A Feed Forward Neural Network Based on Model Output Statistics for Short-Term Hurricane Intensity Prediction

KIRKWOOD A. CLOUD AND BRIAN J. REICH

Department of Statistics, North Carolina State University, Raleigh, North Carolina

CHRISTOPHER M. ROZOFF AND STEFANO ALESSANDRINI

National Security Applications Program, National Center for Atmospheric Research, Boulder, Colorado

WILLIAM E. LEWIS

Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin–Madison, Madison, Wisconsin

LUCA DELLE MONACHE

Center for Western Weather and Water Extremes, Scripps Institute of Oceanography, San Diego, California

(Manuscript received 12 October 2018, in final form 25 May 2019)

ABSTRACT

A feed forward neural network (FFNN) is developed for tropical cyclone (TC) intensity prediction, where intensity is defined as the maximum 1-min average 10-m wind speed. This deep learning model incorporates a real-time operational estimate of the current intensity and predictors derived from Hurricane Weather Research and Forecasting (HWRF; 2017 version) Model forecasts. The FFNN model is developed with the operational constraint of being restricted to 6-h-old HWRF data. Best track intensity data are used for observational verification. The forecast training data are from 2014 to 2016 HWRF reforecast data and cover a wide variety of TCs from both the Atlantic and eastern Pacific Ocean basins. Cross validation shows that the FFNN increasingly outperforms the operational observation-adjusted HWRF (HWFI) in terms of mean absolute error (MAE) at forecast lead times from 3 to 57 h. Out-of-sample testing on real-time data from 2017 shows the HWFI produces lower MAE than the FFNN at lead times of 24 h or less and similar MAEs at later lead times. On the other hand, the 2017 data indicate significant potential for the FFNN in the prediction of rapid intensification (RI), with RI defined here as an intensification of at least 30 kt (1 kt $\approx$ 0.51 m s^{-1}) in a 24-h period. The FFNN produces 4 times the number of hits in HWFI for RI. While the FFNN has more false alarms than the HWFI, Brier skill scores show that, in the Atlantic, the FFNN has significantly greater skill than the HWFI and probabilistic Statistical Hurricane Intensity Prediction System RI index.

1. Introduction

The prediction of tropical cyclone (TC) intensity, typically defined by operational forecast centers as the 1-min averaged maximum 10-m wind speed, is a challenge that has received significant attention from the research and operational forecast communities in recent decades. Rapid intensification (RI), commonly defined as the 95th percentile of 24-h intensity change prediction (Kaplan et al. 2010), has been a particularly daunting problem for operational intensity prediction. Computing resources are increasingly able to resolve the various scales of motion in TCs, handle more advanced data assimilation, and support dynamical ensembles to improve forecast accuracy and the estimation of forecast uncertainty. However, improvements in intensity forecast skill have lagged compared to many other important forecast problems such as TC track prediction (DeMaria et al. 2014). The primary hindrances to intensity prediction include uncertainty in the observations of 1-min sustained wind speed estimates (e.g., Torn and Snyder 2012; Landsea and Franklin 2013; Nolan 2014) and also the conditional high sensitivity of TC intensity to subtle changes in the initial conditions and

Corresponding author: Brian J. Reich, bjreich@ncsu.edu

DOI: 10.1175/WAF-D-18-0173.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

its environment and track at later forecast lead times (e.g., Zhang et al. 2014; Judt et al. 2016; Emanuel and Zhang 2016). Intensity prediction error growth is most rapid in the first 48 h, commencing at the smallest spatial scales, and with increasing time, at the larger scales. With this in mind, reducing the error growth rate over the first two days of a forecast is a worthwhile endeavor in reducing the overall intensity forecast error.

Despite relatively slow progress, numerical weather prediction (NWP) model forecasts of intensity and RI are improving (DeMaria et al. 2014; Cangialosi 2017). Moreover, postprocessing techniques can be applied to NWP and statistical model output to create more accurate forecasts of intensity. A simple multimodel ensemble of intensity forecasts is a widely used technique in operational forecasting. An equally weighted average of intensity predictions from independent models, known as a consensus, often outperforms the constituent models (Sampson et al. 2008; Goerss and Sampson 2014). The multimodel superensemble technique of Krishnamurti et al. (1999) has also demonstrated the value in applying unequal weights in its consensus of several bias-corrected models (e.g., Williford et al. 2003; Simon et al. 2018). Most recently, Ghosh and Krishnamurti (2018) illustrated the promising use of a generalized regression neural network in deriving a weighted multimodel consensus forecast for intensity.

Another promising approach in the realm of postprocessing is to calibrate current forecasts from a NWP or statistical model using past forecasts from the same model along with accompanying historical observational data. Some examples of these types of postprocessing techniques for a deterministic model or a single-model ensemble include simple bias correction, multivariate linear regression, Bayesian model averaging (Raftery et al. 2005), nonhomogeneous Gaussian regression (Feldmann et al. 2015), quantile mapping (Déqué 2007; Alessandrini et al. 2018, hereafter A18), and analog ensembles (e.g., A18). For an example in TC intensity prediction, Bhatia and Nolan (2017) provided a postprocessing method that improves intensity predictions by using a stepwise multiple linear regression formula on large-scale meteorological predictors, along with information describing the atmospheric flow stability and the uncertainty in initial conditions, to predict forecast intensity error in operational prediction schemes. A18 also addressed intensity prediction with the analog ensemble method. More recently, machine learning has gained increasing prominence in postprocessing. Evolutionary programming, simple neural networks, and deep learning have shown significant promise as postprocessing tools (e.g., Gagne et al. 2014; McGovern et al. 2017; Roebber 2018; Rasp and Lerch 2018). All postprocessing techniques, of course, benefit from more extensive training datasets. Therefore, it is imperative to have a sufficiently large archive of previous forecasts or reforecasts, ideally with the same model configuration as used in current operations. Postprocessing techniques also vary in the complexity of input predictors, ranging from a single predictor that is the predicted quantity of interest itself (in our case, TC intensity) to a larger set of more complex spatiotemporal predictors describing various features in the model forecast fields (e.g., A18).

An area in machine learning–based postprocessing for TC intensity prediction that has not yet been deeply explored is incorporating physical details from the model output fields into deep learning methods. In A18's analog ensemble approach for TC intensity prediction, exploiting information about a TC's inner core and environment from an NWP model for TCs helped produce a more skillful forecast of intensity than using an analog ensemble based solely on the NWP model's intensity itself. In the analog ensemble of A18, a calibrated ensemble prediction of TC intensity was developed for a deterministic TC model by seeking a fixed number of analog forecasts from historical data that resemble the current forecast. This was accomplished using a set of predictors that includes the model's predicted intensity itself and a couple of other predictors typically describing the kinematic or thermodynamic aspects of a storm's inner core or environment. Given the model's success in improving intensity forecasts, it is of interest to apply machine learning to a similarly wide ranging set of physical predictors related to TC intensity change.

This paper describes a feed forward neural network prediction tool for TC intensity derived from an operational, full-physics, high-resolution TC model. Section 2 describes the full-physics model that is the basis of this study, along with the various datasets used. The feed forward neural network postprocessing tool and methods for design and testing are described in section 3. Results highlighting the neural network's ability to predict intensity and RI are presented in section 4, followed by the conclusions in section 5.

2. Data sources

a. HWRF

The U.S. Environmental Modeling Center (EMC)'s 2017 operational configuration of the Hurricane Weather Research and Forecasting (HWRF) Model (Biswas et al. 2017) (version 3.9a; hereafter referred to as H217) forms the basis of the data used to construct a neural network postprocessing tool. This full-physics model is a primary tool used at the U.S. National Weather Service (NWS)'s National Hurricane Center (NHC) for the prediction of TC track, intensity, and wind structure. HWRF is a primitive equation, nonhydrostatic, coupled atmosphere–ocean model with a parent domain and two inner TC-following nested grids. The parent domain possesses 18-km horizontal grid spacing and covers about 80° × 80° on a rotated latitude–longitude E-staggered grid centered on a 72-h forecasted TC position, while the inner grids have 6- and 2-km grid spacing and domain sizes of 24° × 24° and 7° × 7°, respectively. For the Atlantic and eastern Pacific Ocean basins examined in this study, the model top is 10 hPa and 75 vertical levels are used. More details on the HWRF physics schemes, ocean coupling, and initialization procedures can be found in Biswas et al. (2017).

Each year, the EMC updates the operational HWRF to take advantage of improved computing capabilities, initialization procedures, and model components. In the annual process of improving the HWRF, reforecasts of storms from a small set of previous seasons are carried out in the Atlantic and eastern Pacific Ocean basins for HWRF testing and evaluation purposes. These reforecast data are also essential for postprocessing and forecast calibration techniques (e.g., A18). For the H217 configuration, reforecast data for TCs in the years 2014–16 are available. A set of physically based predictors are derived from the reforecast output fields to construct the deep-learning model. To subject the model to the rigors of a quasi-operational environment, the same set of predictors are obtained from H217 real-time forecasts during the 2017 hurricane season to evaluate the model's real-time performance in 2017.

It is important to note that the initialization for the reforecasts includes observation-based estimates of the storm's current intensity and position from the hurricane database (HURDAT2; Landsea et al. 2015), which is the final estimate of observations available after a hurricane season ends. In an operational environment, HWRF forecasts are based on a preliminary set of observations, implying the reforecasts will likely possess better error characteristics than real-time data. In addition, in the operational environment, while the HWRF is carried out at the synoptic times of 0000, 0600, 1200, and 1800 UTC, each HWRF run is completed after the NHC has already issued its intensity forecast. The constraints of the operational environment require the NHC to use a 6-h-old HWRF forecast in each of its intensity forecasts. This older run of the HWRF is colloquially referred to as an "early model." Given the seasoned age of the forecast at the time of the NHC forecast, the NHC applies a simple correction to the HWRF intensity prediction that accounts for the current estimate of intensity. This correction is typically applied over the first 96 h of the forecast. The correction is linearly tapered in time from a full correction at a lead time of 0 h to no correction at a lead time of 96 h (Goerss et al. 2004; Sampson et al. 2006, 2008). This corrected ("interpolated") HWRF intensity prediction is referred to at NHC as the HWFI. To be consistent with the operational environment, the prediction technique described in this paper is developed using the "early," 6-h-old version of the HWRF. This approach also provides a more realistic validation of the feed forward neural network by allowing a fair comparison with operational tools that are used in official NHC intensity forecasts.

b. Constructed covariates

Eighteen predictors are incorporated into the deep learning-based postprocessing method for intensity prediction. These predictors are derived directly from HWRF output fields that are available at 3-h increments for each 126-h forecast. These forecast fields are used to derive storm-centered predictors describing thermodynamic and kinematic aspects of the storm's large-scale environment and inner core. Most predictors are axisymmetric with respect to the storm center while a few additional predictors related to azimuthal asymmetries are included. The choice of predictors is based on physical characteristics linked to intensity change in previous studies (e.g., Schubert and Hack 1982; Emanuel 1986; Miyamoto and Takemi 2013; Rogers et al. 2013; Kaplan et al. 2015; A18). In addition to the HWRF-based predictors, the HWFI intensity prediction itself, the initial operational estimate of intensity, and a binary indicator representing whether a storm is in the eastern Pacific or Atlantic Ocean basin are also included as model predictors. The list of predictors is described in detail in Table 1. The mean and standard deviation of each predictor (except for the ocean basin indicator) are provided to give the reader a sense of typical predictor values.

c. Verification dataset

The HURDAT2 database (Landsea and Franklin 2013), which contains multiplatform-based estimates of a TC's intensity throughout its lifetime, is used as the verifying observational benchmark in which to validate the deep learning technique described in this paper. To further evaluate the performance of the deep learning model, we compare the deep learning model

TABLE 1. A complete list of all 18 predictors incorporated into the neural network. All predictors are derived from HWRF output unless stated otherwise. Below, r, p, and T represent radius, pressure, and temperature, respectively. The mean and standard deviation of the predictors are provided in the two rightmost columns as well.

Description of predictor class	No. of predictors for this class	Mean	Std dev
Latitude of storm center (°N)	1	24.0	9.9
Longitude of storm center (°W)	1	107.1	3.9
"Interpolated" maximum 1-min 10-m wind speed (HWFI) (kt)	1	49.7	23.9
Minimum sea level pressure (hPa)	1	991.9	19.9
850–200-hPa vertical wind shear magnitude averaged over $0 \le r < 500$ km (kt)	1	20.7	13.6
Storm translation speed (kt)	1	11.1	6.7
Sea surface T averaged over $0 \le r < 50$ km (K)	1	296.7	9.1
Relative humidity ($200 \le r < 800$ km) averaged over the layer $850 \le p < 700$ hPa (%)	1	66.7	8.8
Convective available potential energy averaged over $0 \le r < 100$ or $200 \le r < 500$ km (J kg^{-1})	2	1132.0/1059.9	836.8/681.9
Surface turbulent sensible heat fluxes averaged over $100 \le r < 200$ km (W m^{-2})	1	8.6	16.4
Total condensate averaged over $100 \le r < 250$ km (g kg^{-1})	1	12.3	9.8
Two inertial stability-based parameters averaged over $850 \le p < 500$ hPa and $0 \le r < 100$ or $100 \le r < 250$ km (10^{-6} s^{-2})	2	4.2/0.2	5.2/0.2
Symmetry parameter (as defined in Miyamoto and Takemi 2013) for total condensate and the coupling of inertial stability/vertical motion, over $850 \le p < 500$ hPa and $0 \le r < 100$ and $100 \le r < 250$ km, respectively (%)	2	31.9/48.8	14.9/21.8
Operational estimate of the maximum 1-min 10-m wind speed at the initial time (kt)	1	57.6	28.8
A binary indicator specifying whether a storm is in the Atlantic or eastern Pacific basin	1	—	—

with the performance of the HWRF prediction for intensity itself, along with the National Oceanic and Atmospheric Administration (NOAA)'s official NHC forecast of TC intensity.

3. Deep learning model

a. *General model*

The model used is a standard feed forward neural network with L layers. Given an input vector of features $\mathbf{x}$ (we use all of the variables in Table 1), the first layer is simply $a_j^1 = x_j$, the jth input predictor. Subsequent layers $l \in \{2, 3, \ldots, L\}$ form linear combinations of outputs from the previous layer and apply a possibly nonlinear activation function:

$$a_j^l = \sigma\left(\sum_{k=1}^{K} w_{jk}^l a_k^{l-1} + b_j^l\right),$$

where w_{jk}^l is the weight from the kth neuron in the $(l-1)$th layer to the jth neuron in the lth layer and $\sigma(x)$ is the activation function. The predicted intensity is $\hat{y}(\mathbf{x}) = a_1^L$, which is a nonlinear function of $\mathbf{x}$ determined by the weights w_{jk}^l and biases b_j^l. A depiction of this architecture is given in Fig. 1.

The weights and biases must be estimated from the training data. The L_p regularization was applied to all weights in the network with a penalty factor of $\lambda > 0$, tuned during cross validation in order to shrink the weights of the network. Regularization stabilizes the parameter estimates by shrinking them toward zero to avoid over fitting. The cost function used was mean squared error, along with the L_p-regularization term:

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{y}(\mathbf{x}_i) - y_i]^2 + \lambda \sum_{l=1}^{L}\sum_{j,k}|w_{jk}^l|^p.$$

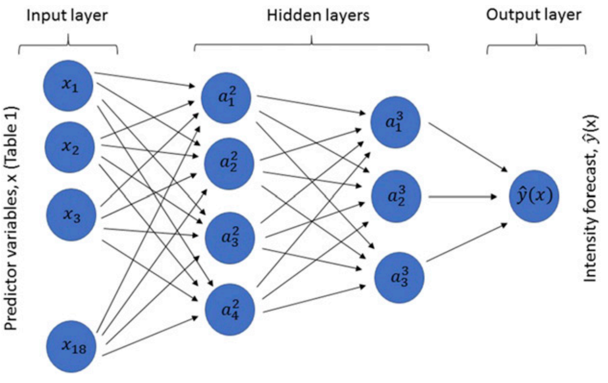

FIG. 1. Neural network architecture: feed forward neural network model that maps the predictors x in Table 1 to the forecast $\hat{y}(\mathbf{x})$. This example network has two hidden layers with four and three neurons, respectively.

All input variables were normalized before training the neural network by subtracting the mean and dividing by the standard deviation of each predictor across the training set [i.e., $x_{i,\text{norm}} = (x_i - \overline{x}_{\text{train}})/s_{\text{train}}$].

b. Treatment of missing data

The HWRF model data output includes missing values that come from different sources, including corrupted data and intentionally omitted data where relevant (e.g., if the storm crosses over land or the storm in a weaker stage gets lost by the vortex tracker). Every feature has at least 5% missing values. Some had up to 19%. We tried a few methods for handling missing values, including indicator (''dummy'') terms in the model for each predictor and data replacement (in particular, multiple imputation by linear regression), but none of these provided a meaningful performance improvement over the naive approach of simply replacing each missing value for a given predictor with the mean across all nonmissing values of that predictor in the training dataset (i.e., $x_{i,\text{norm}} = 0$ if x_i is missing).

c. Tuning and model selection

The model training assumed 6-h-old HWRF data at the forecast time to be consistent with operational constraints. Model tuning was performed using tenfold cross validation on the 2014–16 reforecast data, with the same random partition used for all runs. Combinations of the following model configurations were attempted: between $L = 1$ and $L = 4$ hidden layers, from $K = 100$ to $K = 4000$ neurons per layer, sigmoid $\{\sigma(x) = 1/[\exp(-x) + 1]\}$ versus rectified linear unit [ReLu; $\sigma(x) = x$ if $x > 0$ and $\sigma(x) = 0$ if $x \leq 0$] activation functions, and L_1 versus L_2 regularization methods. These combinations were evaluated based on test set performance across all 10 sets. This tuning process was performed on 3-h lead time data, with occasional checks for performance across lead times 3–75 h. Once strong model candidates were selected on the basis of the 3-h lead time performance, we ran tenfold cross validation separately for each of the 3-, 6-, . . . , 75-h lead times and computed the mean absolute error (MAE) across test sets for each lead time.

The final architecture selected used three layers: the input layer, a densely connected hidden layer with $K = 1300$ sigmoid neurons, and a single linear output layer. We also applied a small L_1 penalty ($\lambda = 0.001$) to all weights in the network. Results from this model are plotted versus lead time in Fig. 2. Based on these comparisons, we tuned the aforementioned configurations in order to minimize cross-validated MAE for 2014–16 storms across all lead times.

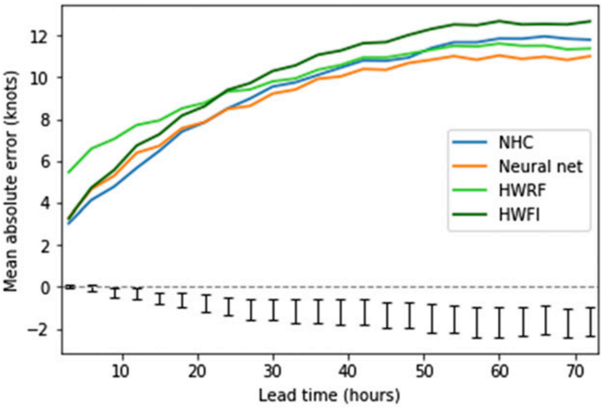

FIG. 2. Cross-validation assessment of forecast accuracy: comparison of the MAE for the baseline HWRF (light green), HWFI (dark green), NHC (blue), and neural network (orange) intensity forecasts across different prediction lead times. Error bars show bootstrapped confidence interval lengths for error terms between the neural network and HWFI forecasts, computed by storm-level resampling of observed absolute errors. The length of the interval is the difference between the 95th and 5th percentile of values for MAE that were calculated under resampling. Note that these interval lengths correspond to symmetric intervals about the values given in the curve but are shifted down on the plot for visual clarity.

d. Cross-validation design

To assess model performance, we once again employed tenfold cross validation, maintaining total separation of testing–training sets for each iteration. Therefore, the model is tuned using only the training data to avoid bias caused by tuning and testing on the same test set observations. To create the partition, we randomly assigned each of the 94 unique storms to a set such that the sets were approximately equal sized. We chose to partition by storms rather than individual observations because observations within a storm are highly correlated and allowing them to be assigned to multiple sets might result in overfitting (Arlot and Celisse 2010).

Within each training–testing pair, neural net weights were reinitialized so that no learning from the previous step carried over. Normalization and imputation were then applied to the nine training sets, and the mean and standard deviation of each predictor in the nine training sets was used to apply normalization and imputation to the test set.

At the end of this process, each observation had one test prediction $\hat{y}_{i,\text{test}}$ and model performance was evaluated using MAE across those values:

$$\frac{1}{n}\sum_{i=1}^{n} |\hat{y}_{i,\text{test}} - y_i|.$$

These results are presented in the next section.

TABLE 2. Cross-validation results for 2014–16 storms, shown for all lead times and competitor models, measured in mean absolute error.

Lead time	No. of obs	No. of storms	NHC	NN	HWRF	HWFI
3	2512	120	3.044	3.311	5.464	3.267
6	2510	120	4.137	4.652	6.591	4.726
9	2480	119	4.774	5.293	7.053	5.566
12	2480	119	5.669	6.391	7.718	6.740
15	2424	119	6.484	6.723	7.939	7.276
18	2424	119	7.411	7.569	8.516	8.175
21	2382	117	7.849	7.860	8.782	8.619
24	2382	117	8.503	8.487	9.314	9.393
27	2314	115	8.983	8.633	9.420	9.721
30	2314	115	9.557	9.221	9.805	10.301
33	2267	112	9.757	9.416	9.944	10.567
36	2267	112	10.112	9.929	10.362	11.075
39	2174	106	10.461	10.042	10.588	11.279
42	2174	106	10.807	10.398	10.950	11.632
45	2131	104	10.791	10.361	10.952	11.682
48	2131	104	10.941	10.685	11.126	12.025
51	1972	100	11.400	10.838	11.316	12.312
54	1972	100	11.675	11.015	11.504	12.525
57	1934	98	11.673	10.840	11.475	12.491
60	1934	98	11.854	11.046	11.616	12.686
63	1881	98	11.845	10.882	11.507	12.529
66	1881	98	11.958	10.986	11.514	12.546
69	1826	98	11.846	10.831	11.337	12.528
72	1826	98	11.799	11.009	11.376	12.675

4. Results

a. *Cross-validation prediction accuracy*

Model performance was evaluated using the MAE across observations in all testing sets in tenfold cross validation, reported in Table 2 for all lead times, including sample size information. We also plot this information in Fig. 2, showing that the neural network (NN) model offers a substantial performance increase over the base, uncorrected HWRF model for 2014–16 reforecast cases. The NN has similar performance to the HWFI at lead times of 3–6 h, but becomes increasingly better at later lead times, with statistically significant differences at the 95% level according to bootstrap confidence intervals. The NN also has slightly improved performance over the NHC official intensity forecast for certain lead times.

b. *Out-of-sample prediction accuracy*

To gain insight into how the NN model would actually perform in an operational environment, we conducted out-of-sample validation on a dataset of 2017 storms that was not available during the model fitting process. The basis of the 2017 validation is real-time HWRF forecasts. As stated earlier, the real-time HWRF is initialized with operational estimates of various vortex parameters such as the intensity. On the other hand, the reforecast data used for the model fitting process employ the more accurate, refined best track data for initialization. To assess true model performance, we simulated operational conditions as closely as possible: the 6-h-old (''early'') operational HWRF was used, predictions were made sequentially in time, and the model was refitted on new storms once they had ended. If two storms overlapped in time, the model would not be refitted until both storms had expired. This ensured that the model only ever utilized information that would have been available in real world usage. The specific algorithm used is as follows.

For each lead time (3, . . . , 72):

1) Initialize training dataset as all observations for storms in 2014–16 (at the given lead time).
2) Reset neural net weights to the initial settings, and then fit neural net on the current training dataset.
3) Predict intensity (VMAX) for each observation in the first storm in 2017 that has not been tested yet.
4) If the storm tested in step 3 ends before the next storm in 2017 begins, add it to the training set and return to step 2. Otherwise, store it in memory until it can be added to the training set, and return to step 3. Repeat until all storms in 2017 have been tested.

There were 21 storms in the 2017 data (Table 3), corresponding to 13 training batches of TCs. The updated 2017 configuration of HWRF used in this study was not

TABLE 3. A list of the 2017 tropical cyclones used in the out-of-sample validation set, along with the basin in which each storm occurred and the maximum intensity (kt) of the storm.

Tropical cyclone name	Basin	Max intensity (kt)	Dates active
Franklin	Atlantic	75	7–10 Aug
Gert	Atlantic	95	12–17 Aug
Harvey	Atlantic	115	17 Aug–1 Sep
Nontropical disturbance 10	Atlantic	40	27–29 Aug
Irma	Atlantic	155	30 Aug–12 Sep
Jose	Atlantic	135	5–22 Sep
Katia	Atlantic	90	5–9 Sep
Lee	Atlantic	100	14–30 Sep
Maria	Atlantic	150	16–30 Sep
Nate	Atlantic	80	4–8 Oct
Ophelia	Atlantic	100	9–15 Oct
Rina	Atlantic	50	4–9 Nov
Tropical depression 11	Eastern Pacific	30	4–8 Aug
Jova	Eastern Pacific	35	11–13 Aug
Kenneth	Eastern Pacific	115	18–23 Aug
Lidia	Eastern Pacific	55	29 Aug–3 Sep
Max	Eastern Pacific	80	13–15 Sep
Norma	Eastern Pacific	65	14–19 Sep
Otis	Eastern Pacific	100	11–19 Sep
Pilar	Eastern Pacific	45	22–25 Sep
Selma	Eastern Pacific	35	26–28 Oct

implemented operationally until the beginning of August 2017. Therefore, the out-of-sample validation dataset does not include numerous storms from 2017. Nonetheless, the Atlantic portion of the dataset contains a significant number of major hurricanes, with 6 TCs exceeding 100-kt intensity ($1\,\mathrm{kt} \approx 0.51\,\mathrm{m\,s^{-1}}$). On the other hand, the eastern Pacific data sample is smaller and only includes two major hurricanes.

Results for the real-time validation are shown in Fig. 3. For the validation set, the NN never greatly outperforms the HWFI, and in fact performs slightly worse in the first 24 h, although it outperforms the baseline HWRF model in the first 18 h and performs better than the official NHC forecast after 27 h.

We also assessed model accuracy by converting the models' continuous wind speed predictions into binary predictions of rapid intensification (RI), where RI is defined in this paper as a 30-kt increase or greater in VMAX from the time of forecast to 24 h in the future. We say the model predicts RI if its 24-h prediction for VMAX is at least 30 kt greater than operational VMAX at the time of prediction. This follows the conventional RI threshold being defined as the 95th percentile of intensity change (e.g., Kaplan and DeMaria 2003; Kaplan et al. 2010).

Given this definition, we evaluated the skill of RI prediction for the NHC, HWFI, and NN according to the Gilbert skill score (equitable threat score), a function of hits, misses, and false alarm rate defined as

$$\frac{\mathrm{hits} - \mathrm{hits}_{\mathrm{random}}}{\mathrm{hits} - \mathrm{hits}_{\mathrm{random}} + \mathrm{misses} + \mathrm{false\ alarms}},$$

with $\mathrm{hits}_{\mathrm{random}} = (\mathrm{hits} + \mathrm{misses})(\mathrm{hits} + \mathrm{false\ alarms})/\mathrm{total}$. Therefore, this score measures skill in forecasting RI events compared to random guessing based on the observed frequency of RI events. This is a common metric for evaluating forecast accuracy that emphasizes true positives over true negatives (Wilks 2006).

Table 4 shows the results on the out-of-sample 2017 dataset, which are promising. With 5 hits, 44 misses, and

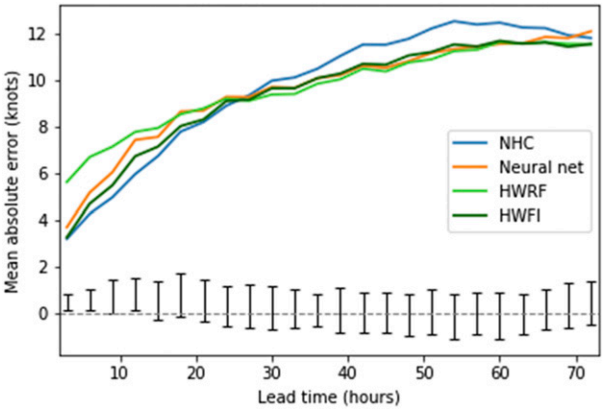

FIG. 3. Out-of-sample assessment of forecast accuracy: comparison of baseline HWRF (light green), HWFI (dark green), NHC (blue), and neural network (orange) performance by MAE on the out-of-sample 2017 data. Error bars showing bootstrapped confidence intervals are as described in Fig. 2.

TABLE 4. Rapid intensification detection on out-of-sample validation set, by prediction (n = 452 predictions per model across 23 storms).

Index	Total	Hits	Misses	False alarms	Gilbert skill score
NHC	452	5	44	0	0.092
HWFI	452	6	43	2	0.102
NN	452	24	25	13	0.345

0 false alarms, the NHC is conservative when it comes to detection of RI events. The HWFI also had a significant number of misses (43), but now also had 2 false alarms. In contrast, the NN successfully identified the most RI events with 24 hits, with the least number of misses (25), although it had an elevated false alarm rate (13). Nonetheless, the NN's high hit rate resulted in a significantly higher Gilbert skill score than the NHC and HWFI.

Another way to evaluate the predictive skill of RI forecasting methods is through the Brier skill score (BSS). The BSS is defined here with respect to the training data's baseline climatological probability of RI. The climatological frequency of RI events in the Atlantic and eastern Pacific basins combined is 11.4% for this dataset, slightly higher than a climatology defined over a longer period of time (e.g., Kaplan et al. 2010). The BSS allows for comparison of the deterministic aids, which either predict RI (100% probability) or do not (0% probability), with operational probabilistic RI models (which have a continuous distribution of probabilities between 0 and 100%). Figure 4 shows the 2017 real-time, out-of-sample BSS for the NN, the probabilistic Statistical Hurricane Prediction System (SHIPS) RI index (RII) (SHIPS-RII; Kaplan et al. 2015), HWFI, and the NHC, for both the Atlantic and eastern Pacific Ocean basins combined, and also for each basin individually. This comparison is homogeneous in that the BSS only considers forecast times in which forecasts are produced by all of the methods. Because there was some missing SHIPS-RII data, the sample here is slightly smaller than the sample used in Table 4 for the Gilbert skill score calculations.

For both basins combined, the sample size is 423 and contains 48 RI cases. The BSS values shown in Fig. 4 are consistent with the earlier Gilbert skill scores. The NN has a BSS of 22.5% while the HWFI is −12.8%. Because the NHC has a lower probability of detection, it is also negative in this particular sample (−8.1%). The SHIPS-RII is competitive with the NN with a BSS of 18.5%. The majority of the 2017 sample is from the Atlantic, with a sample size $N = 318$ and 39 RI cases. All methods perform better here than in the combined basin sample. Moreover, the BSS of the NN is now substantially higher than the next best performing SHIPS-RII (38.7% vs 22.7%). On the other hand, the eastern Pacific shows drastically different results, with a very negative BSS for the NN (−44.4%), which is even lower than the HWFI. The SHIPS-RII is uncharacteristically poor compared to larger samples examined in Kaplan et al. (2015), with a barely skillful model, but it is still significantly better than all other forecasts. It is important to note the eastern Pacific has a small sample size of $N = 105$ and only had 9 RI events in this period of study.

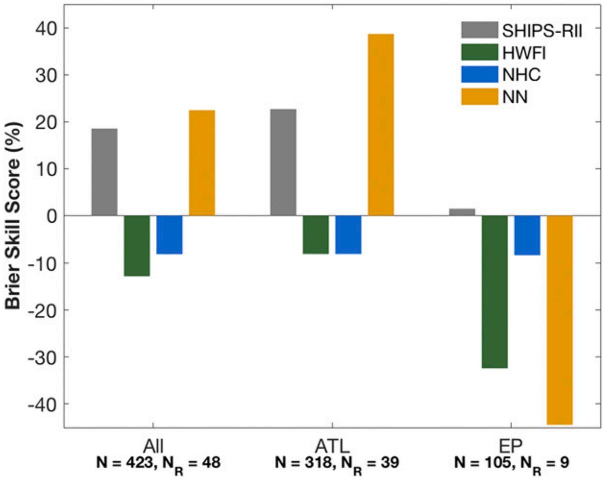

FIG. 4. Brier skill score: the 2017 out-of-sample BSSs with a baseline of the climatology for 24-h RI probabilities predicted by the SHIPS-RII (gray), HWFI (dark green), NHC (blue), and NN (orange). The BSS is shown for 2017 including predictions from both the Atlantic and eastern Pacific Ocean basins and also by single basin. The total sample size N and the total number of observed RI events N_R is shown along the x axis.

A small sample size limits the conclusions that can be drawn from the negative eastern Pacific BSS values, but it is even more mysterious why the NN is worse in that basin than the HWFI in terms of BSS. To gain further insight into the BSS of the NN and operational HWFI that the NN incorporates, Fig. 5 shows all of the NN and HWFI 24-h intensity change predictions for the 2017 samples for the individual ocean basins. The data here are sorted in ascending order. Moreover, forecasts that are at times in which RI actually verified in the following 24 h are highlighted in cyan in Fig. 5. In the Atlantic, Figs. 5a and 5b show that the NN tends to bias correct the HWFI toward higher intensity change values. As a result, more NN predictions correctly predict RI. The NN simultaneously reduces the number of false alarms in the Atlantic. The higher HWFI and NN intensity change predictions both tend to correlate with verifying observations of RI, but clearly the NN appears to be improving upon the HWFI. Figures 5c and 5d show some similar patterns in the eastern Pacific, in that

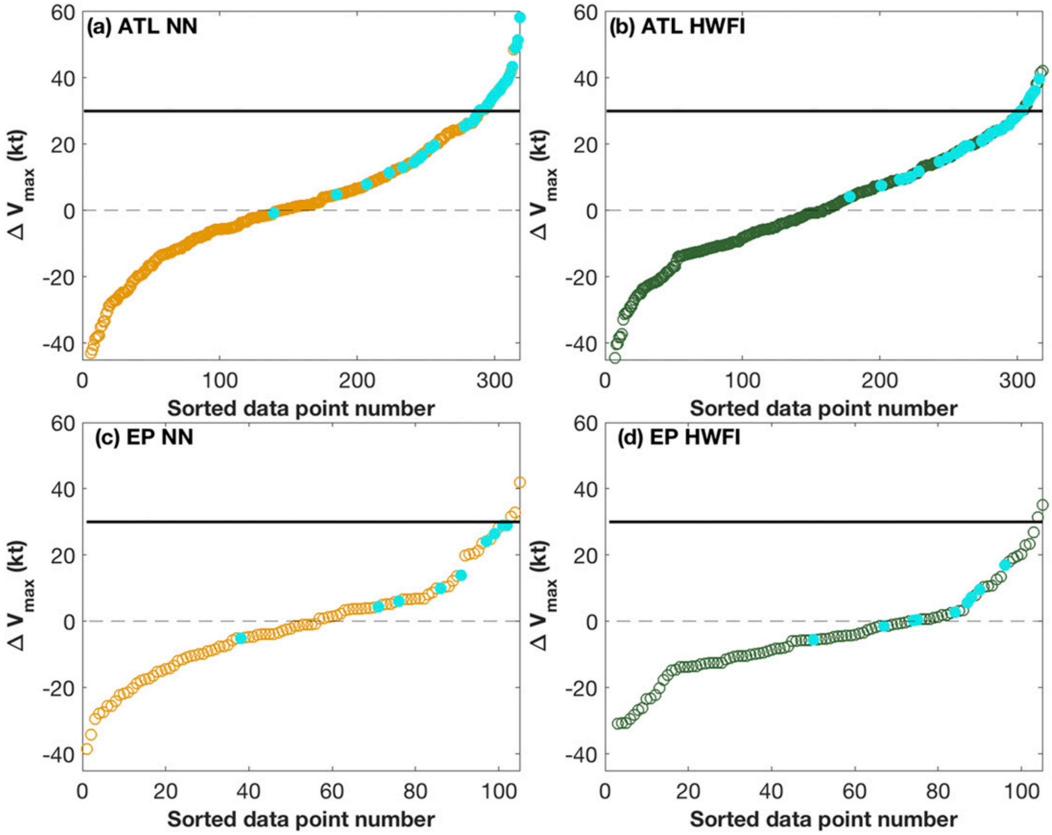

FIG. 5. Basin-specific distributions of 24-h intensity change: distributions of the 24-h intensity change $\Delta v_{\max}$ predicted by NN (orange) in (a) the Atlantic and (c) the eastern Pacific, and also by the HWFI (dark green) in (b) the Atlantic and (d) the eastern Pacific. Cyan highlighted circles indicate where the observed best-track 24-h intensity change was at least 30 kt. The dotted and solid horizontal lines demarcate the 0- and 30-kt values of intensity change.

there is some positive correlation between verifying RI observations and more positive intensity change predictions. In addition, the correlation is better for the NN. Thus, the NN appears closer to making correct RI predictions, but the NN intensity change predictions are still not quite high enough for the times in which RI is actually occurring. Now, the NN makes three false RI predictions versus the two false predictions by the HWFI. Because the sample size is small, this produces a significant degradation in the BSS for the NN despite its overall improved intensity change predictions for higher end intensity change values. Thus, the NN appears to improve upon the upper end 24-h intensity change predictions in both basins, despite the poor BSS in the eastern Pacific. Additionally, the more data rich Atlantic produces substantially better BSS than the operational SHIPS-RII prediction.

These results suggest that neural networks have a role to play in the detection of RI events, even in cases where their wind speed predictions are not better than other models' on average.

c. *Variable importance*

To gain insight into why the neural network was able to improve performance over the HWRF model, we ranked predictors according to the Garson variable importance score. This is a simple method that compares products of neural network weights to obtain a ''relative importance'' score for each predictor, as described in Goh (1995). In the case of a three layer network, variable importance of predictor i of p is calculated as

$$g_i = \frac{c_i}{\sum_{k=1}^{p} c_k}, \quad \text{with} \quad c_i = \sum_j |w_{j,i}^2 w_{1,j}^3|.$$

Figure 6 shows the top five variables in the neural network according to their Garson variable importance score. The HWFI VMAX was by far the most important predictor at all lead times, although its importance slowly decreased by lead time. The current operational VMAX value is scored as the second most influential

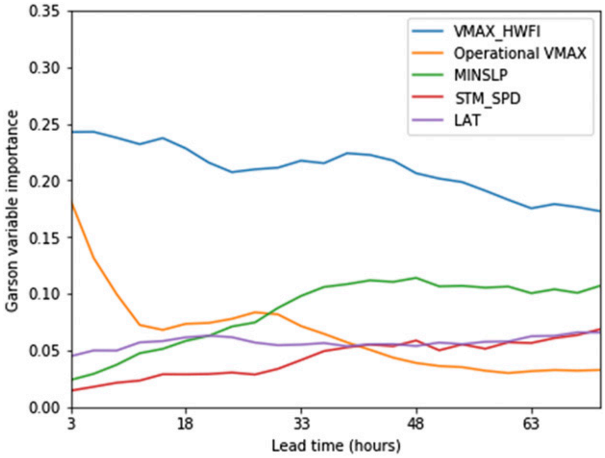

FIG. 6. Variable importance by lead time: importance of the overall top five most important predictors as determined by Garson variable importance measure, measured across different lead times.

predictor for short lead times. For lead times greater than 30 h, other HWRF-based predictors began to take on increasing prominence. This makes sense, as one would expect the current observed maximum wind speed to be a worse predictor as lead times increase. Minimum sea level pressure, another common metric of TC intensity, becomes increasingly important and is the second-most crucial variable by 30 h. At all lead times, the latitude (LAT) of the storm center maintains a relatively constant importance and varies between the third and fourth most important variable. The variable LAT determines the probability space of important environmental parameters and it also sets the Coriolis force. The next most important variable the storm speed, which can directly impact intensity by modifying the feedback of storm-induced sea surface temperature reduction (e.g., Mei et al. 2012) or also indicate environmental impacts on the TC (e.g., a fast-moving TC being impacted by a midlatitude trough). Storm motion also creates an azimuthal wavenumber-1 asymmetry in the boundary layer that may impact intensity (e.g., Shapiro 1983; Jorgensen 1984).

d. Case studies

To better highlight the real-time performance of the NN, we now consider some case studies of TCs in 2017 that experienced RI. With the strength of the NN residing in RI prediction, we focus now on 24-h forecasts of intensity. Figure 7 shows 24-h forecasts of intensity for eastern Pacific Kenneth, and Atlantic Hurricanes Harvey, Lee, and Maria from the NHC, HWFI, and the NN. Also plotted are the verifying best track observations of intensity. This figure also highlights periods of RI over the previous 24 h for the observations and for the 24-h intensity predictions that qualify as RI.

In terms of 24-h forecasts of intensity, each storm indicates a mixture of successes and failures for the NHC, HWFI, and NN 24-h intensity forecasts. There is some tendency for all 24-h predictions to underestimate the initial intensification period and for the predictions to be too slow in weakening the storm. The case studies also provide some insight into the RI prediction of each of the forecasts. In the case of Hurricane Kenneth (Fig. 7a), none of the models perform particularly well at predicting its period of RI on 19–21 August and actually produce some false alarms for the 24-h period ending at 1800 UTC 19 August (HWFI) and 0000 UTC 20 August (HWFI and the NN). While the NN failed to achieve the maximum intensity that was observed in Kenneth, its peak intensity was slightly better timed than the HWFI and NHC. Hurricane Harvey (Fig. 7b) yielded some of the best RI predictions for the NN, where it correctly captured 6 of 7 24-h periods of RI compared to 3 and 2 correct RI predictions by the NHC and from HWFI, respectively. Both the HWFI and NN performed fairly well for the RI period in Lee (Fig. 7c). Finally, the NN performed respectably in RI prediction for Maria, having one false alarm and missing RI at 3 of the 7 times of verifying RI (Fig. 7d). Overall, the case studies are consistent with the more general Gilbert and Brier skill score results described earlier.

5. Discussion and conclusions

In conclusion, a feed forward neural network has been developed for the prediction of tropical cyclone (TC) intensity. This deep learning model incorporated both the real-time operational estimate of the current intensity and predictors derived from Hurricane Weather Research and Forecasting (HWRF; 2017 version) model forecasts. The model was developed with operational constraints in mind, where numerical weather prediction (NWP) model output produced at synoptic times is typically not yet available at the same synoptic times that the operational centers issue TC intensity forecasts. Therefore, 6-h-old HWRF data were used in making a forecast as would be done in an operational environment. Best track intensity data were used for observational verification. The forecast training data were from 2014–16 HWRF reforecasts and covered a wide variety of TCs from both the eastern Pacific and Atlantic Ocean basins. Tenfold cross validation showed that the neural network outperforms the HWRF in terms of mean absolute error (MAE) at all lead times and was increasingly better than the observation-adjusted HWRF (HWFI) with increasing lead time. Out-of-sample testing on real-time data from 2017 showed the neural network produces lower MAE than the HWRF in the first 15 h of a forecast but performed similarly at later lead times. In this

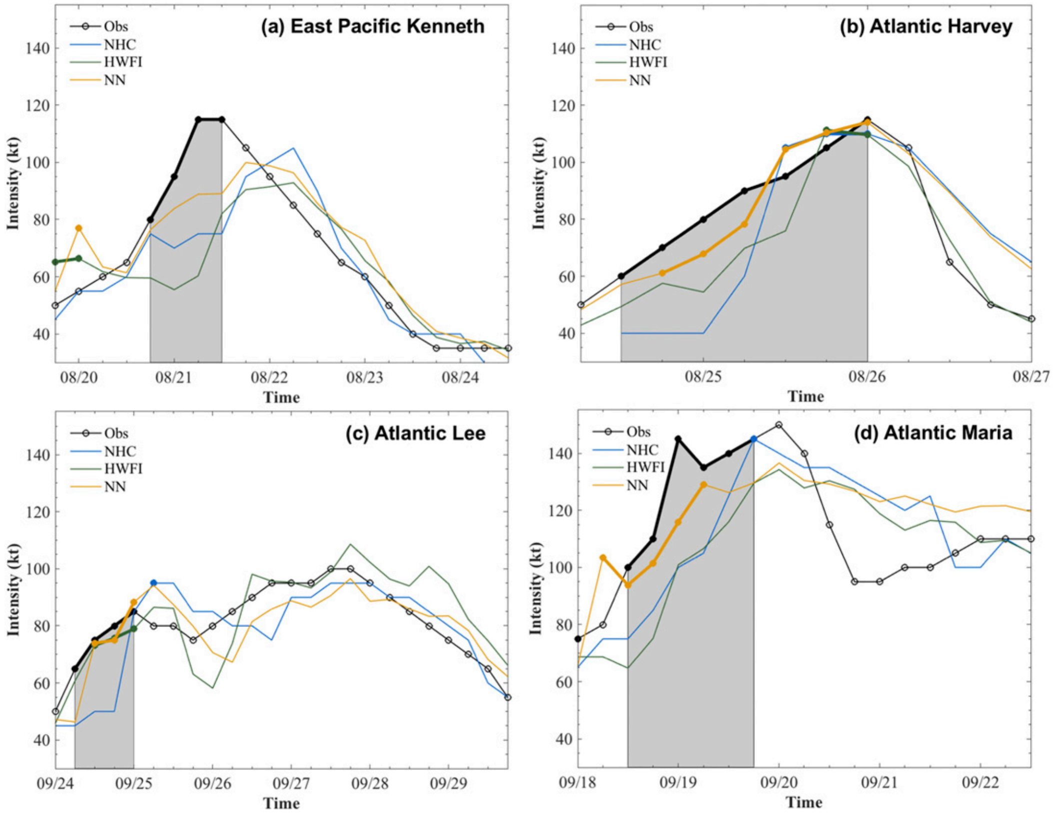

FIG. 7. Case studies for out-of-sample storms: best track intensity (black) and the predicted intensities as forecast 24 h earlier from the NHC (blue), HWFI (green), and NN (orange) for four different 2017 TCs that experienced RI: (a) eastern Pacific Kenneth, (b) Atlantic Harvey, (c) Atlantic Lee, and (d) Atlantic Maria. The bold dots and bold segments indicate times in which RI had occurred over the previous 24 h in the observations and predictions. The shaded gray regions highlight the times where the verifying best track intensity indicated RI in the previous 24 h.

out-of-sample test, the HWFI actually performed even better than the neural network at earlier lead times.

The most notable strength of the neural network is in its ability to predict 24-h rapid intensification (RI) in the out-of-sample testing period from year 2017. Gilbert skill scores were 0.35 for the neural network, while they were 0.10 and 0.09 for the HWFI and official National Hurricane Center forecasts, respectively. In comparison with the operational SHIPS-RII probabilistic model, the neural network produces superior skill in the combined Atlantic and eastern Pacific basins for the out-of-sample 2017 data. The neural network particularly excelled in the busy 2017 Atlantic hurricane season, which featured numerous RI events between several major hurricanes.

Overall, the feed forward neural network for intensity prediction is a computationally inexpensive postprocessing method that can be applied to any operational NWP TC model. The neural network excels in the prediction of RI. Reliable RI prediction has been a particularly challenging prediction problem. Hence, any postprocessing scheme that can significantly enhance the predictive skill of RI over that produced by its baseline NWP model would be a great asset to operational forecast centers. While the neural network is computationally efficient, it is important to note that there is a significant upfront cost in that a sufficiently large reforecast dataset must be generated in order to produce a robust and reliable postprocessing system. Fortunately, each time a model is updated and improved, it is not uncommon for operational forecast centers to generate reforecasts of past events. The encouraging results from postprocessing methods, particularly in the advent of machine learning methods for NWP, motivate the continued practice of generating large samples of reforecasts.

Acknowledgments. We thank Dr. John Knaff and two anonymous reviewers for their insightful comments that helped improve this manuscript. We also acknowledge the support of the NOAA Hurricane Forecast Improvement Program Grant NA16NWS4680027 in supporting this research. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

REFERENCES

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

Arlot, S., and Z. Celisse, 2010: A survey of cross-validation procedures for model selection. *Stat. Surv.*, **4**, 40–79, https://doi.org/10.1214/09-SS054.

Bhatia, K. T., and D. S. Nolan, 2017: Improving tropical cyclone intensity forecasts with PRIME. *Wea. Forecasting*, **32**, 1353–1377, https://doi.org/10.1175/WAF-D-17-0009.1.

Biswas, M., L. Carson, K. Newman, L. Bernardet, E. Kalina, E. Grell, and J. Frimel, 2017: Community HWRF users' guide v3.9a. NOAA Tech. Memo. OAR GSD-51, 163 pp., https://doi.org/10.7289/V5/TM-OAR-GSD-51.

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

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, https://doi.org/10.1175/BAMS-D-12-00240.1.

Déqué, M., 2007: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global Planet. Change*, **57**, 16–26, https://doi.org/10.1016/j.gloplacha.2006.11.030.

Emanuel, K., 1986: An air–sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. *J. Atmos. Sci.*, **43**, 585–604, https://doi.org/10.1175/1520-0469(1986)043<0585:AASITF>2.0.CO;2.

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

Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir, 2015: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.*, **143**, 955–971, https://doi.org/10.1175/MWR-D-14-00210.1.

Gagne, D., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

Ghosh, T., and T. N. Krishnamurti, 2018: Improvements in hurricane intensity forecasts from a multimodel superensemble utilizing a generalized neural network technique. *Wea. Forecasting*, **33**, 873–885, https://doi.org/10.1175/WAF-D-17-0006.1.

Goerss, J. S., and C. R. Sampson, 2014: Prediction of consensus tropical cyclone intensity forecast error. *Wea. Forecasting*, **29**, 750–762, https://doi.org/10.1175/WAF-D-13-00058.1.

——, ——, and J. M. Gross, 2004: A history of western North Pacific tropical cyclone track forecast skill. *Wea. Forecasting*, **19**, 633–638, https://doi.org/10.1175/1520-0434(2004)019<0633:AHOWNP>2.0.CO;2.

Goh, A. T. C., 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.

Jorgensen, D. P., 1984: Mesoscale and convective-scale characteristics of mature hurricanes. Part II: Inner core structure of Hurricane Allen (1980). *J. Atmos. Sci.*, **41**, 1287–1311, https://doi.org/10.1175/1520-0469(1984)041<1287:MACSCO>2.0.CO;2.

Judt, F., S. S. Chen, and J. Berner, 2016: Predictability of tropical cyclone intensity: Scale-dependent forecast error growth in high-resolution stochastic kinetic-energy backscatter ensembles. *Quart. J. Roy. Meteor. Soc.*, **142**, 43–57, https://doi.org/10.1002/qj.2626.

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

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

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

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved skills for weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, https://doi.org/10.1126/science.285.5433.1548.

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.

——, ——, and J. Beven, 2015: The revised Atlantic hurricane database (HURDAT2). NOAA/NHC, 6 pp., https://www.nhc.noaa.gov/data/hurdat/hurdat2-format-atlantic.pdf.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Mei, W., C. Pasquero, and F. Primeau, 2012: The effect of translation speed upon the intensity of tropical cyclones over the tropical ocean. *Geophys. Res. Lett.*, **39**, L07801, https://doi.org/10.1029/2011GL050765.

Miyamoto, Y., and T. Takemi, 2013: A transition mechanism for the spontaneous axisymmetric intensification of tropical cyclones. *J. Atmos. Sci.*, **70**, 112–129, https://doi.org/10.1175/JAS-D-11-0285.1.

Nolan, D. S., J.A. Zhang, and E.W. Uhlhorn, 2014: On the limits of estimating the maximum wind speeds in hurricanes. *Mon. Wea. Rev.*, **142**, 2814–2837, https://doi.org/10.1175/MWR-D-13-00337.1.

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

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

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

Rogers, R., P. Reasor, and S. Lorsolo, 2013: Airborne Doppler observations of the inner-core structural differences between intensifying and steady-state tropical cyclones. *Mon. Wea. Rev.*, **141**, 2970–2991, https://doi.org/10.1175/MWR-D-12-00357.1.

Sampson, C. R., J. S. Goerss, and H. C. Weber, 2006: Operational performance of a new barotropic model (WBAR) in the western North Pacific basin. *Wea. Forecasting*, **21**, 656–662, https://doi.org/10.1175/WAF939.1.

——, 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.

Schubert, W. H., and J. J. Hack, 1982: Inertial stability and tropical cyclone development. *J. Atmos. Sci.*, **39**, 1687–1697, https://doi.org/10.1175/1520-0469(1982)039<1687:ISATCD>2.0.CO;2.

Shapiro, L. J., 1983: The asymmetric boundary layer flow under a translating hurricane. *J. Atmos. Sci.*, **40**, 1984–1998, https://doi.org/10.1175/1520-0469(1983)040<1984:TABLFU>2.0.CO;2.

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.

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.

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Williford, C. E., T. N. Krishnamurti, R. C. Torres, S. Cocke, Z. Christidis, and T. S. Vijaya Kumar, 2003: Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Wea. Rev.*, **131**, 1878–1894, https://doi.org/10.1175//2571.1.

Zhang, Y., Z. Meng, F. Zhang, and Y. Weng, 2014: Predictability of tropical cyclone intensity evaluated through 5-yr forecasts with a convection-permitting regional-scale model in the Atlantic basin. *Wea. Forecasting*, **29**, 1003–1022, https://doi.org/10.1175/WAF-D-13-00085.1.