 seasons are representative of a wide range of TC origination locations, tracks, and intensity evolutions, as objectively assessed using the performance of no-skill climatology and persistence-based track and intensity models for each TC relative to their respective long-term means (Cangialosi 2018, 2019), such that the EP model performance reported on herein is not believed to be specific to only the 2017–18 testing data. However, verifying the model's performance for a lengthier independent period for which the data do not yet exist is needed to evaluate this statement.

The SHIPS dataset contains numerous predictors, but when more predictors are kept, the solution space that must be explored grows larger. This potentially compromises the skill of the algorithms that result from the training and validation process, as it may be hard to search the solution space completely due to a lack of training information (Bellman 1961). While there is no preferred method to determine when the solution space is of optimal size, it is generally desirable to reduce the number of predictors. Here, this reduction of variables is done through a combination of linear correlation analysis, where we require that no two variables be correlated above 0.8, and domain expertise, which we use to remove variables with a presumed lesser influence on TC intensity. This process initially results in a selection of 34 variables (not shown). However, based on initial performance evaluations of the model, and given that only ~6000 cases are available for training, we concluded that the dimensionality of the problem was still too large. Therefore, the 34 predictor variables are separated into groupings of similar properties (e.g., thermodynamics, moisture, shear) and domain expertise is used to subjectively select a single representative variable from each group. The resulting selection of eight variables (Table 1) yields improved performance over that derived from the larger dataset (not shown). We note, however, that we did not attempt selecting different combinations of eight predictors from the variable groupings [indeed, there are $O(10^6)$ potential combinations, such that assuming equal likelihood to any given combination being the most skillful, it is simply untenable to search through this entire combination space], so we cannot assert that this is the most optimal model obtainable using this approach.

Of the retained variables, all but one are converted into standardized anomalies (Grumm and Hart 2001; here computed relative to the predictor value means over the full 2000–16 dataset) to aid direct comparisons between variables with dissimilar units. However, one predictor, the 0–600 km-averaged symmetric tangential wind at 850 hPa from the National Centers for Environmental Prediction (NCEP) analysis (TWAC) is notably non-Gaussian (not shown) and is instead converted to a linear scaling from −1 to 1, with the extremes representing the maximum and minimum values of TWAC in the training dataset (Roebber 2010, 2013, 2015a,b). Last, a constant value of 10 is provided as a ninth potential predictor, the purpose of which is explained in section 2b when discussing the algorithm structure.

Once the desired variables are chosen, the dataset is processed to remove cases that fall into either of two categories: the case features missing predictor information, or the case initializes or verifies over land. There are several reasons why it is beneficial to remove the cases with the missing predictor information. First, climatological values in terms of standard anomalies are zero, thus using such values to replace missing data may significantly alter a forecast. Because the algorithms generally feature nonlinear relationships between variables, even small changes in inputs can lead to large changes in the forecast. Last, the model cannot be run operationally with missing predictor information, and thus removing these cases ensures consistency with operational practice (and the formal verification described in section 2e). Meanwhile, cases where the TC forecast initializes or verifies over land are removed, since an inland decay model is used in operational practice to

postprocess the intensity forecast over land and account for inland wind decay (Kaplan and DeMaria 1995, 2001; DeMaria et al. 2006).

With the culling of problematic cases from the dataset complete, the remaining TC cases are assigned to one of three datasets: training (two-thirds of the data), validation (one-sixth), or independent testing (one-sixth). However, the dynamical and empirical relationships between predictors may vary with intensity. Therefore, if the training dataset is biased toward TCs of a particular intensity relative to climatology, the potential exists for the algorithms to be calibrated toward only a subset of all TC intensities. Consequently, to mitigate against such an intensity bias, each TC in the dataset is separated into three intensity classes based on its lifetime maximum-achieved intensity: tropical depressions and tropical storms, weak hurricanes (lifetime maximum-achieved sustained 10-m winds of 33–49 $m\,s^{-1}$ or 64–95 kt; $1\,kt \approx 0.5144\,m\,s^{-1}$), or major hurricanes (lifetime maximum-achieved sustained 10-m winds of $>49\,m\,s^{-1}$ or 95 kt). TCs and all their respective forecasts are then pulled from each of these three intensity classes to form the training, validation, and independent testing datasets, with the relative proportions of cases from each intensity class being identical between datasets. Lifetime maximum intensity is used instead of the instantaneous best track intensity to populate the three intensity classes primarily for ease, with the result being that all forecasts for a given TC are contained within a single class. However, successive forecasts for an individual TC are serially correlated, thus reducing the *effective* sample sizes for each of the training, validation, and testing datasets through this methodology. Future research is planned to evaluate the potential benefit from using the instantaneous best track intensity to populate the three intensity classes.

Last, while the training dataset contains analysis values of the predictors at all future lead times, in reality the future values of these variables are unknown and must be forecast. This produces uncertainty and inaccuracy in the real-time input variables as compared to the analysis variables used in training, and consequently real-time performance can be expected to be to be worse than training performance. Therefore, to prevent overfitting of the idealized relationships between the analysis variables and to simulate uncertainty in the forecast values during the training process, noise is added to the analysis values. The magnitude of this noise is specified by comparing the differences between analysis values and archived real-time 12-h forecast values for homogenous training and validation cases across the 2010–16 seasons (based on archived data availability). Since the 12-h forecast error distributions are approximately normal about means of zero (not shown), the applied perturbations are randomly drawn from a Gaussian distribution centered on zero that has a standard deviation equal to an empirically derived value of one-quarter of the observed standard deviation in the differences between the analysis and forecast values across homogeneous cases. This noise is dynamic, meaning that each time the algorithms forecast for a new case during the training process, the added noise is changed. However, noise is not added to the validation forecasts to ensure that the algorithms themselves do not overfit to the noise. A comparison of model performance across independent testing cases when utilizing perturbed analysis variables as model inputs versus real-time, 12-h GFS-predicted forecast variables showed similar performance between the two sets of forecasts, suggesting that adding noise is having the desired effect (not shown).

b. Evolutionary programming

From this curated dataset, a large ensemble of algorithms is generated via a perfect-prognostic approach using the evolutionary principles of cloning, mutation, and selective pressure to determine the empirical relationships between the selected predictor variables and TC intensity. These algorithms are trained to forecast a 12-h adjustment to a persistence forecast using predictor values valid at the end of the specified 12-h interval. The exception to this is the DELV predictor value, which is specified as the change in intensity over the 12 h prior to the forecast interval.

As in previous studies in which EP is applied to weather forecasts (e.g., Roebber 2010, 2013, 2015a,b), the basic genetic architecture of a single algorithm is a summation of if-then equations, which can be written most generally in the following form:

$$F = \varepsilon + \sum_{i=1}^{5} \text{IF} \quad (V_{i1} R_{i1} V_{i2}) \quad \text{THEN} \\ (C_{i1} V_{i3}) O_{i2} (C_{i2} V_{i4}) O_{i3} (C_{i3} V_{i5}), \tag{1}$$

where V_{ij} is any of the predictor variables in Table 1, R_{i1} is a relational operator ($\leq$ or $>$), C_{ij} are real-valued constants ranging from $[-1, 1]$, and O_{ij} are either of the arithmetic operators $+$ or $\times$; ε is the bias-correction factor, which is zero through the training process and set during postprocessing (section 2c). While conditional statements and the potential for both linear and nonlinear predictor combinations allow for flexible algorithms, the imposed structure maintains interpretability since the logic can be connected to dynamical processes familiar to forecasters (section 2d). While earlier studies used a summation of 10 if–then statements (Roebber 2010, 2013, 2015a,b), the use of 5 statements here is an

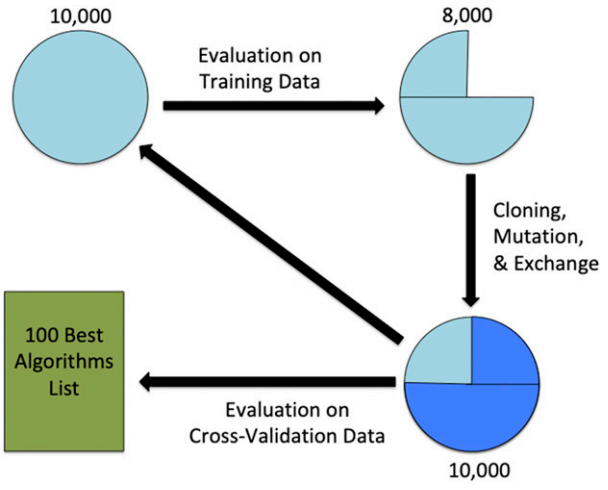

FIG. 2. Schematic overview of the EP training process. (top left) An initial population of 10 000 algorithms is randomly generated, (top) which then forecast on training data. (top right) Their performance is evaluated, at which time the top 2000 are left unchanged, the bottom 2000 are replaced by cloned and mutated versions of the top 2000, and (right) the middle 6000 undergo an information exchange and mutation. (bottom right) The resulting population of 10 000 algorithms then forecasts on the validation dataset, and (bottom) the 100 best-performing algorithms are retained to generate the initial best-performing algorithms list. The process then repeats for 300 iterations and five randomly initialized populations, with the best-performing algorithms list updated rather than entirely replaced at subsequent stages. Please refer to section 2b for more details.

empirical choice to balance computational expense and model skill since many algorithms in the previous studies featured multiple conditional statements that never executed.

Previously, it was mentioned that one of the input variables could be a constant value of 10. This value allows the EP process to generate lines within the algorithm that always or never execute, as is deemed necessary by the evolutionary process, since no variable in the training dataset exceeds and no variable in theory should exceed ±10 standard deviations from its climatological mean. Additionally, this value of 10 provides an additive or multiplicative scaling factor, if deemed necessary by the training process, to use when calculating the adjustment based on one or more predictors.

The EP training process for both the North Atlantic and eastern and central North Pacific basins starts with a randomly initialized population of 10 000 algorithms (top left of Fig. 2). While the population size is somewhat arbitrary and could be increased, prior experimentation has shown that the improved skill from larger populations is minimal and does not compensate adequately for the increase in computational time (Roebber 2016). The algorithms then perform an initial forecast on the training dataset to determine their fitness/skill (top middle of Fig. 2), and the worst 2000 performing algorithms [as determined by root-mean-square error (RMSE), which is chosen as the performance criterion to better address large-error cases in the initial algorithm development process] are eliminated (top right of Fig. 2).

The next generation of algorithms is then produced in what is referred to as the "evolutionary step" (right side of Fig. 2). The process starts by cloning the 2000 best-performing algorithms (also determined by RMSE), which returns the population to its full capacity of 10 000 algorithms. The 2000 clones each then undergo a mutation, wherein one of its five lines is randomly selected, completely erased, and refilled with randomly selected coefficient, predictor, arithmetic, and relational operators (all subject to the rules described above). The 6000 middle-performing algorithms undergo a process of swapping genetic information in which each algorithm swaps the entirety of one of its five lines with another randomly selected algorithm. After swapping genetic information, these middle-performing algorithms also undergo a mutation in the same manner as the cloned algorithms. The best 2000 performing algorithms are left untouched in order to provide a source of good genetic information for future generations. At this point, the evolutionary step is complete, and the population of algorithms is in its second generation (bottom right of Fig. 2).

This new generation of algorithms then forecasts for the validation dataset. The 100 best-performing algorithms from this generation are used for the initial listing of the "best algorithms list" (bottom of Fig. 2). The process described above then repeats for 300 generations, after each of which the best algorithm list is updated to include any new algorithms with RMSE below that of the worst performers on the top-100 list (with those poorer performers being removed). This method ensures that the best-performing algorithms are kept, no matter the generation in which they occur, rather than simply selecting the best-performing algorithms from the final generation.

Although the performance of the algorithms improves rapidly in the first few generations, the rate of improvement eventually plateaus with only small improvements found in the worst-performing algorithms toward the end of 300 generations (not shown). Therefore, after 300 generations, an altogether new population of 10 000 algorithms is randomly initialized, from which the same training and validation process described above is followed. The algorithms from this second population are considered for inclusion on the same "best algorithms list" from the previous population. Altogether, five different

populations of 10 000 algorithms are run for 300 iterations each to produce a final set of 100 algorithms on the "best algorithms list." Note that it is theoretically possible for two or more identical algorithms to appear on the best algorithms list; however, in practice, this is extremely unlikely (and did not occur in this study) given the randomness inherent to the initialization and mutation evolutionary stages and the size of the parameter space considered.

c. *Bayesian model combination*

While each individual algorithm on the final 100 best-algorithms list constitutes a TC intensity model, statistical postprocessing techniques can be used to achieve improved performance and reliability relative to any individual algorithm. Here, performance on the training and validation cases is first used to bias correct the 100 best-performing algorithms, which enables setting of the value of ε in (1) for each algorithm. Since the training and validation cases are not ordered chronologically (recall that cases were sorted based on maximum TC intensity), we use a simple bias correction (average error of the training and validation cases) rather than weight decay, as used in other EP studies (e.g., Roebber 2018). The correction calculated for those cases is then applied uniformly to all test cases.

Next, Bayesian model combination (BMC) is used to assign weights to an ensemble of multiple algorithms, such that their weighted combination results in a forecast that is superior in skill to that from any individual algorithm (Monteith et al. 2011; Roebber 2018 and references therein). However, a limitation of BMC is that it is computationally expensive (which scales with n^x, where n is the number of possible weights and x is the number of algorithms considered), and therefore members must be subselected from the overall population (e.g., Hoeting et al. 1999), even as the "best algorithms list" is already a subset of the wider population of all algorithms. In this study, 10 algorithms from the bias-corrected 100-best-performing algorithm list, as chosen to minimize mean absolute error (MAE) and maximize mean absolute difference (MAD) across the set of bias-corrected forecasts for training and validation cases, are selected for processing by BMC. This gives a subset of algorithms that are both skillful and diverse.

This subselection is performed as follows. First, the 100 best algorithms are ranked according to MAE. Then, starting from the best performer by MAE and moving down the 100-best-performing-algorithm list toward the worst performer in sequential order, the MAD of the algorithm under consideration is compared against that of all other algorithms that have been added to what will be the final list of 10 algorithms. If that algorithm has a MAD with any other algorithm on that list that is below a certain specified threshold (i.e., it is in some sense too similar to another algorithm; here, this difference is arbitrarily set to 0.9 kt), it is rejected. This process is followed until 10 algorithms are obtained.

After the 10 algorithms have been identified, the BMC process loops through all possible combinations of the 10 members using four raw weights (0, 1, 2, 3), with the sum of the weights normalized to equal one, under the requirement that at least one algorithm receives a nonzero raw weight. Thus, the minimum nonzero normalized weight that can be obtained is $1/(1 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3)$ or 1/28. In that instance, the other nine algorithms would have normalized weights of 3/28. Similarly, the maximum weight that is not unity (which occurs when nine algorithms have raw weights of 0 and the tenth is given any nonzero weight) is 3/4. A wide range of algorithm weights are thus obtainable through this procedure. A discussion of the Bayesian attributes of the weight-determination process is given by Monteith et al. (2011). Their procedure is followed in this study, but under the condition that a model combination is estimated to be correct provided the forecast provided is better than or equal to a persistence forecast.

The final weighting used for the model is the one that minimizes MAE across the validation dataset, with MAE chosen over RMSE to mirror National Hurricane Center (NHC) operational performance evaluation metrics for TC intensity forecasts (Cangialosi 2019). Although we acknowledge the discrepancy that this causes with respect to the deterministic EP model training and validation process, wherein RMSE is used as the performance criterion, the deterministic forecast skill of a version of the EP model in which RMSE is used as the BMC performance criterion is less than that of the MAE-based version (not shown). As the goal at this stage is to produce a deterministic model that minimizes the operational forecast skill metric of MAE over large forecast samples, many of which are dominated by small-error cases (especially at short lead times), we believe that using MAE as the performance criterion is warranted. The selection of only 10 algorithms is deemed sufficient as, in practice, multiple algorithms typically receive a weighting of zero, indicating that the BMC process identifies that more algorithms are present than are necessary. In fact, this is the case for both the North Atlantic and eastern and central North Pacific models herein: only seven algorithms are retained with nonzero weights for the Atlantic, whereas only two algorithms are

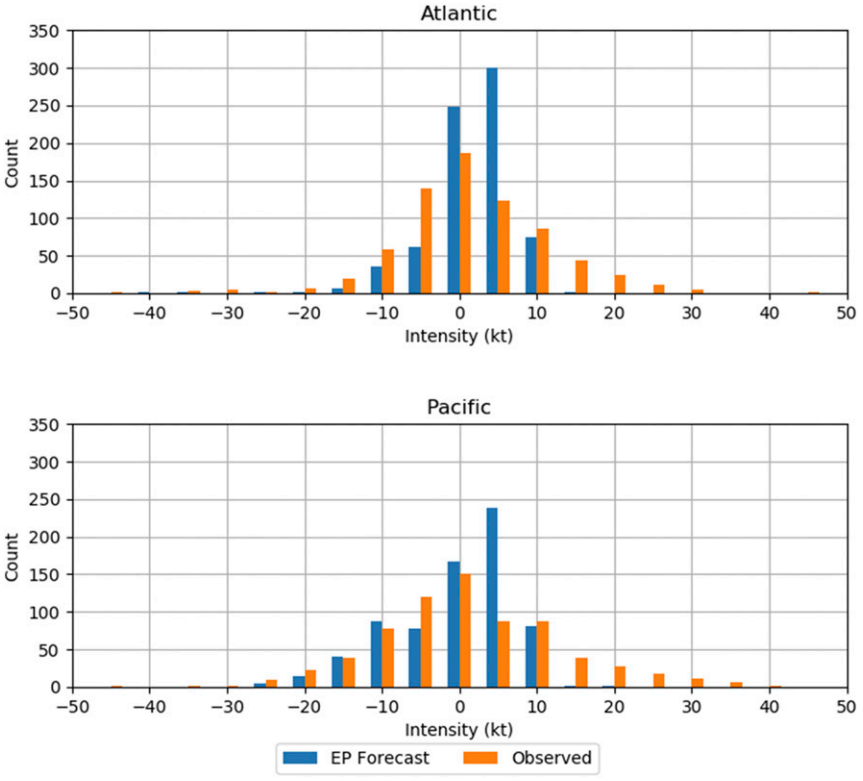

FIG. 3. Histogram of observed (orange) and forecast (blue) 12-h intensity changes (rounded to the nearest 5 kt, consistent with operational classification practices) across the 2000–16 training data for the (top) North Atlantic and (bottom) eastern and central North Pacific basins.

retained with nonzero weights for the eastern and central North Pacific basins (see the appendix). With the weightings of each algorithm obtained, the final deterministic forecast is comprised of a simple weighted sum of the retained bias-corrected algorithms' individual forecasts. A discussion of the Bayesian attributes of the BMC process is given by Monteith et al. (2011).

BMC can also be used to generate a PDF from which probabilistic forecasts of RI and RW can be obtained. To do so, Gaussian distributions with mean equal to the forecast intensity change from each unweighted algorithm and standard deviations determined from the PDF of observed 0–12 h intensity change across the training and validation cases are generated and normalized to obtain PDFs for each algorithm's forecast. These individual PDFs are then weighted by the BMC-derived weightings and summed to give the total normalized forecast intensity-change PDF. The fraction of the PDF that exceeds any of the standard RI/RW intensity-change thresholds provides the uncalibrated forecast RI/RW probability (section 2e). This formulation leads to the probabilistic model's performance characteristics being directly reliant on those of the deterministic model, which is likely a useful yet suboptimal approach for predicting intensity changes on the tails (RI and RW) of the intensity-change distribution depicted in Fig. 3. Alternative approaches that use a probabilistic performance criterion such as Brier skill score (BSS; Brier 1950; Murphy 1973) or the continuous ranked probability score (an extension of the Brier score to multiple thresholds of a continuous predictand; Hersbach 2000) are likely to produce superior probabilistic skill (e.g., Raftery et al. 2005), and future research is planned to explore their viability for RI/RW forecasts.

d. Algorithm interpretability

Predictive models generated using evolutionary programming have the desirable characteristic of interpretability: unlike other machine-learning techniques such as neural networks, it is straightforward to diagnose what predictors form the model and what weights are given to each and thus to quantify the relative contributions of each predictor and algorithm. This subsection demonstrates these attributes, first through

an analysis of the final model for each basin followed by how the relative contributions from each predictor are diagnosed.

In the Atlantic basin, seven algorithms received a nonzero weighting from the BMC process (see the appendix). Of the seven algorithms, six of them include two if-then statements that are always true, such that the equation that follows will always be calculated. Meanwhile, algorithm 53, which has the largest weighting of any of the algorithms, has three if-then statements (algorithm lines 1, 4, and 5) that are always true and a fourth (algorithm line 2) that, so long as D200 is not more than 10 standard deviations above climatology, is also always true. Otherwise, the outcomes of the remaining if-then statements in the Atlantic model are conditioned on the values of the respective predictors. No if-then statements in this model are always false. In the Pacific basin, only two algorithms received a nonzero weighting from the BMC process (see the appendix). As for the Atlantic basin, two if-then statements in both algorithms are always true, and there are no if-then statements that are always false.

The physical interpretability of each algorithm is best illustrated through an example. Consider the scenario of an intensifying east Pacific TC and focus on the first line in algorithm 69 of the Pacific model (see the appendix). The if-then statement contains DELV and U20C. For an intensifying TC, DELV is large and positive, whereas U20C is likely slightly negative (e.g., easterly shear) to near zero. Thus, this line is likely to execute. The intensity adjustment itself depends on the product of maximum potential intensity (VMPI, which is typically positive, or above the basinwide climatology, for intensifying TCs) and the negative of U20C, which is positive. This positive value is then added to the product of a negative coefficient with the 850–200-hPa vertical wind shear magnitude (SHDC), which is typically negative (i.e., below the basinwide climatology) for an intensifying TC. Consequently, the net contribution from this line to the intensity forecast from this algorithm is positive, indicating a forecast of continued intensification (albeit potentially offset by the remaining lines of this algorithm and contributions from the other algorithm). The magnitude of this positive forecast intensity increment depends on the extent to which DELV and VMPI are above climatology and U20C and SHDC are below climatology. This increment is subsequently added to those from the other lines within this algorithm, the result of which is then bias corrected and weighted using the BMC-determined weight applicable to that algorithm.

When considering individual forecasts from the complete model, and not just a single line within a single algorithm as described above, it is useful to know the extent to which certain predictors contributed to the overall intensity-change forecast. Here, the relative contribution from each predictor to the overall forecast is obtained by rerunning the forecast with the variable of interest set to an input value of zero (i.e., a climatological value). The direction and magnitude of the change in the intensity forecast as compared to the original forecast, each as bias-corrected and BMC-weighted as described above, provides a measure of the impact that predictor has on the forecast. For example, if a predictor is set to zero and the resulting intensity forecast decreases by 5 kt, that predictor is said to have had a +5 kt contribution to the original forecast. Conversely, if zeroing out a predictor results in an increased intensity forecast, that predictor is said to have a negative contribution to the original forecast. Since the algorithms forecast for a 12-h intensity change, these relative contributions are calculated only over a 12-h interval. An estimation of the relative contribution at for example, 36 h (as in the operational model application; section 2e), still presumes an accurate 24-h forecast with no zeroing of the variable at the earlier lead times. Thus, the relative contribution of that variable at later lead times is estimated by summing up its individual relative contributions over each 12-h interval.

e. Operational implementation and verification

While the training process produces algorithms that forecast a 12-h intensity change, multiple successive forecasts are required to obtain intensity forecasts beyond 12 h in duration (e.g., every 12 h out to 120 h, currently the longest lead time in NHC operational forecasts). Although the same model is used in each successive 12-h intensity-change forecast, the input values change. Each 12-h adjustment is calculated using predictor values derived from the most-recent GFS forecast fields at the end of each specified 12-h interval, when the intensity forecast verifies. The exception to this again is the DELV predictor; the DELV predictor is an observed value from the NHC working best track analysis for the first 0–12 h intensity-change forecast, whereas the DELV predictor is calculated from EP model outputs at subsequent lead times (e.g., for 12–24 h, DELV is defined as the EP model's predicted intensity change from 0 to 12 h).

To obtain the probabilistic forecasts at each lead time, a new PDF is generated around each of the successive intensity-change forecasts and the probability of RI/RW as compared to the 0-h intensity is calculated at the standard thresholds of $\pm 20\,\text{kt}\ (12\,\text{h})^{-1}$, $\pm 25\,\text{kt}\ (24\,\text{h})^{-1}$, $\pm 30\,\text{kt}\ (24\,\text{h})^{-1}$, $\pm 35\,\text{kt}\ (24\,\text{h})^{-1}$, $\pm 40\,\text{kt}\ (24\,\text{h})^{-1}$, $\pm 45\,\text{kt}\ (36\,\text{h})^{-1}$, $\pm 55\,\text{kt}\ (48\,\text{h})^{-1}$, and $\pm 65\,\text{kt}\ (72\,\text{h})^{-1}$ (Kaplan and DeMaria 2003; Kaplan et al. 2015; Wood and Ritchie 2015).

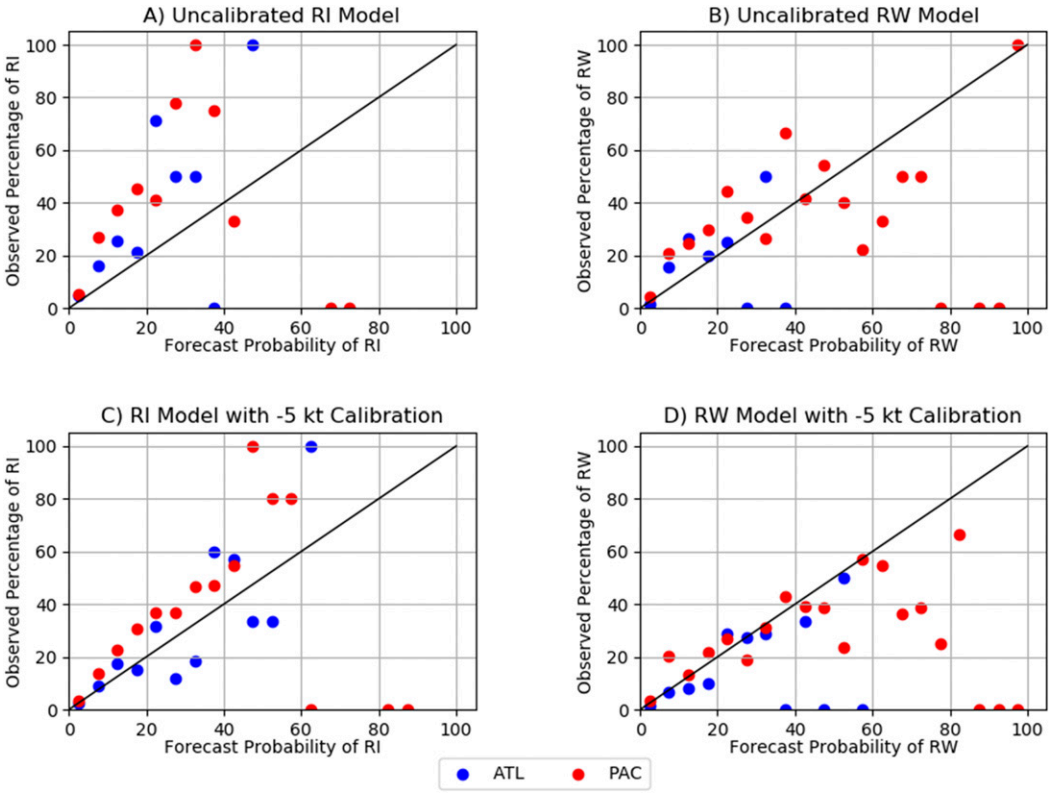

FIG. 4. Reliability of the EP model for the North Atlantic (blue) and east and central North Pacific basins (red) at the 30 kt $(24\,h)^{-1}$ threshold for the (a) uncalibrated RI model, (b) uncalibrated RW model, (c) RI model with a −5 kt calibration applied, and (d) RW model with a −5 kt calibration applied.

The probabilities obtained this way form the *uncalibrated* model; however, the EP models' intensity-change PDFs are insufficiently wide (i.e., underdispersive; Fig. 3), with aggregate forecast RI probabilities that are lower than the verifying observations. Therefore, RI probabilities are subjectively calibrated using probability matching, which is akin to the quantile mapping approach described by Alessandrini et al. (2018). This is best illustrated by an example. Consider the 30 kt $(24\,h)^{-1}$ RI threshold for which the EP model underpredicts RI probabilities in both basins (Fig. 4a). Probability matching involves subjectively determining what model-predicted 24-h intensity-change threshold results in a perfectly reliable (i.e., identical forecast and observed probabilities) forecast at the 30 kt $(24\,h)^{-1}$ RI threshold. In this instance, a model-predicted 24-h intensity change of 25 kt results in much improved forecast reliability relative to the 30 kt $(24\,h)^{-1}$ observed RI threshold (Fig. 4c). This process is then repeated for all RI thresholds. For the RI models in both basins, reliability at each threshold improves when the EP model probabilities are calibrated using a model threshold 5 kt lower than the observed threshold (as in the example described above). The same calibration, however, was not applied to the RW model as the model is generally reliable when uncalibrated (Fig. 4b) and the same −5 kt calibration causes forecast probabilities to be higher than observed (Fig. 4d).

Performance of the deterministic EP models for the Atlantic and eastern and central North Pacific basins is evaluated across forecast fields from the independent 2017–18 seasons. Following standard NHC practice (e.g., Cangialosi 2019), performance is evaluated in terms of MAE and skill normalized relative to that of the no-skill Statistical Hurricane Intensity Forecast model accounting for overland decay [Decay SHIFOR Model Intensity Forecast (DSHF); here referred to as blended-intensity Operational CLIPER 5 (CLP5) and 120-h DSHF (OCD5) since DSHF forms the intensity component of the combined-track-intensity OCD5 forecast, which is based on TC time, position, movement, and intensity and its 12-h change; Knaff et al. 2003]. To place the results into an appropriate context, model performance is compared to that of several of the most-skillful operational intensity models: the 6-h interpolated version of the Hurricane Weather Research and Forecasting

TABLE 2. Climatologies of RI and RW occurrences, and their rates relative to their respective samples of all TCs, across the 2000–16 training cases used to calculate the BSS.

	ATL			PAC		
RI/RW threshold (kt h^{-1})	No. of RI/RW cases	Total No. of cases	Climatological RI/RW rates	No. of RI/RW cases	Total No. of cases	Climatological RI/RW rates
20/12	70/25	3029	2.31%/0.83%	105/74	3318	3.16%/2.23%
25/24	177/57	2624	6.74%/2.17%	291/240	2928	9.93%/8.20%
30/24	100/31	2624	3.81%/1.18%	187/146	2928	6.38%/4.99%
35/24	66/16	2624	2.51%/0.61%	128/84	2928	4.37%/2.87%
40/24	44/7	2624	1.67%/0.27%	94/52	2928	3.21%/1.78%
45/36	78/14	2305	3.38%/0.61%	146/91	2565	5.69%/3.60%
55/48	73/9	2028	3.59%/0.44%	134/74	2241	5.97%/3.30%
65/72	86/3	1583	5.43%/0.19%	98/68	1624	6.03%/4.19%

Model (HWFI; Tallapragada et al. 2014), LGEM, SHIPS, and official (OFCL) and 6-h interpolated (OFCI) NHC official forecasts. Note that the samples for all models are homogenized (i.e., only synoptic times at which all models provided a forecast are retained) and the evaluation considers only overwater cases.

The performance of the probabilistic EP RI and RW models is determined using BSS, which is calculated as a percent improvement over a climatological forecast, defined here as the climatological probabilities of RI and RW at each threshold over the training dataset. These climatological probabilities are given in Table 2. Likewise, performance of the probabilistic models is compared to the SHIPS-RII, logistic model, Bayesian model, and a consensus of the three models' forecasts using a homogeneous forecast set featuring only overwater forecasts across the 2017–18 season.

3. Results

a. Deterministic model performance

Across the 2017–18 seasons, the EP model skill is 5%–19% higher than that of OCD5 through 96 h in the Atlantic basin (Fig. 5a). However, the model fails to exhibit the characteristic plateau in intensity errors beyond 96 h (Fig. 5c), and consequently, EP model skill becomes 11% worse than OCD5 at 120 h. Although model performance lags the best-performing HWFI model and the NHC official forecast, performance through 96 h is statistically indistinguishable from that of both SHIPS and LGEM (Fig. 5c). Although EP model skill closely mirrors that of SHIPS, with which it shares some predictors and conceptual underpinnings, only 25%–50% of the variance (most at shorter lead times, least at longer lead times) in the EP model forecasts can be explained by the corresponding SHIPS model forecasts (not shown), suggesting that the EP model provides independent predictions to those from SHIPS. Meanwhile, the EP model's bias over the 2017–18 Atlantic seasons is comparable to that of the other models considered, with a small negative bias at all forecast lead times (not shown).

In the eastern and central North Pacific basin, the EP model is less skillful than the no-skill OCD5 model prior to 72 h (Figs. 5b,d). At later forecast times, the EP model's MAE is statistically indistinguishable from those of OCD5 and LGEM, albeit over small forecast samples (Fig. 5d). Further, although all models considered are negatively biased at all forecast lead times, the EP model is slightly more negatively biased than other models (not shown). The largest negative bias of −10 kt at the 48-h lead time coincides with the largest MAE and, in general, the bias and MAE of the model mirror each other at all lead times. Insight into model performance for forecasts with particularly large MAE is provided in section 3c.

b. Probabilistic model performance

For RI, the skill (as measured by BSS) of the *uncalibrated* EP model in the Atlantic is approximately equal to that of a climatological forecast at all except the 25–30 kt $(24\,h)^{-1}$ thresholds, at which it is marginally more skillful than climatology (Fig. 6, top). Calibration adds 10%–50% skill to the uncalibrated model skill at all thresholds except the 20 kt $(12\,h)^{-1}$ threshold, at which a significant skill reduction (for unknown reasons) is noted. Atlantic calibrated RI model skill is competitive (here characterized by overlapping 5th–95th percentile forecast ranges determined using bootstrapping) with that of most operational RI models at the 25–40 kt $(24\,h)^{-1}$ thresholds. In the eastern and central North Pacific, *uncalibrated* EP model skill is approximately equal to that of a climatological forecast at all thresholds (Fig. 6, bottom). Calibration again adds 10%–50% skill to the uncalibrated model at all thresholds except the 20 kt

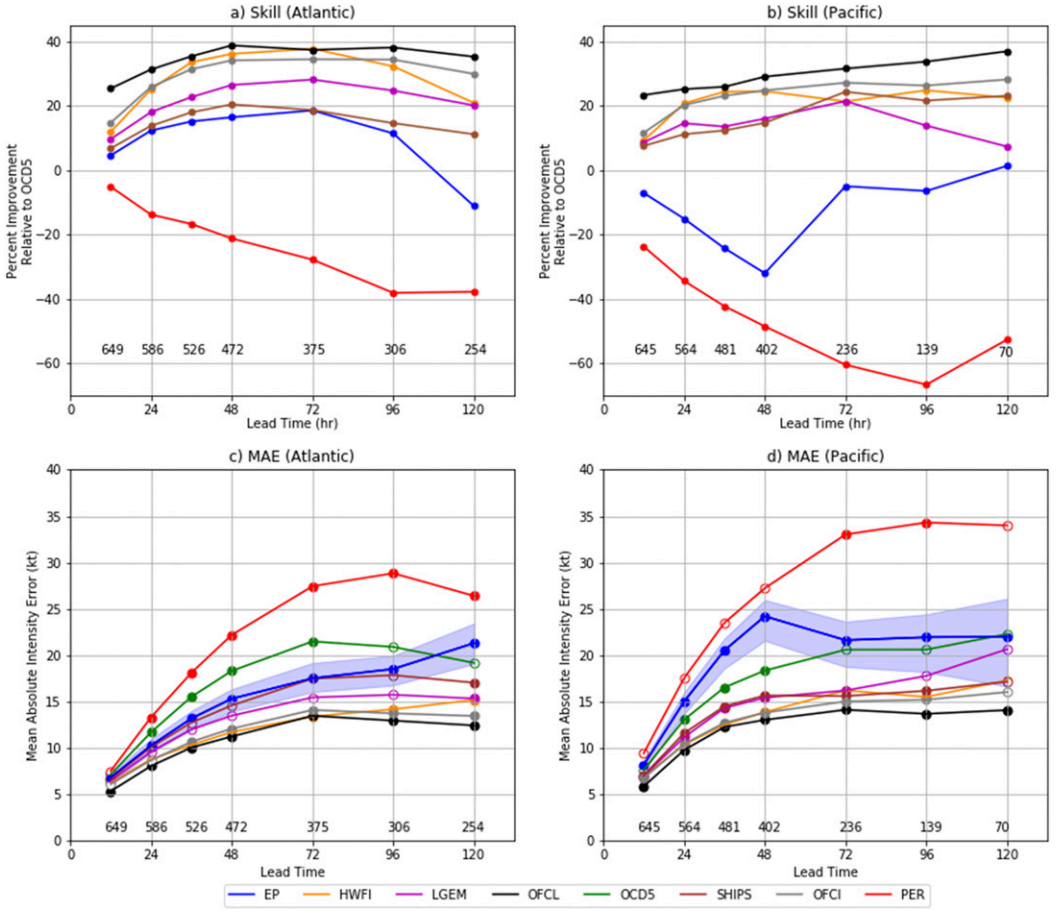

FIG. 5. (a),(b) Percentage improvement relative to the OCD5 model and (c),(d) MAE for the deterministic EP model (blue), HWFI (yellow), LGEM (magenta), official NHC forecast (OFCL, black), SHIP (brown), 6-h interpolated NHC forecast (OFCI, gray), and persistence (Per, red) forecasts across the 2017–18 seasons for the (a),(c) North Atlantic and (b),(d) eastern and central North Pacific basins. The number of forecast cases in each homogeneous forecast set is given at the bottom of the each panel. The blue-shaded region in (c),(d) represents the 5th–95th percentile range of the EP model forecasts, as determined via Monte Carlo bootstrapping using 1000 samples (with sample sizes equivalent to the number of forecasts at each lead time) with replacement. Filled circles denote that the given model's MAE at that lead time is significantly different from that of the EP model to at least 95% confidence, whereas open circles denote that the given model's MAE at that lead time is not significantly different from that of the EP model to at least 95% confidence.

$(12\,\text{h})^{-1}$ threshold, at which a significant skill reduction is again noted. Pacific calibrated RI model skill is competitive (characterized in the same fashion as for the Atlantic model) only with the operational Bayesian RI model at all thresholds.

While the calibrated EP RI model is generally more skillful than a climatological forecast, the same is not true for the RW model. In the North Atlantic basin, model performance is slightly worse than that of a climatological forecast at all thresholds (Fig. 7, top). In the eastern and central North Pacific basins, the EP RW model's skill is worse than that of a climatological forecast at all thresholds (Fig. 7, bottom).

While most RI/RW models are specifically developed to forecast a percent chance of RI/RW, the EP RI/RW models are developed from the deterministic model and seek to transform the forecast intensity change into a percent chance for RI/RW. Although forming an RI/RW model around this transformation from intensity change to probability of RI/RW is logical, it may not be the best way to form a probabilistic model. With many more non-RI/RW cases than RI/RW cases, we speculate the relationship between forecast intensity change and the probability of RI/RW is nonlinear and thus cannot be accounted for in the current model formulation. Consequently, independent RI/RW probabilistic models,

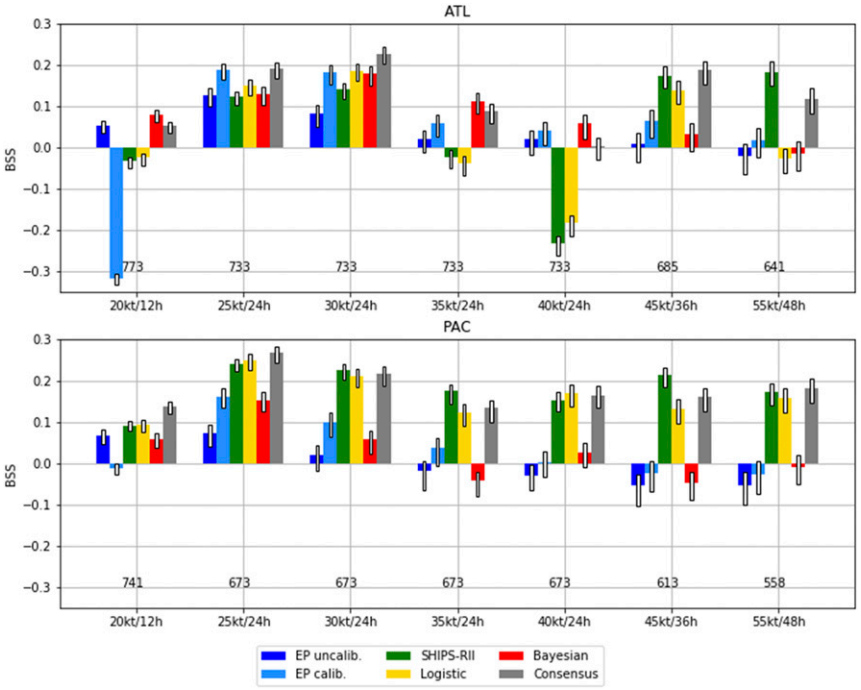

FIG. 6. BSS (as a function of RI threshold) for the uncalibrated EP RI model (light blue), calibrated EP RI model (blue), SHIPS-RII model (green), logistic model (yellow), Bayesian model (red), and a consensus of the SHIPS-RII, logistic, and Bayesian models (gray) across the 2017–18 TC seasons for the North Atlantic (ATL; top) and eastern and central North Pacific (PAC; bottom) basins. Note that no verifying forecasts for the 65 kt $(72\,h)^{-1}$ RI threshold were available from the non-EP models, and thus performance at this threshold is not shown. The number of forecast cases in each homogeneous forecast set is given at the bottom of the figure. The narrow white bars for each model at each lead time represent the 5th–95th percentile ranges of each model's forecasts, as determined via Monte Carlo bootstrapping using 1000 samples (with sample sizes equivalent to the number of forecasts at each lead time) with replacement.

such as the SHIPS-RII, logistic, and Bayesian models described above, can be expected to be more skillful than the deterministic-based probability forecasts described herein. Thus, the competitive performance of the EP RI models relative to these operational RI models, particularly in the Atlantic basin at the 25–40 kt $(24\,h)^{-1}$ RI thresholds, suggests that applying the EP methodology specifically to RI has great promise to provide forecasts with skill superior to existing operational RI guidance.

c. *Case studies*

As is true for TC intensity forecasts in general (Rappaport et al. 2012; Kaplan et al. 2010), RI/RW cases provide a major contribution to model errors in both the North Atlantic and eastern and central North Pacific basins. For example, RI/RW cases [here defined by a change in intensity of 30 kt $(24\,h)^{-1}$] comprise just 8.2% of cases across the North Atlantic basin for the 2017–18 seasons but contribute 16.5% to the total sum of forecast intensity errors for the EP model over the same period. Likewise, in the eastern and central North Pacific basins, RI/RW cases comprise 14.1% of forecasts but are responsible for 23.9% of the total sum of forecast intensity errors for the EP model over the same period. Below, two representative case studies are discussed to provide further insight into deterministic model performance for these cases.

1) MARIA—0000 UTC 18 SEPTEMBER 2017

Maria began as a tropical depression around 1200 UTC 16 September 2017 over the tropical Atlantic, but it rapidly intensified as it moved toward the Lesser Antilles, reaching hurricane intensity by 1800 UTC 17 September 2017. Aided by warm waters and weak vertical wind shear, Maria continued to rapidly intensify, reaching an intensity of 145 kt just prior to landfall on Dominica at 0115 UTC 19 September 2017 (Pasch et al. 2019).

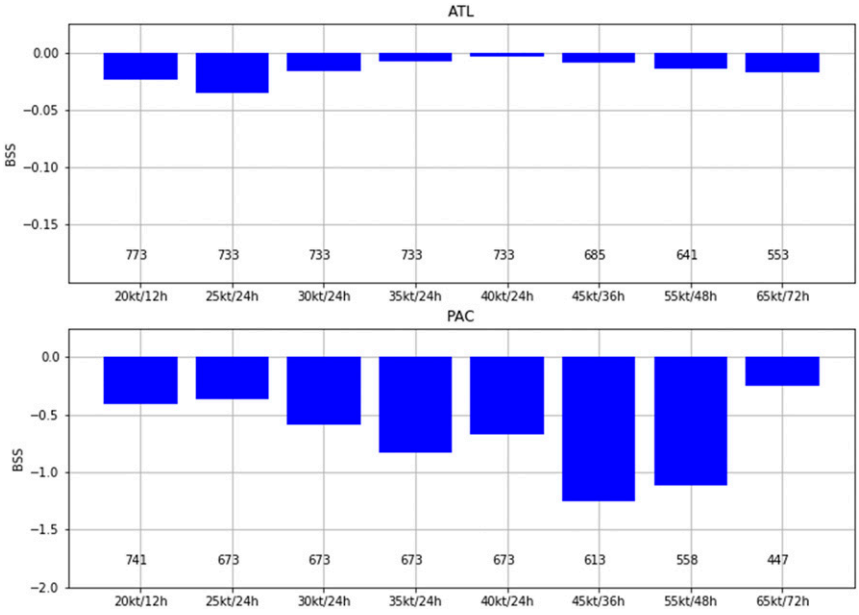

FIG. 7. BSS (as a function of RW threshold) for the EP RW model across the 2017–18 TC seasons for the (top) North Atlantic (ATL) and (bottom) eastern and central North Pacific (PAC) basins. The number of forecast cases in each forecast set is given at the bottom of the figure.

While the deterministic EP model accurately forecast that Maria would intensify (Fig. 8, top), it failed to predict the extreme intensification rate, resulting in one of the largest 24-h forecast errors by the EP model over the 2017–18 seasons. For example, the forecast initialized 0000 UTC 18 September 2017 verified at the conclusion of the RI event on 0000 UTC 19 September 2017, just before Maria struck Dominica. While the EP model forecast Maria to intensify from 75 to 94 kt, Maria intensified to 145 kt. This 70 kt $(24\,h)^{-1}$ intensification rate more than doubled the 30 kt $(24\,h)^{-1}$ intensification rate seen over the previous 24 h.

The predictors for the 12 and 24 h lead times indicate a favorable environment for RI, with input values for CD26, VMPI, and SHDC suggesting climatologically warm waters to depth and weak vertical wind shear (Table 3). Additionally, U20C and D200 show anomalous easterly 200-hPa zonal wind and above-normal upper-level divergence, the latter of which has been shown to aid TC intensification (DeMaria and Kaplan 1999). Despite their large input values, however, CD26 had little to no impact on the intensity-change forecast, while U20C and D200 had minor contributions to the forecast. This is largely because CD26 is often given little weight and features sparingly in determining which lines get executed, with the same being generally true for U20C and D200 as well (see the appendix). VMPI, while having a lesser contribution, contributed positively to the forecast. Meanwhile, SHDC was one of two primary positive forecast contributors to the forecast, contributing 5.6 kt to the total 19-kt $(24\,h)^{-1}$ forecast over the two lead times. The other predictor with a positive forecast contribution is the DELV predictor. The large contribution from DELV is primarily a function of its frequent appearance in the model algorithms, while the contribution from SHDC is a mixture of its weighting in the model algorithms and in determining which lines get executed (see the appendix). Consequently, the model responds to Maria's ongoing intensification in a low-shear environment.

The observed intensity change of 70 kt $(24\,h)^{-1}$ for this example is 5.6 standard deviations above the average 24-h intensity-change forecast by the EP model and similarly lies within the upper 1% of all 24-h observed intensity changes within the Atlantic basin. Although the forecast intensity change of +19 kt $(24\,h)^{-1}$ is well below what occurred, it is in the 98th percentile of all 24-h intensity-change forecasts by the deterministic EP model (not shown). The forecast then fits within what the model considers RI, following how RI was initially defined (from the 95th percentile of the intensity-change distributions In other words, the forecast under consideration is unrepresentative of the larger set of forecasts to which

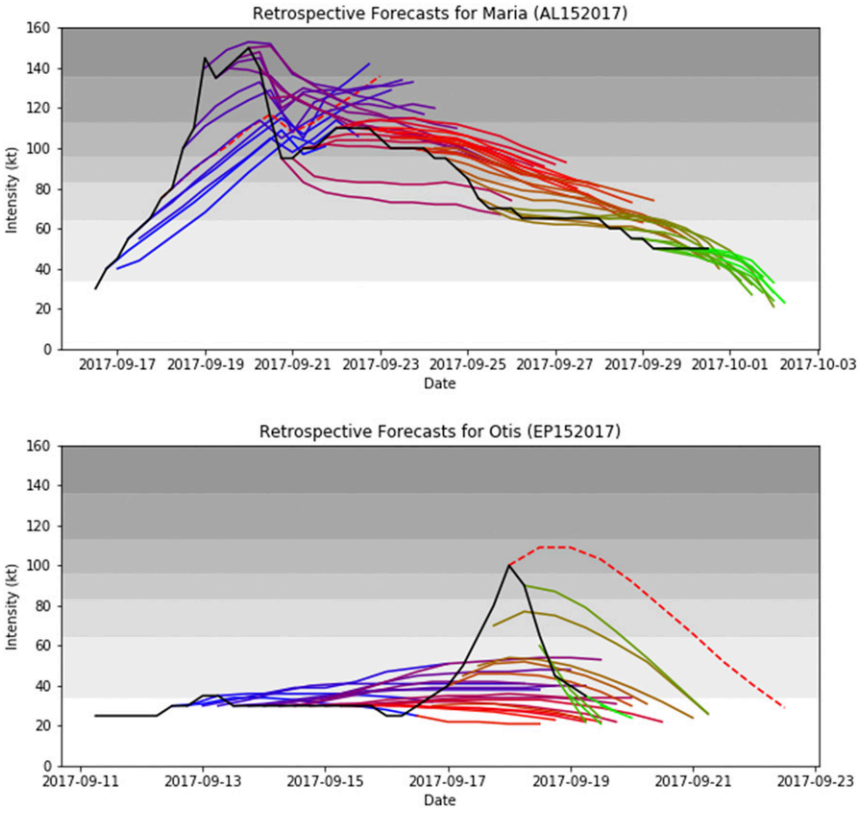

FIG. 8. Retrospective intensity forecasts (colored lines; blues and purples denote earlier-issued forecasts, whereas yellows and greens denote later-issued forecasts) and observed intensity (black line) for (top) TC Maria (AL152017) and (bottom) TC Otis (EP152017). The forecasts discussed in section 3c are shown in dashed red lines.

the model is trained to forecast. The deterministic EP model correctly indicates that this forecast scenario is atypical—albeit within the context of its training data rather than observations. Similarly, the calibrated EP RI model forecast a 41% chance of RI at the 25 kt $(24\,h)^{-1}$ threshold, and a 23% chance of RI at the 30 kt $(24\,h)^{-1}$ threshold. While both percentages are low, they both are in the 98th percentile of all probabilistic forecasts by the EP model for the given thresholds. Thus, while the probabilistic EP model also discerns the correct forecast scenario, forecasting the exact intensification magnitude is difficult due to the intensity change being on the far tail of the model's training data.

2) OTIS—0000 UTC 18 SEPTEMBER 2017

Although Otis (Blake 2018) was hindered by strong wind shear and associated dry-air intrusion for much of its lifespan, the TC turned northward and moved into a weaker vertical wind shear environment on 17 September 2017. This helped mitigate the intrusion of dry air into its center and, consequently, Otis underwent RI, intensifying by 60 kt in 24 h to reach a peak intensity of 100 kt. As Otis continued northward, however, it again encountered stronger vertical wind shear, reestablishing the intrusion of dry air into the storm center. This brought about RW, with Otis' intensity decreasing 60 kt over the subsequent 24 h to return to an intensity of 40 kt.

TABLE 3. List of predictor values, in standard anomaly form, and their relative contribution to the 12- and 24-h intensity forecasts of TC Maria for the forecast initialized at 0000 UTC 18 Sep 2017.

Predictor	12-h predictor variable value (std dev)/contribution (kt)	24-h predictor variable value (std dev)/contribution (kt)
DELV	1.3/3.5	0.8/2.1
CD26	1.2/0.0	1.3/0.0
U20C	−1.0/0.1	−1.0/0.0
D200	1.4/0.4	0.9/0.1
TWAC	0.0/0.0	−0.1/0.0
SHDC	−1.5/2.8	−1.5/2.8
VMPI	0.7/1.4	0.8/1.3
CFLX	0.2/−0.4	0.1/−0.3

TABLE 4. List of predictor values, in standard anomaly form, and their relative contribution to the 12- and 24-h intensity forecasts of TC Otis for the forecast initialized at 0000 UTC 18 Sep 2017.

Predictor	12-h predictor variable value (std dev)/contribution (kt)	24-h predictor variable value (std dev)/contribution (kt)
DELV	3.6/15.4	0.8/6.8
CD26	−0.4/−0.1	−0.2/0.0
U20C	1.0/0.7	0.6/0.6
D200	0.4/0.0	0.7/0.0
TWAC	0.3/1.1	0.1/0.2
SHDC	−0.8/2.2	−0.8/2.5
VMPI	−0.6/−0.9	−0.9/−1.6
CFLX	3.2/−3.6	2.9/−3.2

The EP model generally forecast Otis to strengthen slightly through its lifespan (Fig. 8, bottom), but it failed to forecast RI in any forecast verifying on 17 September 2017. Of notable concern is the forecast initializing at the transition from RI to RW on 0000 UTC 18 September 2017 (red dotted line in Fig. 8, bottom), as this forecast is associated with the largest 24-h forecast error by both the North Atlantic and eastern and central North Pacific EP models over the 2017–18 seasons. Whereas Otis weakened by 60 kt over the following 24 h, the EP model forecast Otis to strengthen to 110 kt by 12 h and to 111 kt by 24 h before plateauing in intensity and rapidly weakening over the subsequent three days. However, the forecast RW probability is zero at all thresholds because RW is defined as a magnitude change from the *initial* intensity and not over a moving window (i.e., 0–48 h not 24–48 h).

When looking at the relative contributions from each input variable at the 12- and 24-h forecast times, a clear explanation for the EP model's poor performance emerges (Table 4). Although the input value of DELV for the 12-h lead time is well above normal, corresponding to Otis' just-completed RI, the 12-h value of the dry-air predictor (CFLX) is also large, indicating that a large amount of dry air is being mixed back into the TC's circulation. However, despite their similar magnitudes, the two variables have distinctly different contributions to the forecast, with the DELV predictor having a much larger impact over the first 12 h. At the 24-h lead time, the value of the DELV parameter drops notably, whereas the value of the CFLX parameter remains high. However, the positive contribution from DELV is still more than double the negative contribution of CFLX. This greater contribution of DELV not only stems from the weighting of the parameter in the calculations, but also its role in the conditional statements and thus in determining how many lines get executed. Meanwhile, the other variables feature only modest deviations from climatology and have a mixed impact on the forecast. As a result, rather than forecasting a sharp change in intensity, the EP model forecasts are more subdued in their predicted rates of intensity change given the importance of the DELV parameter to the forecast.

4. Summary and conclusions

This paper describes the development, application, and evaluation of two TC deterministic and probabilistic intensity forecast models, one for the North Atlantic and another for the eastern and central North Pacific basins, from a large ensemble of evolutionary algorithms. These algorithms utilize an if-then structure as well as linear and nonlinear predictor combinations to forecast a change in intensity over a 12-h period. Run iteratively, these algorithms produce a deterministic forecast for TC intensity every 12 h out to 120 h and probabilistic forecasts for RI and RW at specified thresholds out to 72 h. A set of eight predictors from the SHIPS developmental dataset, which are converted to standard anomalies to aid comparison between variables of dissimilar units, provide the input data for model training, application, and evaluation. After being randomly initialized, the EP process involving cloning, mutation, and selective pressure drives the algorithms toward skillful predictor combinations. In total, five populations with 10 000 algorithms are run over 300 iterations, from which the 100 best-performing algorithms over all populations and iterations are retained. Bias correction is then applied to all retained algorithms before 10 skillful yet diverse algorithms are selected to be combined through BMC. Finally, BMC is used to determine the weighting for each of the 10 members to produce the final deterministic model, from which a PDF is obtained to generate RI/RW probabilities.

Each model is tested on independent cases from the 2017 and 2018 TC seasons. In the North Atlantic basin, the deterministic model is 10%–20% more skillful than the ''no skill'' OCD5 forecast at all leads except 120 h, with skill comparable to that of the operational SHIPS and LGEM models at these times. Conversely, for the eastern and central North Pacific basin, the deterministic model is less skillful than the ''no skill'' OCD5 forecast and all operational models at all lead times except 120 h. Calibrated RI forecasts

are skillful relative to climatology and are competitive with operational RI forecasts at the 25–40 kt $(24\,h)^{-1}$ intensity-change thresholds in the North Atlantic and for the 25–30 kt $(24\,h)^{-1}$ intensity-change thresholds in the eastern and central North Pacific basins. However, calibrated RW forecasts in both basins are not skillful for any intensity-change threshold, and there are no operational RW forecast models to provide context for these results. The mixed performance of the RI and RW models is likely due to the underlying probabilistic model being derived from a deterministic model that is trained on all intensity-change cases (of which there are many more non-RI/RW cases than there are RI/RW cases) rather than on only RI/RW cases. Probability calibration has mixed impacts on forecast skill. An alternative model formulation (e.g., one developed only for rapid intensity change, as with the SHIPS-RII, logistic, and Bayesian models described in section 3b) is likely necessary to achieve further skill increases for RI and RW forecasts. Altogether, the results suggest that the EP method holds great promise, with substantial room for further (and in some cases necessary) improvements, for both deterministic and probabilistic TC intensity-change predictions.

Selected case studies demonstrate that the model forecast often contain large contributions from the DELV predictor (i.e., intensity-change persistence), which led to difficulties in producing accurate deterministic forecasts for RI and RW cases (as each are often associated with abrupt intensity changes from persistence). One might therefore conclude that the DELV predictor is detrimental to model performance. While this conclusion is correct in part for RI and RW cases, it is not true over the larger set of all forecast cases. The EP process selected the predictor to be meaningful and therefore weighted it heavily to increase model skill, based on its training data. Consequently, over the training data, which are representative of the full TC populations in each basin, persistence is a reliable intensity predictor. Further support for DELV's inclusion, especially in the Pacific model, comes from the fact that improvements over the OCD5 model tend to be smaller in the Pacific basin than in the Atlantic (Cangialosi 2019, their Figs. 11 and 13). This suggests that it is harder to beat climatology and persistence in the Pacific basin, and as such climatological and persistence parameters should be given more weight.

However, as noted above, this type of a persistence forecast can lead to large errors, such as at the onset or ending of RI and RW. While the importance of keeping DELV is also supported by its use in both the OCD5 and SHIPS models (Knaff et al. 2003; DeMaria and Kaplan 1994; Shimada et al. 2018), it is worth investigating whether other information can be leveraged so that the model has a priori knowledge of when a persistence forecast may not be warranted. One method to do this would be to introduce a variable such as the difference between the intensity of the storm and its maximum potential intensity, as is done in the LGEM model (DeMaria 2009). This would help inform the model on whether a TC is located toward the higher or lower end of the climatological intensity distribution and thus know when a TC is more or less capable of undergoing RI or RW. Structural information derived from geostationary and polar-orbiting sensors operating at infrared and microwave wavelengths, in which structures that reliably distinguish between RI and non-RI events can be identified (e.g., Jiang and Ramirez 2013; Rozoff et al. 2015; Fischer et al. 2018), also offers promise for a priori discernment of cases when a persistence forecast is less warranted.

The quasi-Gaussian nature of intensity change and the bias toward TCs of weaker intensities in the historical record is challenging to TC intensity forecasting, as the cases that are of highest interest (RI, RW, and intense hurricanes) are the least prevalent across all forecast times. This is a particular challenge to training machine-learning algorithms, which may sacrifice performance on these select few cases to perform well across all cases as a whole. Different cost functions (e.g., using RMSE rather than MAE at the BMC weight-determination stage, or by weighting errors from stronger TCs more heavily in the training process or at the BMC weight-determination stage) and/or different model formulations [e.g., using an alternative skill metric such as BSS or continuous ranked probability score to derive the BMC weights for probabilistic applications, developing altogether separate deterministic and probabilistic RI/RW models, or using more advanced versions of the EP method such as the coevolution predator–prey ecosystem described by Roebber and Crockett (2019)] may hold promise for addressing these challenges, and future work aims to consider these approaches for TC intensity-change prediction.

Acknowledgments. This work was supported by funding from the National Oceanic and Atmospheric Administration's Joint Hurricane Testbed (JHT) under Award NA17OAR4590137 to the second and third authors. We acknowledge fruitful conversations with Alan Brammer, Mark DeMaria, and Jason Sippel during this research. Constructive reviews from Rachel Mauk and

two anonymous reviewers provided invaluable insight that improved the quality of this paper.

APPENDIX

Model Algorithms

After bias correction and BMC weighting are performed (section 2c), seven algorithms are retained with nonzero weight for the Atlantic basin and two algorithms are retained with nonzero weight for the eastern and central North Pacific basin. These algorithms, including their BMC-determined weights and bias-correction factors, are in Table A1 for the Atlantic basin and Table A2 for the North Pacific basin. The algorithm number for each is included only for completeness; it has no specific meaning.

Algorithms are given by the form prescribed by Eq. (1) in section 2b. The individual line numbers at left refer to the i within (1). The V_{in} refer to predictor

TABLE A1. Atlantic model.

		V_{i1}	R_{i1}	V_{i2}		$C_{i1} \times V_{i3}$	O_{i1}	$C_{i2} \times V_{i4}$	O_{i2}	$C_{i3} \times V_{i5}$
					Algorithm 6: weighting = 0.166 67; bias (ε) = 0.52					
1	IF	SHDC	≤	TWAC	THEN	0.145 98 × VMPI	+	−0.447 44 × U20C	×	−0.1582 × D200
2	IF	SHDC	≤	DELV	THEN	0.361 27 × DELV	×	−0.0746 × 10	×	0.236 45 × DELV
3	IF	SHDC	≤	SHDC	THEN	−0.954 43 × DELV	+	0.954 13 × DELV	+	0.023 58 × 10
4	IF	DELV	≤	DELV	THEN	−0.188 35 × SHDC	+	0.408 03 × DELV	+	−0.247 38 × CFLX
5	IF	VMPI	>	TWAC	THEN	−0.947 45 × DELV	×	0.181 54 × VMPI	×	0.849 04 × D200
					Algorithm 8: weighting = 0.083 33; bias (ε) = −0.57					
1	IF	CFLX	≤	DELV	THEN	0.902 16 × VMPI	×	0.653 79 × D200	×	0.216 44 × DELV
2	IF	SHDC	≤	DELV	THEN	0.361 27 × DELV	×	−0.0746 × 10	×	0.236 45 × DELV
3	IF	SHDC	≤	SHDC	THEN	−0.954 43 × DELV	+	0.954 13 × DELV	+	0.023 58 × 10
4	IF	DELV	≤	DELV	THEN	−0.188 35 × SHDC	+	0.408 03 × DELV	+	−0.247 38 × CFLX
5	IF	TWAC	≤	U20C	THEN	−0.325 57 × TWAC	+	−0.205 41 × VMPI	×	−0.385 64 × CFLX
					Algorithm 9: weighting = 0.083 33; bias (ε) = 0.28					
1	IF	CFLX	>	U20C	THEN	−0.263 81 × U20C	×	0.149 71 × VMPI	+	0.2113 × VMPI
2	IF	SHDC	≤	DELV	THEN	0.361 27 × DELV	×	−0.0746 × 10	×	0.236 45 × DELV
3	IF	SHDC	≤	SHDC	THEN	−0.954 43 × DELV	+	0.954 13 × DELV	+	0.023 58 × 10
4	IF	DELV	≤	DELV	THEN	−0.188 35 × SHDC	+	0.408 03 × DELV	+	−0.247 38 × CFLX
5	IF	SHDC	≤	CD26	THEN	−0.427 66 × TWAC	×	0.361 28 × CFLX	×	0.033 89 × SHDC
					Algorithm 34: weighting = 0.083 33; bias (ε) = 0.21					
1	IF	TWAC	≤	CFLX	THEN	0.317 31 × CFLX	×	−0.905 71 × D200	×	−0.217 76 × SHDC
2	IF	SHDC	>	CFLX	THEN	0.252 37 × TWAC	+	−0.363 17 × TWAC	×	0.279 41 × CFLX
3	IF	DELV	>	TWAC	THEN	0.033 56 × CD26	×	0.098 53 × DELV	+	0.038 53 × 10
4	IF	VMPI	≤	VMPI	THEN	0.335 92 × DELV	×	−0.212 64 × TWAC	+	0.157 55 × DELV
5	IF	SHDC	≤	SHDC	THEN	−0.182 06 × SHDC	+	0.1172 × VMPI	+	−0.176 64 × CFLX
					Algorithm 35: weighting = 0.083 33; bias (ε) = 0.10					
1	IF	CFLX	≤	SHDC	THEN	0.862 28 × CD26	+	0.413 23 × TWAC	+	−0.853 29 × CD26
2	IF	CD26	>	D200	THEN	−0.109 33 × DELV	+	0.543 57 × TWAC	×	−0.287 23 × CD26
3	IF	DELV	>	TWAC	THEN	0.033 56 × CD26	×	0.098 53 × DELV	+	0.038 53 × 10
4	IF	VMPI	≤	VMPI	THEN	0.335 92 × DELV	×	−0.212 64 × TWAC	+	0.157 55 × DELV
5	IF	SHDC	≤	SHDC	THEN	−0.182 06 × SHDC	+	0.1172 × VMPI	+	−0.176 64 × CFLX
					Algorithm 49: weighting = 0.166 67; bias (ε) = 0.19					
1	IF	D200	≤	VMPI	THEN	0.323 67 × TWAC	+	−0.156 24 × D200	×	0.118 85 × CFLX
2	IF	D200	≤	SHDC	THEN	−0.242 29 × TWAC	×	0.0833 × DELV	+	−0.074 26 × DELV
3	IF	DELV	>	TWAC	THEN	0.033 56 × CD26	×	0.098 53 × DELV	+	0.038 53 × 10
4	IF	VMPI	≤	VMPI	THEN	0.335 92 × DELV	×	−0.212 64 × TWAC	+	0.157 55 × DELV
5	IF	SHDC	≤	SHDC	THEN	−0.182 06 × SHDC	+	0.1172 × VMPI	+	−0.176 64 × CFLX
					Algorithm 53: weighting = 0.25; bias (ε) = −0.67					
1	IF	CD26	≤	CD26	THEN	−0.595 28 × 10	×	−0.831 68 × TWAC	×	−0.1173 × TWAC
2	IF	D200	≤	10	THEN	−0.789 33 × VMPI	×	0.264 22 × TWAC	×	−0.782 23 × CFLX
3	IF	DELV	>	TWAC	THEN	0.033 56 × CD26	×	0.098 53 × DELV	+	0.038 53 × 10
4	IF	VMPI	≤	VMPI	THEN	0.335 92 × DELV	×	−0.212 64 × TWAC	+	0.157 55 × DELV
5	IF	SHDC	≤	SHDC	THEN	−0.182 06 × SHDC	+	0.1172 × VMPI	+	−0.176 64 × CFLX

TABLE A2. Pacific model.

		V_{i1}	R_{i1}	V_{i2}		$C_{i1} \times V_{i3}$	O_{i1}	$C_{i2} \times V_{i4}$	O_{i2}	$C_{i3} \times V_{i5}$
					Algorithm 31: weighting = 0.25; bias (ε) = −0.07					
1	IF	TWAC	>	VMPI	THEN	0.366 79 × CFLX	×	0.559 76 × TWAC	+	−0.037 05 × DELV
2	IF	CFLX	≤	DELV	THEN	0.167 84 × CFLX	×	0.839 09 × DELV	×	0.581 32 × TWAC
3	IF	SHDC	>	D200	THEN	−0.122 43 × VMPI	+	0.313 32 × TWAC	+	0.018 71 × CD26
4	IF	D200	≤	D200	THEN	−0.890 92 × TWAC	×	0.289 28 × TWAC	+	−0.1396 × CFLX
5	IF	VMPI	≤	VMPI	THEN	0.6716 × VMPI	+	−0.443 36 × VMPI	+	0.420 04 × DELV
					Algorithm 69: weighting = 0.75; bias (ε) = −0.09					
1	IF	DELV	>	U20C	THEN	0.178 81 × VMPI	×	−0.737 21 × U20C	+	−0.363 76 × SHDC
2	IF	DELV	>	CFLX	THEN	−0.145 89 × DELV	+	0.0649 × TWAC	×	0.8098 × CD26
3	IF	SHDC	>	D200	THEN	−0.122 43 × VMPI	+	0.313 32 × TWAC	+	0.018 71 × CD26
4	IF	D200	≤	D200	THEN	−0.890 92 × TWAC	×	0.289 28 × TWAC	+	−0.1396 × CFLX
5	IF	VMPI	≤	VMPI	THEN	0.6716 × VMPI	+	−0.443 36 × VMPI	+	0.420 04 × DELV

variables, a full listing of which is given in Table 1. R_{i1} refers to one of the allowable relational operators of ≤ or >. The C_{in} refer to coefficients in the range of [−1, 1]. The O_{in} refer to one of the allowable arithmetic operators of + or ×. For each, n denotes the nth instance of the parameter on a given line.

REFERENCES

Alessandrini, S., L. Delle Monache, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Mon. Wea. Rev.*, **146**, 1723–1744, https://doi.org/10.1175/MWR-D-17-0314.1.

Bellman, R. E., 1961: *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 255 pp.

Blake, E. S., 2018: National Hurricane Center tropical cyclone report: Hurricane Otis. NOAA Rep. EP152017, 14 pp., https://www.nhc.noaa.gov/data/tcr/EP152017_Otis.pdf.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Cangialosi, J. P., 2018: National Hurricane Center Forecast verification report: 2017 hurricane season. NOAA, 73 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2017.pdf.

——, 2019: National Hurricane Center Forecast verification report: 2018 hurricane season. NOAA, 73 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2018.pdf.

Cloud, K. A., B. J. Reich, C. M. Rozoff, S. Alessandrini, W. E. Lewis, and L. Delle Monache, 2019: A feed forward neural network based on model output statistics for short-term hurricane intensity prediction. *Wea. Forecasting*, **34**, 985–997, https://doi.org/10.1175/WAF-D-18-0173.1.

DeMaria, M., 2009: A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.*, **137**, 68–82, https://doi.org/10.1175/2008MWR2513.1.

——, and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2.

——, and ——, 1999: An updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **14**, 326–337, https://doi.org/10.1175/1520-0434(1999)014<0326:AUSHIP>2.0.CO;2.

——, M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543, https://doi.org/10.1175/WAF862.1.

——, J. Knaff, and J. Kaplan, 2006: On the decay of tropical cyclone winds crossing narrow landmasses. *J. Appl. Meteor. Climatol.*, **45**, 491–499, https://doi.org/10.1175/JAM2351.1.

——, C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, https://doi.org/10.1175/BAMS-D-12-00240.1.

Emanuel, K., and F. Zhang, 2016: On the predictability and error sources of tropical cyclone intensity forecasts. *J. Atmos. Sci.*, **73**, 3739–3747, https://doi.org/10.1175/JAS-D-16-0100.1.

——, and ——, 2017: On the role of inner-core moisture in tropical cyclone predictability and forecast skill. *J. Atmos. Sci.*, **74**, 2315–2324, https://doi.org/10.1175/JAS-D-17-0008.1.

——, C. DesAutels, C. Holloway, and R. Korty, 2004: Environmental control of tropical cyclone intensity. *J. Atmos. Sci.*, **61**, 843–858, https://doi.org/10.1175/1520-0469(2004)061<0843:ECOTCI>2.0.CO;2.

Fischer, M. S., B. H. Tang, K. L. Corbosiero, and C. M. Rozoff, 2018: Normalized convective characteristics of tropical cyclone rapid intensification events in the North Atlantic and eastern North Pacific. *Mon. Wea. Rev.*, **146**, 1133–1155, https://doi.org/10.1175/MWR-D-17-0239.1.

Fogel, L. J., 1999: *Intelligence Through Simulated Evolution: Forty Years of Evolutionary Programming*. John Wiley, 162 pp.

Gopalakrishnan, S. G., F. Marks Jr., X. Zhang, J.-W. Bao, K.-S. Yeh, and R. Atlas, 2011: The experimental HWRF system: A study on the influence of horizontal resolution on the structure and intensity changes in tropical cyclones using an idealized framework. *Mon. Wea. Rev.*, **139**, 1762–1784, https://doi.org/10.1175/2010MWR3535.1.

Grumm, R. J., and R. Hart, 2001: Standardized anomalies applied to significant cold season weather events: Preliminary findings. *Wea. Forecasting*, **16**, 736–754, https://doi.org/10.1175/1520-0434(2001)016<0736:SAATSC>2.0.CO;2.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.*, **14**, 382–417.

Jiang, H., and E. M. Ramirez, 2013: Necessary conditions for tropical cyclone rapid intensification as derived from 11 years

of TRMM data. *J. Climate*, **26**, 6459–6470, https://doi.org/10.1175/JCLI-D-12-00432.1.

Kaplan, J., and M. DeMaria, 1995: A simple empirical model for predicting the decay of tropical cyclone winds after landfall. *J. Appl. Meteor.*, **34**, 2499–2512, https://doi.org/10.1175/1520-0450(1995)034<2499:ASEMFP>2.0.CO;2.

——, and ——, 2001: On the decay of tropical cyclone winds after landfall in the New England area. *J. Appl. Meteor.*, **40**, 280–286, https://doi.org/10.1175/1520-0450(2001)040<0280:OTDOTC>2.0.CO;2.

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

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

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

Kieu, C. Q., and Z. Moon, 2016: Hurricane intensity predictability. *Bull. Amer. Meteor. Soc.*, **97**, 1847–1857, https://doi.org/10.1175/BAMS-D-15-00168.1.

Knaff, J. A., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting*, **18**, 80–92, https://doi.org/10.1175/1520-0434(2003)018<0080:SDTCIF>2.0.CO;2.

McGovern, A., R. Lagerquist, D. J. Gagne II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Monteith, K., J. Carroll, K. Seppi, and T. Martinez, 2011: Turning Bayesian model averaging into Bayesian model combination. *Proc. 2011 Int. Joint Conf. on Neural Networks*, San Jose, CA, Institute of Electrical and Electronics Engineers, 2657–2663, https://doi.org/10.1109/IJCNN.2011.6033566.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

NCEP, 2016: The Global Forecast System (GFS)–Global Spectral Model (GSM). NOAA, accessed 4 March 2019, https://www.emc.ncep.noaa.gov/GFS/doc.php.

Pasch, R. J., 2015: National Hurricane Center annual summary: 2014 Atlantic hurricane season. NOAA, accessed 3 April 2020, https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php.

——, A. B. Penny, and R. Berg, 2019: National Hurricane Center tropical cyclone report: Hurricane Maria. NOAA Rep. AL152017, 48 pp., https://www.nhc.noaa.gov/data/tcr/AL152017_Maria.pdf.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rappaport, E. N., J.-G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The Joint Hurricane Test Bed: Its first decade of tropical cyclone research-to-operations activities reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371–380, https://doi.org/10.1175/BAMS-D-11-00037.1.

Roebber, P. J., 2010: Seeking consensus: A new approach. *Mon. Wea. Rev.*, **138**, 4402–4415, https://doi.org/10.1175/2010MWR3508.1.

——, 2013: Using evolutionary programming to generate skillful extreme value probabilistic forecasts. *Mon. Wea. Rev.*, **141**, 3170–3185, https://doi.org/10.1175/MWR-D-12-00285.1.

——, 2015a: Evolving ensembles. *Mon. Wea. Rev.*, **143**, 471–490, https://doi.org/10.1175/MWR-D-14-00058.1.

——, 2015b: Using evolutionary programming to maximize minimum temperature forecast skill. *Mon. Wea. Rev.*, **143**, 1506–1516, https://doi.org/10.1175/MWR-D-14-00096.1.

——, 2016: Development of a large member ensemble forecast system for heavy rainfall using evolutionary programming (EP). DTC, 10 pp., https://dtcenter.org/sites/default/files/visitor-projects/Roebber-DTCFinalReport.pdf.

——, 2018: Using evolutionary programming to add deterministic and probabilistic skill to spatial model forecasts. *Mon. Wea. Rev.*, **146**, 2525–2540, https://doi.org/10.1175/MWR-D-17-0272.1.

——, and J. Crockett, 2019: Using a coevolutionary postprocessor to improve skill for both forecasts of surface temperature and nowcasts of convection occurrence. *Mon. Wea. Rev.*, **147**, 4241–4259, https://doi.org/10.1175/MWR-D-19-0063.1.

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

——, C. S. Velden, J. Kaplan, J. P. Kossin, and A. J. Wimmers, 2015: Improvements in the probabilistic prediction of tropical cyclone rapid intensification with passive microwave observations. *Wea. Forecasting*, **30**, 1016–1038, https://doi.org/10.1175/WAF-D-14-00109.1.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1058, https://doi.org/10.1175/2010BAMS3001.1.

Shimada, U., H. Owada, M. Yamaguchi, T. Iriguchi, M. Sawada, K. Aoniashi, M. DeMaria, and K. D. Musgrave, 2018: Further improvements to the Statistical Hurricane Intensity Prediction Scheme using tropical cyclone rainfall and structural features. *Wea. Forecasting*, **33**, 1587–1603, https://doi.org/10.1175/WAF-D-18-0021.1.

Stewart, S., 2014: National Hurricane Center annual summary: 2012 Atlantic hurricane season. NOAA, 11 pp., http://www.nhc.noaa.gov/data/tcr/summary_atlc_2012.pdf.

——, 2016: National Hurricane Center annual summary: 2015 Atlantic hurricane season. NOAA, 16 pp., http://www.nhc.noaa.gov/data/tcr/summary_atlc_2015.pdf.

Tallapragada, V., C. Kieu, Y. Kwon, S. Trahan, Q. Liu, Z. Zhang, and I.-H. Kwon, 2014: Evaluation of storm structure from the operational HWRF during 2012 implementation. *Mon. Wea. Rev.*, **142**, 4308–4325, https://doi.org/10.1175/MWR-D-13-00010.1.

Wood, K. M., and E. Ritchie, 2015: A definition for rapid weakening of North Atlantic and eastern North Pacific tropical cyclones. *Geophys. Res. Lett.*, **42**, 10 091–10 097, https://doi.org/10.1002/2015GL066697.