May 2019

Using Evolutionary Programming to Generate a Tropical Cyclone Intensity Model

Jesse Schaffer
University of Wisconsin-Milwaukee

USING EVOLUTIONARY PROGRAMMING TO GENERATE A TROPICAL CYCLONE INTENSITY MODEL

by

Jesse Schaffer

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Atmospheric Science

at

The University of Wisconsin – Milwaukee

May 2019

ABSTRACT

USING EVOLUTIONARY PROGRAMMING TO GENERATE A TROPICAL CYCLONE INTENSITY MODEL

by

Jesse Schaffer

The University of Wisconsin-Milwaukee, 2019
Under the Supervision of Professor Paul Roebber and Professor Clark Evans

An innovative 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 through selective pressure a population of algorithms with skillful predictor combinations. From this process the 100 most skillful algorithms as determined by root-mean square error on cross-validation data is kept and bias corrected. Bayesian model combination is then used to assign individual weights to a subset of ten algorithms from the 100 best algorithms list, which are chosen to minimize mean-absolute error (MAE) and maximize mean-absolute difference across the selected algorithms. This results in combining both skillful and diverse algorithms, which together produce a forecast that is superior in skill to that from any individual algorithm. Using these methods and a perfect-prognostic approach, two similar but distinctly separate TC intensity models are developed to forecast for TC intensity every 12 h out to 120 h, with one forecasting TC intensity for the North Atlantic basin and the other for the east/central North Pacific basins. Results show improvements as defined by MAE over the "no skill" Decay Statistical Hurricane Intensity Forecast (OCD5) climatology/persistence model in the North Atlantic basin out to 96 h. In the east/central Pacific basins performance over the 12-24 h lead-time is similar to the OCD5,

while at later lead times performance drops below that of OCD5. Specific case studies are analyzed to give more insight into the behavior and performance of the models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First I would like to thank my advisors Dr. Paul Roebber and Dr. Clark Evans for their tremendous help, guidance, and support throughout my thesis research. I would also like to thank Dr. Jon Kahl who served on my committee and offered constructive feedback. Additionally, I would like to thank NOAA and the Joint Hurricane Tesbed (JHT) for their funding and support. Lastly, I would like to thank my friends and family for their support and their role in the the great memories I've made throughout this process.

1. INTRODUCTION

Tropical cyclone intensity forecasting is recognized as being particularly challenging with only slow improvements in recent years. But this sentiment and the problem itself is nothing new. DeMaria and Kaplan (1994) mentioned this idea roughly 25 years ago when developing the Statistical Hurricane Intensity Prediction Scheme (SHIPS) and since then, the challenge of forecasting TC intensity has persisted over the years (Emanuel et al., 2003, Rappaport et al., 2012, DeMaria and Kaplan, 2014, Emanuel and Zhang, 2016). Over the past 25 years improvement rates across the 24-72 h range have averaged only 1%-2% yr^{-1} (DeMaria et al. 2014; Fig. 1 & 2). This lack of improvement is even more dramatic when the time series is placed alongside track errors, which are improving at three times the rate of intensity errors over the same 24-72 h range (DeMaria et al. 2014). Over these shorter lead times intensity errors are dominated by the mischaracterization of the storm's initial intensity and of inner-core and eyewall processes (Emanuel and Zhang 2016, 2017; Kieu and Moon 2016). Furthermore, the challenge of forecasting the occurrence, timing, and magnitude of rapid intensification (RI) and rapid weakening (RW) significantly contributes to large absolute forecast errors and overall forecast difficulty over the shorter lead times (Rappaport et al. 2012, Kaplan et al. 2010).

Despite these challenges, notable improvements in TC intensity forecast have occurred. DeMaria et al. (2014) demonstrated that while improvement rates of 1%-2% seem negligible (especially when compared to track improvements) they are nonetheless statistically significant at the majority of lead times. Furthermore, at lead times longer than 72 h significant improvements have occurred, with improvement rates averaging 2%-4% yr^{-1} (DeMaria et al. 2014). However, this improvement rate is largely attributed to improvements in track forecasts, which have had similar improvement rates over the same time period (DeMaria et al. 2014,

Emanuel and Zhang 2016). Despite these improvements, the sentiment that TC intensity forecasts have not improved quickly enough is still indicative of the idea of that these small improvements may not be meaningful enough to produce significant practical improvements when it comes to better aiding emergency managers in preparation, planning, and decision-making.

While there is no significant relationship between TC wind risk and the likelihood of evacuations (Lazo et al. 2010), and even though TC winds alone account for only 8% of fatalities attributed to Atlantic TCs in the United States (Rappaport 2014), TC intensity forecasts are nonetheless critically important. TC intensity play an important role in the inundation of water and storm surge that is of primary concern when a TC makes landfall, with this being the leading cause of Atlantic TC related fatalities in the U.S., accounting for nearly half of all direct deaths (Linn et al. 2013, Rappaport 2014). Additionally, TC intensity forecasts are an important input to storm surge and inundation models such as the Sea, Lake, and Overland Surges from Hurricanes model (SLOSH; Jelesnianski, 1992), which is consulted heavily during evacuation planning and decision-making (Glahn 2009, Linn et al. 2013). Furthermore, Sheets (1990) noted that as populations grow in TC-prone coastal areas, longer lead times are needed for communities to adequately prepare. This idea is then amplified in the case of evacuations as a variety of factors lead to heavy traffic flow and prolonged evacuation times. These factors include, but are not limited to, the propensity for households to take multiple cars, the propensity for people to leave at similar times after evacuations have been ordered, and the propensity to evacuate on interstate routes rather than smaller highways and roads (Dow and Cutter 2002). More recently, Klotzbach et al. (2018) noted that while the number of landfilling hurricanes in the continental US since the year 1900 has held steady, inflation-adjusted hurricane-related damage has shown rapid growth.

They too attribute this to a growing coastal population and mention that as the population of coastal areas continues to climb higher this will be a problem well into the future. Thus, improved TC intensity forecasts are and will continue to be important for early evacuation planning and decision-making.

In 2001 NOAA, in conjunction with the U.S. Weather Research Program (USWRP), established the Joint Hurricane Testbed (JHT) to improve TC forecasting and expedite the transfer of research advances to operations. While intensity forecasts have been a top priority, the projects funded by the JHT through 2010 had struggled to produce significant improvements demonstrating the difficulty of this task (Rappaport et al. 2012). Despite this lack of improvement, Emanuel and Zhang (2016) suggested that quite a bit of improvement could still be made at all lead times out to 120 h. They suggested that these improvements would result from better characterizations of the initial TC, better models, and large ensembles of diverse models with heterogeneous vortex and environmental states, which would be valuable for better quantifying uncertainty in intensity forecasts.

TC intensity forecasts can be generated from three different types of models: dynamical, statistical-dynamical, and consensus models. Dynamical (or numerical weather prediction - NWP) models obtain a TC intensity forecast by solving the governing equations of motion for the atmosphere and appropriately parameterizing other processes such as moisture, radiation transfer, boundary-layer turbulence, surface energy fluxes, etc. These model domains can be global or limited in area, the latter of which allows for a smaller, convection-permitting, horizontal grid spacing. The result of this smaller grid spacing is that these models are better able to resolve the TC's inner core, resulting in more skillful forecasts so long as the inner-core structure can be initialized accurately through the assimilation of inner-core observations.

Currently, the most skillful limited-area dynamical models in the Atlantic and eastern North Pacific basins are the Hurricane Weather Research and Forecasting (HWRF; Tallapragada et al. 2014) and the Hurricanes in a Multiscale Ocean Coupled Nonhydrostatic Model (HMON, Mehra 2017), which replaced the Geophysical Fluid Dynamics Laboratory (GFDL; Bender et al. 2007) model in 2017.

Statistical-dynamical models, meanwhile, use statistical methods to assign appropriate weights for empirical relationships between environmental and structural TC characteristics, which themselves are obtained from dynamical models and/or observations. In the Atlantic and eastern North Pacific, the best-performing statistical dynamical models are the Statistical Hurricane Intensity Prediction Scheme (SHIPS; DeMaria and Kaplan 1994, 1999; DeMaria et al. 2005), Decay-SHIPS (DSHIPS) and the Logistic Growth Equation Model (LGEM; DeMaria 2009). SHIPS is an ever-changing multiple linear regression model that relates a large number of input parameters to TC intensity while DSHIPS pairs this with an empirical inland wind decay model (Kaplan and DeMaria 1995, Kaplan and DeMaria 2001, DeMaria et al. 2006). Meanwhile, LGEM uses a relatively small number of predictors to form a growth parameter that forecasts for TC intensity using a logistic growth equation, wherein TC intensity is confined between zero and an upper bound determined by a maximum potential intensity parameter.

Lastly, consensus models combine intensity forecasts from multiple models and use a variety of methods to derive the weights for the selected models. These models can be dynamical, statistical-dynamical, or a combination thereof. To date, consensus models have outperformed the other model types in the Atlantic and eastern North Pacific basins, but they are followed closely by statistical-dynamical models and recently by the best-performing dynamical models (Stewart 2016; Blake 2014; Pasch 2015).

Here we propose and develop an innovative way to forecast for TC intensity, RI, and RW by using a statistical-dynamical model derived from a large ensemble of algorithms which are generated through evolutionary programming (EP, Fogel 1999). While convective-permitting grid scales have brought increased skill to TC forecasts by dynamical models (Gopalakrishnan 2011), obtaining large and diverse ensembles at such grid spacing is not easily feasible using the traditional initial condition and model physics perturbation strategies. Consequently, skillful and reliable probability density functions (PDFs) should not be expected to stem from NWP models. In contrast, EP was specifically developed in the 1960s to produce large-member ensemble forecasts by utilizing the evolutionary the principles of reproduction and mutation to develop through selective pressure predictor combinations that maximize forecast skill (Fogel 1999). These EP generated predictor combinations have shown superior performance over dynamical models in 500-hPa heights forecasts (Roebber 2013), as well as statistical-dynamical models like Model Output Statistics (MOS) in minimum temperatures forecasts (Roebber 2010, 2015ab). Furthermore, algorithms generated through the EP process provide forecast PDFs superior in probabilistic and deterministic skill than many traditional models, particularly at the tails of the distribution (Roebber 2013). This is not only the result of an EP process that can create more skillful forecasts, but also of a process that allows for more heterogeneity amongst the algorithms. Additionally, while EP crosses over into machine-learning, the genetic structure of EP algorithms can still be formed in such a way that the algorithms are readily interpretable. Lastly, this flexible process can be structured to use existing predictors already in use by other statistical-dynamical models forecasting for TC intensity, RI, and RW.

Section 2 provides a more complete, yet still generalizable, conceptual overview of EP before the rest of the paper goes on to describe in detail the development of a TC intensity model

for the North Atlantic and eastern/central North Pacific basins with Section 3 discussing the data and methods. It fully discusses the specific EP process as well as post-processing techniques used to generate the final structure of the algorithms that constitute the models. The performance of the two models with enlightening case examples is presented in section 4 followed by a discussion of the results more generally in Section 5. Section 6 offers a summary of the presented research before the paper concludes with a discussion of future work in Section 7.

2. CONCEPTUAL OVERVIEW OF EP

EP has been used in a variety of contexts across the meteorological community over the years. For example, Bakhshaii and Stull (2009) applied gene-expression programming to forecasting precipitation over complex terrain, while Roebber (2010) constructed a series of IF-THEN nonlinear equations to forecast for minimum temperature at a site in Ohio. While a general overview of EP is discussed in each of these papers and in Ferreira (2006), a conceptual overview of EP is provided here as well.

The EP process is applicable when there exists a well-defined problem with a clear measure of success, as an algorithm can then be constructed to forecast a solution by combining inputs. The structure of the algorithm can vary based on the problem, but for simplicity let us consider an EP process that generates a multiple linear regression (MLR), where the solution is the result of summing input variables that are weighted by a coefficient. In EP, the coefficients and predictors are randomly initialized and, in the case of more complex algorithms, some of the operators themselves are randomly initialized. After a population of algorithms is generated, the performance, or fitness, of each algorithm is then measured, hence the need for a clear success measure. This attribute of fitness, with some members of the population exhibiting greater/better forecast skill than others, is reminiscent of fitness in biology and its role in survival and

consequently, reproductive success. By repeating this process (as has happened over the course of history in nature) the successive pressure of fitness drives the algorithms towards better and better (i.e., more fit) solutions. Natural evolution, however, is more complicated than what is presented here as different sexes, tribal groupings, diseases, etc., all affect gene transmission and complicate the evolutionary process. Some of these dynamics may be beneficial to mimic in EP as well, particularly those that might lead to a more genetically diverse population thus helping to generalize the obtained algorithms to a wide range of forecasts. A more complete description of the specific formatting of the EP process used for this research follows in the methods section.

3. DATA AND METHODS

a. Data

Two similar, but distinctly separate, TC intensity models are developed using a perfect-prognostic approach, with one forecasting TC intensity for the North Atlantic basin and the other for the east/central North Pacific basins. The process of developing each model is identical with the only difference being the Atlantic model is trained on data from the north Atlantic basin while the Pacific model is trained on data from the central and eastern North Pacific basins. TC intensity and predictor data is sourced from the SHIPS developmental database, which contains 0-h reanalysis data every 6 h for all TCs from 1982 to present for the North Atlantic and eastern/central North Pacific basins (DeMaria and Kaplan 1994). Despite this long time series, only storms from the year 2000-onward are utilized for model development and evaluation as the reanalysis data prior to 2000 is derived from the Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) rather than the Global Forecast System (GFS; NWS 2016) model as is used for TCs since 2000. Meanwhile, storms through the year 2016 are included, coinciding with the start of this project. While 2017 cases could have been included as well, it was elected to leave this

out for multiple reasons. First, while a subset of TCs from the 2000-2016 dataset is reserved for independent testing, setting aside an entire season provides a wholly independent testing dataset without the chance for storm-to-storm interactions across datasets. Secondly, the 2017 season is ideal for testing as the North Atlantic saw above-normal TC activity and challenging forecasts, particularly at shorter lead times, while the eastern North Pacific basin also saw above normal activity. Specifically, for North Atlantic basin 407 official forecasts were issued in 2017, while intensity errors for the official National Hurricane Center (NHC) forecasts over the 12-48 h lead-times were 10-15% higher than their 5-year means (Cangialosi 2017). Lastly, archived real-time predictor information from 2017 allows for retrospective testing of the operational (i.e., real-time) performance of the models. Thus, having both analysis and archived real-time predictor information for the 2017 season allows for an estimation of the degradation in performance that results from transitioning away from the analysis data used in training and towards the forecast (i.e., imperfect) data used in operations.

For each forecast period, over 80 predictor variables are available from the SHIPS dataset with variants of these variables pushing the total number of predictors to up over 100. A variant of a predictor variable, as defined in this dataset, is the same variable but averaged over a different radial distance, for example precipitable water averaged over 0-200 km from the storm center versus from 0-800 km. However, when more predictors are kept, the solution-space that must be explored grows larger, thus increasing the time and information needed to adequately search the space. This potentially compromises performance because it may be hard to search the solution-space completely and furthermore much of the solution-space may contain expansive deserts of unskillful solutions. This is especially true when the dataset is small, as in this study, as compared to what is typical for machine-learning. While there is no good method to know at

what point the solution space is of optimal size, it is often desirable (as in this study) to reduce the number of predictor variables through the application of domain expertise and selecting only the most meaningful variables.

From the full predictor list 88 potentially meaningful variables are initially selected and the 0-h linear correlation between each paring is tabulated. These correlations are then consulted when using domain expertise to select a subset of variables that are relatively independent of each other, with no two variables having a linear correlation above 0.8. This process results in 47 variables, which are then further refined to 34 variables to eliminate variables deemed to be of lesser importance. However, the dimensionality of the search space resulting form 34 variables is believed to still be too large, especially for the relatively small amount of data that was available. For reference, the total number of forecasts available in the North Atlantic basin is 6110, whereas for the east/central North Pacific it is 6443, and this is before processing the data to remove undesirable cases such as those with missing predictor information. Thus, the remaining 34 variables are separated into groupings of similar dynamical properties (e.g. thermodynamics, moisture, shear, etc.) and a single variable is chosen based on domain expertise to be representative of the general dynamical characteristics of all the variables in that particular group. This results in eight variables (Table 1), which are then passed to the EP for processing. The majority of these variables are converted to standard anomalies (Grumm and Hart 2001) to better identify the relative magnitude and rarity of each predictor value as well as to aid a direct comparison between variables of dissimilar units. However, the 0-600 km average symmetric tangential wind at 850 hPa from NCEP (National Center for Environmental Prediction) analysis (TWAC) is notably non-Gaussian (not shown) and therefore is converted to a linear scaling from (-1 to 1), with the extremes representing the maximum and minimum values of TWAC in the

training dataset. Lastly, a constant value of 10 is provided as a potential predictor variable as well. The purpose of this constant is explained later when discussing the algorithm structure.

Once the desired variables are chosen, the dataset is then processed to remove cases with missing predictor information. Even if only one of the chosen predictor variables is missing it is still more advantageous to remove the case rather than to fill the gap with climatological values or neighboring values. This is because the potential for non-linear relationships between variables as well as the fact that climatological values in terms of standard anomalies are zero, could lead to large changes in the training forecasts. Furthermore, operationally the model cannot be run with missing predictor information so removing these cases ensures consistency with operational practice. Lastly, cases where the forecast initializes or verifies over land are removed since an inland decay model (Kaplan and DeMaria 1995, Kaplan and DeMaria 2001, DeMaria et al. 2006) is used operationally to post-process the intensity forecast and account for any weakening that occurs due to the TCs interaction with land.

With the processing of the data complete, the remaining cases are assigned to either a training, cross-validation, or independent testing dataset. However, the dynamical and empirical relationship between predictor variables likely varies as a function of intensity. Therefore, it would be undesirable if for example the algorithms were trained on TCs of a weak intensity but tested on TCs of a strong intensity, as this would result in the algorithms attempting to forecast for something they have not been calibrated to. Consequently, each TC in the dataset is categorized into three intensity classes based on the maximum-achieved intensity in its lifetime; tropical storms, weak hurricanes (category 1 or 2), and major hurricanes (category 3, 4, or 5). TCs and all their respective forecasts are then pulled from each of these three intensity classes to form the training, cross-validation, and independent testing datasets so that the intensity bias

across the three datasets is mitigated. At this point the dataset has been cultivated and the EP process can begin.

b. Evolutionary Programming

From the cultivated dataset, a large ensemble of algorithms is generated using a perfect prognostic approach, with the EP process of cloning, mutation, and selective pressure used to determine the empirical relationships between the predictor variables and TC intensity. As in previous studies (e.g. Roebber 2010, 2013, 2015ab), the basic genetic architecture of an algorithm is a summation of IF-THEN equations, which in this model are structured to forecast a 12-h adjustment to a persistence forecast. Operationally, this adjustment is calculated using values derived from the forecast fields of the GFS at the end of the specified 12-h interval, when the intensity forecast verifies, except for the DELV predictor, which is the change in intensity over the previous 12-h. So initially the change in intensity over the previous 12 h and the 12-h forecast fields are used to calculate an adjustment, which gets added to the observed 0-h intensity to become the 12-h forecast. Then the process iterates forward and the 24-h forecast fields are used to calculate an adjustment from the 12-h forecast, which becomes the 24-h forecast. The DELV parameter meanwhile would be the forecast change in intensity from the 0-h observation to the 12-h forecast. This process is then repeated iteratively to obtain intensity forecasts in 12-h intervals out to 120 h. Conversely, in training, the algorithms use only the 12-h analysis fields of the chosen predictors to make a 12-h forecast.

The structure of a single algorithm can be written most generally in the following form:

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

An algorithm consists of five lines that sum together to forecast for a change in intensity over a 12-h interval. V_{ij} can be any of the input variables including the constant value of 10; R_{i1} is a relational operator (≤ or >); C_{ij} are real-valued constants ranging from [-1,1]; and O_{ij} are operators (+ or *). The reason for adding a constant value of 10 to the list of input variables is to allow the EP process to generate lines within the algorithm that always execute, or never execute, as is deemed necessary by the EP, since no variable theoretically should exceed +/- 10 standard deviations. Additionally, the constant value of 10 provides a scaling factor for the EP to use when calculating the adjustment if necessary. This algorithm structure with the conditional statements and the potential for linear and non-linear predictor combinations allows for a flexible yet interpretable equation that can be linked back to logical and dynamical processes familiar to a forecaster.

Figure 3 provides a schematic overview of the EP process and can be used as a visual aid for the explanation that follows. The EP training process starts with a population of 10,000 algorithms of the aforementioned form that are randomly initialized from the eight variables listed in Table 1 as well as the constant value of 10 (top left of Fig. 3). While the size of the population 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). Once a population of algorithms is initiated, they forecast on the training dataset to determine their fitness/skill, from which the 2,000 worst-performing algorithms, as based on root mean square error (RMSE), are eliminated leaving only 8000 algorithms (top right of Fig. 3).

Next comes the "evolutionary step," which is where the next generation of algorithms is produced through cloning and mutation, the result of which is increased exploration of the

solution space. The evolutionary step starts by cloning the 2,000 best-performing algorithms (again as determined by RMSE) thus returning the population to its full capacity of 10,000 algorithms. The 2,000 cloned algorithms additionally undergo a mutation where a randomly selected line is completely and randomly reinitialized. Meanwhile, middle-performing algorithms – those ranked from 2,001 to 8,000 – then undergo a process of swapping genetic information. In this step, each algorithm swaps a line of genetic information with another randomly selected algorithm. If an algorithm is selected multiple times by the other algorithms to swap information it can result in more than one line being altered or the same line being altered multiple times or a combination thereof. Conversely, because the swapping of genetic information and cloning of algorithms leads to similar genetic information across the population, there is a chance that the line selected for exchange is identical in form resulting in no change to the algorithm, although the odds are against it. After this process these middle-performing algorithms also undergo a mutation in the same manner as the cloned algorithms. Consequently, even if an algorithm did select an identical line for genetic exchange it still undergoes some change. At this point the evolutionary step is complete and the population of equations is in its second generation (bottom-right of Fig. 3). Meanwhile the best 2,000 performing algorithms are left untouched in order to provide a source of good genetic information for future generations.

This new generation of algorithms then forecasts on the cross-validation dataset and the 100 best-performing algorithms (as determined by RMSE) are used to populate the "best algorithms list". The purpose of this list is to ensure that we keep the best-performing algorithms no matter the generation in which they occur, rather than simply selecting the best algorithms from the final generation. The EP process then repeats for 300 iterations and after each new generation the best algorithm list is updated to include any new, particularly skillful algorithms

that emerge, while the less-skillful algorithms get bumped off. While the performance of the algorithms improves rapidly in the first few generations, the rate of improvement eventually plateaus. Small improvements may be made to the skill of worst performing algorithms on the list, but it may not notably improve upon the most skillful algorithms or the skill of the algorithms on the list as a whole (Fig. 4). Therefore, a new population of algorithms is then randomly initialized and it goes through the same EP process. The algorithms this second population produces are also considered for inclusion on the same best algorithms list and must out-perform those already present to be included. Even if this population does not provide much improvement in overall performance it may still help add more diversity to the best algorithm list (Fig. 4). In total, five different populations of 10,000 algorithms are run for 300 iterations to produce the final set of 100 algorithms on the best algorithms list.

As mentioned previously a perfect-prognostic approach is utilized, with the EP model being developed on analysis data, but run operational using forecast data. Despite using a cross-validation dataset to prevent over fitting of the model, this serves to mostly prevent over fitting to the specific selection of training cases. However, the model would still be training on the ideally perfect empirical relationships amongst the predictors, which in operation would not be so perfect. Thus to prevent over fitting of the relationships between variables, noise is added to the analysis variables used in the training. The magnitude of the perturbations have a Gaussian distribution centered on zero, with a standard deviation that is a quarter of the observed standard distribution in the differences between the analysis and forecast values across both the independent and cross-validation datasets from 2010-2016. Furthermore, this noise is dynamic, meaning that each time the algorithms forecast on the testing data the noise added to the analysis variables is changed. However, cross-validation performance and selection for the best algorithm

list is still done using analysis variables without the added noise. This ensures that the EP developed algorithms don't start to train to any anomalous relationships that appear as a result of noise and instead remains connected to real-world relationships.

c. Bayesian Model Combination

After a model is developed, statistical post-processing techniques can be used to bring about improved performance and reliability. The most common techniques involve removing model bias and assimilating other model data and/or climatological data. Here, bias correction, along Bayesian model combination (BMC) is used to post-process the algorithms on the best algorithms list. Bias-correction is applied first, resulting in the final generic form of an algorithm as:

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

where ε is the bias correction factor. Then BMC assigns weights to multiple ensemble members, which taken together produce a forecast that is superior in skill to that from any individual ensemble member (Monteith 2011). The BMC process also produces a PDF, which allows a probabilistic forecast of RI and RW to be produced. While Bayesian model averaging (BMA, Raftery et al. 2005) is more commonly used in the meteorological community over BMC there is a notable distinction that makes BMC a superior post-processing choice. Although BMA also provides a weighting across multiple ensemble members, it actually attempts to select the ensemble member that gives the best individual forecast, after which it applies weights to other members to account for the uncertainty in its selection. Conversely, with BMC there is no assumption that a single ensemble member individually produces the best forecast. Instead, BMC

assumes that the best forecast is obtained through a combined weighting of multiple ensemble members. Consequently, BMC provides superior skill through a more effective weighting of ensemble members (Monteith et al. 2011). However a limitation of this technique is that it is computationally expensive and therefore members must be sub-selected from the overall population (e.g., Hoeting et al. 1999). In this case ten members are sub-selected from the best algorithms list as evaluating ten members with four possible raw weights requires an evaluation of only 4^{10} or 1,048,576 possible combinations while evaluating all 100 members would require evaluating 4^{100} or $1.6x10^{60}$ possible combinations.

For BMC, the ten sub-selected members are chosen such that the mean absolute error (MAE) of the forescasts from the chosen bias-corrected algorithms is minimized while the mean absolute difference (MAD) between their forecasts is maximized, resulting in a subset of algorithms that are both skillful and diverse. Once the ten members are selected the BMC processes then steps through all possible raw weights (including zero) of the individual members, which are then normalized to sum to one. The selected weighting used in the model is the one that minimizes MAE across the cross-validation dataset, with MAE used here over RMSE to mirror the way in which the NHC evaluates the performance between different TC intensity models. With the weightings of each ensemble member obtained, the deterministic forecast is a simple weighted sum of the individual members' forecast. Meanwhile, to obtain a probabilistic forecast, a normal curve is fitted about each ensemble member with the standard deviation being determined from the PDF of 0-12 h intensity change. These individual PDFs are then weighted by the BMC weightings and summed to give the total forecast PDF, from which the resulting RI/RW probabilities at any specified intensity-change interval (e.g. 30kt in 24 h) are calculated.

While the model is producing probabilistic forecasts, they have not been archived or examined yet. This is because the complete PDF that is outputted from the probabilistic model is a weighted sum of PDFs centered on the deterministic forecast from each individual algorithm. As such, the deterministic forecast forms the underpinnings of the probabilistic forecasts. Consequently, focus has primarily been on getting a satisfactory deterministic model first and foremost. That being said, the development of the deterministic model was done with consideration of the probabilistic model and some experimentation with the probabilities has been performed. For example, rather than fitting a Gaussian PDF to each individual algorithm's deterministic forecast a mixed-normal (MN) curve was fitted. This MN curve was a function of storm intensity and accounted for the fact that a strong (weak) TC is more likely to under go RW (RI) than RI (RW) owing to the fact that it is close to (far from) its maximum potential intensity. This MN PDF was also used to post-process the deterministic model by averaging over the MN PDF to also help the deterministic forecasts with RI/RW forecasting. Initial results, however demonstrated that a slight decrease in performance (not shown) and thus it was determined to keep with a normal PDF and try to continue to probe the formulation of the EP process for better performance.

4. RESULTS

a. Model Performance

Operational performance of the North Atlantic (EPA) and eastern/central North Pacific (EPP) models is evaluated using archived real-time predictor information across reserved independent test cases from the 2010-2016 season as well as across the entirety of the 2017 season. While the EP model is paired with the inland decay model for operational purposes, cases where the forecast initializes overland or verifies overland are removed to mirror the way

the NHC evaluates TC model performance. As described previously, the model is developed using analysis data around which noise was added. Thus, to compare the effects of using analysis data versus forecast data the EP models are run with archived real-time predictor data (denoted by "-R") as well as analysis data without adding noise (denoted by "-A"). The performance of these two models is then compared to the Statistical Hurricane Intensity Forecast model accounting for decay over land (OCD5; Knaff et al. 2002), official NHC forecasts, and a persistence model using homogeneous forecast cases.

The OCD5 model uses climatology and persistence to forecast for future track and TC intensity and is therefore often used as a baseline for model skill (Knaff et al. 2002; Cangialosi 2017) with persistence errors instead demonstrating the average magnitude of observed intensity change. When comparing to the other models the performance of the EP models will always be given relative to the model being discussed, however figures will only show overall MAE performance as well as performance relative to the OCD5 model.

From the ten sub-selected members chosen for the BMC process, seven algorithms received a non-zero weighting to become the EPA model (Appendix). The EPA-A and EPA-R models show mixed performance, sometimes performing better than the baseline OCD5 model and at other times performing worse (Fig. 5 and 7). Across the 2010-2016 dataset the 12-h forecast errors from the EPA-A model are 30.9% better than OCD5 and only 3.8% worse than official NHC forecasts (Fig. 6). However, EPA-R performance averages about 1 kt worse than the EPA-A model over the 12-h lead time and therefore performance drops to become 22.0% worse than official NHC forecasts, while retaining a 18.8% improvement over OCD5. The EPA-R model continues to outperform the OCD5 model through the 24-72 h range with improvements of 5.9%-10.2%. However, both the EPA-A and the EPA-R model fail to see a plateau in TC

intensity errors at the later lead times, as is seen in the NHC and OCD5 forecasts (Fig. 5). Consequently, model skill drops off with the EPA-R model becoming 4.6% worse than OCD5 at 96 h and 39.5% worse at 120 h. It should also be noted that the effect of real-time predictor data on model performance varies across the lead times, with the largest degradation in performance coming at 120 h, where the MAE of the EPA-R model is 2.7 kt larger than the EPA-A model. However, at 72 h the EPA-R model performs notably better than the EPA-A model with a MAE that is 1.6 kt less. Across the 2017 North Atlantic season, both the EPA-A and EPA-R models see more consistent performance across all lead times, with the EPA-R model showing consistent improvements over the OCD5 in the 12.7%-20.0% range over the 24-96 h lead times (Fig. 8). Meanwhile, the EPA-A forecasts between 12-96 h are 8.6% -18.8% less skillful than official NHC forecasts over the 2017 season.

In the eastern/central North Pacific basins, two algorithms received a non-zero weighting from the BMC process to become the EPP model (Appendix). Across the 2010-2016 independent test cases, the performance of the EPP-R model starts with a 13.0% improvement over OCD5 at 12 h, but this performance becomes 7.1% worse than OCD5 at 36 h (Fig. 9 and 10). Through the 48-96 h lead times, the EPP-R model ranges from 10.2-12.3% worse than OCD5 before coming within 1% of OCD5 at 120 h. Meanwhile, the use of real-time predictor variables seems to have little impact on the forecasts as EPP-A and EPP-R model performance is similar over all lead times. Across the 2017 season, performance of the EPP-A and EPP-R models roughly matches OCD5 to within 2% over the 12-36 h range, while the EPP-R performance at 48 h drops to be 5.8% worse than OCD5 (Fig. 11 and 12). As the lead time continues to grow, the EPP-R model then becomes better than OCD5 as errors decrease leading to improvements of 8.8%, 13.6%, and 10.6% at the 72, 96, and 120-h lead times respectively.

Meanwhile, the EPP-A errors continue to grow throughout the lead times leading to performance errors that are much worse than OCD5 peaking at 40.0% at the 120-h lead time. The difference in performance between the EPP-A and EPP-R models shows that using forecast predictor variables over analysis variables can have a large impact on the forecast and it does not always have to be a negative impact. Looking at certain cases studies, specifically Joaquin from the Atlantic basin, will demonstrate how this can be.

b. Case Studies

As for TC intensity forecasting in general (Rappaport et al. 2012 and Kaplan et al. 2010), RI and RW cases provide a major contribution to model errors in both the North Atlantic and east/central North Pacific basins. For example, if we sum the performance errors across at the 24-h lead time over the 2010-2016 independent dataset in the North Atlantic basin, we see that the 16 RI/RW cases, which comprise only 6.0% of the forecasts, account for 17.4% of the errors. Over the 2017 season, the 40 RI/RW cases, which comprise only 15.2% of the forecasts, account for 34.7% of the forecast errors in the North Atlantic basin. Meanwhile, in the eastern/central North Pacific basins the 16 RI/RW cases from 2010-2016, comprising 7.0% of the forecasts, account for 14.9% of the errors. Over the 2017 season, the 28 RI/RW cases, which comprise only 14.7% of the cases, account for 39.0% of the errors.

To analyze certain RI/RW cases, it is helpful to know just how much each predictor contributes to the overall intensity forecast. However, with the algorithms containing conditional statements and non-linear predictor combinations getting a direct measurement of this is not straightforward. Here, the relative contribution from each variable is obtained by re-running the forecast with the variable of interest set to an input value of zero (i.e., a climatological value) for

the given lead time. The direction and magnitude of the change in the forecast intensity as compared to the original forecast then tells us the impact that variable has on the forecast. For example, if a predictor is set to zero and the resulting intensity forecast decreases by 5 kt, then it is said to have had a 5 kt contribution to the original forecast. Conversely, if zeroing out a predictor leads to an increase in the forecast, then 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 a12-h interval as well. An estimation of the relative contribution of a variable at for example, 36 h, still presumes an accurate 24-h forecast with no zeroing of the variable at the earlier lead times. Thus, the relative contribution at later lead times is estimated by summing up the contribution of the variable over each 12-h interval.

Maria – 0000 UTC 18 September 2017

Maria started out as a tropical depression around 1200 UTC 16 September 2017 over the tropical Atlantic. The depression then moved westward toward the Lesser Antilles and quickly strengthened into a tropical storm and then by 1800 UTC 17 September 2017, a hurricane (Fig. 13). Aided by warm waters and weak shear the storm continued to rapidly intensify, becoming a 100-kt major hurricane 24-h later at 1800 UTC 18 September 2017. Maria then made landfall on the island of Dominica around 0115 UTC 19 September 2019 as a 145-kt hurricane (Pasch et al. 2017). Maria's intensity weakened slightly after crossing Dominica, but it quickly reorganized and went on to reach a peak intensity of 150-kt before striking Puerto Rico. After crossing Puerto Rico and dropping to an intensity of 95 kt, Maria was able to retain its strength for a few days as it reached a relative peak of 110 kt and held onto the classification of a major hurricane.

However, as Maria progressed northward and into the mid-latitudes, it slowly weakened into a weak hurricane before being swept up by the westerlies without ever making landfall again.

One of the largest 24-h forecast errors from the EPA-R model is a 51 kt under forecast of TC Maria given by the forecast initializing 0000 UTC 18 September 2017 (Fig. 14). At this time Maria was in the middle of a nearly unprecedented RI event, wherein it strengthened 115 kt over 60 h, and furthermore was at a time when the rate of intensification was beginning to increase. The 24-h verification then occurs at the conclusion of the RI event on 0000 UTC 19 September 2017, just before Maria crossed Dominica. While the EPA-R model forecast Maria to strengthen from 75 kt to 94 kt over the 24 h, Maria instead strengthened much more rapidly reaching 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 period. With a forecast of 100 kt and a 45 kt error the NHC had a slightly better forecast but also underestimated the rate of intensification. With the 24-h forecast from the EPA-R and the EPA-A model being so similar only the later will be analyzed as the input variables truthfully represent the environment.

At the 24-h verification time the depth of the 26˚C isocline (CD26) was 1.2 standard deviations (σ) above climatology while vertical wind shear (SHDC) was 0.9 σ below climatology (Table 2), indicating a favorable environment for RI with warm waters and weak vertical wind shear. Despite the warm waters, CD26 has little to no impact on the forecast contributing only 0.1 kt to the forecast intensity change of 18-kt from the EPA-A model. This is the result of the CD26 predictor lacking presence in the algorithms and receiving a relatively small weighting when it does appear. In algorithm 6 and 8 the CD26 parameter is absent all-together while in algorithm 9 and 34 it only appears in the conditional statement. Furthermore, while CD26 appears in algorithms 34 and 49, it only appears once and with a small coefficient that is made

smaller by being multiplied by another variable with a small coefficient. Only in algorithm 35 does CD26 have a good chance to make a solid contribution to the forecast. Meanwhile, SHDC has only a modest impact on the forecast contributing 2.4 kt to the 12-h forecast and 1.7 kt to the 24-h forecast. Combined over both lead times SHDC contributes 3.1 kt in total to the 18-kt (24 h)$^{-1}$ intensity-change forecast. The primary predictor contributing to the forecast is the DELV parameter, which was 1.3 σ above climatology to begin with and then forecast to be 0.8 σ above climatology over the next 12-h interval. This resulted in a 5.4 kt and 4.5 kt contribution to each 12-h forecast interval thus having a 9.9 kt contribution to the total 18-kt (24 h)$^{-1}$ intensity-change forecast. While the 200-hPa divergence averaged over 0-1000 km of the storm center (D200) was 1.8 σ above climatology at 24 h this too had little impact on the intensity forecast, contributing only 0.2 kt because of the relatively sparse appearance of the variable in the algorithms. While these variables indicate a favorable environment for RI, it is reasonable to say that they do not necessarily show any indication that such impressive RI would be seen. Consequently, even though we would like to improve upon this, it seems reasonable that a first-attempt machine-learning algorithm such as the EP model used here might underforecast this event with the large error from the NHC forecast seeming to verify this as well.

Otis – 0000 UTC 18 September 2017

For much Otis' lifespan, it was hindered by strong wind shear and associated dry air infiltration and therefore was unable to develop into a hurricane as it traveled westward across the eastern North Pacific (Fig. 15). But, on 17 September 2017, the storm turned northward into a more-weakly sheared environment, which helped to slow the intrusion of dry air into its center. Consequently, Otis underwent RI, increasing 60 kt over 24 h to reach a peak intensity of 100 kt.

However, as the northward progression of the storm continued it brought the storm back into a more-highly sheared environment, reestablishing the intrusion of dry air into the storm center. As a result, Otis experienced RW, decreasing 60 kt over the subsequent 24 h to return to an intensity of 40 kt, before continuing to weaken and eventually dissipating 48 h later. The forecast initiating on 0000 UTC 18 September 2017 when the storm was at its peak intensity of 100 kt and transitioning from RI to RW featured the largest 24-h forecast error by either the EPP or the EPA model (Fig. 16).

When looking at the relative contribution from each input variable a clear explanation emerges for why this was the case (Table 3). At the analysis time, DELV was 3.1 σ above normal, corresponding to the RI that Otis just underwent. However, the dry-air predictor (CFLX) is also well-above-normal at 3.5 σ, indicating that a lot of dry air is being mixed back into the storm. Despite their similar magnitudes, the two variables have distinctly different magnitudes of impact on the forecast. While the DELV predictor contributes 15.7 kt to the 12-h forecast, the CFLX predictor contributes only -0.9 kt. This greater contribution of DELV not only stems from the weighting of the parameter in the calculations, but also its role in the conditional statement and determining which lines get executed. For example, given the analysis values from the 12-h lead time (Table 3), line 2 of algorithm 31 does not execute because the conditional statement is false (Appendix). When the CFLX predictor is zeroed the statement becomes true, but then the calculated adjustment given by that line equals zero. Therefore, CFLX is said to have no impact on the forecast adjustment given by that line thus helping DELV to have a larger contribution as compared to CFLX. At the 24-h lead time the value of the DELV parameter drops to 0.8 σ, while the CFLX parameter remains high at 3.4 σ. However, because the DELV parameter has a larger relative contribution to the algorithms, the relative contribution from the two variables turns out

to be roughly equal and opposite, with DELV contributing roughly 5.3 kt to the intensity-change forecast and CFLX contributing -4.6 kt to the intensity-change forecast. Meanwhile, other variables feature only modest deviations from climatology and a mixed impact on the forecast. As a result, rather than forecasting for a sharp change in intensity, the EPP model forecasts a more rounded change as the DELV parameter needs eroded before the effect of the CFLX predictor can take hold.

Joaquin – 1200 UTC 3 October 2015

While the previous cases demonstrate the challenge of forecasting for RI and RW, they also demonstrate that generally the use of analysis data versus real-time data has little impact on the forecast. This is because for most variables the errors in the real-time values are small compared to their climatological range and consequently when converted to standard anomalies the errors are minimal and contribute little to differences between the forecasts. However, sometimes post-season analysis finds times where the intensity of the storm was mischaracterized. Not only does this mean that the real-time operational model uses the wrong starting point from which the intensity-change forecast adjusts, but it also means that the initial DELV parameter that gets calculated from the observed and forecast intensities is incorrect. Furthermore, given that the NHC bins intensities in 5 kt intervals, the smallest possible non-zero error in the input value is 0.52 σ. Thus, the resulting errors in the DELV parameter can be relatively large compared to its climatological range, which may lead to large impacts on the forecast. In the forecast for Joaquin initializing1200 UTC 3 October 2015 though, this mischaracterization actually led to an improved intensity forecast.

Joaquin began as a surface low about 355 n mi east-northeast of San Salvador Island in the Bahamas on 1800 UTC 26 September 2015, and by 0000 UTC 28 September 2015 it was classified as a tropical depression (Fig. 17). Joaquin then drifted to the southwest over warm waters and began to intensity. By 0000 UTC 1 October 2015, Joaquin was a major hurricane located 90 n mi to the east of San Salvador Island. However, as a trough deepened over the eastern United States, Joaquin slowed down, made a hairpin turn, and started to move the northeastward toward where it originated. Aided by the warm waters Joaquin continued to be classified as a major hurricane, and as the trough weakened Joaquin was able to reach a peak intensity of 135 kt on 1200 UTC 3 October 2015. Thereafter, vertical wind shear increased and the category-4 hurricane weakened to a category 1 in just 36 h. From there, Joaquin was picked up by the westerlies and carried off to the east.

While Joaquin was noted to have reached a peak intensity of 135 kt at 1200 UTC 3 October 2015 before undergoing RI and decreasing 60 kt over the next 36 h, in real-time the storm was thought to be at 115 kt. With both the EPA-A and EPA-R models failing to pick up on the RW of the Joaquin, this underforecast allowed the EPA-R model to have a much more accurate forecast than the EPA-A model (Fig. 18). Still, it is interesting to note that over the 0-12 h lead time the EPA-A model notably increases Joaquin's intensity, while the EPA-R model does so only slightly. This is primarily due to the difference in the DELV parameter as a result of the mischaracterization of the storm's intensity. The analysis data shows the DELV parameter to be 1.8 σ for the 12-h forecast, resulting in a 6.3 kt contribution (Table 4). Therefore, the behavior of the forecast is similar to what was discussed with Otis and the EPP model with the DELV predictor having a greater impact than other variables and leading to a forecast increase in intensity. In the real-time data though the DELV parameter was only 0.3 σ as a result of a much

lower intensity and, thus DELV contributed only -0.7 kt to the forecast allowing for a flatter intensity curve forecast. This combined with a lower initial intensity allowed the EPA-R model to have a more accurate forecast than the EPA-A model.

Harvey – 1800 UTC 24 August 2017

Harvey began its life as a tropical storm out in the tropical Atlantic, but while moving westward across the Caribbean Sea strong vertical wind shear caused the system to devolve into a tropical depression and then further to a tropical wave (Figure 19). After crossing the Yucatan Peninsula, convection became more persistent and Harvey became a tropical storm again. Then, on 23 August 2017 and continued to rapidly intensify as it encountered warm waters, high mid-level moisture, and weak vertical wind shear. Harvey became a hurricane just after 1200 UTC on 24 August 2017, and continued to rapidly intensify up until landfall reaching a maximum intensity of 115 kt. Once Harvey moved overland, it underwent RW and just 12 h later was again a tropical storm again as it sat over the Texas coast giving record breaking rainfall to the area.

The case studies examined so far have only looked at times when the EPA or EPP models performed poorly, however there are cases where the model accurately picked up on RI/RW. The best of example of this is with the forecast of Harvey initializing 1800 UTC 24 August 2017, just after Harvey became a hurricane. However, the accuracy of the forecast is aided by the fact that Harvey had already been intensifying at a rate of 10 kt (12 h)$^{-1}$ when the forecast initialized and continued to intensify at roughly the same rate. Consequently, DELV that was a hindrance to model skill in previous forecasts was more appropriate in this one.

Initializing with an intensity of 70 kt, the EPA-A model forecast Harvey to strengthen 20.9 kt over 24 h to reach an intensity of 90.9 kt. While the EPA-A model picked up on the RI

event continuing the forecast 90.9 kt was still 14.1 kt under the observed intensity of 105 kt. The forecast by the EPA-R model though was aided by a 5 kt overestimation (compared to post-season analysis) of the initial intensity of Harvey. As a result, the input value of the DELV parameter to the EPA-R model was larger, resulting in a slightly larger intensity-change forecast of 25.2 kt (c.f. Tables 5 and 6). This combined with the 5 kt head start in intensity allowed for a more accurate prediction of the RI event with a 100.2 kt intensity forecast at 24 h, 4.8 kt below what was observed. Beyond 30 h Harvey moved overland triggered the inland decay model, which did a decent job capturing the decrease in intensity of the storm.

5. DISCUSSION

The multiple case studies presented in the previous section highlight the high relative contribution of the DELV predictor in the EP model and perhaps an overreliance on it. One might jump to the conclusion that this is a bad predictor and that it should not be used, but it is just the opposite. The EP process selected this predictor to be one of the most important, and therefore weighted it more heavily to increase the skill of the model across training data. The importance of keeping the DELV predictor is supported by its used in both the OCD5 and SHIPS model (Knaff et al. 2002, DeMaria and Kaplan 1994, Shimada 2018). The parameter represents persistence, however, persistence is not always a great baseline for a forecast, as in the case of RI and RW. Instead, future work will need to focus on ways to lessen the reliance on the DELV predictor and help the model identify times in which maybe it should be weighted less. Ideas for how to do this are discussed further in the future work section.

One may also wonder if the selection of only two algorithms by the BMC process to be the basis of the eastern/central North Pacific model is too few and if the weighting of more algorithms would be better. However, of the ten members that were sub-selected to undergo the

BMC process this weighting of the two algorithms produced the best results. As such, even some small non-zero weighting of the other algorithms to help aid diversity would act only to degrade the forecast over the cross-validation dataset without anyway to know whether this would be beneficial to the operational performance of the model. Therefore, their inclusion is not as beneficial as it may seem. However, one could rightly question whether the ten members selected to undergo the BMC process were the best selection. Maybe there is a subset of algorithms that at their face value might have worse skill, but when combined via BMC might produce superior results. Yet it is not obvious how to identify the selection of such members. Furthermore, owing to the similarity in performance amongst the 100 best algorithms any improvements from a different sub-selection of members for the BMC process would likely be minimal with no guarantee that the performance would be distinctly better that the current model.

6. SUMMARY

In this paper the development of two TC intensity models, one for the North Atlantic and another for the eastern/central North Pacific basins, were developed from a large ensemble of algorithms. These algorithms forecasted a change in intensity over a 12-h period using an IF-THEN structure as well as linear and non-linear predictor combinations. Run iteratively, these algorithms produced a deterministic forecast for TC intensity every 12 h out to 120 h. Each algorithm had access to eight predictors from the SHIPS developmental dataset, which were converted to standard anomalies to aid comparison between variables of dissimilar units. The algorithms also could use a constant value of 10 as a scaling factor in calculations or to create conditionals statements that always had a certain outcome. While the algorithms were trained via a perfect-prognostic approach noise was added to the training dataset to prevent overfitting of the

analysis data as forecast data would be used in operations. After being randomly initialized, the EP process involving cloning, mutation, and selective pressure drove the algorithms towards skillful predictor combinations. In total five populations with 10,000 algorithms were run over 300 iterations to produce algorithms that competed to get on the 100 best algorithms list. Bias correction was then applied to all the algorithms on the list before ten algorithms were sub-selected to undergo BMC. These members were sub-selected based on minimizing the MAE amongst the algorithms but maximizing the MAD, so that the ten members were both skillful and diverse. Finally, BMC was used to determine the weighting for each of the ten members to produce the final model, which would have an overall skill better than any single algorithm.

Each model was tested on reserved independent storms across the 2010-2016 seasons as well as across the entirety of the 2017 season. In the North Atlantic basin, the model showed improvements over OCD5 at all but the latest lead times over both the 2010-2016 independent dataset and the 2017 season. Meanwhile, in the eastern/central North Pacific basin the model struggled to produce more skillful results than OCD5. Case-studies further demonstrated that both models struggled to forecast for RI and RW and that this was likely the result of an overreliance on the DELV predictor. Furthermore, mischaracterization of the storm's initial intensity at times lead to large errors in the DELV predictor value, which could significantly affect a forecast. However, the use of forecast values amongst the other predictor variables had less of an impact on the forecast. This is because the errors in the forecast values were small relatively compared to their climatological range, unlike the case for DELV. Consequently, when converted to standard anomalies the errors are minimal and contribute little to differences between the analysis forecasts and the real-time forecasts.

While the current model shows only small improvements over the baseline OCD5 model it is believed that there is still unrealized potential in the EP process. In EP, as with other machine learning methods, many and almost countless tweaks and changes can be made to the structure of the process. Uncertainty though lies in knowing to what effect certain changes may have on algorithm performance. While a certain tweak may be thought to help algorithm performance it may have unforeseen consequences that result in just the opposite. Other times, a change to the process may lead to an increased computational load without necessarily producing any significant improvement. There are many other ways this model could be constructed going forward, and this will be contemplated to figure out how to extract the maximum performance from the EP process, as well as ways in which the algorithms could be post-processed to extract even more performance. Some ideas for how to improve the model are be discussed in the next section.

7. FUTURE WORK

As was discussed earlier, it seems that there may be an overreliance on the DELV predictor in the current model formulation, but it is not yet clear how to change this. One method would be to return to a larger number of input predictors. However, if not done thoughtfully this can be a somewhat imprecise approach to the problem. This is because while increasing the search space with more predictors diminishes the prevalence of the DELV predictor, if one does not know which predictors to add then the EP process will spend time trying to search through unskillful predictors and incorporate less skillful predictor combinations. However, going forward this seems to make the most sense as getting rid of the DELV predictor seems like a worse option. As discussed previously, the DELV predictor clearly exhibits skill, so it should be included. As of right now though there is no good method to know at what point the solution

space is of optimal size so trying to balance the DELV parameter with other variables is a somewhat inexact science.

One variable that may be worthwhile to include is the intensity of the storm at the start of the forecast period (VMAX). While initially the inclusion of another persistence variable brings to mind the problems that came with the DELV predictor, this likely would not be seen. While the magnitude of errors in VMAX would equal those of DELV the former has a larger climatological range, thus the relative size of the errors in the VMAX predictor would be smaller. Consequently, there should not be as large impacts of to the forecast if the initial intensity of the storm is mischaracterized. Furthermore, while DELV is strictly a persistence variable, VMAX does bring with it more dynamical meaning. A direct measure of the intensity of a storm can be contextually important to other parameters such as VMPI and DELV, as well as vertical wind shear and moisture variables. For example, consider the case of an intense hurricane where the VMPI is above normal but the observed intensity is close to the VMPI. In this case the storm is unlikely to strengthen much further and rather is actually more likely to undergo RW than RI. For this reason VMAX is actually negatively correlated with intensification as was found by DeMaria (1994) and independently through experimentation in this study (not shown). Consequently, even if the DELV parameter is well above climatology the inclusion of a VMAX predictor may allow the model to choose more accurately when to incorporate it. Lastly, VMAX can also give information on the organization of the storm and how resilient it is to shear and the entrainment of dry air.

Looking holistically at the SIHPS parameter list, the majority are environmental and synoptic predictors. Consequently, there is a lack of predictor information with a concentrated focus on the inner-core of the storm where dynamical processes controlling RI and RW are most

prevalent. The inclusion of variables representing inner-core properties would then likely provide useful to the model. This data would likely have to be sourced from other datasets, but microwave satellite imagery for example might be able to give more information about the convection occurring in the inner-core of a TC.

Beyond looking at which predictors are chosen one could also question the choice of using predictors from the verification time to forecast for a change in intensity. The current thinking is that the initial intensity of the storm at the start of each interval will capture how favorable the environmental conditions and dynamics are and, consequently, using forecast variables at the end of that interval demonstrates where the environmental conditions are headed. However, using forecast predictor information at the end of a forecast interval is not necessarily guaranteed to be representative of the environment across most of the forecast interval. Instead, it may be advantageous to average the value of predictor variables from the start of the interval time with those at the end to have a better representation of the conditions throughout the interval over which the storm is changing. Or, since the data are available in 6-h intervals, one could simply take the forecast valuables of the predictor variables from the middle of the forecast interval. How much impact this will have, though, is not yet clear.

Transitioning away from the variables and towards the EP process generally, a fundamental problem of forecasting for TC intensity using a machine-learning technique is the quasi-Gaussian nature of TC intensity change. Across all 12-h intervals there are many more periods where a TC's intensity does not change, or changes very little, as compared to when it changes a lot (not shown). Consequently, poor model performance on these relatively few cases, which may feature different environmental and dynamical characteristics, may be outweighed by the necessity to fit all cases. While using RMSE in the training process helps to inflate the large

errors that might typically be seen in a RI and RW forecast, it seems more needs to be done to account for this. One idea is to use bootstrapping to increase the prevalence of RI and RW forecasts in the dataset. What cost this might have on overall performance as well as false alarms of RI and RW, though, is unknown.

Lastly, little attention has been paid here to the probabilistic performance of this model. With the deterministic forecast forming the basis for the probabilistic model, it is important to first have a deterministic model that performs well and promotes confidence before attempting to develop and analyze the probabilistic model. This has not yet occurred and, consequently, not much is known about the performance of the probabilistic forecasts. This will be the next major area of research and model development.

FIGURES

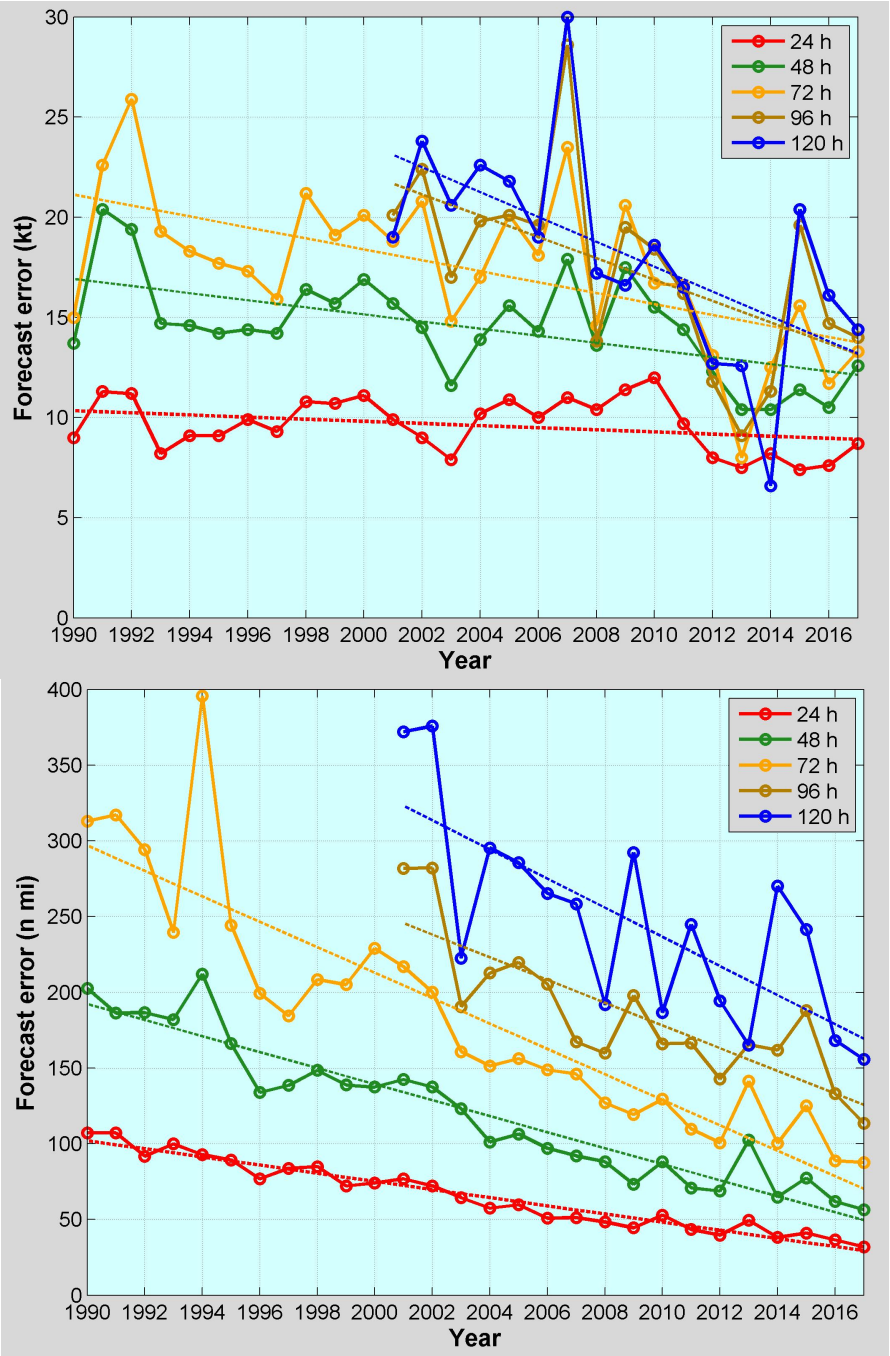

Figure 1: Annual average official NHC intensity errors (top) and track errors (bottom), for the North Atlantic basin for the period 1990-2017 with least-squares lines superimposed (Cangialosi 2017).

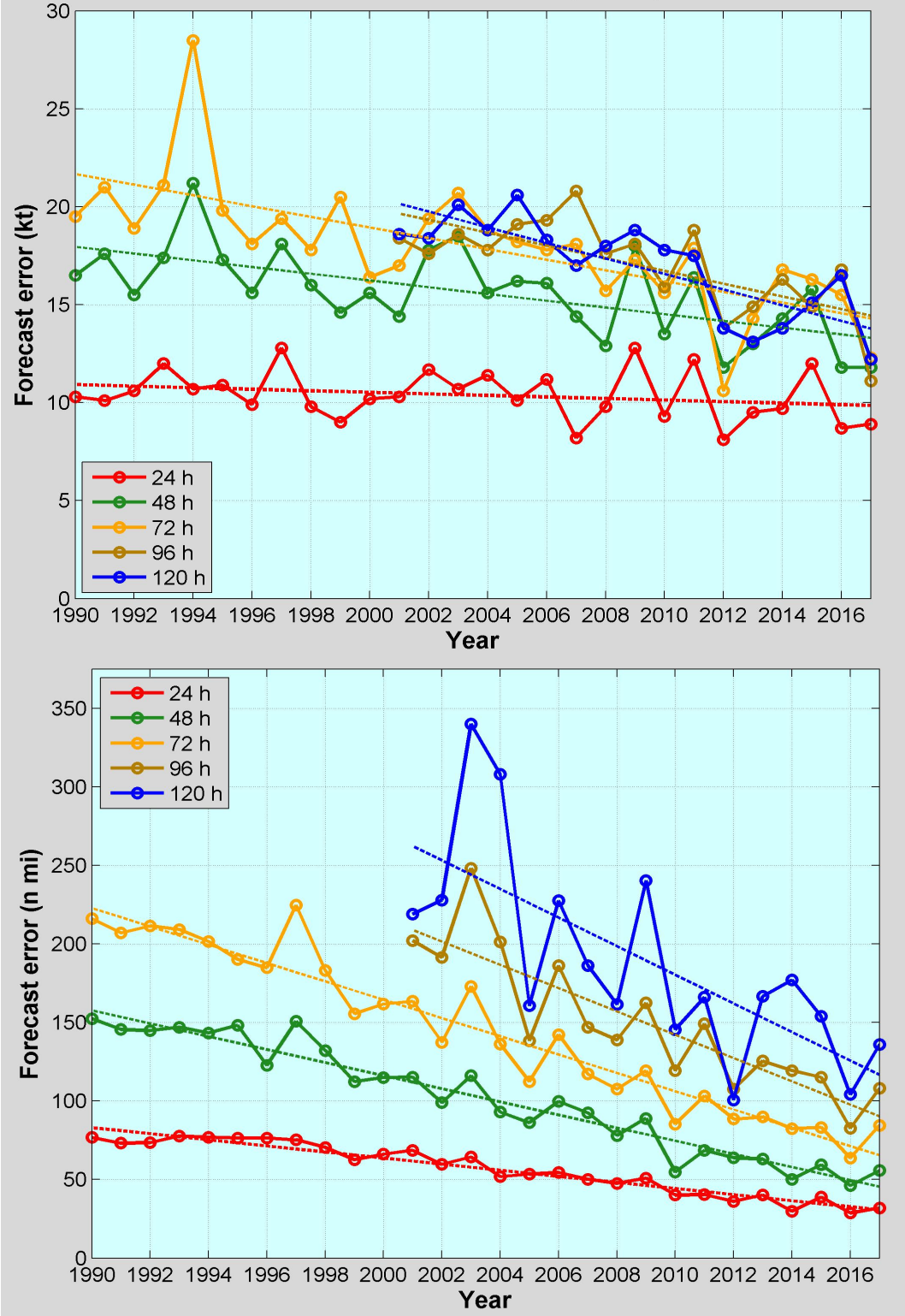

Figure 2: Same as Figure 1 but for basin in eastern North Pacific TCs (Cangialosi 2017).

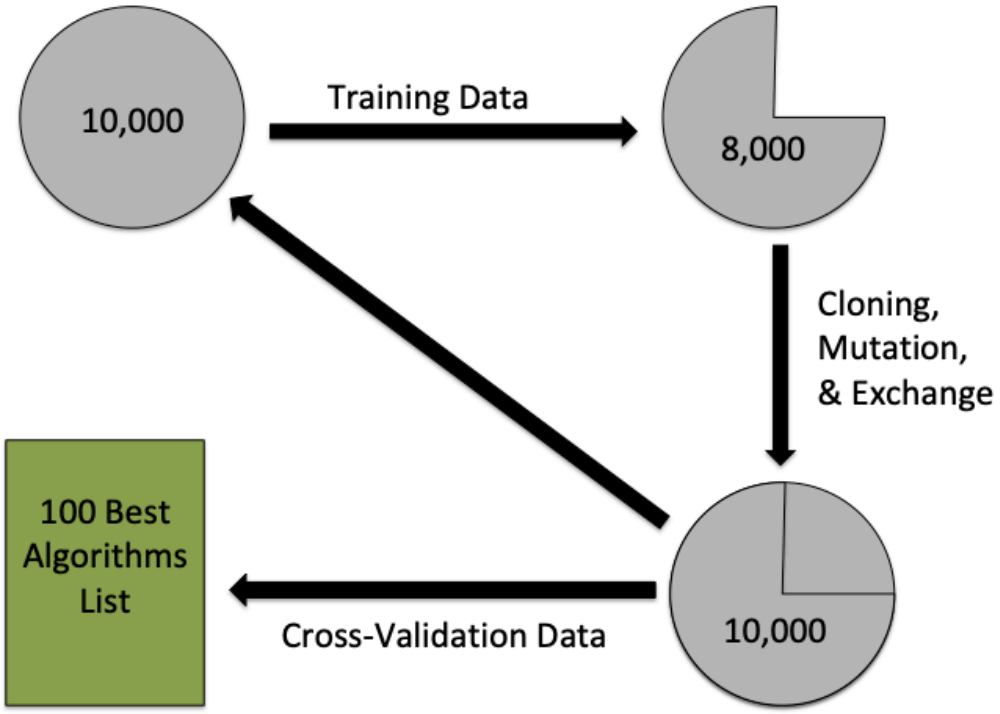

Figure 3: Schematic Overview of the EP process used to train the algorithms.

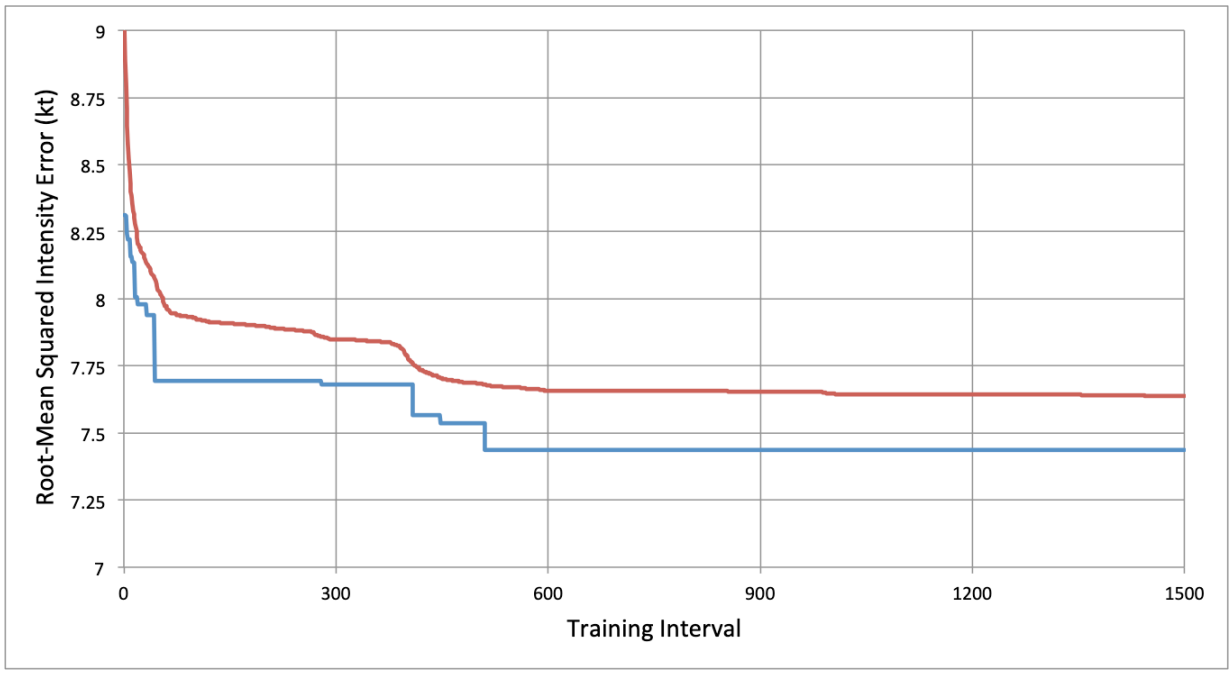

Figure 4: Range of performance of the algorithms on the best algorithm list as given by the best algorithm (blue) and worst algorithm (red) through the eastern/central North Pacific training process covering the 300 intervals of each of the 5 populations.

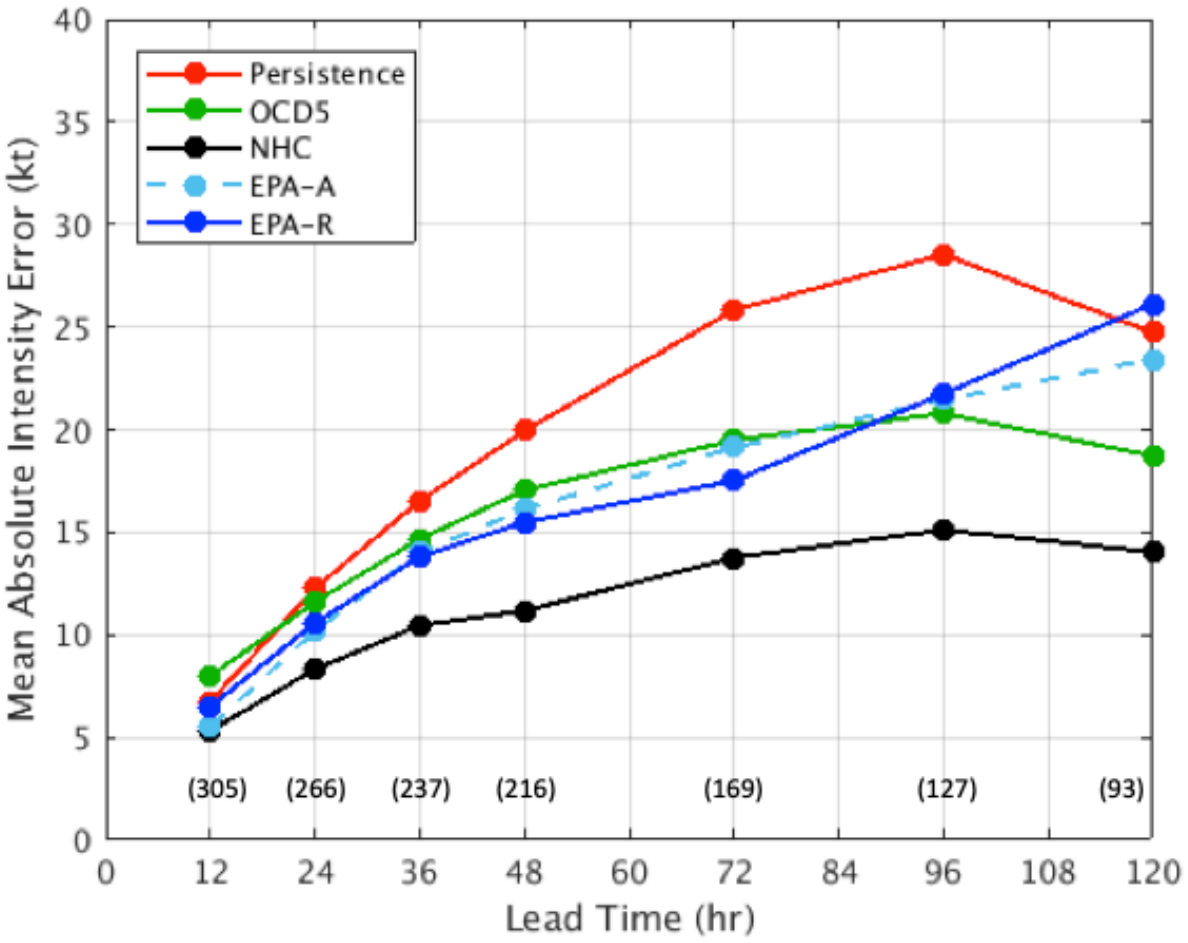

Figure 5: Mean absolute intensity errors across independent testing cases from the 2010-2016 North Atlantic TC season for the EP model with real time predictor variables (EPA-R, solid blue), EP model with analysis predictor variables (EPA-A, dashed light blue), as well as from the SHIFOR model (OCD5, green), official NHC forecasts (black), and a persistence model (red). Sample sizes are indicated along the bottom of the figure in parenthesis.

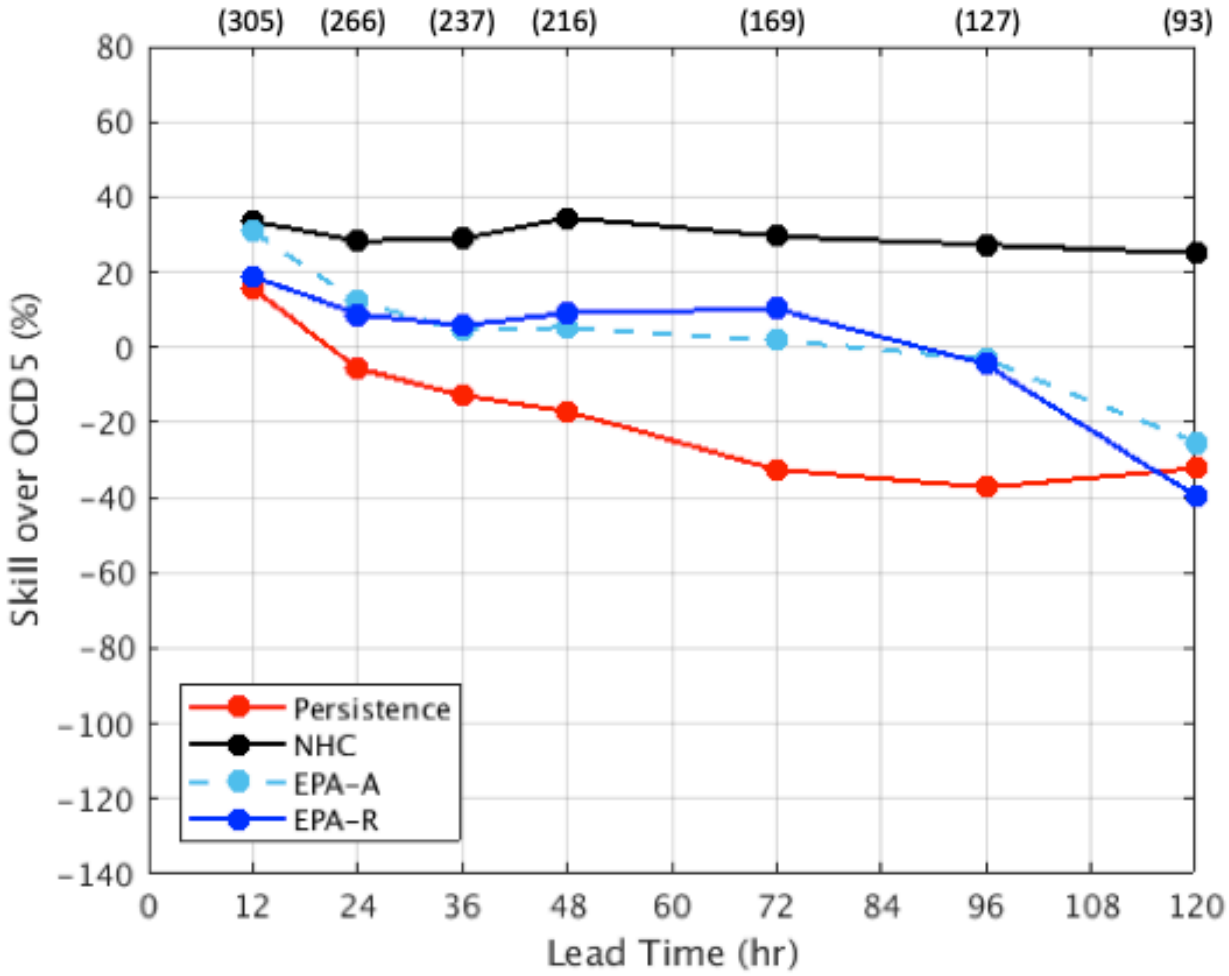

Figure 6: Error Relative to OCD5 model across the independent cases for the 2010-2016 North Atlantic TC season for the EP model with real time predictor variables (EPA-R, solid blue), EP model with analysis predictor variables (EPA-A, dashed light blue), as well as from the official NHC forecasts (black), and a persistence model (red). Sample sizes are indicated along the top of the figure in parenthesis.

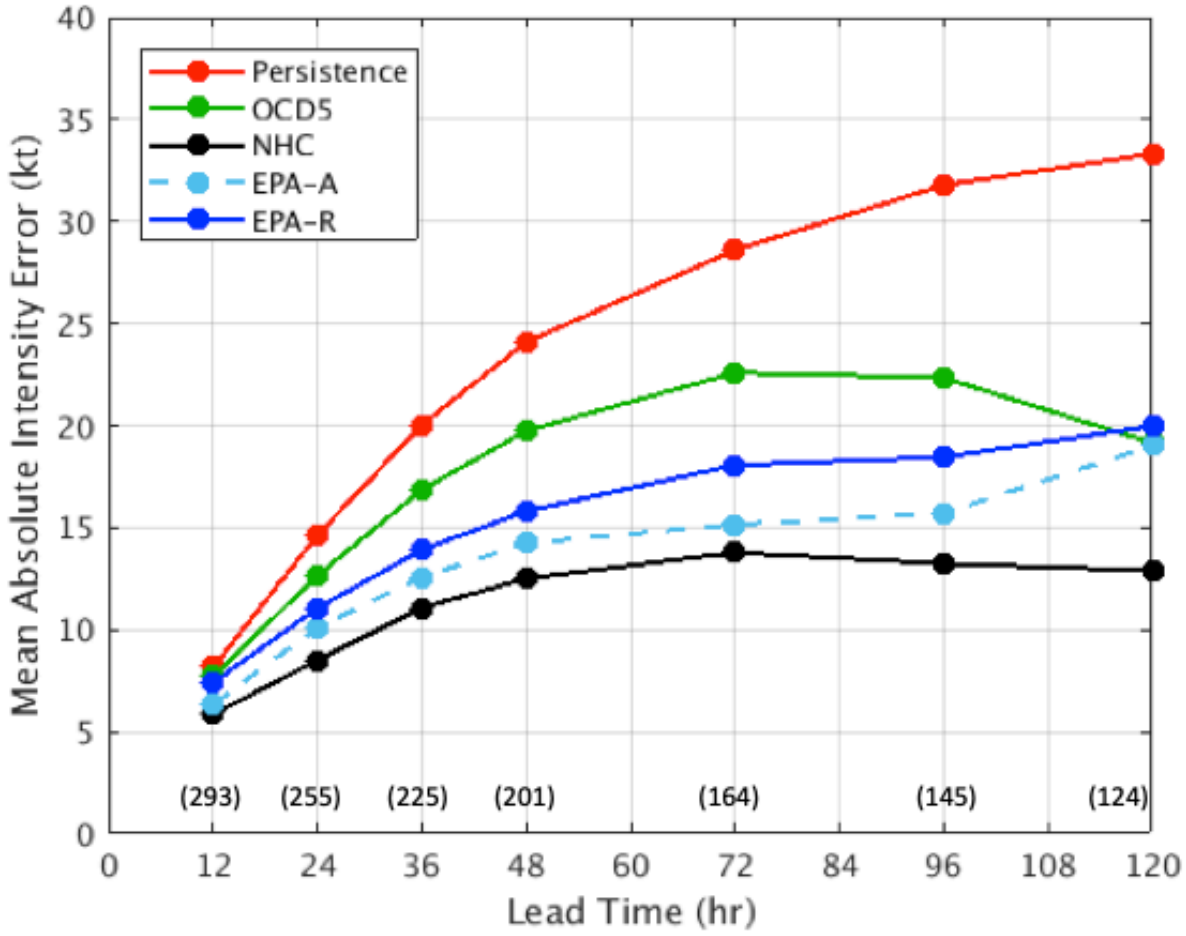

Figure 7: Same as Figure 5, but for the 2017 North Atlantic TC season.

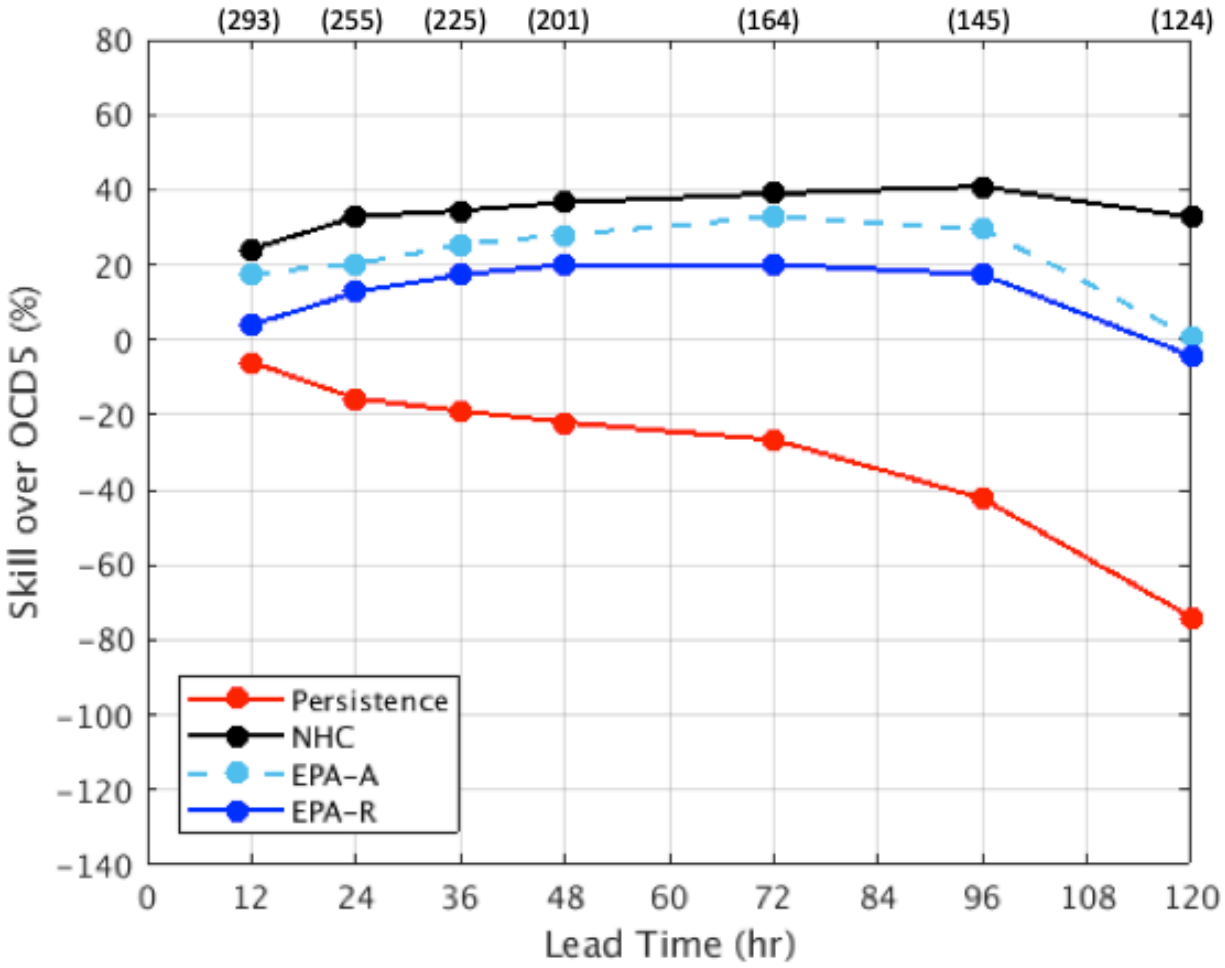

Figure 8: Same as Figure 6, but for the 2017 North Atlantic TC season.

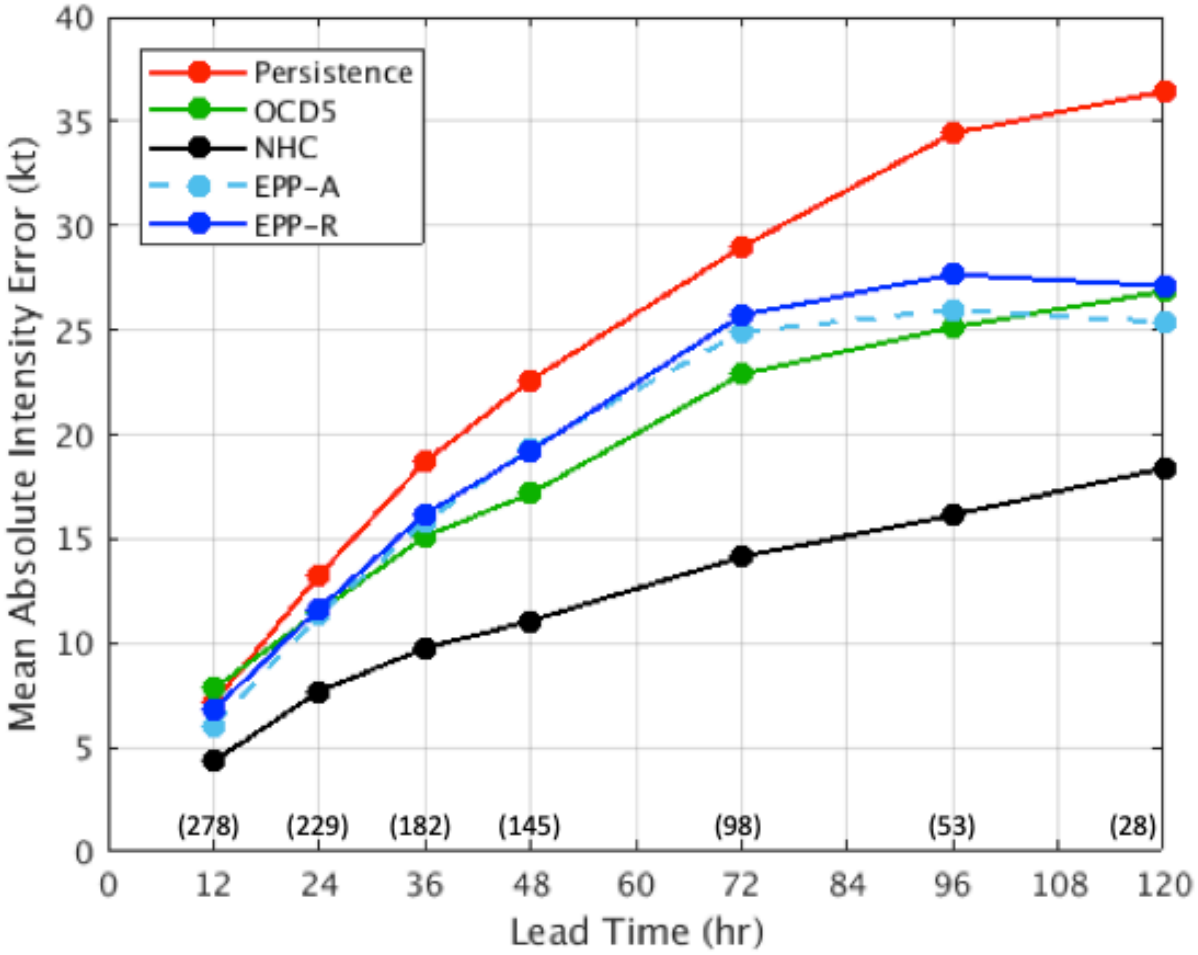

Figure 9: Mean absolute intensity errors across independent testing cases from the 2010-2016 eastern/central North Pacific TC season for the EP model with real time predictor variables (EPP-R, solid blue), EP model with analysis predictor variables (EPP-A, dashed light blue), as well as from the SHIFOR model (OCD5, green), official NHC forecasts (black), and a persistence model (red). Sample sizes are indicated along the bottom of the figure in parenthesis.

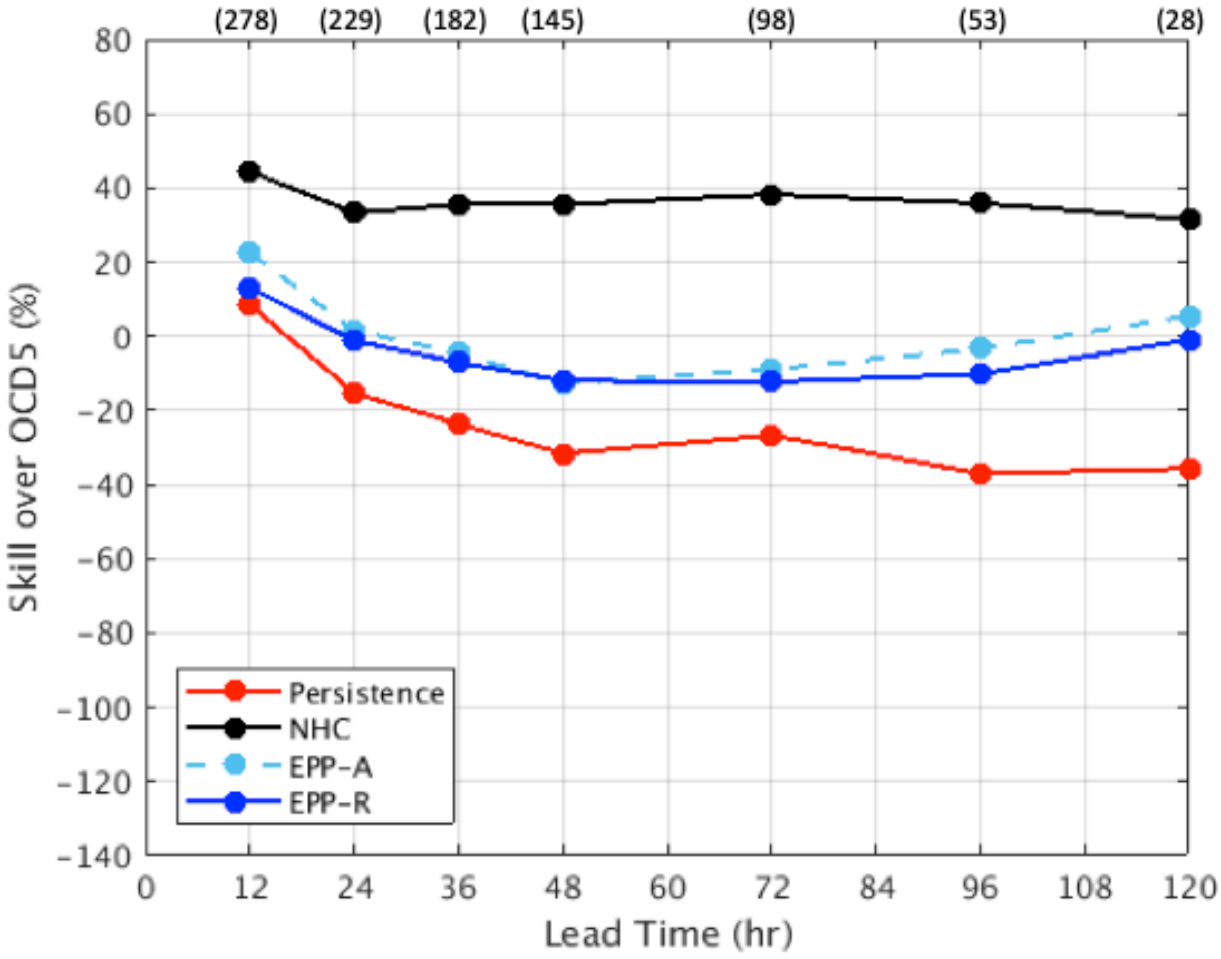

Figure 10: Error Relative to OCD5 model across the independent cases for the 2010-2016 eastern/central North Pacific TC season for the EP model with real time predictor variables (EPP-R, solid blue), EP model with analysis predictor variables (EPP-A, dashed light blue), as well as from the official NHC forecasts (black), and a persistence model (red). Sample sizes are indicated along the top of the figure in parenthesis.

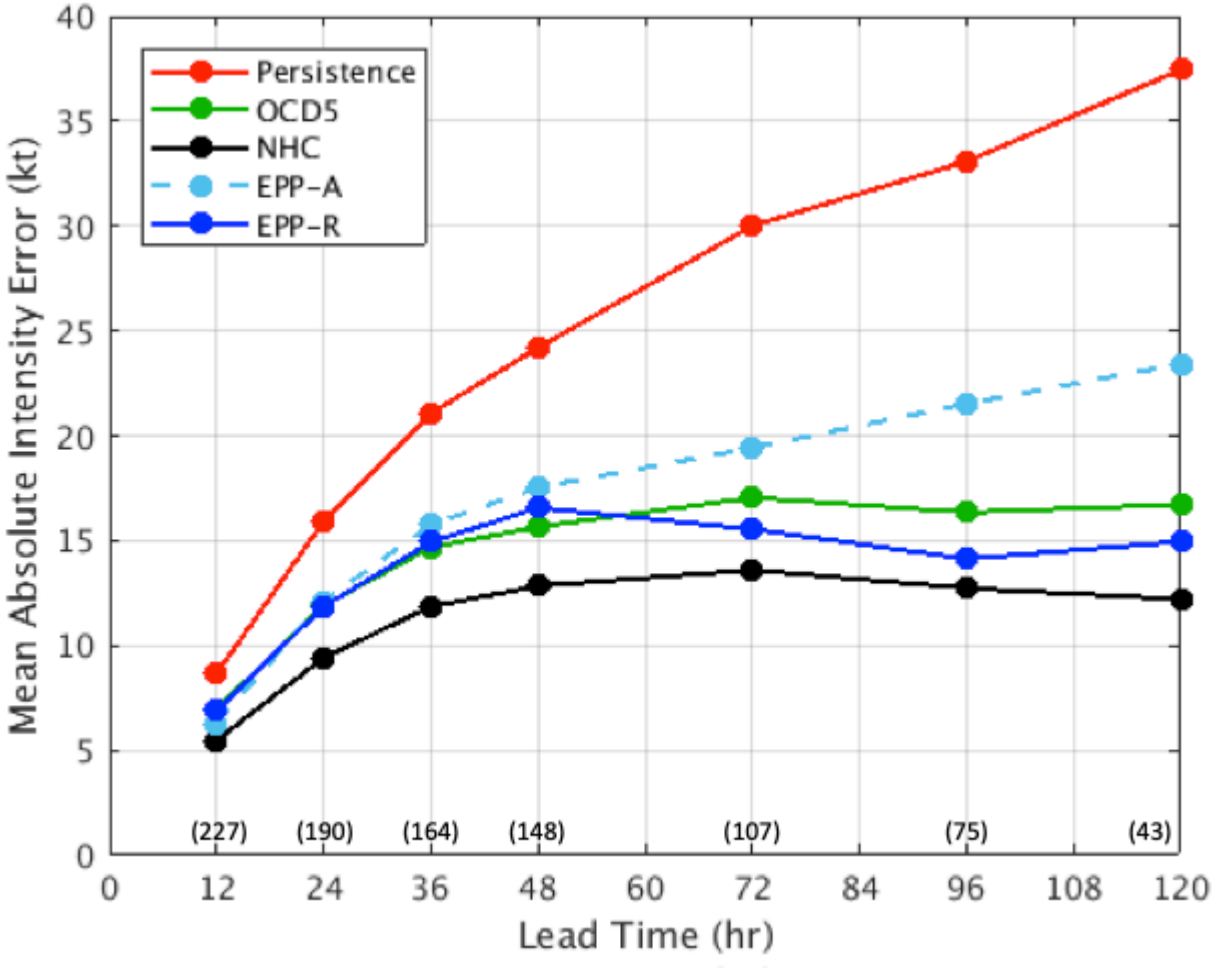

Figure 11: Same as Figure 9, but for the 2017 eastern/central North Pacific TC season.

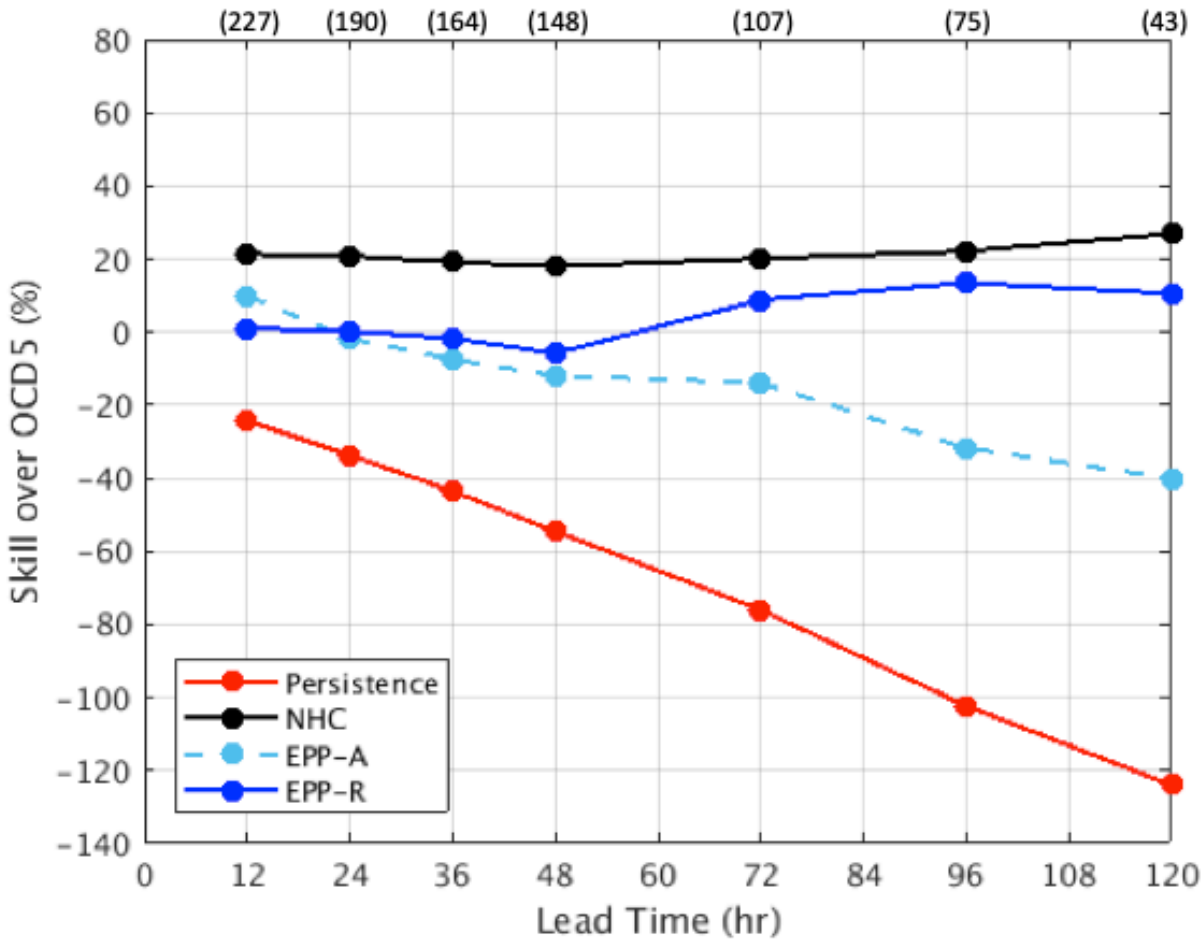

Figure 12: Same as Figure 10, but for the 2017 eastern/central North Pacific TC season.

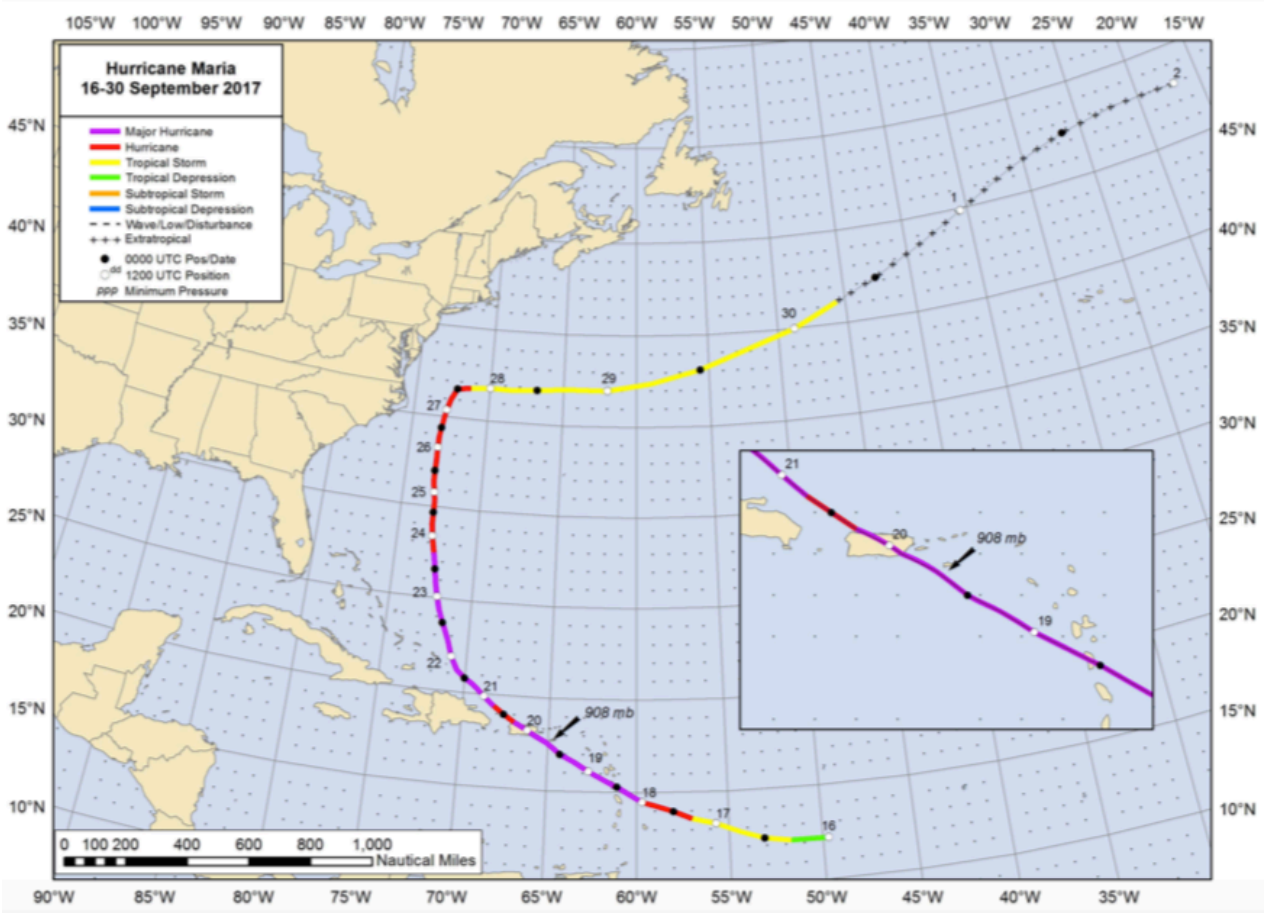

Figure 13: Best-track positions and TC intensity categories for TC Maria (Pasch et al. 2017).

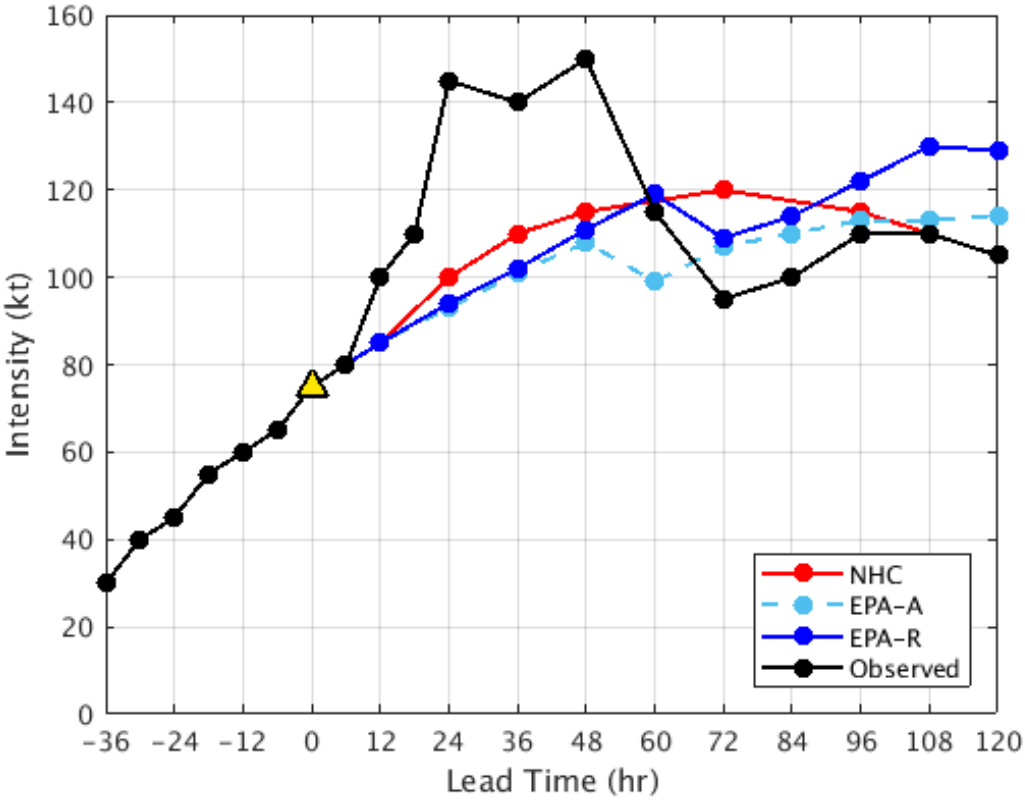

Figure 14: The observed intensity (black) and the 0000 UTC 18 September 2017 intensity forecasts from the NHC (red), EPA-A model (dashed light blue) and from the EPA-R model (solid blue) for TC Maria 2017. The initial intensity and 0-h forecast time is highlighted with a yellow triangle.

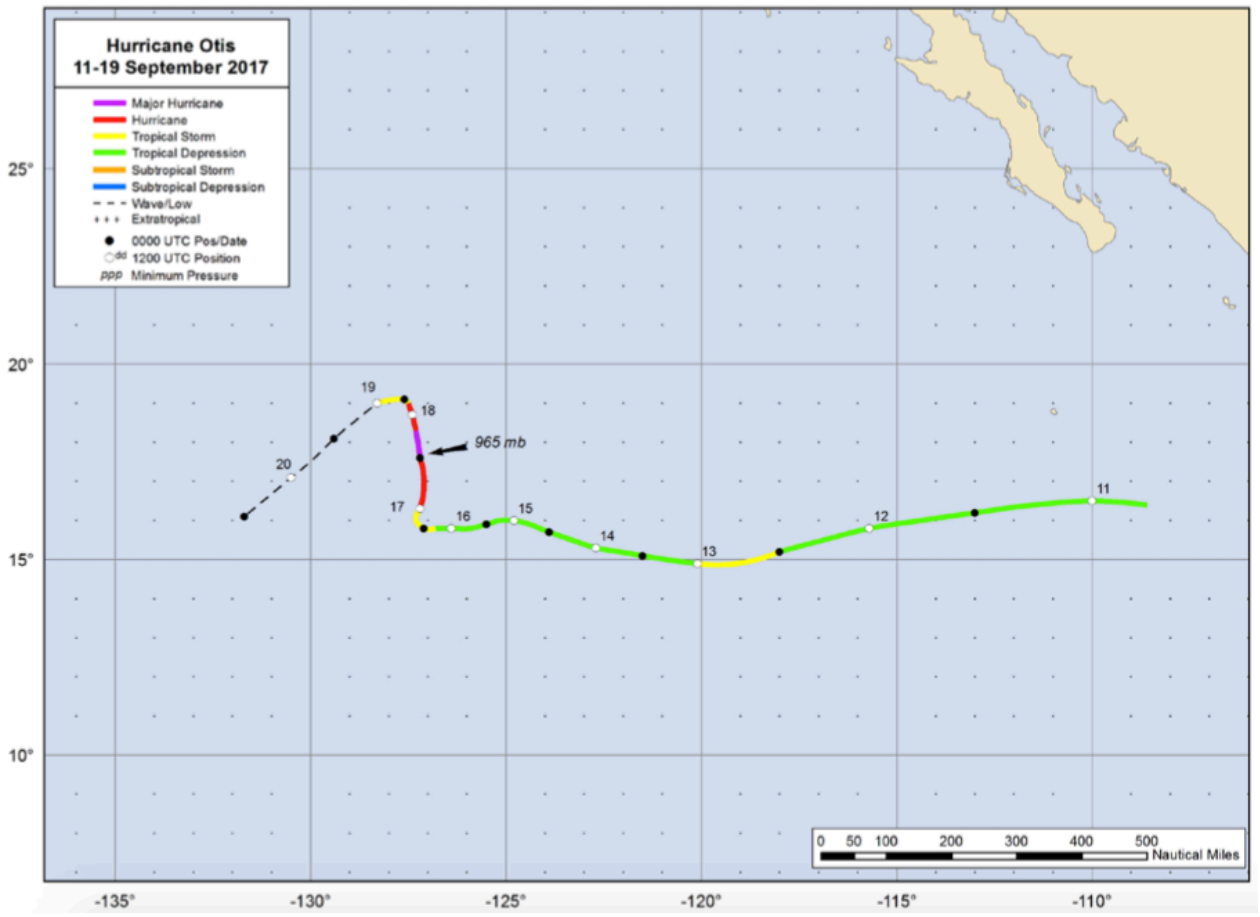

Figure 15: Best-track positions and TC intensity categories for TC Otis (Blake 2018b).

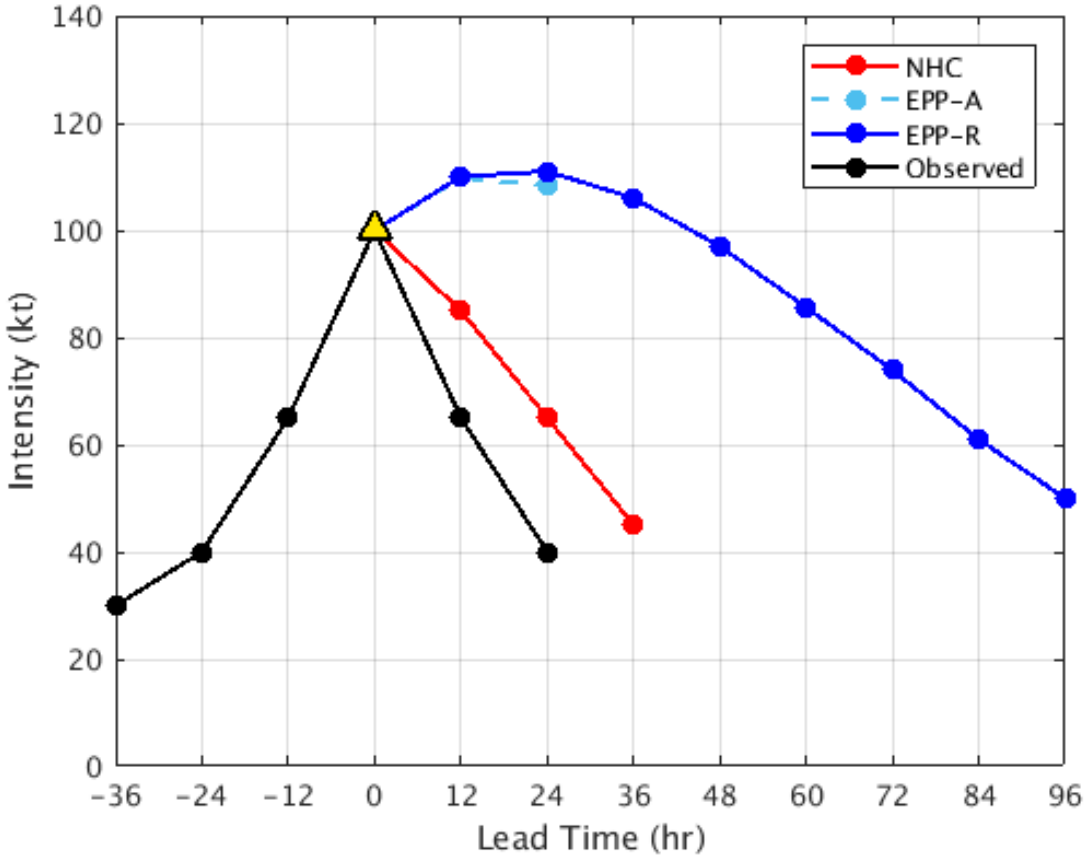

Figure 16: The observed intensity (black) and the 0000 UTC 18 September 2017 intensity forecast from the NHC (red), EPP-A model (dashed light blue) and from the EPP-R model (solid blue) for TC Otis 2017. The initial intensity and 0-h forecast time is highlighted with a yellow triangle.

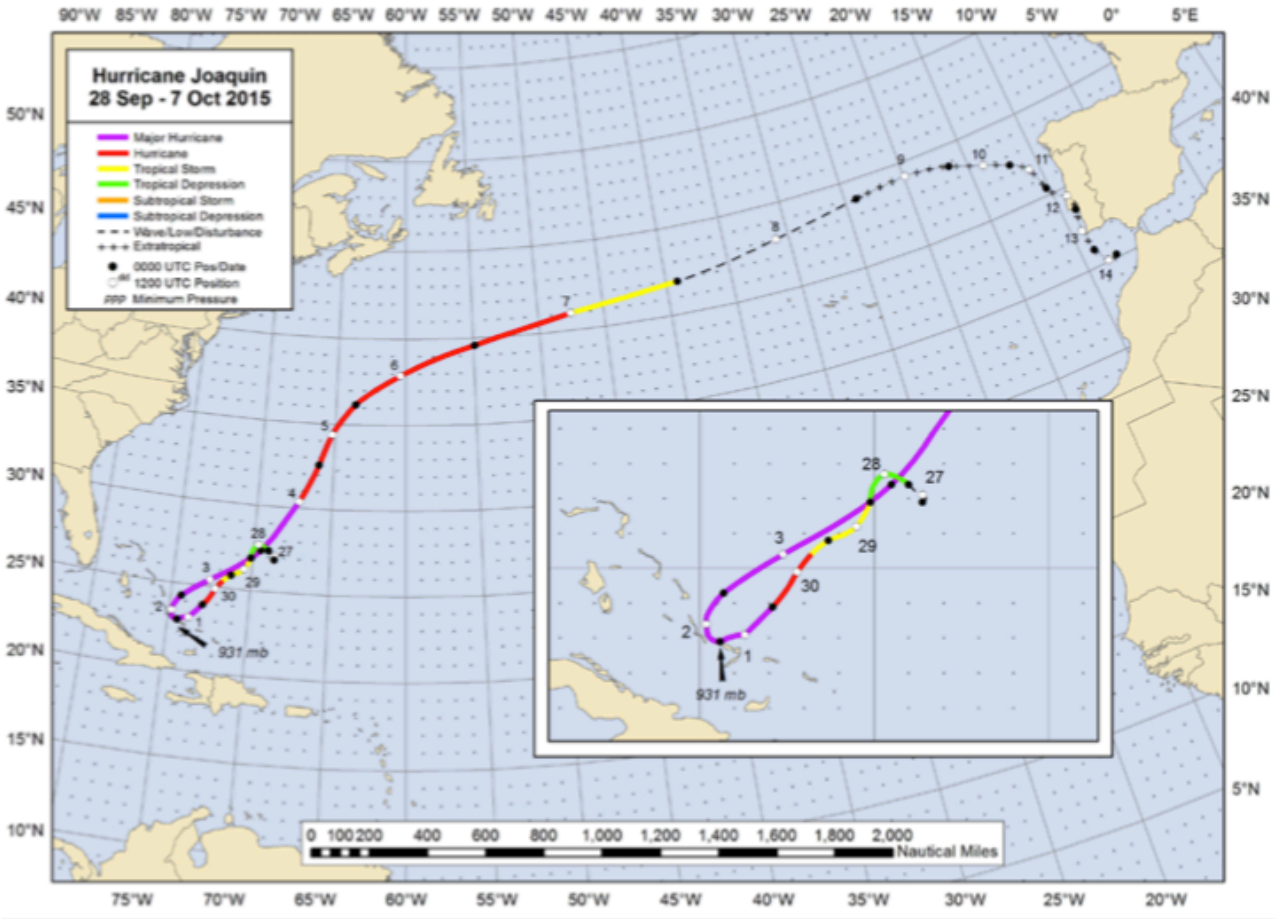

Figure 17: Best-track positions and TC intensity categories for TC Joaquin (Berg 2016).

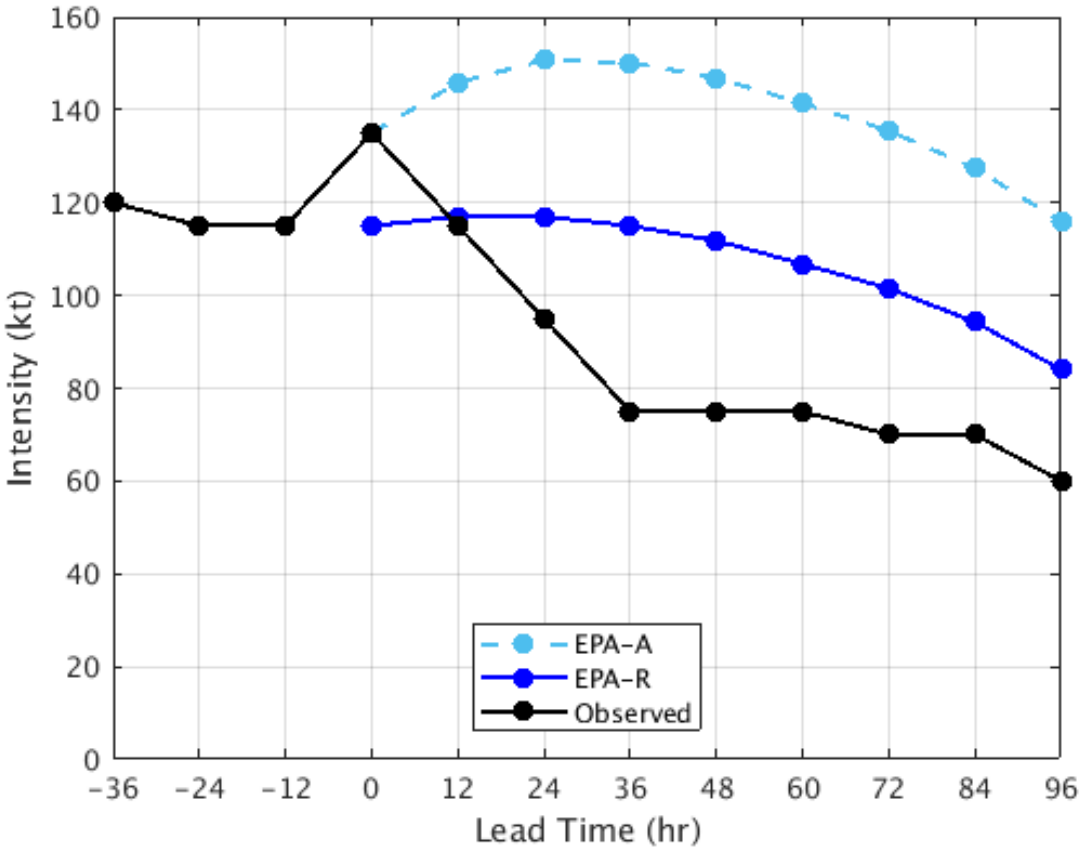

Figure 18: The observed intensity (black) and the 1200 UTC 3 October 2015 intensity forecast from the EPA-A model (dashed light blue) and from the EPA-R model (solid blue) for TC Joaquin 2015.

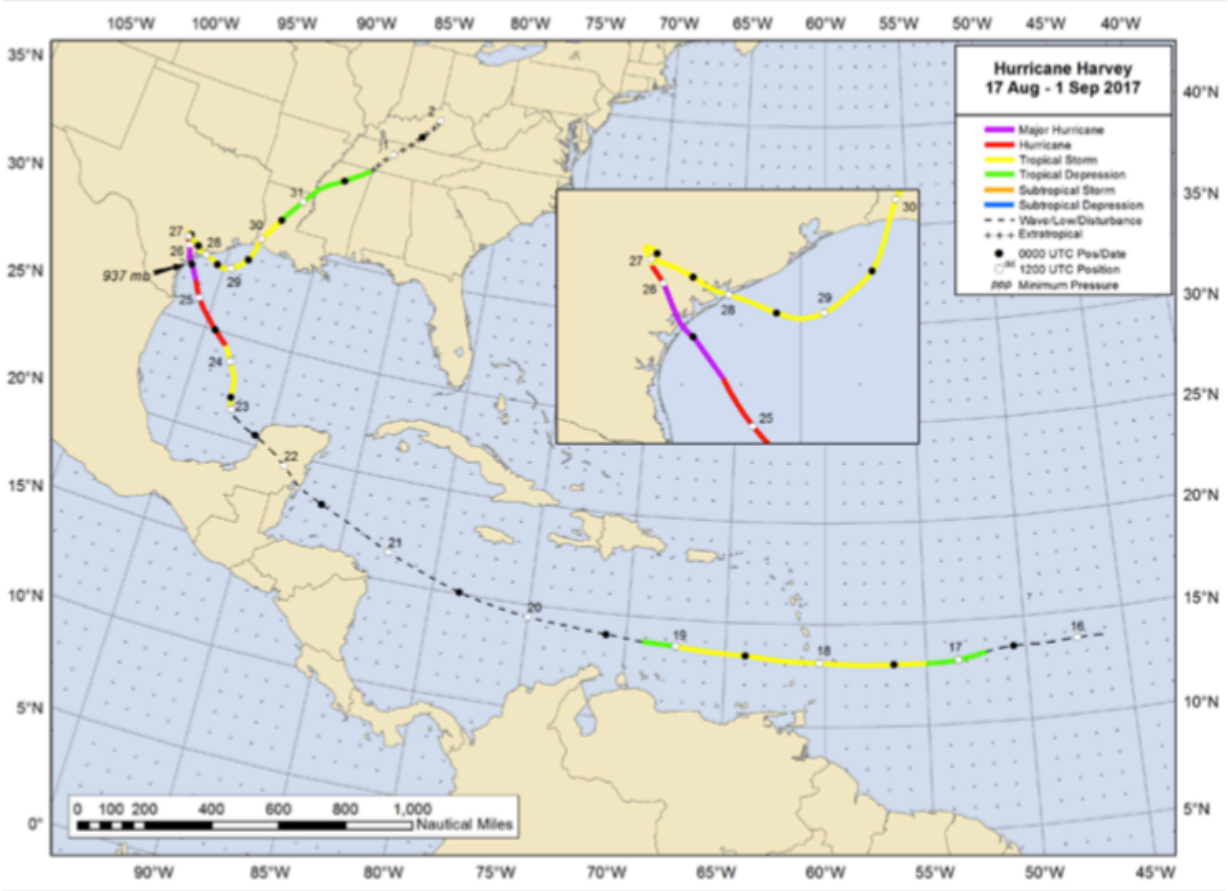

Figure 19: Best-track positions and TC categories for TC Harvey (Blake 2018a).

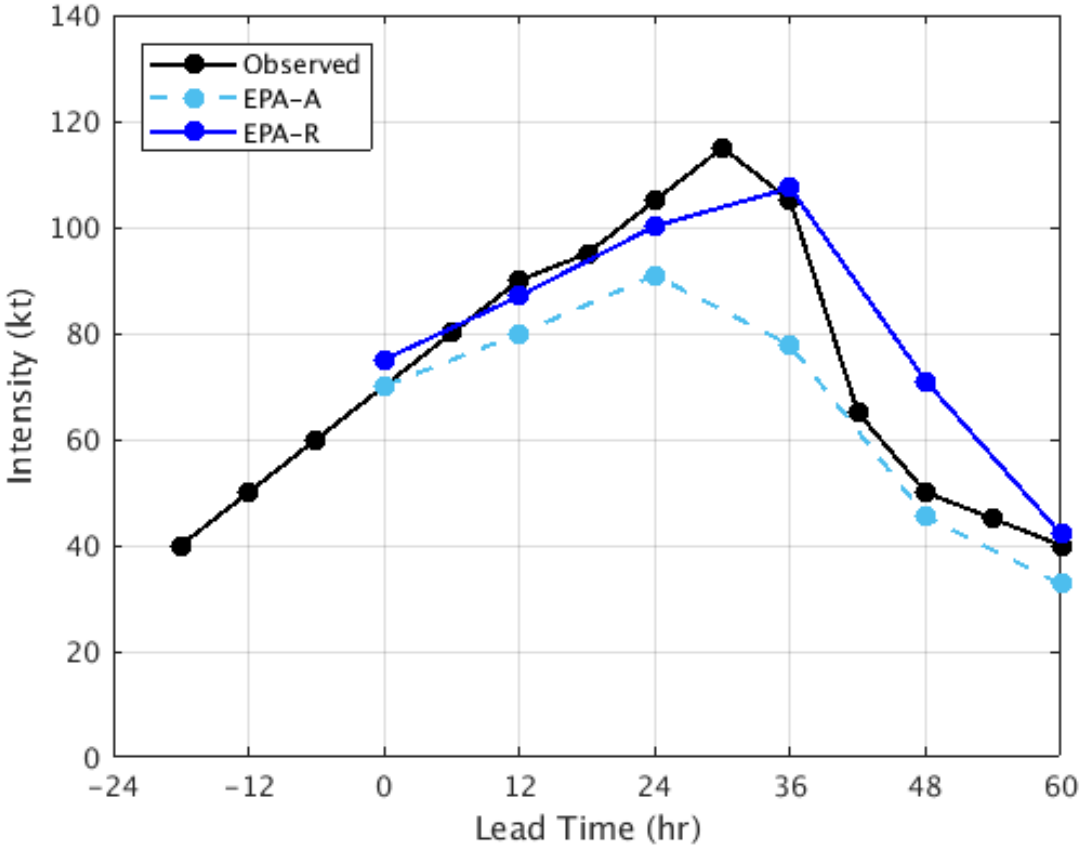

Figure 20: The observed intensity (black) and the 1800 UTC 24 August 2017 intensity forecast from the EPA-A model (dashed light blue) and from the EPA-R model (solid blue) for TC Harvey 2017.

TABLES

DELV	Change in intensity over the prior 12h
CD26	Climatological depth of 26˚C Isotherm from 2005-2010 NCODA analysis
U20C	200 hPa zonal wind (r=0-500 km
D200	200 hPa divergence (r=0-1000 km)
TWAC	0-600 km average symmetric tangential wind at 850 hPa from NCEP analysis
SHDC	850-200 hPa shear magnitude (kt *10) (200-800 km) with vortex removed and averaged from 0-500 km relative to 850 hPa vortex center
VMPI	Maximum potential intensity from Kerry Emanuel 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.
CONS	Constant value of 10.

Table 1: List of chosen predictor variables used in EP model.

Predictor	12-h Analysis Predictor Values (Std. Anom.) / Contribution (kt)	24-h Analysis Predictor Values (Std. Anom.) / Contribution (kt)
DELV	1.3 / 5.4	0.8 / 4.5
CD26	1.2 / -0.1	1.2 / 0.0
U20C	-0.9 / 0.0	-1.1 / -0.2
D200	0.8 / 0.2	1.8 / 0.2
TWAC	0.0 / 0.0	0.0 / 0.0
SHDC	-1.3 / 2.4	-0.9 / 1.7
VMPI	0.8 / 1.4	0.8 / 1.5
CFLX	0.0 / 0.0	0.1 / -0.1

Table 2: List of analysis predictor values in standard anomaly form and their relative contribution to the 12 and 24-h forecast of TC Maria initiating 0000 UTC 18 September 2017.

Predictor	12-h Analysis Predictor Values (Std. Anom.) / Contribution (kt)	24-h Analysis Predictor Values (Std. Anom.) / Contribution (kt)
DELV	3.1 / 15.7	0.8 / 5.3
CD26	-0.3 / 0.0	-0.2 / 0.0
U20C	1.1 / 0.9	0.5 / 0.5
D200	0.8 / 0.0	0.2 / 0.0
TWAC	0.3 / 0.3	-0.2 / -0.5
SHDC	-0.3 / 0.9	-0.4 / 1.3
VMPI	-0.7 / -0.9	-0.9 / -2.2
CFLX	3.5 / -0.9	3.4 / -4.6

Table 3: List of analysis predictor values in standard anomaly form and their relative contribution to the 12 and 24-h forecast of TC Otis initiating 0000 UTC 18 September 2017.

Predictor	12-h Analysis Predictor Values (Std. Anom.) / Contribution (kt)	12-h Real-Time Predictor Values (Std. Anom.) / Contribution (kt)
DELV	1.8 / 6.3	0.3 / 0.7
CD26	-0.1 / 0.0	-0.3 / 0.0
U20C	1.4 / 0.1	1.1 / 0.1
D200	0.6 / 0.6	1.0 / 0.0
TWAC	0.1 / 0.3	0.4 / -2.2
SHDC	0.0 / 0.0	0.6 / 1.4
VMPI	0.6 / 0.9	0.6 / 0.5
CFLX	-1.2 / 2.5	-0.3 / 0.7

Table 4: List of analysis and real-time predictor values in standard anomaly form and their relative contribution to the 12-h forecast of TC Joaquin initiating 1200 UTC 3 October 2015.

Predictor	12-h Analysis Predictor Values (Std. Anom.) / Contribution (kt)	24h Analysis Predictor Values (Std. Anom.) / Contribution (kt)
DELV	1.8 / 3.1	0.8 / 4.2
CD26	-0.1 / 0.0	-0.3 / 0.0
U20C	-0.3 / 0.1	-0.6 / 0.0
D200	-0.6 / -0.9	0.4 / 0.1
TWAC	-0.1 / -0.1	0.0 / 0.0
SHDC	-0.4 / 1.0	-0.5 / 0.9
VMPI	1.1 / 0.6	0.8 / 0.9
CFLX	-0.8 / 1.1	-1.9 / 3.7

Table 5: List of analysis predictor values in standard anomaly form and their relative contribution to the 12 and 24-h forecast of TC Harvey initiating 1800 UTC 24 August 2017.

Predictor	12-h Real-Time Predictor Values (Std. Anom.) / Contribution (kt)	24h Real-Time Predictor Values (Std. Anom.) / Contribution (kt)
DELV	3.4 / 6.3	1.0 / 4.6
CD26	-0.1 / 0.0	-0.3 / 0.0
U20C	-0.2 / 0.1	-0.3 / 0.0
D200	0.2 / 0.2	0.3 / 0.1
TWAC	-0.1 / 0.0	0.1 / 0.3
SHDC	-0.3 / 0.7	-0.9 / 1.8
VMPI	1.2 / 1.2	0.8 / 1.0
CFLX	-0.4 / 0.3	-2.2 / 4.5

Table 6: List of real-time predictor values in standard anomaly form and their relative contribution to the 12 and 24-h forecast of TC Harvey initiating 1800 UTC 24 August 2017.

REFERENCES

Bender, M. A., I. Ginis, R. Tuleya, B. Thomas, and T. Marchok, 2007: The Operational GFDL Coupled Hurricane-Ocean Prediction System and a Summary of Its Performance. *Mon. Wea. Rev.*, **135**, 3965-3989.

Berg, R., 2016: National Hurricane Center Tropical Cyclone Report: Hurricane Joaquin (AL112015). [Available online at https://www.nhc.noaa.gov/data/tcr/AL112015_Joaquin.pdf]

Blake, E. S., 2014: National Hurricane Center annual summary: 2013 Atlantic hurricane season. [Available online at http://www.nhc.noaa.gov/data/tcr/summary_atlc_2013.pdf].

Blake, E. S., and D. A. Zelinsky, 2018: National Hurricane Center Tropical Cyclone Report: Hurricane Harvey (AL092017) [Available online at https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf]

Blake, E S., 2018: National Hurricane Center Tropical Cyclone Report: Hurricane Otis (EP152017). [Available online at https://www.nhc.noaa.gov/data/tcr/EP152017_Otis.pdf]

Cangialosi, J. P., 2017: National Hurricane Center Forecast Verification Report: 2017 Hurricane Season. [Available online at https://www.nhc.noaa.gov/verification/pdfs/Verification_2017.pdf]

DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209-220.

DeMaria, M., and J. Kaplan, 1999: An Updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and Eastern North Pacific Basins. *Wea. Forecasting*., **14**, 326-337.

DeMaria, M., 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.

DeMaria, M., J. Knaff, and J. Kaplan, 2006: On the Decay of Tropical Cyclone Winds Crossing Narrow Landmasses. *J. Appl. Meteor.,* **45**, 491-499.

DeMaria, M., 2009: A Simplified Dynamical System for Tropical Cyclone Intensity Prediction. *Mon. Wea. Rev.*, **137**, 68-82.

DeMaria, M., C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387-398.

Dow, K., and S. L. Cutter, 2002: Emerging Evacuation Issues: Hurricane Floyd and South Carolina, *Nat. Hazards Rev.*, **3(1)**, 12-18.

Emanuel, K., C. DesAutels C. Holloway, and R. Korty, 2003: Environmental Control of Tropical Cyclone Intensity. *J. Atmos. Sci.*, **61**, 843-858.

Emanuel, K., and F. Zhang, 2016: On the predictability and error sources of tropical cyclone intensity forecasts. *J. Atmos. Sci.*, **73**, 3739-3747.

Emanuel, K., and F. Zhang, 2017: On the role of Inner-Core Moisture in Tropical Cyclone Predictability and Forecast Skill. *J. Atmos. Sci*., **74**, 2315-2324.

Ferreira, C., 2006: *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. 2nd ed. Springer, 478 pp.

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

Glahn, B., A. Taylor, N. Kurkowski, and W. A. Shaffer, 2009: The Role of the SLOSH Model in National Weather Service Storm Surge Forecasting. Naitonal Weather Diges, **33**, 3-14.

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.

Grumm, R. J., and R. Hart, 2001: Standardized Anomalies Applied to Significant Cold Season Weather Events: Preliminary Findings. *Wea. Forecasting.*, **16**, 736-754.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian Model Averaging: A Tutorial. *Stat. Sci.*, **14**, 382-417.

Jelesnianski, C. P., J. Chen, and W. A. Shadder, 1992: SLOSH: Sea, Lake, and Overland Surges from Hurricanes. NOAA Tech. Report NWS 48, 71pp [Available from NOAA/AOML Library, 4301 Rickenbacker, Csw., Miami, FL 33149. Or online at https://slosh.nws.noaa.gov/sloshPub/pubs/SLOSH_TR48.pdf]

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.

Kaplan, J., and M. DeMaria, 2001: On The Decay of Tropical Cyclone Winds after Landfall in the New England Area. *J. Appl. Meteor.*, **40**, 280-286.

Kaplan, J., M. DeMaria, 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.

Kieu, C. Q., and Z. Moon, 2016: Hurricane intensity predictability. *Bull. Amer. Meteor. Soc.*, **97**, 1847-1858.

Klotzbach, P. J., S. G. Bowen, R. Pielke Jr., and M. Bell, 2018: Continental U.S. Hurricane Landfall Frequency and Associated Damage: Observations and Future Risks. *Bull. Amer. Meteor. Soc.*, **99**, 1359-1376.

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. Forecasing*, **18**, 80-92

Lazo, J. K., D. M. Waldman, B. H. Morrow, and J. A. Thacher, 2010: Household Evacuation Decision Making and the Benefits of Improved Hurricane Forecasting: Developing A Framework for Assessment. *Wea. Forecasting*, **25**, 207-219

Lin, I.I., G. J. Goni, J. A. Knaff, C. Forbes, and M. M. Ali, 2013: Ocean Heat Content For Tropical Cyclone Intensity Forecasting and its Impact on Storm Surge. *Nat. Hazards*, **66**, 1481-1500

Mehra, A., 2017: 2017 Hurricane Model Implementations Briefing to NCEP Director: Much improved operational forecast guidance for global tropical cyclones [Available online at https://www.emc.ncep.noaa.gov/gc_wmb/vxt/HMON/web/doc/FY17_HMON_OD_brie_042817.pdf]

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, IEEE, 2657-2663

NCEP, 2016: The Global Forecast System (GFS) – Global Spectral Model (GSM), Accessed 4 March 2019, [Available online at https://www.emc.ncep.noaa.gov/GFS/doc.php]

Pasch, R. J., 2015: National Hurricane Center annual summary: 2014 Atlantic hurricane season. [Available online at http://www.nhc.noaa.gov/data/tcr/summary_atlc_2014.pdf].

Pasch, R. J., and A. B. Penny, and R. Berg, 2017: National Hurricane Center Tropical Cyclone Report: Hurricane Maria (AL152017). Accessed 12 April 2019, [Available online at https://www.nhc.noaa.gov/data/tcr/AL152017_Maria.pdf]

Rappaport, E. N., J.-G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The Joint Hurricane Test Bed: It's First Decade of Tropical Cyclone Research-To-Operations Activities Reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371-380

Rappaport, E. N., 2014: Fatalities in the United States from Atlantic Tropical Cyclones: New Data and Interpretation. *Bull. Amer. Meteor. Soc.*, **95**, 341-346

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian Model Averaging to Calibrate Forecasts Ensembles. *Mon. Wea. Rev.,* **133**, 1155-1174,

Roebber, P. J., 2010: Seeking Consensus: A New Approach. *Mon. Wea. Rev.*, **138**, 4402-4415

Roebber, P. J., 2013: Using Evolutionary Programming to Generate Skillful Extreme Value Probabilistic Forecasts. *Mon. Wea. Rev.*, **141**, 3170-3185

Roebber, P. J., 2015: Evolving Ensembles. *Mon. Wea. Rev.,* **143**, 471-490

Roebber, P. J., 2015: Using Evolutionary Programming to Maximize Minimum Temperature Forecast Skill. *Mon. Wea. Rev.*, **143**, 1506-1516

Roebber, P. J., 2016: Development of a Large Member Ensemble Forecast System for Heavy Rainfall using Evolutionary Programming (EP), 10pp.

Saha, S., and Coauthors: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**,1015-1057

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

Stewart, S., 2014: National Hurricane Center annual summary: 2012 Atlantic hurricane season. [Available online at http://www.nhc.noaa.gov/data/tcr/summary_atlc_2012.pdf].

Stewart, S. R., 2016: National Hurricane Center annual summary: 2015 Atlantic hurricane season. [Available online at 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

APPENDIX

EPA MODEL

Algorithm 6:

Bias = 0.52

Weighting = 0.16667

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{i3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	SHDC	<=	TWAC	THEN	0.14598*VMPI	+	-0.44744*U20C	*	-0.1582*D200
2	IF	SHDC	<=	DELV	THEN	0.36127 *DELV	*	-0.0746*10	*	0.23645*DELV
3	IF	SHDC	<=	SHDC	THEN	-0.95443 *DELV	+	0.95413*DELV	+	0.02358*10
4	IF	DELV	<=	DELV	THEN	-0.18835*SHDC	+	0.40803*DELV	+	-0.24738*CFLX
5	IF	VMPI	>	TWAC	THEN	-0.94745*DELV	*	0.18154*VMPI	*	0.84904*D200

Algorithm 8:

Bias = -0.57

Weighting = 0.08333

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{i3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	CFLX	<=	DELV	THEN	0.90216*VMPI	*	0.65379*D200	*	0.21644 *DELV
2	IF	SHDC	<=	DELV	THEN	0.36127 *DELV	*	-0.0746*10	*	0.23645*DELV
3	IF	SHDC	<=	SHDC	THEN	-0.95443 *DELV	+	0.95413*DELV	+	0.02358*10
4	IF	DELV	<=	DELV	THEN	-0.18835*SHDC	+	0.40803*DELV	+	-0.24738*CFLX
5	IF	TWAC	<=	U20C	THEN	-0.32557*TWAC	+	-0.20541 *VMPI	*	-0.38564 *CFLX

Algorithm 9:

Bias = 0.28

Weighting = 0.08333

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{i3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	CFLX	>	U20C	THEN	-0.26381*U20C	*	0.14971*VMPI	+	0.2113 *VMPI
2	IF	SHDC	<=	DELV	THEN	0.36127 *DELV	*	-0.0746*10	*	0.23645*DELV
3	IF	SHDC	<=	SHDC	THEN	-0.95443 *DELV	+	0.95413*DELV	+	0.02358*10
4	IF	DELV	<=	DELV	THEN	-0.18835*SHDC	+	0.40803*DELV	+	-0.24738*CFLX
5	IF	SHDC	<=	CD26	THEN	-0.42766*TWAC	*	0.36128 *CFLX	*	0.03389 *SHDC

Algorithm 34:
Bias = 0.21
Weighting = 0.08333

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{i3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	TWAC	<=	CFLX	THEN	0.31731 *CFLX	*	-0.90571 *D200	*	-0.21776 *SHDC
2	IF	SHDC	>	CFLX	THEN	0.25237 *TWAC	+	-0.36317*TWAC	*	0.27941*CFLX
3	IF	DELV	>	TWAC	THEN	0.03356 *CD26	*	0.09853*DELV	+	0.03853*10
4	IF	VMPI	<=	VMPI	THEN	0.33592 *DELV	*	-0.21264 *TWAC	+	0.15755 *DELV
5	IF	SHDC	<=	SHDC	THEN	-0.18206 *SHDC	+	0.1172 *VMPI	+	-0.17664 *CFLX

Algorithm 35:
Bias = 0.10
Weighting = 0.08333

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{i3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	CFLX	<=	SHDC	THEN	0.86228 *CD26	+	0.41323*TWAC	+	-0.85329 *CD26
2	IF	CD26	>	D200	THEN	-0.10933 *DELV	+	0.54357 *TWAC	*	-0.28723 *CD26
3	IF	DELV	>	TWAC	THEN	0.03356 *CD26	*	0.09853*DELV	+	0.03853*10
4	IF	VMPI	<=	VMPI	THEN	0.33592 *DELV	*	-0.21264 *TWAC	+	0.15755 *DELV
5	IF	SHDC	<=	SHDC	THEN	-0.18206 *SHDC	+	0.1172 *VMPI	+	-0.17664 *CFLX

Algorithm 49:
Bias = 0.19
Weighting = 0.16667

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{i3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	D200	<=	VMPI	THEN	0.32367 *TWAC	+	-0.15624 *D200	*	0.11885 *CFLX
2	IF	D200	<=	SHDC	THEN	-0.24229*TWAC	*	0.0833 *DELV	+	-0.07426 *DELV
3	IF	DELV	>	TWAC	THEN	0.03356 *CD26	*	0.09853*DELV	+	0.03853*10
4	IF	VMPI	<=	VMPI	THEN	0.33592 *DELV	*	-0.21264 *TWAC	+	0.15755 *DELV
5	IF	SHDC	<=	SHDC	THEN	-0.18206 *SHDC	+	0.1172 *VMPI	+	-0.17664 *CFLX

Algorithm 53:
Bias = -0.67
Weighting = 0.25

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{i3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	CD26	<=	CD26	THEN	-0.59528 *10	*	-0.83168*TWAC	*	-0.1173 *TWAC
2	IF	D200	<=	10	THEN	-0.78933 *VMPI	*	0.26422 *TWAC	*	-0.78223 *CFLX
3	IF	DELV	>	TWAC	THEN	0.03356 *CD26	*	0.09853*DELV	+	0.03853*10
4	IF	VMPI	<=	VMPI	THEN	0.33592 *DELV	*	-0.21264 *TWAC	+	0.15755 *DELV
5	IF	SHDC	<=	SHDC	THEN	-0.18206 *SHDC	+	0.1172 *VMPI	+	-0.17664 *CFLX

EPP MODEL

Algorithm 31:
Bias = -0.07
Weighting = 0.25

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{I3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	TWAC	>	VMPI	THEN	0.36679*CFLX	*	0.55976*TWAC	+	-0.03705*DELV
2	IF	CFLX	<=	DELV	THEN	0.16784*CFLX	*	0.83909*DELV	*	0.58132*TWAC
3	IF	SHDC	>	D200	THEN	-0.12243*VMPI	+	0.31332*TWAC	+	0.01871*CD26
4	IF	D200	<=	D200	THEN	-0.89092*TWAC	*	0.28928*TWAC	+	-0.1396*CFLX
5	IF	VMPI	<=	VMPI	THEN	0.6716*VMPI	+	-0.44336*VMPI	+	0.42004*DELV

Algorithm 69:
Bias= -0.09
Weighting = 0.75

		V_{i1}	R_{i1}	V_{i2}		$C_{i1}*V_{I3}$	O_{i1}	$C_{i2}*V_{i4}$	O_{i2}	$C_{i3}*V_{i5}$
1	IF	DELV	>	U20C	THEN	0.17881*VMPI	*	-0.73721*U20C	+	-0.36376*SHDC
2	IF	DELV	>	CFLX	THEN	-0.14589*DELV	+	0.0649*TWAC	*	0.8098*CD26
3	IF	SHDC	>	D200	THEN	-0.12243*VMPI	+	0.31332*TWAC	+	0.01871*CD26
4	IF	D200	<=	D200	THEN	-0.89092*TWAC	*	0.28928*TWAC	+	-0.1396*CFLX
5	IF	VMPI	<=	VMPI	THEN	0.6716*VMPI	+	-0.44336*VMPI	+	0.42004*DELV