

Deep Learning Experiments for Tropical Cyclone Intensity Forecasts

WENWEI XU,^a KARTHIK BALAGURU,^a ANDREW AUGUST,^b NICHOLAS LALO,^b NATHAN HODAS,^b
MARK DEMARIA,^c AND DAVID JUDI^d

^a Marine and Coastal Research Laboratory, Pacific Northwest National Laboratory, Seattle, Washington

^b Computing and Analytics Division, Pacific Northwest National Laboratory, Richland, Washington

^c Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado

^d Earth Systems Science, Pacific Northwest National Laboratory, Richland, Washington

(Manuscript received 30 June 2020, in final form 25 May 2021)

ABSTRACT: Reducing tropical cyclone (TC) intensity forecast errors is a challenging task that has interested the operational forecasting and research community for decades. To address this, we developed a deep learning (DL)-based multilayer perceptron (MLP) TC intensity prediction model. The model was trained using the global Statistical Hurricane Intensity Prediction Scheme (SHIPS) predictors to forecast the change in TC maximum wind speed for the Atlantic basin. In the first experiment, a 24-h forecast period was considered. To overcome sample size limitations, we adopted a leave one year out (LOYO) testing scheme, where a model is trained using data from all years except one and then evaluated on the year that is left out. When tested on 2010–18 operational data using the LOYO scheme, the MLP outperformed other statistical–dynamical models by 9%–20%. Additional independent tests in 2019 and 2020 were conducted to simulate real-time operational forecasts, where the MLP model again outperformed the statistical–dynamical models by 5%–22% and achieved comparable results as HWFI. The MLP model also correctly predicted more rapid intensification events than all the four operational TC intensity models compared. In the second experiment, we developed a lightweight MLP for 6-h intensity predictions. When coupled with a synthetic TC track model, the lightweight MLP generated realistic TC intensity distribution in the Atlantic basin. Therefore, the MLP-based approach has the potential to improve operational TC intensity forecasts, and will also be a viable option for generating synthetic TCs for climate studies.

SIGNIFICANCE STATEMENT: Scientists have been searching for decades for breakthroughs in tropical cyclone intensity modeling to provide more accurate and timely tropical cyclone warnings. To this end, we developed a deep learning (DL)-based predictive model for North Atlantic 24- and 6-h intensity forecast. We simulated 2019 and 2020 tropical cyclones as if in an operational forecast mode, and found that the model's 24-h intensity forecast outperformed some of the most skillful operational models by 5%–22%. Also, the 6-h intensity model produced realistic intensity labels for synthetic tropical cyclone tracks. These results highlight the potential for using deep neural network–based models to improve operational hurricane intensity forecasts and synthetic tropical cyclone generation.


KEYWORDS: Tropical cyclones; Forecast verification/skill; Forecasting techniques; Operational forecasting; Short-range prediction; Statistical forecasting; Deep learning; Neural networks

1. Introduction

Tropical cyclones (TCs) pose a significant socioeconomic threat in the global tropics and subtropics. Accurate prediction of TC tracks and intensities to provide actionable information on TC hazards to mitigate TC damages and loss of life therefore offers high societal benefits. While significant progress has been made over the past few decades in TC track forecasting, TC intensity forecasting has only improved modestly (Rappaport et al. 2012; DeMaria et al. 2014; Balaguru et al. 2018). According to the U.S. National Weather Service's National Hurricane Center (NHC) official error trend (<https://www.nhc.noaa.gov/verification/verify5.shtml>), the average 24-h track forecast error decreased from 85.4 to 38.9 n mi (1 n mi = 1.852 km) (a 54%

reduction) between 1990–99 and 2010–18. In contrast, the 24-h intensity mean absolute error (MAE) over the Atlantic region has only decreased from 9.8 kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$) in 1990–99 to 8.7 kt in 2010–18 (an 11% reduction). Intensity forecast improvements have accelerated in the past 5 years (Cangialosi et al. 2020), but consistent forecast skill is still lacking, especially for rapidly intensifying TCs (Knaff et al. 2020; Torn and DeMaria 2021). Further improvement in TC intensity forecast has been partially hindered by the relatively large uncertainty in the best track intensity records (Combot et al. 2020; Landsea and Franklin 2013; Torn and Snyder 2012). At the same time, there are urgent needs for TC intensity modeling breakthroughs to substantially improve our ability to forecast intensity at 1–7 days horizon.

Broadly, TC intensity forecast models can be divided into three categories: dynamical models or physics-based models, statistical models, and statistical–dynamical models, which blend dynamical model output with statistics. In this study we compare our results to several skillful operational models and consensus as listed in Table 1. Improvements in intensity

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

Corresponding author: Xu, Wenwei, wenwei.xu@pnnl.gov

DOI: 10.1175/WAF-D-20-0104.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

TABLE 1. Operational TC intensity models and consensus for intercomparisons with MLP.

Name	Acronym	Formation	Reference
Statistical Hurricane Intensity Prediction Scheme	SHIPS	Statistical–dynamical	DeMaria et al. (2005)
Decay-Statistical Hurricane Intensity Prediction Scheme	DSHP	Statistical–dynamical	DeMaria et al. (2006)
Statistical Hurricane Intensity Prediction Scheme Logistic Growth Equation Model	LGEM	Statistical–dynamical	DeMaria (2009)
Operational Hurricane Weather Research and Forecasting	HWFI	Dynamical	Tallapragada et al. (2014)
NHC official forecast	OFCL	Consensus	Simon et al. (2018)

forecasts during recent years have mostly been driven by 1) higher resolution dynamical models, which take advantage of increased computational power, to resolve inner-core dynamics and physics as well as air–sea interactions, 2) enhanced observations that enable better initializations of dynamical models, 3) consensus methods that combine forecasts from multiple dynamical and statistical–dynamical models (Sampson et al. 2008; Sampson and Knaff 2009; Simon et al. 2018), and 4) improvements in track forecasts (Cangialosi et al. 2020). At the same time, dynamical intensity models still have some limitations, such as an incomplete understanding of air–sea interaction processes (Lloyd and Vecchi 2011), the underestimation of convection and convective cloud processes (Fovell and Bu 2015), and the lack of real-time inner core TC observations (Shimada et al. 2018). Hence, statistical models and statistical–dynamical models remain competitive and are included in official intensity forecasts worldwide (Cangialosi 2019; Courtney et al. 2019).

Historically, multiple linear regression has been the most commonly used method in statistical and statistical–dynamical intensity forecast models. As suggested by DeMaria and Kaplan (1994), the intensity change for a fixed time interval follows an approximate normal distribution with a mean close to zero, hence multiple linear regression may be a reasonable choice. Despite this, air–sea interaction and other processes in the TC core are highly nonlinear, which suggests that nonlinear methods have the potential to improve statistically based intensity forecast models. In particular, deep learning (DL), or deep neural networks, are capable of learning highly complex and nonlinear relationships involving many predictors. DL has achieved remarkable success in computer vision and pattern recognition in recent years. Yet its application to TC intensity prediction has been rare. A few pioneer studies that applied neural networks to forecast intensity (Sharma et al. 2013 for the western North Pacific; Chaudhuri et al. 2013 for the north Indian Ocean) were conducted before the advent of graphics processing units (GPUs). Hence, the trained networks were much smaller in terms of the number of nodes and the number of trainable weights. The few recent studies that used DL to forecast TC intensity in the North Atlantic basin include the following:

- The 2018 Climate Informatic Hackathon (Giffard-Roisin et al. 2018) received solutions from 35 teams to forecast 24-h intensity on global historical storms, including a few DL solutions leveraging convolutional neural networks on feature maps of atmospheric and oceanic conditions. The results revealed a significant overfitting problem in the submitted

DL solutions, indicating the need to thoroughly test for unseen data.

- Cloud et al. (2019) trained a multilayer perceptron (MLP) to predict Atlantic and eastern Pacific 3–72-h TC intensity change using 18 predictors from Hurricane Weather Research and Forecasting (HWRF) (Tallapragada et al. 2014) reforecast data from 2014 to 2016. Their neural network–based model exhibited excellent rapid intensification (RI) forecast skill, but showed no significant improvement over the observation-adjusted HWRF (HWFI) in terms of MAE.

In this study we took a similar approach to that of Cloud et al. (2019) but combined predictors from a statistical–dynamical model, the Statistical Hurricane Intensity Prediction Scheme (SHIPS) (DeMaria et al. 2005), with a MLP model that was extensively optimized in terms of its architecture and hyperparameters. Our contribution to this domain includes the following four elements:

- 1) Use of MLP as an advanced statistical approach to predict 24-h TC intensity change using predictors from SHIPS.
- 2) Use of automated optimization techniques to tune neural network architecture and hyperparameters, and fully explore the potential of MLP in predicting TC intensity.
- 3) Apply the MLP method to climate-scale TC studies by coupling a synthetic track model with an MLP-based intensity model to generate synthetic storms.
- 4) Make the data and method of this study available to promote comparable or more advanced development of machine learning (ML) algorithms for TC intensity forecasting.

2. Data

Global data are used for model development and testing is just conducted in the Atlantic basin. The dataset used for this study consists of SHIPS model inputs as predictors, and 24- or 6-h intensity changes as predictands. The data sources are described below. To compare our model performance to that of other models, we collected operational forecasts from several state-of-the-art models along with the NHC’s official forecast.

a. SHIPS predictors

SHIPS, a statistical–dynamical model employed operationally at NHC, uses a multiple linear regression technique that features climatological, persistence, and synoptic predictors (DeMaria et al. 2005). Storm environment predictors

are derived from the National Centers for Environmental Prediction (NCEP) global forecasting system (GFS), including zonal and meridional wind, shear, vorticity, divergence, etc. Examples of climatological predictors include the climatological sea surface temperature (SST) and the climatological depth of 20°C isotherm at the TC location and time of the year, both of which are derived from the 2005–10 mean of the Navy Coupled Ocean Data Assimilation (NCODA, Cummings 2005) analyses. Oceanic predictors along the path of the TC include SST and oceanic heat content (OHC) derived from daily or weekly NCODA analyses. Several predictors related to brightness temperature are derived from GOES infrared images, which contain information about the strength and organization of convection (DeMaria et al. 2005). The feature selection process for the 24-h model is described in section 2d, and the feature selection process for the 6-h model is described in section 4b. [The short name, long name, mean value, and standard deviation (STD) of the selected predictors for the two models are described in Tables A1 and A2 in the appendix.] As shown in appendix, the SHIPS developmental dataset includes more than 100 predictors, with many of those being available every 6 h, but only about 20% of those are selected for inclusion in the operational SHIPS forecast model, based on standard multiple linear regression significance tests.

The reanalyzed 6-hourly SHIPS predictors, available from 1982 to 2017, are obtained from the National Oceanic and Atmospheric Administration's (NOAA's) Regional and Mesoscale Meteorology Branch (RAMMB) (http://rammb.cira.colostate.edu/research/tropical_cyclones/ships/index.asp, last accessed 29 October 2020). The operational 6-hourly SHIPS predictors from 2010 to 2020 were acquired directly from NHC, but the same dataset is available from RAMMB as well. Because different types of predictors affect the model evaluation results, it is worthwhile to distinguish the differences between the reanalysis and operational datasets:

- Reanalysis data (or reforecast data): Reanalysis is a scientific method that combines earth system observations and a numerical model to generate a synthesized estimate of the state of the system. The reanalyzed TC datasets are often released annually or after significant improvements have been made in the numerical models and data assimilation schemes, thereby incorporating more accurate observed track location and environmental conditions. A “perfect prog” approach is used for the SHIPS developmental dataset, in which the final best track positions are used and analysis fields valid at future times are used instead of model forecast fields.
- Operational data (or real-time data): Unlike the reanalysis dataset, the operational predictors are estimates that are available when the NHC official forecasts are made. The operational dataset may contain inaccurate track information, inaccurate initial intensity and trends of intensity, along with other biases in environmental conditions due to the use of global model forecast fields. As expected, with the same statistical or statistical–dynamical model, TC intensity predictions based on operational data may have larger errors than those based on reanalysis data.

Although operational predictors are less accurate than reanalysis predictors, it is important to use the former for training to make sure that the data distribution in the model development environment and the final deployment environment are consistent. Due to the small sample size of operational data, for training and validation purposes we included all reanalysis samples (1982–2017) in addition to the operational samples (2010–18). To evaluate the impacts of the use of both reanalysis and operational data for an overlapping set of cases (for years 2010–17), two controlled experiments were conducted with one including reanalysis data, and the other excluding reanalysis data. The testing results in Year 2020 showed that including overlapping reanalysis cases improves the MLP predictive skills, indicating that the small sample size is still a limiting factor for the MLP's performance. To further overcome the sample size limitations, we adopted a leave one year out (LOYO) testing strategy, as described in detail in section 3c.

b. Intensity records

The TC track and intensity data for 1982–2019 were obtained from Kerry Emanuel's Global Tropical Cyclone Data in Network Common Data Form format (ftp://texmex.mit.edu/pub/emanuel/HURR/tracks_netcdf/; last accessed 29 October 2020), which combines data from the NHC and the U.S. Navy's Joint Typhoon Warning Center. The TC track and intensity data for 2020 were obtained from NOAA's International Best Track Archive for Climate Stewardship (IBTrACS) (Knapp et al. 2010, 2018). The intensity change, calculated as the target time-step intensity minus current intensity, is used as the predictand in this study. Predictands are matched to predictors based on location and time.

c. SHIPS, DSHP, LGEM, HWFI, and OFCL

To provide a comparison with the MLP forecasts, the operational forecasts from SHIPS, Decay-Statistical Hurricane Intensity Prediction Scheme (DSHP) (DeMaria et al. 2006), Logistic Growth Equation Model (LGEM) (DeMaria 2009), HWFI, and NHC Official Forecasts (OFCL) (Simon et al. 2018) for the Atlantic basin were acquired from the NHC's operational forecast archive (<https://ftp.nhc.noaa.gov/atcf/archive/>; last accessed 26 October 2020) (Sampson and Schrader 2000). The DSHP is based on SHIPS but has an inland decay component. LGEM uses a dynamical prediction system, whereby the wind speed growth rate is determined by a subset of SHIPS predictors. OFCL is subjectively determined by NHC forecasters and is based on all the model predictions that are available within about 3 h after synoptic time.

d. Feature selection and data preprocessing

In the standard release of the SHIPS developmental dataset, more than 500 predictors are related to the 24-h intensity change. We first removed the predictors that are only available in the reanalysis, and then removed predictors that are dependent on TC intensity, such as normalized ocean age by the maximum wind (NAGE). Most of the remaining features were time-dependent predictors (such as vertical shear) that are provided at each of the time steps of 0, 6, 12, 18, and 24 h.

TABLE 2. Summary of the availability of data for the 24-h intensity model.

Type	Basin	Available years	No. of cases	No. of overland cases
Reanalysis	Atlantic	1982–2017	10 710	828
	Central Pacific	25 years between 1982 and 2016	853	0
	Eastern North Pacific	1982–2017	13 397	89
	North Indian Ocean	1990–2017	2381	145
	Southern Hemisphere	1998–2017	10 836	539
	Western North Pacific	1990–2017	18 046	686
	Total	—	56 223	2287
Operational	Atlantic	2010–18	2772	152
	Atlantic, reserved for independent testing	2019–20	1349	173

SHIPS's practice is to average the time-dependent predictors along the track (DeMaria et al. 2005). We conducted linear regression experiments using only 24-h predictors and using only 0–24-h average predictors, and found that the 24-h predictors resulted in models with smaller predictive errors. This result is consistent regardless whether reanalysis data or operational data are used for training and testing. As a result, we retained only the 24-h time-step predictors for the time-dependent variables. For TC persistence and trend, we retained the current intensity and the last 12-h intensity change (DELV-12), following common practices (Knaff et al. 2003; DeMaria et al. 2005) reported in the literature, based on intensity records described in section 2b. (The 121 final predictors used in this study along with statistics are listed in the appendix, Table A1.) Missing values for the distance to the nearest landmass (DTL) are filled based on the projected track location. For the rest of the missing values, we used the reanalysis mean value to fill rows with missing predictors. The final 24-h intensity model predictors include a total sample size of 56 223 in the reanalysis data from 1982 to 2017, and a sample size of 2772 in the operational data from 2010 to 2018. The predictors are normalized by removing the reanalysis data mean and dividing by the reanalysis data STD for each of the predictors.

The features for the 6-h lightweight model are selected differently and are described in the discussion under the experiments and results section. The final 6-h intensity model predictors include a total sample size of 62 590 in the reanalysis data from 1982 to 2017, and a sample size of 3005 in the operational data from 2010 to 2018. The summary of final data availability for the training-validating-testing at basin-level is

described in Table 2 (the 24-h model) and Table 3 (the 6-h model).

3. Methods

a. Multilayer perceptron

DL is a part of a broader set of ML methods based on artificial neural networks, with architectures including MLP, deep belief networks, recurrent neural networks, convolutional neural networks, etc. Here, we start with a relatively simple architecture, MLP, which also requires tabular data as inputs like the predictors of SHIPS. MLP refers to a DL model that has multiple feed-forward and fully connected hidden layers between an input layer and an output layer. Each hidden layer computes a linear combination of the outputs from the previous layer and applies a nonlinear activation function. Thus, each layer transforms its input data into a slightly more abstract and composite representation, making it possible for MLP to learn highly complex and nonlinear relationships involving many predictors. MLP can be used for both classification and regression.

b. Neural networks optimization

To optimize the hyperparameters of our model architecture, we use a Bayesian search algorithm called tree of parzen estimators (TPE) (Bergstra et al. 2011, 2013b), as implemented in the hyperopt python package (Bergstra et al. 2013a). TPE is a sequential model-based optimization algorithm that uses a history of hyperparameter settings $\{x_0, x_1, \dots, x_t\}$ and corresponding objective function evaluations $\{f(x_0), f(x_1), \dots, f(x_t)\}$ to construct an approximation of f , denoted \tilde{f} , that is computationally easier to evaluate and optimize. The \tilde{f} is then

TABLE 3. Summary of data availability for the 6-h intensity model.

Type	Basin	Available years	No. of cases	No. of overland cases
Reanalysis	Atlantic	1982–2017	11 575	1132
	Central Pacific	25 years between 1982 and 2016	942	0
	Eastern North Pacific	1982–2017	14 756	216
	North Indian Ocean	1990–2017	2771	351
	Southern Hemisphere	1998–2017	12 300	816
	Western North Pacific	1990–2017	20 246	1333
	Total	—	62 590	3848
Operational	Atlantic	2010–18	3005	201

TABLE 4. Hyperoptimization search space.

Activation function	No. of layer nodes	No. of hidden layers	Learning rate	Batch size
relu, sigmoid	4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096	1, 2, 3, 4, 5	0.01, 0.001, 0.0001, 0.000 01	4, 8, 16, 32, 64, 128

optimized with respect to an improvement criterion (Jones 2001), which finds x_{t+1} such that $f(x_{t+1})$ is likely to be near a minimum and $f(x_{t+1})$ is likely to produce information that, when used to update f , will reduce the dissimilarity between it and f . With x_{t+1} , f is evaluated, and the value $f(x_{t+1})$ is used to update f . This process continues for a predefined number of iterations or until a stopping condition is reached. In our application, we specified the number of iterations to be 500 and allocated 7 days of computing time on a GPU. The optimization algorithm iterated through 257 combinations before reaching the 7-day time limit, and the best performing model was subsequently selected in the search record.

When TPE was used in Experiment I, we searched the number of hidden layers, the number of nodes in each layer, activation functions in each layer, batch size, learning rate, and regularization coefficient. As an objective function, we used the validation MAE on the epoch when early stopping was reached. Our early stopping criterion occurs when the best validation MAE does not decrease by more than 0.05 kt after 10 epochs. The values of the hyperparameters we search over are shown in Table 4.

While TPE was used in Experiment I, the model in Experiment II was lightweight and allowed for a more thorough and yet computationally demanding search algorithm—Grid Search, which exhaustively examines each possible combination of hyperparameter choices. The implementation details for the two search algorithms can be found in the experiments and results section.

c. LOYO testing scheme

As shown in Fig. 1, to overcome the challenge of the observation period being too short for training a DL model, we designed a LOYO testing scheme that allows for testing of individual years while using the majority of the data. The idea of LOYO can be illustrated in an example of testing the 24-h intensity model for operational forecasts for the Atlantic 2017 season: we trained and validated the model with all the data in all the years and all the basins available, including both reanalysis and operational data, except for the 2017 Atlantic reanalysis or operational data; then the model was tested on the 2017 Atlantic operational data. To test the model on a different year, e.g., 2018, the model was reinitiated and trained from scratch. In this way, the model was only evaluated on unseen testing data, and this provided a fair way to evaluate model performance for different years, while leveraging the limited number of observations. Training and validation data follow a 90%–10% ratio, where validating data are used for early stopping purposes in training.

Although the LOYO testing scheme solved the data limitation problem, it gave the MLP an advantage over other operational models being compared, because the MLP was

exposed to more years of training data. For example, to test model performance for Year 2010, the MLP was exposed to a training dataset that includes 2011–18 data, while the operational models back in 2010 did not have access to such data. To address this problem, additional independent tests were conducted for 2019 and 20 in the first experiment to further validate the intermodel comparisons.

4. Experiments and results

a. 24-h intensity model for operational forecasts

The first experiment trains an MLP model on the 121 selected SHIPS predictors to predict the change in wind speed intensity during a 24-h period. The model training follows the LOYO strategy, where the model is always tested on an unseen year's operational data for the tested basin. The best architecture and hyperparameters are found by conducting a Bayesian TPE algorithm as described in section 3b. The optimal network was found to have two hidden layers and 2048 nodes in each layer. To evaluate the MLP model performance, we compared its 24-h intensity change error statistics with that of SHIPS, DSHP, LGEM, HWFI, and OFCL. For a homogeneous comparison, we only kept the events for which operational forecast results are available from all models.

Figure 2 shows that the MLP model outperformed the statistical-dynamical models SHIPS, DSHP, and LGEM on most LOYO tested years; the 2010–18 mean MAE was lower by 1.91, 1.01, and 0.87 kt compared to the three models, respectively. The 2010–18 multiyear average MAE of the MLP model is slightly lower (by 0.46 kt) than HWFI and slightly higher (by 0.18 kt) than OFCL. To simulate how the MLP may possibly perform in a real-time operational forecast mode, additional independent tests were conducted using 2019 and 20 Atlantic operational data. Again, the MLP outperformed all three statistical-dynamical models. In 2019 the MLP has slightly higher MAE than HWFI (by 5%) but performed comparably with HWFI in 2020. While Fig. 2 uses MAE as the only metric, Table 5 included four more metrics to evaluate MLP: mean error (ME), root-mean-square error (RMSE), R^2 , and prediction STD. As an error metric, RMSE is more sensitive to large errors than MAE. The fact that the MLP has higher MAE but lower RMSE than HWFI in 2019–20 independent tests, indicates that the MLP is more skillful than HWFI in correctly predicting extreme intensity changes. According to Na et al. (2018), the 24-h intensity change STD in OFCL is only two-thirds of that in reality, indicating that not enough cases of large intensity change are forecasted, which is a potential area of improvement. The MLP has the highest STD among all the models in both the LOYO tests (75% of observed STD) and the independent tests (87% of observed

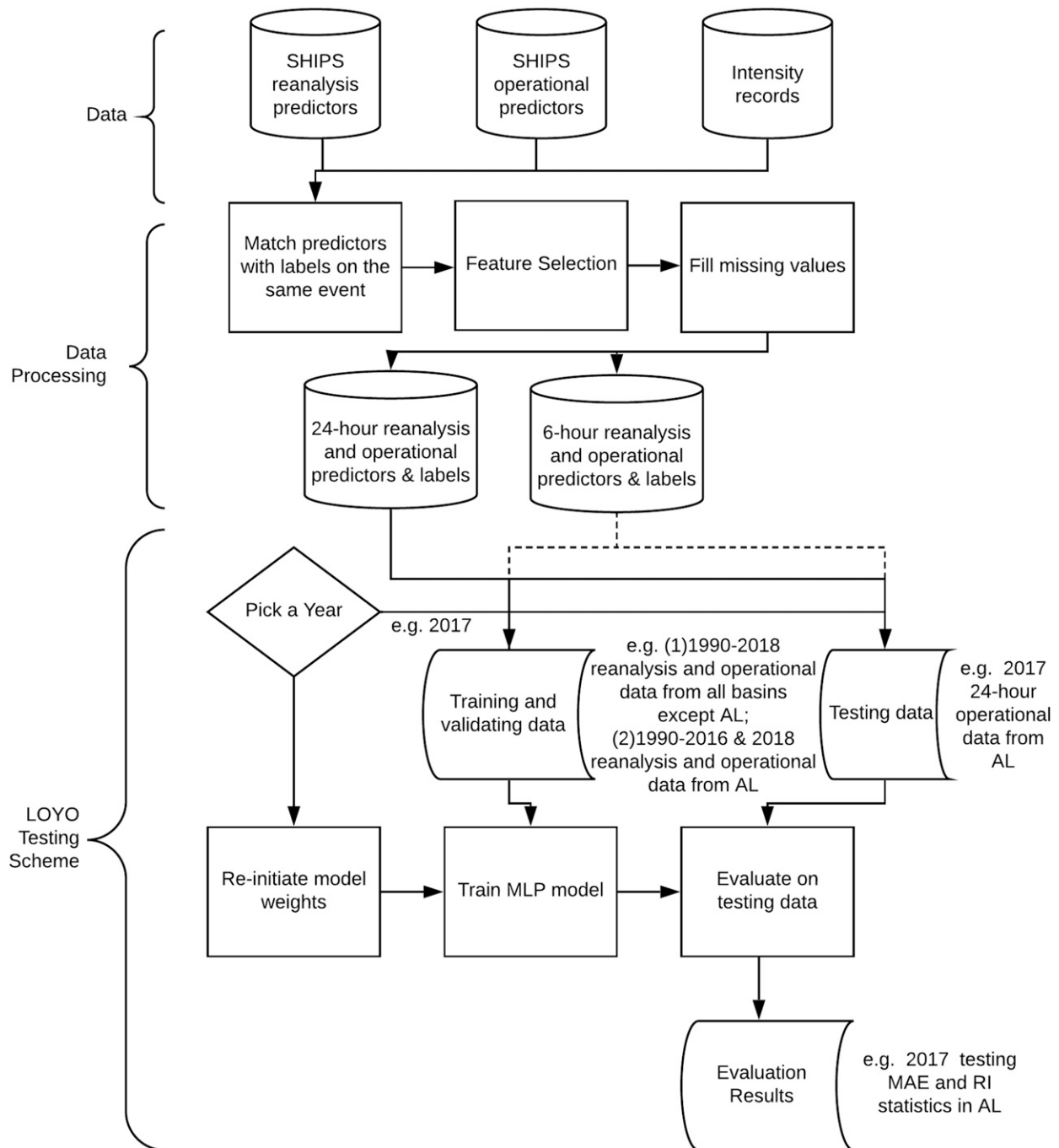


FIG. 1. Data processing and model training flowchart.

STD), suggesting that the MLP can predict a wider range of intensity changes than any other models compared. In terms of bias, all models have fluctuating MEs year-to-year, which is linked to the fact that all models suffer from systematic biases when forecasting extreme intensity changes, as further explained in the error distribution analysis below.

Figure 3 compares the 24-h intensity forecast error distribution from the MLP, SHIPS, DSHP, LGEM, HWFI, and

OFCL of the 2010–18 Atlantic TCs. Na et al. (2018) suggested that OFCL is strongly anticorrelated with TC intensity change, particularly tending to produce negative errors (underforecast) for rapid intensification (RI; 30-kt or more intensity increase over a 24-h period) events, while tending to produce positive errors (overforecast) for rapid weakening (RW, 30-kt or more intensity decrease over a 24-h period) events. In our analysis, OFCL made 26 large overforecasts (>30 kt) for RW events

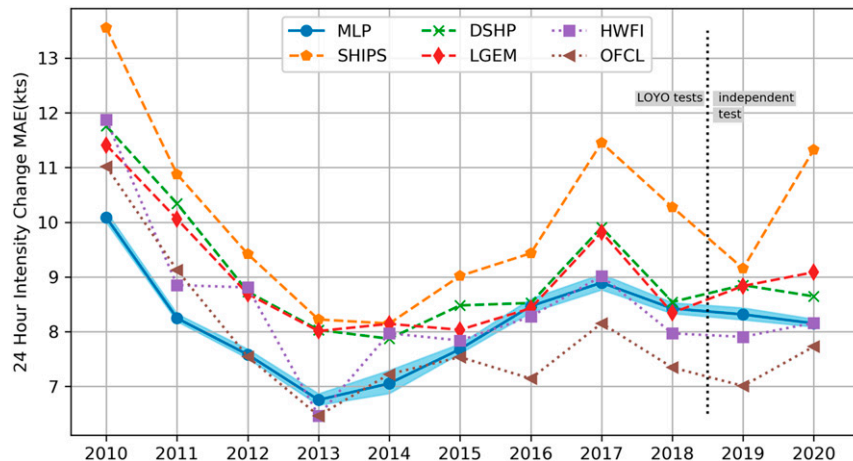


FIG. 2. The 24-h intensity change MAE (kt) from MLP, SHIPS, DSHP, LGEM, HWFI, and OFCL, tested on 2010–20 Atlantic TCs over the same 6-hourly locations. While the MLP model was tested using the LOYO method in 2010–18, it was tested truly independently in 2019 and 20 as if in a real-time mode. The shading around the MLP line denotes the 95% confidence interval based on the bootstrap method with sample size equal to 40 and the number of repetitions equal to 10 000.

and 34 large underforecasts (< -30 kt) for RI events. With knowledge of where land is located, DSHP and HWFI made significantly less overforecasts for RW events (21 and 9, respectively) comparing to SHIPS (50). By comparison, the MLP model only made 6 large overforecasts during the same period, the best among all models. The MLP model made the same number of large underforecasts for RI events as OFCL, better than all other models.

Besides the assessment of model error distribution, it is also important to examine the degree of freedom or independence (Sampson et al. 2008) among forecasting models and consensus. As suggested by Sampson et al. (2008), larger degrees of independence will result in larger improvements of the multi-model consensus. Here we calculated correlations for biases between different models and consensus for 2010–18, as shown in Fig. 4. The correlation between the MLP and other models remains below 0.8, while the correlations between SHIPS, DSHP and LGEM are above 0.8, indicating that the MLP

maintains a relatively higher degree of independence from the SHIPS-related models despite using the same predictors. The correlation between the MLP and HWFI (0.69) is also relatively low, further indicating that the MLP has the potential to improve the consensus.

Although the MLP model is not trained specifically for RI detections, we also assessed its robustness from a different perspective by converting continuous intensity change predictions to binary RI classifications. Table 6 shows RI detection statistics from LOYO tests conducted for the 2017–18 season, as well as from independent tests conducted for the 2019–20 season. For the 54 observed RI events in the Atlantic basin in 2017–18, the MLP made the most correct predictions of all the models and the least missed predictions. Although the MLP produces a slightly higher number of false alarms (FA, 5 versus 3 in HWFI and OFCL), it has the highest Gilbert skill score (GSS) and Peirce skill score (PSS) of all models. The GSS is a commonly used metric in natural disaster prediction that

TABLE 5. The 24-h intensity change model error statistics and model prediction STDs. The MLP model was tested using the LOYO testing scheme during the 2010–18 period and was tested independently in 2019 and 20. Data about the rest of the models (SHIPS, DSHP, LGEM, HWFI, OFCL) come from the NHC's operational forecast archive. During the 2010–18 testing period, there are 2464 6-hourly locations tested, and the observed 24-h intensity change STD is 17.17 kt. During the 2019–20 independent testing period, there are 828 6-hourly locations tested, and the observed 24-h intensity change STD was 19.83 kt.

Model	2010–18 tests statistics					2019–20 test statistics				
	ME	MAE	RMSE	R^2	STD	ME	MAE	RMSE	R^2	STD
MLP	−0.25	8.37	11.42	0.75	13.35	−1	8.22	11.07	0.83	17.28
SHIPS	2.76	10.42	14.35	0.58	11.1	2.05	10.51	15.05	0.66	12.44
DSHP	1.03	9.37	12.45	0.69	12.76	−1.15	8.69	11.61	0.81	17.03
LGEM	−0.35	9.21	12.38	0.69	12.16	−2.62	8.98	11.89	0.81	16.12
HWFI	−0.61	8.83	12.04	0.71	12.79	−1.35	8.08	11.2	0.83	16.86
OFCL	1.51	8.16	11.39	0.75	12.22	1.02	7.45	10.29	0.86	16.56

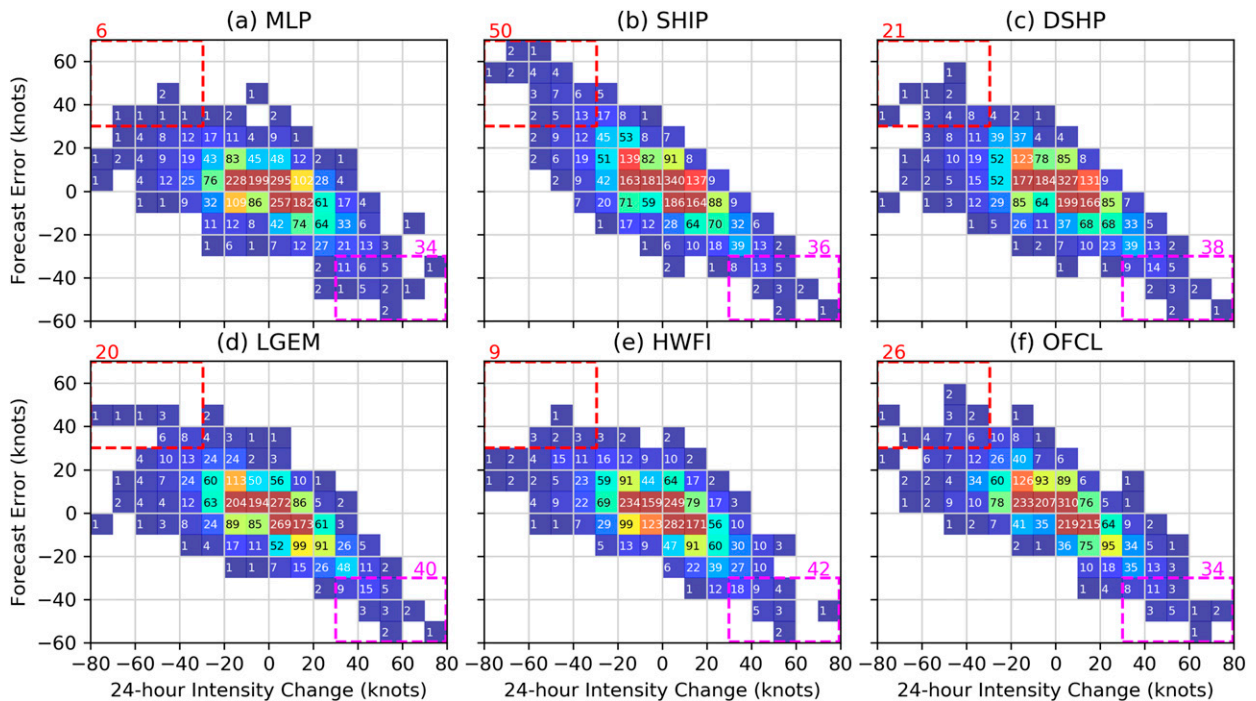


FIG. 3. Distribution of 24-h intensity forecast error frequencies from different models: (a) MLP, (b) SHIPS, (c) DSHP, (d) LGEM, (e) HWFI, and (f) OFCL, with respect to the observed TC 24-h intensity change on 2010–18 Atlantic TCs. All frequencies are labeled by numbers in each grid. The red bounding boxes highlight large overforecast (>30 kt) for rapid weakening (RW) events, and the magenta bounding boxes highlight large underforecast (<-30 kt) for rapid intensification (RI) events. The numbers on top of the red and magenta boxes represent the event counts in bounding boxes.

answers the question of how well did the model detect RI events corresponding to the observed RI events (Wilks 2006). The PSS is calculated as the difference of the probability of detection and the probability of false detection (Wilks 2006), and answers the question of how well did the model separate the RI events from the non-RI events. For the 50 observed RI events in the Atlantic basin in 2019–20, the MLP made the

most correct predictions and, has the highest GSS and PSS of all the models except for OFCL. It is also interesting to note that, SHIPS, DSHP, LGEM and OFCL make significantly more correct RI detections for the 2019 and 20 seasons compared to the 2017 and 18 seasons. HWFI makes the same number of correct RI detections for 2019–20 as in 2017–18; however, it raises significantly more FAs (9 in 2019–20

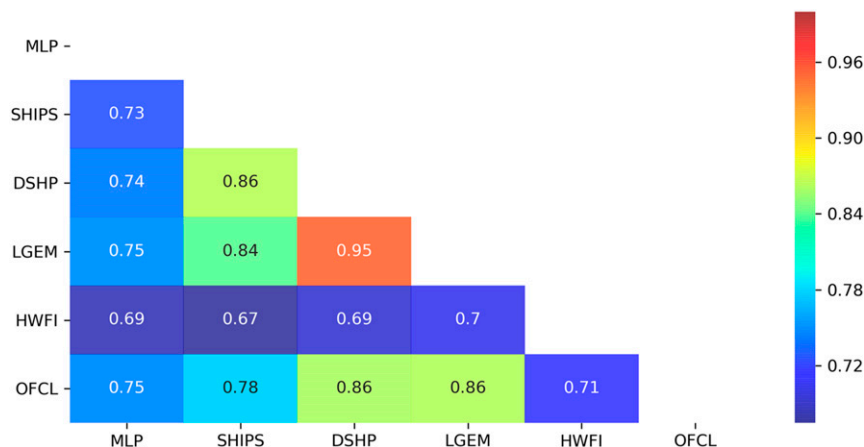


FIG. 4. Heat map of correlation coefficient on 24-h model biases among different models and consensus for 2010–18. The smaller correlation coefficient indicates the higher level of degree of independence between models and the consensus.

TABLE 6. RI detection tested in the Atlantic basin during 2017–18 (LOYO test) and 2019–20 (independent test). The MLP RI statistics come from a majority voting schemes. The numbers in parentheses for the MLP row denote the 95% confidence interval based on the bootstrap method, and sample size is equal to 40 and 10 000 bootstrap repetitions.

	2017–18 RI statistics						2019–20 RI statistics					
	Total	Hits	Misses	FA	GSS	PSS	Total	Hits	Misses	FA	GSS	PSS
MLP	681	10 (7–10)	44 (44–47)	5 (4–6)	0.152 (0.104–0.156)	0.177 (0.122–0.179)	827	11 (10–12)	39 (38–40)	5 (5–6)	0.186 (0.164–0.203)	0.214 (0.192–0.234)
SHIPS	681	1	53	0	0.017	0.019	827	7	43	2	0.125	0.137
DSHP	681	1	53	0	0.017	0.019	827	7	43	2	0.125	0.137
LGEM	681	2	52	0	0.034	0.037	827	6	44	1	0.110	0.119
HWFI	681	7	47	3	0.110	0.125	827	7	43	9	0.104	0.128
OFCL	681	8	46	3	0.127	0.143	827	13	37	4	0.226	0.255

compared to 3 in 2017–18). The year-to-year variability in an individual model's performance may possibly be attributed to its development and interannual variability in RI characteristics. Again, MLP's cross-validation uncertainty is small, with even the lower bound of the 95% confidence level scenario for the MLP achieving a performance comparable to or better than HWFI, indicating that the MLP's superior skill in performing RI classification is statistically significant. These results suggest that the MLP is a powerful tool for detecting RI events, consistent with the conclusions of [Cloud et al. \(2019\)](#). However, even though the skill of the MLP was statistically significant, the detection rate is still around 20%, so considerable improvement is needed to make it a viable guidance model for RI.

We conclude that the MLP has significant skill in predicting the 24-h intensity change. It outperformed the operational statistical-dynamical models and is comparable to the leading dynamical model. The model's cross-validation uncertainty is small, suggesting that it is robust and can produce reliable real-time operational forecasts. The improved prediction power of the MLP model is due to three factors: 1) the nonlinear activation functions enable the MLP to simulate highly nonlinear and complex processes, which probably better describe the TC intensification processes than linear functions; 2) the Bayesian-based hyperparameter search, which optimized the architecture and hyperparameters; and 3) the global reanalysis SHIPS predictors since 1982 provided sufficient training samples for proper training of a DL algorithm.

b. Lightweight 6-h intensity model for climate studies

In addition to forecasting changes in wind intensity for operational forecasts, the community also wants to understand the evolution of long-term TC risk associated with climate change. In such applications, environmental predictors from global climate models are often fed into a synthetic TC model, as described by [Emanuel et al. \(2006\)](#) and [Lee et al. \(2018\)](#), to generate synthetic TC tracks along with intensities. Here, the lightweight model is designed to only depend on the large-scale environmental conditions that would be available from a global climate model. In this setting, since the intensity is updated every 6 h, the target becomes a 6-h intensity change. Also, because the interest here is to better predict TC intensities in climate studies instead of in operational forecasting, the testing is performed only on the reanalysis dataset.

We narrowed the list of predictors based on the literature. [Lin et al. \(2017\)](#) demonstrated that reducing the number of predictors from 26 to 11 yields similar R^2 values with linear and mixture models for 6-h intensity change predictions. Here, we adopted seven important variables from [Lin et al. \(2017\)](#): the current intensity, last 6-h intensity change, vertical shear, 200-hPa zonal wind, maximum potential intensity, latitude, and longitude. We added two variables based on univariate feature importance analysis: 1000-hPa equivalent potential temperature, and DTL. (The details of the nine features and associated statistics are listed in the [appendix, Table A2](#)).

Because the lightweight MLP uses only nine predictors and runs considerably faster than the 24-h model, we were able to

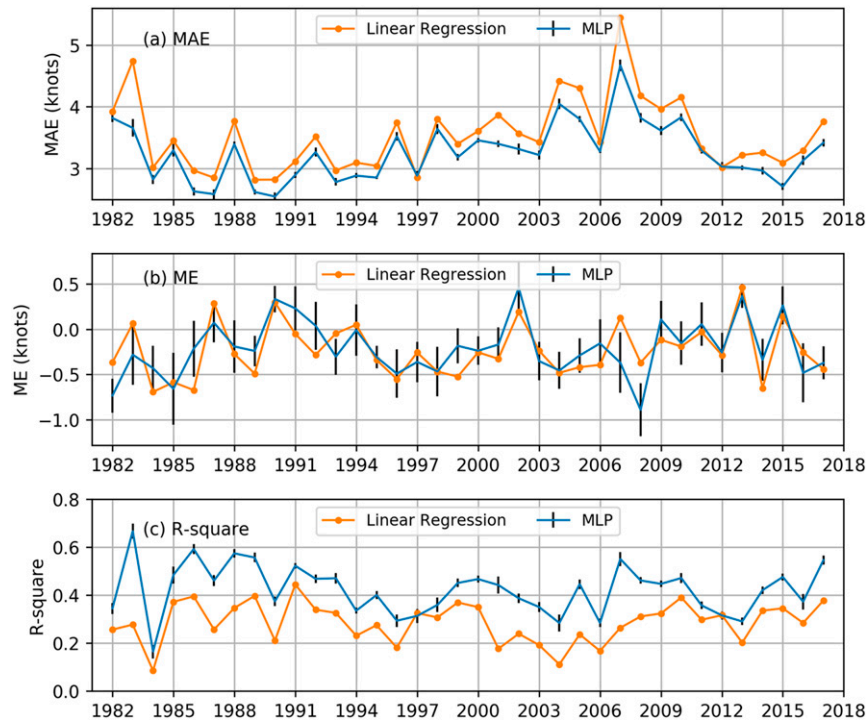


FIG. 5. The 6-h intensity change (a) MAE (kt), (b) ME (kt), and (c) R^2 from the MLP model for each year of 1982–2017 tested on North Atlantic TCs. The black error bars indicate the STD from 10 experiments with different training and validation splits and different random weight initializations.

conduct an exhaustive grid search instead of a Bayesian-based search to determine the best neural network architecture. We used GridSearchCV in the Scikit-learn library (Pedregosa et al. 2011) to perform the grid search. In addition to searching hyperparameters, we searched the number of hidden layers with choices ranging from two to five, and the number of neurons in each layer with choices from 8, 16, 32, 64, 128, or 256. Threefold cross validation was conducted during the grid search. The optimal network was found to have five hidden layers, with number of nodes as 128, 256, 128, 256, and 256, respectively, for each layer. The rest of the hyperparameters are specified as follows: Adam optimization, adaptive learning rate, “ReLU” activation, L2 regularization with alpha as 0.0005, maximum epochs of 200, and without early stopping.

Figure 5 shows the 6-h intensity change testing MAE, ME and R^2 on the 1982–2017 Atlantic basin using the LOYO testing scheme. The MLP consistently performs better than the simple linear regression (LR) model: the MAE from the MLP is lower and the R^2 is higher across all years. The 1982–2017 mean MAE from the MLP is 3.26 kt, compared to 3.63 kt from the LR. The 1982–2017 mean R^2 from the MLP is 0.42, compared to 0.29 from the LR. Both the MLP and LR models have a negative bias when averaged over testing years (mean ME for MLP: -0.21 kt; mean MAE for LR: -0.24 kt); however, the biases are considerably smaller compared to observed 6-hourly intensity change STD (6.38 kt). Because the NHC forecast archive does not provide 6-h intensity

predictions, here our model comparisons are only limited to the MLP and the LR using the same predictors. The MAE STD from 10 experiments generated by changing the training and validation dataset and initializing the weight differently is small, as shown in the error bars (mean STD is 0.065 kt), indicating that the MLP model is robust. Again, the improved predictive skills of MLP for 6-hourly intensity is due to the superiority of the DL algorithm, the automated neural networks optimization, and the abundance of more than 60 000 entries of reanalyzed SHIPS predictors available for model training.

To demonstrate the usefulness and possible applications of the 6-hourly MLP intensity model, we coupled it with a synthetic track model to generate synthetic Atlantic TCs of the current climate. The track model is based on the method of Emanuel et al. (2006) and includes additional improvement of a spatially varying beta drift, as described by Zhao et al. (2009). For a detailed description of the track model application, see Kelly et al. (2018). We generated 50 000 synthetic tracks with MLP-calculated intensities. While storm longitude, latitude, and large-scale wind directly come from the track model, maximum potential intensity and equivalent potential temperature are calculated from the ERA-Interim monthly reanalysis dataset (Berrisford et al. 2011). To generate realistic and yet random environmental conditions, for each storm a random year between 1979 and 2018 is selected, and the ERA-Interim monthly environmental variables from that year and month are used for the MLP intensity model. Tracks are

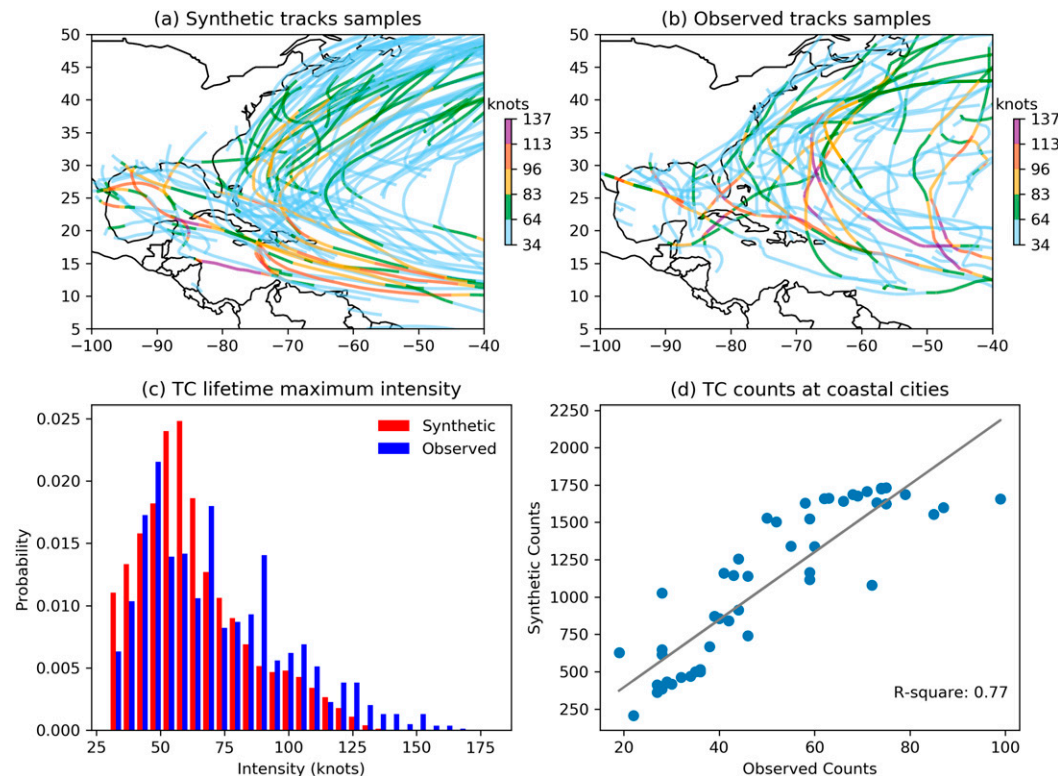


FIG. 6. (a) 100 synthetic tracks with intensity shown by line color, randomly selected from 50 000 synthetic tracks. (b) Randomly selected 100 observed tracks after 1970. (c) TC lifetime maximum intensity distribution comparisons between 50 000 synthetic tracks and 718 observed tracks (1970–2013) in the North Atlantic basin. (d) TC landfall counts at 51 major coastal cities on the U.S. East Coast and the Gulf of Mexico, comparing 50 000 synthetic tracks and 1753 observed tracks (1851–2013). A TC is counted once if the track passes within a 1° radius from the city center coordinates.

terminated when the intensity falls below 25 kt, as described by Emanuel et al. (2006). Figures 6a and 6b show that the overall distribution of synthetic TC tracks resembles that of the observed events. Most synthetic tracks intensify in the Caribbean Sea and the Gulf of Mexico, and tracks are effectively terminated when TCs reach the continent of North America, which is consistent with the physics of TC intensification. In Fig. 6c,

the lifetime maximum intensity distribution from the synthetic TCs is largely similar to the observed distribution. Synthetic TCs can reach a maximum intensity of 148 kt, suggesting that the model can produce storms of category-5 strength. Despite this, the model slightly underestimates the number of intense category-3 and above TCs compared to observations. This is likely because monthly mean values, instead of daily or 6-hourly

TABLE 7. The top 10 most important predictors for 24-h MLP. Relative feature importance scores are determined by the Garson variable importance measure. The feature importance scores of all 121 predictors add up to 1.

Predictors	Long name	Feature importance	Rank
vs0	Initial maximum 1-min sustained wind speed at 10 m (kt)	0.101	1
DTL_t24	Distance to nearest major landmass (km)	0.048	2
TWXC_t24	Maximum 850-hPa symmetric tangential wind at 850 hPa from NCEP analysis ($\text{m s}^{-1} \times 10$)	0.033	3
CD20_t24	Climatological depth (m) of 20°C isotherm from 2005 to 2010 NCODA analyses	0.029	4
VVAC_t24	Average (0–15 km) vertical velocity ($\text{m s}^{-1} \times 100$) of a parcel lifted from the surface, with soundings from 0 to 500 km with GFS vortex removed	0.024	5
T250_t24	200–800-km area average 250-hPa temperature ($^{\circ}\text{C} \times 10$)	0.024	6
V000_t24	Tangential wind ($\text{m s}^{-1} \times 10$) azimuthally averaged at $r = 500$ km	0.023	7
IR00_v16	Minimum GOES brightness temperature from 20- to 120-km radius ($^{\circ}\text{C} \times 10$)	0.021	8
T200_t24	200–800-km area average 200-hPa temperature ($^{\circ}\text{C} \times 10$)	0.020	9
U200_t24	200–800-km area average 200-hPa zonal wind ($\text{kt} \times 10$)	0.019	10

TABLE A1. Complete list of all 121 engineered features used in the 24-h intensity model in Experiment I.

Predictor description	Predictor acronym	Mean	Std dev
Initial maximum 1-min sustained wind speed at 10 m (kt)	vs0	58.9	26.8
Pressure of the center of mass (hPa) of the layer where storm motion best matches environmental flow	PSLV_v2	616.1	77.2
The observed zonal storm motion component ($\text{m s}^{-1} \times 10$)	PSLV_v3	-20.6	40.6
The observed meridional storm motion component ($\text{m s}^{-1} \times 10$)	PSLV_v4	23.1	23.6
As in PSLV_v2, but for the 1000–100-hPa mass weighted deep layer environmental wind ($\text{m s}^{-1} \times 10$)	PSLV_v5	-6.1	40.7
As in PSLV_v3, but for the 1000–100-hPa mass weighted deep layer environmental wind ($\text{m s}^{-1} \times 10$)	PSLV_v6	16.4	22.5
As in PSLV_v2, but for the optimally weighted deep layer mean flow ($\text{m s}^{-1} \times 10$)	PSLV_v7	-14.6	39.2
As in PSLV_v3, but for the optimally weighted deep layer mean flow ($\text{m s}^{-1} \times 10$)	PSLV_v8	18.9	21.0
The parameter alpha that controls the constraint on the weights from being not too “far” from the deep layer mean weights (nondimensional $\times 100$)	PSLV_v9	40.0	0.0
The optimal vertical weights for $p = 100$ hPa (nondimensional $\times 1000$)	PSLV_v10	25.6	9.2
The optimal vertical weights for $p = 150$ hPa (nondimensional $\times 1000$)	PSLV_v11	31.7	31.3
The optimal vertical weights for $p = 200$ hPa (nondimensional $\times 1000$)	PSLV_v12	38.1	29.1
The optimal vertical weights for $p = 250$ hPa (nondimensional $\times 1000$)	PSLV_v13	41.2	23.0
The optimal vertical weights for $p = 300$ hPa (nondimensional $\times 1000$)	PSLV_v14	55.6	39.6
The optimal vertical weights for $p = 400$ hPa (nondimensional $\times 1000$)	PSLV_v15	82.0	46.7
The optimal vertical weights for $p = 500$ hPa (nondimensional $\times 1000$)	PSLV_v16	149.5	76.9
The optimal vertical weights for $p = 700$ hPa (nondimensional $\times 1000$)	PSLV_v17	250.4	92.3
The optimal vertical weights for $p = 850$ hPa (nondimensional $\times 1000$)	PSLV_v18	226.8	92.5
The optimal vertical weights for $p = 1000$ hPa (nondimensional $\times 1000$)	PSLV_v19	99.0	41.8
0–200-km average total precipitable water (TPW) at $t = 0$ from the GFS analysis ($\text{mm} \times 10$)	MTPW_v2	585.2	67.5
0–200-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v3	30.7	13.9
200–400-km average TPW ($\text{mm} \times 10$)	MTPW_v4	529.2	66.3
200–400-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v5	43.0	16.1
400–600-km average TPW ($\text{mm} \times 10$)	MTPW_v6	484.2	67.1
400–600-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v7	55.5	21.8
600–800-km average TPW ($\text{mm} \times 10$)	MTPW_v8	456.2	64.9
600–800-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v9	67.3	25.7
800–1000-km average TPW ($\text{mm} \times 10$)	MTPW_v10	439.2	62.6
800–1000-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v11	76.3	27.3
0–400-km average TPW ($\text{mm} \times 10$)	MTPW_v12	542.2	65.3
0–400-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v13	48.7	17.3
0–600-km average TPW ($\text{mm} \times 10$)	MTPW_v14	509.5	64.8
0–600-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v15	62.1	20.1
0–800-km average TPW ($\text{mm} \times 10$)	MTPW_v16	486.0	63.5
0–800-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v17	71.4	22.0
0–1000-km average TPW ($\text{mm} \times 10$)	MTPW_v18	469.0	62.0
0–1000-km TPW standard deviation ($\text{mm} \times 10$)	MTPW_v19	78.0	23.1
Percent TPW less than 45 mm, $r = 0$ –500 km in 90° azimuthal quadrant centered on up-shear direction	MTPW_v20	228.4	302.0
0–500-km averaged TPW ($\text{mm} \times 10$) in 90° up-shear quadrant	MTPW_v21	513.4	72.1
0–500-km average TPW ($\text{mm} \times 10$)	MTPW_v22	524.5	65.2
Time (hhmm) of the GOES image	IR00_v2	12.9	10.8
Average GOES channel-4 brightness temperature (BT) ($^{\circ}\text{C} \times 10$), $r = 0$ –200 km	IR00_v3	-369.6	210.8
Std dev of GOES BT ($^{\circ}\text{C} \times 10$), $r = 0$ –200 km	IR00_v4	158.4	66.1
Average GOES channel 4 brightness temperature (BT) ($^{\circ}\text{C} \times 10$), $r = 100$ –300 km	IR00_v5	-271.4	169.0
Std dev of GOES BT ($^{\circ}\text{C} \times 10$), $r = 100$ –300 km	IR00_v6	212.8	67.2
Percent area $r = 50$ –200 km of GOES channel 4 BT $< -10^{\circ}\text{C}$	IR00_v7	72.4	25.8
Percent area $r = 50$ –200 km of GOES channel 4 BT $< -20^{\circ}\text{C}$	IR00_v8	65.5	27.3
Percent area $r = 50$ –200 km of GOES channel 4 BT $< -30^{\circ}\text{C}$	IR00_v9	58.6	28.0
Percent area $r = 50$ –200 km of GOES channel 4 BT $< -40^{\circ}\text{C}$	IR00_v10	50.9	28.0
Percent area $r = 50$ –200 km of GOES channel 4 BT $< -50^{\circ}\text{C}$	IR00_v11	40.1	26.9
Percent area $r = 50$ –200 km of GOES channel 4 BT $< -60^{\circ}\text{C}$	IR00_v12	24.9	22.9
Max BT from 0- to 30-km radius ($^{\circ}\text{C} \times 10$)	IR00_v13	-354.3	300.6
Avg BT from 0- to 30-km radius ($^{\circ}\text{C} \times 10$)	IR00_v14	-403.6	286.4

TABLE A1. (Continued)

Predictor description	Predictor acronym	Mean	Std dev
Radius of max BT (km)	IR00_v15	15.8	11.5
Min BT from 20- to 120-km radius ($^{\circ}\text{C} \times 10$)	IR00_v16	-488.9	237.9
Avg BT from 20- to 120-km radius ($^{\circ}\text{C} \times 10$)	IR00_v17	-406.7	236.9
Radius of min BT (km)	IR00_v18	54.4	33.7
Variables No. 1 need for storm size estimation	IR00_v19	87.1	78.9
Variables No. 2 need for storm size estimation	IR00_v20	144.9	126.8
Variables No. 3 need for storm size estimation	IR00_v21	126.6	115.9
Climatological SST ($^{\circ}\text{C} \times 10$)	CSST_t24	271.6	24.4
Climatological depth (m) of 20 $^{\circ}\text{C}$ isotherm from 2005 to 2010 NCODA analyses	CD20_t24	128.0	56.1
As for CD20, but for the 26 $^{\circ}\text{C}$ isotherm	CD26_t24	47.1	23.5
As above, but for ocean heat content (kJ cm^{-2})	COHC_t24	31.2	26.7
Distance to nearest major landmass (km)	DTL_t24	782.1	710.7
Reynolds SST ($^{\circ}\text{C} \times 10$)	RSST_t24	273.1	24.0
200-hPa zonal wind ($\text{kt} \times 10$) ($r = 200\text{--}800$ km)	U200_t24	91.5	146.6
As in U200_t24, but for $r = 0\text{--}500$ km	U20C_t24	59.6	151.9
As in U20C_t24, but for the v component of the wind	V20C_t24	72.5	132.0
1000 hPa theta_e ($r = 200\text{--}800$ km) vs time ($\text{K} \times 10$)	E000_t24	3459.4	82.5
The average theta_e difference between a parcel lifted from the surface and its environment (200–800-km average) vs time ($^{\circ}\text{C} \times 10$)	EPOS_t24	86.3	32.5
As in EPOS, but only negative differences are included. The minus sign is not included.	ENEG_t24	20.9	21.5
As in EPOS, but the parcel theta_e is compared with the saturated theta_e of the environment	EPSS_t24	31.9	23.2
As in ENEG, but the parcel theta_e is compared with the saturated theta_e of the environment	ENSS_t24	39.4	30.2
850–700-hPa relative humidity (%) vs time (200–800 km)	RHLO_t24	67.6	7.4
As in RHLO, but for 700–500 hPa	RHMD_t24	54.3	10.9
As in RHLO, but for 500–300 hPa	RHHI_t24	49.1	11.3
850-hPa vorticity ($\text{s}^{-1} \times 10^7$) vs time ($r = 0\text{--}1000$ km)	Z850_t24	23.9	60.1
200-hPa divergence vs time ($r = 0\text{--}1000$ km)	D200_t24	35.4	40.4
Relative eddy momentum flux convergence ($\text{m s}^{-1} \text{ day}^{-1}$, 100–600-km avg)	REFC_t24	3.9	10.0
Planetary eddy momentum flux convergence ($\text{m s}^{-1} \text{ day}^{-1}$, 100–600-km avg)	PEFC_t24	-1.5	3.1
1000-hPa temperature ($^{\circ}\text{C} \times 10$) (200–800-km average)	T000_t24	254.6	24.1
1000-hPa relative humidity (200–800-km average)	R000_t24	76.1	4.6
1000-hPa height deviation (m) from the U.S. standard atmosphere	Z000_t24	2.9	31.2
Latitude of 850-hPa vortex center in NCEP analysis ($^{\circ}\text{N} \times 10$)	TLAT_t24	258.3	81.3
Longitude of 850-hPa vortex center in NCEP analysis ($^{\circ}\text{W} \times 10$)	TLON_t24	640.6	181.0
0–600-km average symmetric tangential wind at 850 hPa from NCEP analysis ($\text{m s}^{-1} \times 10$)	TWAC_t24	99.1	45.1
Maximum 850-hPa symmetric tangential wind at 850 hPa from NCEP analysis ($\text{m s}^{-1} \times 10$)	TWXC_t24	163.7	76.8
Temperature perturbation at 150 hPa due to the symmetric vortex calculated from the gradient thermal wind. Averaged from $r = 200$ to 800 km centered on input lat/lon (not always the model/analysis vortex position). ($^{\circ}\text{C} \times 10$)	G150_t24	-1.1	6.0
As in G150, but at 200 hPa	G200_t24	4.4	5.5
As in G150, but at 250 hPa	G250_t24	7.4	7.0
The tangential wind ($\text{m s}^{-1} \times 10$) azimuthally averaged at $r = 500$ km from (TLAT, TLON). If TLAT, TLON are not available, (LAT, LON) are used.	V000_t24	63.5	32.5
As in V000, but at 850 hPa	V850_t24	78.9	42.2
As in V000, but at 500 hPa	V500_t24	52.9	41.4
As in V000, but at 300 hPa	V300_t24	8.4	44.0
The magnitude of the temperature gradient between 850 and 700 hPa averaged from 0 to 500 km estimated from the geostrophic thermal wind ($^{\circ}\text{C m}^{-1} \times 10^7$)	TGRD_t24	23.9	18.7
The temperature advection between 850 and 700 hPa averaged from 0 to 500 km from the geostrophic thermal wind ($^{\circ}\text{s}^{-1} \times 10^6$)	TADV_t24	3.6	13.7
Azimuthally averaged surface pressure at outer edge of vortex [(hPa - 1000) $\times 10$]	PENC_t24	154.1	28.3
850–200-hPa shear magnitude ($\text{kt} \times 10$) vs time (200–800 km)	SHRD_t24	184.6	99.0
Heading ($^{\circ}$) of above shear vector. Westerly shear has a value of 90 $^{\circ}$.	SHTD_t24	109.7	70.8
850–500-hPa shear magnitude ($\text{kt} \times 10$)	SHRS_t24	73.9	43.7
Heading of above shear vector	SHTS_t24	127.6	74.6

TABLE A1. (Continued)

Predictor description	Predictor acronym	Mean	Std dev
Generalized 850–200-hPa shear magnitude ($\text{kt} \times 10$) vs time (takes into account all levels from 1000 to 100 hPa)	SHRG_t24	256.8	111.1
As in SHRD, but with vortex removed and averaged from 0 to 500 km relative to 850-hPa vortex center	SHDC_t24	168.8	101.9
Heading ($^{\circ}$) of above shear vector. Westerly shear has a value of 90° .	SDDC_t24	121.8	87.7
As in SHRG, but with vortex removed and averaged from 0 to 500 km relative to 850-hPa vortex center	SHGC_t24	265.3	114.0
As in D200, but centered at 850-hPa vortex location	DIVC_t24	35.3	42.8
200–800-km area average 150-hPa temperature ($^{\circ}\text{C} \times 10$)	T150_t24	−662.0	20.6
As above, but for 200-hPa temperature ($^{\circ}\text{C} \times 10$)	T200_t24	−534.0	18.2
As above, but for 250-hPa temperature ($^{\circ}\text{C} \times 10$)	T250_t24	−418.1	21.6
200–800-km average surface pressure [$(\text{hPa} - 1000) \times 10$]	PENV_t24	130.6	36.8
Maximum potential intensity from K. Emanuel equation (kt)	VMPI_t24	118.5	34.5
Average (0–15 km) vertical velocity ($\text{m s}^{-1} \times 100$) of a parcel lifted from the surface where entrainment, the ice phase and the condensate weight are accounted for. Note: Moisture and temperature biases between the operational and reanalysis files make this variable inconsistent in the 2001–07 sample, compared 2000 and before.	VVAV_t24	779.3	479.6
As in VVAV, but a density weighted vertical average.	VMFX_t24	575.1	280.0
As in VVAV, but with soundings from 0 to 500 km with GFS vortex removed	VVAC_t24	825.6	527.1
Storm motion relative helicity ($\text{m}^2 \text{s}^{-2}$) $\times 10$ for $p = 1000$ – 700 hPa, $r = 200$ – 800 km	HE07_t24	−2.8	36.0
As in HE05, but for $P = 1000$ – 500 hPa	HE05_t24	−6.7	52.8
Pressure vertical velocity (hPa day^{-1}) at 500 hPa, averaged from $r = 0$ to 1000 km	O500_t24	−45.3	48.0
As in O500, but at 700 hPa	O700_t24	−40.1	44.7
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	CFLX_t24	319.8	196.5
Last 12-h intensity change (kt)	DELV-12	2.2	10.2

values, are used to represent the large-scale environmental conditions in this particular application, which results in the smoothing out of sharp gradients and extreme conditions that are linked to the strongest TCs. Another possible reason is that the small negative bias (-0.21 kt) in the 6-h intensity model may have caused an overall underestimation of lifetime maximum intensity. In future applications, bias correction may be helpful to generate more realistic future synthetic TCs; however, the application demonstrated here is not bias corrected. Despite this, the model is able to simulate 1534 category-4 and category-5 TCs out of the 50 000 synthetic storms. Figure 6d compares TC landfall counts at 51 major coastal cities, along the U.S. East and Gulf Coasts, from the model and observations. While the observed landfalls are limited in number, the synthetic approach generated an abundant number of landfall

events (20 times more than observations). The synthetic landfalling TC counts correlate very well with the observed landfalling TC counts for the 51 cities with a correlation coefficient of 0.77. This further supports the ability of the model to realistically represent TC track distribution in a given climate.

c. Predictor importance ranking

To gain insight into why the MLP model was able to achieve competitive predictive skills using the same predictors available to SHIPS, we ranked the 121 predictors from Experiment I according to the Garson variable importance score. The Garson variable importance score for a predictor is calculated from the absolute values of the products of all the neural network weights that are connected to the predictor (Goh 1995).

TABLE A2. Complete list of all nine predictors for the lightweight 6-h intensity model in Experiment II.

Predictor description	Predictor acronym	Mean	Std dev
Initial maximum 1-min sustained wind speed at 10 m (kt)	vs0	54.9	26.0
Last 6-h intensity change (kt)	DELV-6	0.5	6.2
850–200-hPa shear magnitude ($\text{kt} \times 10$) (200–800 km)	SHRD_t0	173.4	95.2
200-hPa zonal wind ($\text{kt} \times 10$) ($r = 200$ – 800 km)	U200_t0	70.3	143.4
Maximum potential intensity from K. Emanuel equation (kt)	VMPI_t0	122.5	30.4
1000-hPa theta_e ($r = 200$ – 800 km) ($\text{K} \times 10$)	E000_t0	3464.2	76.6
Distance to the nearest major landmass (km) in the next 6 h	DTL_t6	784.9	700.6
Latitude	LAT_t0	12.3	16.0
Longitude	LON_t0	25.4	112.2

The scores are subsequently scaled so that the scores from all predictors add up to 1.

Table 7 shows the top 10 most important predictors for the 24-h MLP model. Current intensity (vs0), with an importance score of 0.101, is the most importance predictor, which is consistent with Cloud et al. (2019). DTL is the second most important predictor, contributing close to 5% of the predictability. TC's maximum wind speed decreases typically by around 50% during the first 12 h after landfall, and land interaction has been known to be important to intensity forecasting (DeMaria et al. 2006; Cangialosi et al. 2020). Tangential wind related predictors, TWXC and V000, describe the quality of the TC cyclonic structure and are ranked the third and seventh most important predictors. The climatological depth of the 20°C isotherm (CD20) is related to the ocean heat content that controls the energy supply of TCs (Wada and Usui 2007). The uplift vertical velocity of a parcel (VVAC) is related to the convective instability and is an important part of TC intensifying process (Wang 2014). DeMaria and Kaplan (1999) showed that air temperature and zonal wind are significant predictors of SHIPS, and in our MLP feature ranking, temperature at 250 and 200 hPa (T250 and T200) as well as 200-hPa zonal wind (U200) made it to the list of top 10 predictors. Brightness temperature derived from satellite images contains information related to the strength of convection and hence is a good indicator of intensity change (DeMaria et al. 2005; Shimada et al. 2018).

The MLP may rank the significance of a predictor differently from a linear method for two reasons: 1) with 121 predictors, the MLP model leverages more information to make predictions than other statistical–dynamical models, and therefore may use the predictors differently; 2) the MLP is a neural network–based nonlinear model, which allows for more complex relationships between the predictors and the predictand in comparison to single weight coefficient–defined relationships in linear regression. Predictors previously not shown to be important in a linear method may turn out to be useful in a nonlinear approach and vice versa. Despite differences between the MLP and the more traditional multivariate linear regression methods, it may still be surprising that vertical shear is not among the top ten most important predictors for instance. We found that vertical wind shear and U200 (ranked tenth in feature importance) have a significant correlation coefficient of 0.55, which indicates that effects of wind shear are, at least partially, included through the use of U200. Although the predictor importance ranking offers an opportunity to understand how the MLP makes its predictions, we acknowledge that the ranking here does not constitute feature selection recommendations for other modeling practices that are not neural network based.

5. Discussion and conclusions

Two DL experiments were conducted to predict TC intensity change, using a DL model—MLP—trained with predictors that are available from SHIPS, a well-known statistical–dynamical model. In the LOYO tests of the first experiment, the MLP model predicted a 24-h intensity change that had a

20% lower MAE than SHIPS for Atlantic forecasts that used only input that was available in real-time. In the 2019 and 20 independent tests where the MLP model is tested as if in a real-time operational forecast mode, it again outperformed all three statistical–dynamical models, with a MAE lower than those of SHIPS, DSHP and LGEM by 22%, 5%, and 8%, respectively. When compared to the leading dynamical model HWFI, the MLP has lower RMSE (by -1%), higher MAE (by 2%), and the same R^2 indicating that the MLP's forecasting skills are broadly comparable to that of HWFI. The MLP also detected RI events more accurately than other models and had the highest GSS and PSS. In the second experiment, a light-weight MLP model using only nine predictors consistently outperformed the linear regression for 6-hourly predictions and achieved a model-data R^2 correlation coefficient of 0.42 on 1982–2017 data, which is significantly higher than those reported in previous studies (Lin et al. 2017). When coupled with a synthetic track model, the MLP model generated realistic synthetic TCs in the Atlantic basin, which demonstrated the possibility of using this DL-based intensity model to generate a large quantity of synthetic TCs for the current climate and hypothetical climate scenarios.

The MLP has not only demonstrated competitive predictive skills, but also maintained relatively high level of independence from other operational models due to its DL-based modeling framework. As a result, the MLP would potentially be a meaningful addition to the NHC consensus methods to further improve official forecast skills, thereby helping address a task that has challenged scientists for decades. Previous literature indicates that implementing DL for TC intensity predictions has certain drawbacks, including overfitting and optimization challenges. In this study, we overcame the first challenge by using a LOYO testing scheme that keeps the majority of the data for training/validation, while allowing the model to be thoroughly tested on unseen data. To further validate that the MLP model's superior performance is not due to overfitting, we supplemented the LOYO testing with additional independent tests and simulated how the MLP will perform in a real-time operational forecast mode. This study also addressed the optimization problem by implementing two automated optimization methods for DL architecture and hyperparameters. The techniques and methods used in this study could pave the way for future applications of DL in weather and climate-related studies.

The methodology and experiments can be further improved. Although trained on a global dataset, the models have only been tested in the Atlantic basin and not yet on a global scale. It would be interesting to see the predictive skills of the proposed approach in other basins. Also, it would be interesting to see if the MLP can be further improved through the addition of more real-time satellite observations of the upper-ocean and atmosphere (Balaguru et al. 2020; Su et al. 2020). Another major limitation of the current version of the DL model is that it only provides forecasts for 24-h period. NHC and Joint Typhoon Warning Center (JTWC) currently provide intensity forecasts for up to 5 days and are considering extending those to 7 days. The 24-h intensity forecasts rely heavily on persistence, but the longer-range forecasts do not. Applying the DL method beyond 24 h is needed to determine its usefulness for

operational forecasting. Also, because the longer-range forecasts depend more heavily on the time-dependent predictors, the use of time-averaging versus just the value at the forecast time, as used in the current DL model, will need to be reevaluated. The MLP-based intensity models may be further improved by extending the DL architecture search to beyond five hidden layers, which may result in deeper and more powerful models. To incorporate the methodology in real-time operational forecasts, predictors from models other than SHIPS can be added to further improve model performance. Last, the MLP objective function has been defined to be MAE for the first experiment, but the model can be tailored to a specific purpose by modifying the objective function; for example, creating a specialized RI detection model with even higher true detections by rewarding true positives in the objective function.

Acknowledgments. The operational forecast portion of this research was supported by the Deep Science Agile Initiative at Pacific Northwest National Laboratory (PNNL). It was conducted under the Laboratory Directed Research and Development Program at PNNL. PNNL is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830. The synthetic tropical cyclone portion of this research was supported by the Multisector Dynamics program areas of the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research as part of the multi-program, collaborative Integrated Coastal Modeling (ICoM) project. K. B. acknowledges support from the Regional and Global Modeling and Analysis Program of the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) and from NOAA's Climate Program Office, Climate Monitoring Program (Award NA17OAR4310155). The work would not be possible without the data providers; thanks to RAMMB/CIRA for offering the SHIPS predictors. Thanks to NHC and JTWC for the track and intensity records, and thanks to Dr. K. Emanuel for compiling the global TC records in the netcdf format. Also, thanks to NHC for providing the operational forecast archive. We thank Dr. J. Knaff and two anonymous reviewers for their insightful comments that helped improve this manuscript significantly.

Data availability statement. To encourage more development of ML algorithms for TC intensity forecasting using comparable datasets and methods, the processed data and code to make intensity forecast using the MLP model are made available at public domains. The training, validation, and testing data processed during this study can be downloaded at <http://doi.org/10.5281/zenodo.4784610>. The associated code is at https://github.com/DOE-ICoM/tropicalcyclone_MLP.

APPENDIX

Predictors' Details

Table A1 provides a complete list of the 121 predictors used in the 24-h intensity model in Experiment I. **Table A2**

provides a complete list of the nine predictors used in the lightweight 6-h intensity model in Experiment II. All predictor statistics are derived from the SHIPS reanalysis dataset. The predictor description is adapted from RAMMB (http://rammb.cira.colostate.edu/research/tropical_cyclones/ships/docs/ships_predictor_file_2018.doc).

REFERENCES

- Balaguru, K., G. R. Foltz, L. R. Leung, S. M. Hagos, and D. R. Judi, 2018: On the use of ocean dynamic temperature for hurricane intensity forecasting. *Wea. Forecasting*, **33**, 411–418, <https://doi.org/10.1175/WAF-D-17-0143.1>.
- , —, —, J. Kaplan, W. Xu, N. Reul, and B. Chapron, 2020: Pronounced impact of salinity on rapidly intensifying tropical cyclones. *Bull. Amer. Meteor. Soc.*, **101**, E1497–E1511, <https://doi.org/10.1175/BAMS-D-19-0303.1>.
- Bergstra, J. S., R. Bardenet, Y. Bengio, and B. Kégl, 2011: Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, J. Shawe-Taylor et al., Eds., Information Processing Systems Foundation, Inc., 2546–2554.
- , D. Yamins, and D. D. Cox, 2013a: Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *Proc. 12th Python in Science Conf.*, Austin, TX, Citeseer, 13–20.
- , —, and —, 2013b: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proc. 30th Int. Conf. on Machine Learning*, Atlanta, GA, ICML, 115–123.
- Berrisford, P., D. Dee, P. Poli, R. Brugge, K. Fielding, M. Fuentes, and A. Simmons, 2011: The ERA-Interim archive version 2.0. ERA report series 1, Tech. Rep., ECMWF, 23 pp.
- Cangialosi, J. P., 2019: National Hurricane Center forecast verification report: 2018 hurricane season. National Hurricane Center, 73 pp., www.nhc.noaa.gov/verification/pdfs/Verification_2018.pdf.
- , E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Wea. Forecasting*, **35**, 1913–1922, <https://doi.org/10.1175/WAF-D-20-0059.1>.
- Chaudhuri, S., D. Dutta, S. Goswami, and A. Middey, 2013: Intensity forecast of tropical cyclones over North Indian Ocean using multilayer perceptron model: Skill and performance verification. *Nat. Hazards*, **65**, 97–113, <https://doi.org/10.1007/s11069-012-0346-7>.
- Cloud, K. A., B. J. Reich, C. M. Rozoff, S. Alessandrini, W. E. Lewis, and L. Delle Monache, 2019: A feed forward neural network based on model output statistics for short-term hurricane intensity prediction. *Wea. Forecasting*, **34**, 985–997, <https://doi.org/10.1175/WAF-D-18-0173.1>.
- Combrot, C., A. Mouche, J. A. Knaff, Y. Zhao, Y. Zhao, L. Vinour, Y. Quilfen, and B. Chapron, 2020: Extensive high-resolution Synthetic Aperture Radar (SAR) data analysis of tropical cyclones: Comparisons with SFMR flights and best track. *Mon. Wea. Rev.*, **148**, 4545–4563, <https://doi.org/10.1175/MWR-D-20-0005.1>.
- Courtney, J. B., and Coauthors, 2019: Operational perspectives on tropical cyclone intensity change. Part 1: Recent advances in intensity guidance. *Trop. Cyclone Res. Rev.*, **8**, 123–133, <https://doi.org/10.1016/j.tcr.2019.10.002>.
- Cummings, J. A., 2005: Operational multivariate ocean data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131**, 3583–3604, <https://doi.org/10.1256/qj.05.105>.

- DeMaria, M., 2009: A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.*, **137**, 68–82, <https://doi.org/10.1175/2008MWR2513.1>.
- , and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, [https://doi.org/10.1175/1520-0434\(1994\)009<0209:ASHIPS>2.0.CO;2](https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2).
- , and —, 1999: An updated Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic and eastern North Pacific basins. *Wea. Forecasting*, **14**, 326–337, [https://doi.org/10.1175/1520-0434\(1999\)014<0326:AUSHIP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0326:AUSHIP>2.0.CO;2).
- , M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543, <https://doi.org/10.1175/WAF862.1>.
- , J. A. Knaff, and J. Kaplan, 2006: On the decay of tropical cyclone winds crossing narrow landmasses. *J. Appl. Meteor. Climatol.*, **45**, 491–499, <https://doi.org/10.1175/JAM2351.1>.
- , C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398, <https://doi.org/10.1175/BAMS-D-12-00240.1>.
- Emanuel, K., S. Ravela, E. Vivant, and C. Risi, 2006: A statistical deterministic approach to hurricane risk assessment. *Bull. Amer. Meteor. Soc.*, **87**, 299–314, <https://doi.org/10.1175/BAMS-87-3-299>.
- Fovell, R. G., and Y. P. Bu, 2015: Improving HWRF track and intensity forecasts via model physics evaluation and tuning. DTC visitor program final report, Developmental Testbed Center, 28 pp., https://dtcenter.org/sites/default/files/visitor-projects/DTC_report_2013_Fovell.pdf.
- Giffard-Roisin, S., D. Gagne, A. Boucaud, B. K  gl, M. Yang, G. Charpiat, and C. Monteleoni, 2018: The 2018 climate informatics hackathon: Hurricane intensity forecast. *Eighth Int. Workshop on Climate Informatics*, Boulder, CO, Climate Informatics Hackathon, 4 pp.
- Goh, A. T., 1995: Back-propagation neural networks for modeling complex systems. *Artif. Intell. Eng.*, **9**, 143–151, [https://doi.org/10.1016/0954-1810\(94\)00011-S](https://doi.org/10.1016/0954-1810(94)00011-S).
- Jones, D. R., 2001: A taxonomy of global optimization methods based on response surfaces. *J. Global Optim.*, **21**, 345–383, <https://doi.org/10.1023/A:1012771025575>.
- Kelly, P., L. R. Leung, K. Balaguru, W. Xu, B. Mapes, and B. Soden, 2018: Shape of Atlantic tropical cyclone tracks and the Indian monsoon. *Geophys. Res. Lett.*, **45**, 10–746, <https://doi.org/10.1029/2018GL080098>.
- Knaff, J. A., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting*, **18**, 80–92, [https://doi.org/10.1175/1520-0434\(2003\)018<0080:SDTCIF>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<0080:SDTCIF>2.0.CO;2).
- , C. R. Sampson, and B. R. Strahl, 2020: A tropical cyclone rapid intensification prediction aid for the joint typhoon warning center's areas of responsibility. *Wea. Forecasting*, **35**, 1173–1185, <https://doi.org/10.1175/WAF-D-19-0228.1>.
- Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, <https://doi.org/10.1175/2009BAMS2755.1>.
- , H. J. Diamond, J. P. Kossin, M. C. Kruk, C. J. Schreck, 2018: International best track archive for climate stewardship (IBTrACS) project, version 4. NOAA/National Centers for Environmental Information, accessed 20 April 2021, <https://doi.org/10.25921/82ty-9e16>.
- Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, <https://doi.org/10.1175/MWR-D-12-00254.1>.
- Lee, C. Y., M. K. Tippett, A. H. Sobel, and S. J. Camargo, 2018: An environmentally forced tropical cyclone hazard model. *J. Adv. Model. Earth Syst.*, **10**, 223–241, <https://doi.org/10.1002/2017MS001186>.
- Lin, N., R. Jing, Y. Wang, E. Yonekura, J. Fan, and L. Xue, 2017: A statistical investigation of the dependence of tropical cyclone intensity change on the surrounding environment. *Mon. Wea. Rev.*, **145**, 2813–2831, <https://doi.org/10.1175/MWR-D-16-0368.1>.
- Lloyd, I. D., and G. A. Vecchi, 2011: Observational evidence for oceanic controls on hurricane intensity. *J. Climate*, **24**, 1138–1153, <https://doi.org/10.1175/2010JCLI3763.1>.
- Na, W., J. L. McBride, X. H. Zhang, and Y. H. Duan, 2018: Understanding biases in tropical cyclone intensity forecast error. *Wea. Forecasting*, **33**, 129–138, <https://doi.org/10.1175/WAF-D-17-0106.1>.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
- Rappaport, E. N., J. G. Jiing, C. W. Landsea, S. T. Murillo, and J. L. Franklin, 2012: The joint hurricane test bed: Its first decade of tropical cyclone research-to-operations activities reviewed. *Bull. Amer. Meteor. Soc.*, **93**, 371–380, <https://doi.org/10.1175/BAMS-D-11-00037.1>.
- Sampson, C. R., and A. J. Schrader, 2000: The automated tropical cyclone forecasting system (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240, [https://doi.org/10.1175/1520-0477\(2000\)081<1231:TATCFS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<1231:TATCFS>2.3.CO;2).
- , and J. A. Knaff, 2009: Southern Hemisphere tropical cyclone intensity forecast methods used at the Joint Typhoon Warning Center. Part III: Forecasts based on a multi-model consensus approach. *Aust. Meteor. Oceanogr. J.*, **58**, 19–27, <https://doi.org/10.22499/2.5801.003>.
- , J. L. Franklin, J. A. Knaff, and M. DeMaria, 2008: Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecasting*, **23**, 304–312, <https://doi.org/10.1175/2007WAF2007028.1>.
- Sharma, N., M. M. Ali, J. A. Knaff, and P. Chand, 2013: A soft-computing cyclone intensity prediction scheme for the western North Pacific Ocean. *Atmos. Sci. Lett.*, **14**, 187–192, <https://doi.org/10.1002/asl2.438>.
- Shimada, U., H. Owada, M. Yamaguchi, T. Iriguchi, M. Sawada, K. Aonashi, and K. D. Musgrave, 2018: Further improvements to the Statistical Hurricane Intensity Prediction Scheme using tropical cyclone rainfall and structural features. *Wea. Forecasting*, **33**, 1587–1603, <https://doi.org/10.1175/WAF-D-18-0021.1>.
- Simon, A., A. B. Penny, M. DeMaria, J. L. Franklin, R. J. Pasch, E. N. Rappaport, and D. A. Zelinsky, 2018: A description of the real-time HFIP corrected consensus approach (HCCA) for tropical cyclone track and intensity guidance. *Wea. Forecasting*, **33**, 37–57, <https://doi.org/10.1175/WAF-D-17-0068.1>.
- Su, H., L. Wu, J. H. Jiang, R. Pai, A. Liu, A. J. Zhai, P. Tavallali, and M. DeMaria, 2020: Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophys. Res. Lett.*, **47**, e2020GL089102, <https://doi.org/10.1029/2020GL089102>.
- Tallapragada, V., L. Bernardet, M. K. Biswas, S. Gopalakrishnan, Y. Kwon, Q. Liu, and X. Zhang, 2014: Hurricane Weather

- Research and Forecasting (HWRF) model: 2013 scientific documentation. HWRF Development Testbed Center Tech. Rep., 99 pp., http://www.emc.ncep.noaa.gov/gc_wmb/vxt/pubs/HWRFScientificDocumentation2013.pdf.
- Torn, R. D., and C. Snyder, 2012: Uncertainty of tropical cyclone best-track information. *Wea. Forecasting*, **27**, 715–729, <https://doi.org/10.1175/WAF-D-11-00085.1>.
- , and M. DeMaria, 2021: Validation of ensemble-based probabilistic tropical cyclone intensity change. *Atmosphere*, **12**, 373, <https://doi.org/10.3390/atmos12030373>.
- Wada, A., and N. Usui, 2007: Importance of tropical cyclone heat potential for tropical cyclone intensity and intensification in the western North Pacific. *J. Oceanogr.*, **63**, 427–447, <https://doi.org/10.1007/s10872-007-0039-0>.
- Wang, Z., 2014: Characteristics of convective processes and vertical vorticity from the tropical wave to tropical cyclone stage in a high-resolution numerical model simulation of Tropical Cyclone Fay (2008). *J. Atmos. Sci.*, **71**, 896–915, <https://doi.org/10.1175/JAS-D-13-0256.1>.
- Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.
- Zhao, H., L. Wu, and W. Zhou, 2009: Observational relationship of climatologic beta drift with large-scale environmental flows. *Geophys. Res. Lett.*, **36**, L18809, <https://doi.org/10.1029/2009GL040126>.