Nonlinear Wave Ensemble Averaging in the Gulf of Mexico Using Neural Networks

RICARDO MARTINS CAMPOS

Department of Atmospheric and Oceanic Science, University of Maryland,
College Park, College Park, Maryland

VLADIMIR KRASNOPOLSKY

NOAA/NCEP/EMC/Center for Weather and Climate Prediction, College Park, Maryland

JOSE-HENRIQUE G. M. ALVES

Systems Research Group, and NOAA/NCEP/EMC/Center for
Weather and Climate Prediction, College Park, Maryland

STEPHEN G. PENNY

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, and
NOAA/NCEP/EMC/Center for Weather and Climate Prediction, College Park, Maryland

(Manuscript received 15 June 2018, in final form 8 November 2018)

ABSTRACT

Artificial neural networks (ANNs) applied to nonlinear wave ensemble averaging are developed and studied for Gulf of Mexico simulations. It is an approach that expands the conservative arithmetic ensemble mean (EM) from the NCEP Global Wave Ensemble Forecast System (GWES) to a nonlinear mapping that better captures the differences among the ensemble members and reduces the systematic and scatter errors of the forecasts. The ANNs have the 20 members of the GWES as input, and outputs are trained using observations from six buoys. The variables selected for the study are the 10-m wind speed (U10), significant wave height (Hs), and peak period (Tp) for the year of 2016. ANNs were built with one hidden layer using a hyperbolic tangent basis function. Several architectures with 12 different combinations of neurons, eight different filtering windows (time domain), and 100 seeds for the random initialization were studied and constructed for specific forecast days from 0 to 10. The results show that a small number of neurons are sufficient to reduce the bias, while 35–50 neurons produce the greatest reduction in both the scatter and systematic errors. The main advantage of the methodology using ANNs is not on short-range forecasts but at longer forecast ranges beyond 4 days. The nonlinear ensemble averaging using ANNs was able to improve the correlation coefficient on forecast day 10 from 0.39 to 0.61 for U10, from 0.50 to 0.76 for Hs, and from 0.38 to 0.63 for Tp, representing a gain of five forecast days when compared to the EM currently implemented.

1. Introduction

The U.S. National Centers for Environmental Prediction (NCEP) have produced atmospheric forecasts using ensembles since 1992 and wave ensembles since 2005. Kalnay (2003) describes the two main advantages of using ensemble forecasts: the ensemble members tend to smooth out uncertain components, which lead to better skill than single deterministic forecasts; and the spread of the ensemble members provides an estimation of the uncertainty. The mean of the ensemble members is typically more accurate than any deterministic forecast after the first few forecast days, as presented by Zhou et al. (2017) for the NCEP Global Ensemble Forecast System (GEFS). As the wave modeling is strongly dependent on the quality of surface winds (Cavaleri et al. 2007), the benefit of the atmospheric ensemble is transferred to the NCEP Global Wave Ensemble Forecast System (GWES; Chen 2006), which is

Corresponding author: Ricardo Martins Campos, riwave@gmail.com

DOI: 10.1175/JTECH-D-18-0099.1

validated and discussed by Cao et al. (2007) and Alves et al. (2013). Despite the improvement of the operational wave ensemble compared to the deterministic run, the GWES still suffers from shortcomings that limit its skill, especially associated with systematic errors that vary with forecast time and location. The GWES currently uses the conservative arithmetic ensemble mean (EM), as shown in Eq. (1):

$$\mathrm{EM} = \frac{1}{n}\sum_{i=1}^{n} p_i\,, \tag{1}$$

where n is the number of ensemble members and p_i is the state of the ith ensemble member. The major advantage of the conservative approach is that EM can be calculated always without any additional information. However, this measure of central tendency assumes a linear relationship between the EM and ensemble members. Because this relationship may be strongly nonlinear, as stated by Krasnopolsky and Lin (2012), particularly at long lead times, a nonlinear ensemble average calculated using feedforward neural networks is proposed, trained with wind and wave observations at various forecast lead times. The experiments are conducted using a spatial approach for wave simulations in the Gulf of Mexico, a region with intense maritime activity, offshore industry, and coastal vulnerabilities. The prognostic variables selected for analysis are wind speed (U10), significant wave height (Hs), and peak period (Tp).

The use of artificial neural networks (ANNs) for environmental analyses and forecasts has rapidly increased over recent years. Sánchez et al. (2018) presented a mathematical model that uses ANNs for the assessment of wave energy potential. Berbić et al. (2017) applied ANNs and support vector machines for short forecasts of significant wave height. Dixit and Londhe (2016) developed a neuro wavelet technique, combining discrete wavelet transform and ANNs, to explore the predictability of extreme events for five major hurricanes at four locations in the Gulf of Mexico. Deo et al. (2001), Deo and Sridhar Naidu (1998), and Mandal and Prabaharan (2006) developed ANN systems to predict significant wave heights in India, and Tsai et al. (2002) in Taiwan. Krasnopolsky and Lin (2012) developed ANN-based models to produce a nonlinear ANN ensemble forecast for precipitation, and Lo et al. (2015) developed a calibration method using ANN for cyclonic precipitation forecast models. A complete description of the theory of multilayer perceptron neural networks, the basis of our present study, was developed by Haykin (1999), which has been successfully applied to many practical applications in Earth system sciences by Krasnopolsky (2013).

Despite the popularization of ANNs, most of the forecast studies have been aimed at applying ANN models directly to predict wave heights and surface winds as target variables. Campos and Guedes Soares (2016) proposed an alternative methodology using a hybrid model, joining the numerical wave model with ANNs. The numerical model predicts the wave heights while the target of the ANN is to predict the residue (i.e., the difference between the measurement and the model), which is recombined to provide an accurate estimation of wave heights at the Brazilian coast. The final bias was reduced from 0.13 to 0.06 m and the scatter index from 0.12 to 0.03. However, Campos and Guedes Soares (2016) did not consider the ensemble forecasts that significantly reduce the scatter errors when compared to deterministic runs.

In view of the above, we conceived a new approach linking the benefits of ensemble forecast systems with ANNs, which are able to approximate any continuous function. We start by describing our methods in section 2, the observations and preprocessing in section 3, and the neural network models for the Gulf of Mexico in section 4. The results are presented in section 5, while concluding remarks are made in section 6.

2. Multilayer perceptron neural networks

The nonlinear ensemble averaging in the present study is entirely based on the multilayer perceptron (MLP) model (Rumelhart et al. 1986), which is a feedforward artificial neural network that uses supervised learning. Most problems can be mapped using two layers of nodes plus the input layer. Equation (2) describes the model with a hyperbolic tangent as the activation function (Krasnopolsky 2013):

$$\begin{aligned}\mathrm{NN}(x_1, x_2, \ldots, x_n; a, b) &= y_q = a_{q0} \\ &+ \sum_{j=1}^{k} a_{qj}\tanh\left(b_{j0} + \sum_{i=1}^{n} b_{ji}x_i\right); \quad q = 1, 2, \ldots, m,\end{aligned} \tag{2}$$

where x_i is the input; y_q is the output; a and b are the ANN weights; n and m are the number of inputs and outputs, respectively; and k is the number of nonlinear activation functions. Campos et al. (2017) also describes the model, where the first summation on the right-hand side of Eq. (2) represents a linear combination of hyperbolic tangents, while the second summation is the weighted sum of input variables. Haykin (1999) explains the backpropagation training using gradient descent, a simple and powerful optimization able to map highly nonlinear functions. It is based on the idea that searching for a minimum of the error function can be performed

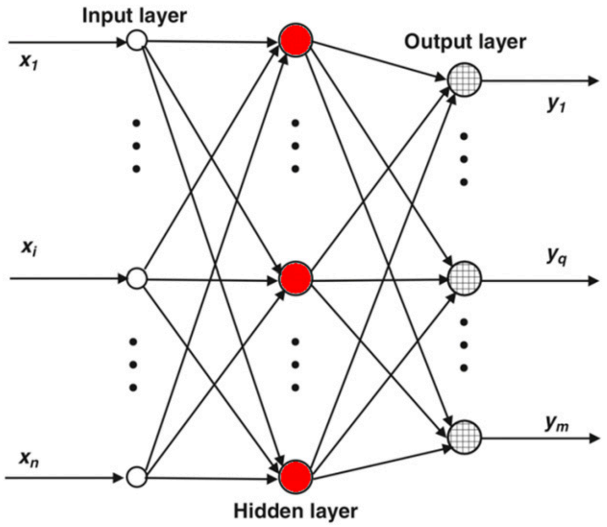

FIG. 1. Illustration of an MLP-NN model containing one hidden layer (red) with sigmoid functions applied to linear combinations of inputs.

step by step iteratively and that at each step there is an increment or decrement of the weights in such a way as to decrease the error function (Krasnopolsky 2013). A supervised learning method, such as backpropagation, relies on a set of historical values considered as target variables (measurements) that must be properly preprocessed—described in the next section. Figure 1 illustrates the MLP neural network model, where in red are the¨neurons¨with sigmoid functions, usually hyperbolic tangent, that are applied to the weighted linear combination of the normalized input variables described by Eq. (2).

Although MLP neural network (MLP-NN) modeling is simple to understand and to code, there are several aspects that must be carefully investigated before and while training and running ANNs, which is described by Krasnopolsky (2013, 2014). Some steps include (i) choice of relevant input variables that contribute to the nonlinear mapping; (ii) proper normalization of input variables; (iii) analysis of optimum ANN architecture and complexity, including the number of layers and neurons; (iv) learning rate and gradient descent scheme; (v) optimum number of iterations during the training process; and (vi) careful assessment of results, dividing the datasets into training, test, and validation sets.

Krasnopolsky (2014) states that the complexity of the ANN should be carefully controlled and kept to the minimum level sufficient for the desired accuracy of the approximation to avoid overfitting. Furthermore, the training set must represent the mapping for the ANN, with a sufficient sample size of properly distributed data points that adequately resolve the functional complexity of the target mapping. For environmental variables, at least one year is necessary in order to properly cover distinct conditions and a full seasonal cycle. Another source of error that must be investigated is the amount of noise in the target variables. Krasnopolsky (2013) draws attention that even a small amount of noise in the data may lead to significant errors in the ANN emulations. An additional discussion of prediction of a time series as a mapping using ANNs can be found in Weigend and Gershenfeld (1994).

A preliminary study, using ANNs to improve GWES at single-buoy locations is presented by Campos et al. (2017), which summarizes the first steps of the current development. Two pairs of National Data Buoy Center (NDBC) buoys were selected for the ANN training and validation: buoys 41004 and 41013 in the Atlantic Ocean and buoys 46047 and 46028 in the Pacific Ocean. The input variables of the MLP-NN consist of the 21 GWES ensemble members (20 plus the control member) associated with the variables U10, Hs, and Tp, as well as the sine and cosine of time (Julian days) to properly include the time and seasonality information. Therefore, a total of 65 input variables x compose the n inputs for the ANN model. The outputs y consist of three variables only (Hs, Tp, and U10; $m = 3$) from the NDBC buoys, targeted by the model. Each ANN addresses one forecast time, with the focus of Campos et al. (2017) on the fifth day, which is approximately the time when ensemble forecasts start to have better performance than deterministic forecasts, according to Alves et al. (2013).

All variables were normalized to the interval between -1 and 1 to run the ANN, and denormalized after training, for the test and validation. It was found that occasional extreme conditions generate sharp peaks that are not properly optimized by the ANN; so, following the suggestion of Krasnopolsky and Lin (2012), a log function was additionally applied to Hs, which leads to better results as a result of a more homogeneous distribution of values, illustrated by Fig. 2.

A cross-validation scheme was implemented, where two-thirds of the data are selected for the training set and one-third for the test set. After the optimization of weights, the ANN models were applied to a nearby buoy, where the data were not included in the ANN training. Several ANNs were tested, changing the number of neurons from 1 to 50 as well as later experiments with more layers. Campos et al. (2017) found that the best ANN architecture has one hidden layer of 11 neurons. It provided an improvement on

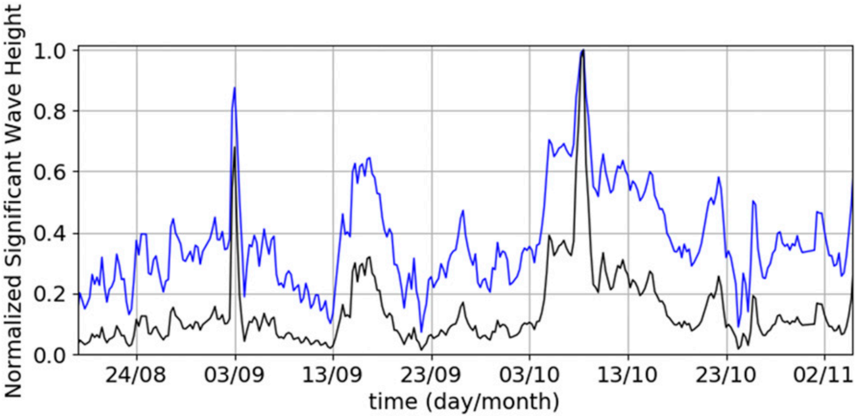

FIG. 2. Example of normalization of time series of Hs applying a natural log function (blue) compared to the original normalized Hs (black). This corresponds to NDBC buoy 41004 in 2016. The highest peak is associated with Hurricane Matthew.

the 5-day forecast of 64% in the bias, 29% in the root-mean-square error (RMSE) and scatter index, and 11% in the correlation coefficient. The successful results from Campos et al. (2017) applied to single-buoy locations supported the current expansion of the methodology to a spatial approach in the Gulf of Mexico.

3. Input data and observations

Following the selection of data in the previous section, the datasets that were used for development and validation of ANNs in the Gulf of Mexico consist of one year (2016) of GWES historical forecasts (input of the ANNs) and NDBC observations for the ANN training. The variables selected are, again, U10, Hs, and Tp. The GWES is run with four cycles per day, and wind and wave models are run at a resolution of 0.5° in space and 3 h in time, extending to 10 days. A total of 20 ensemble members plus a control member compose the GWES, an implementation of the WAVEWATCH III model (Tolman 2016) forced by winds from the GEFS. The interval of GWES and observations selected corresponds to the year of 2016, which is a period without major GEFS or GWES model upgrades, and which composes a complete seasonal cycle. A complete assessment and description of GEFS and GWES can be found at Zhou et al. (2017) and Alves et al. (2013), respectively.

Six NDBC buoys in the Gulf of Mexico moored in deep waters have been selected for the ANNs modeling—42001, 42002, 42003, 42039, 42055, 42360—with locations plotted in Fig. 3. The wind data from buoys were converted to the 10-m level using the wind profile power law (Det Norske Veritas 2007) with a friction coefficient of 0.10, as suggested by Hsu et al. (1994), for lakes and oceans.

A quick assessment of GWES is plotted in Fig. 4, where observations from the buoys are bundled to form a single dataset with a length of 7913, which are paired with the ensemble results. The top row of plots (Figs. 4a–c) compares the ensemble members with the deterministic run and the arithmetic EM. Figure 4a shows the success of GEFS and the methodology described by Zhou et al. (2017), where the RMSE of the ensemble mean of U10 performs significantly better than the deterministic run after the fourth forecast day. This difference reaches nearly 30% of

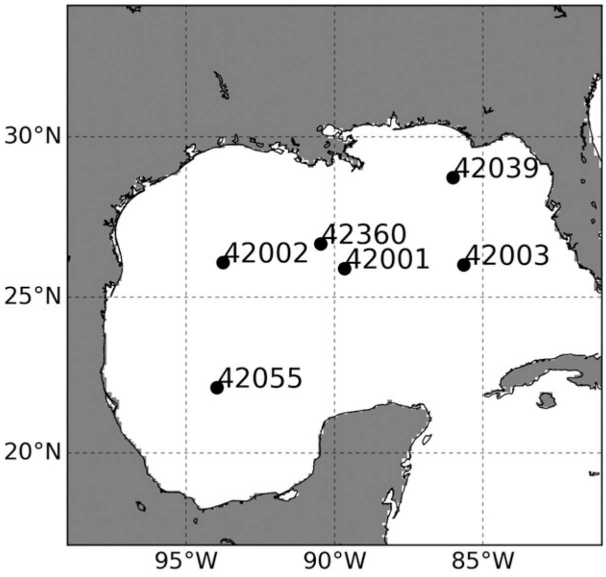

FIG. 3. Location of the six NDBC metocean buoys moored in deep waters.

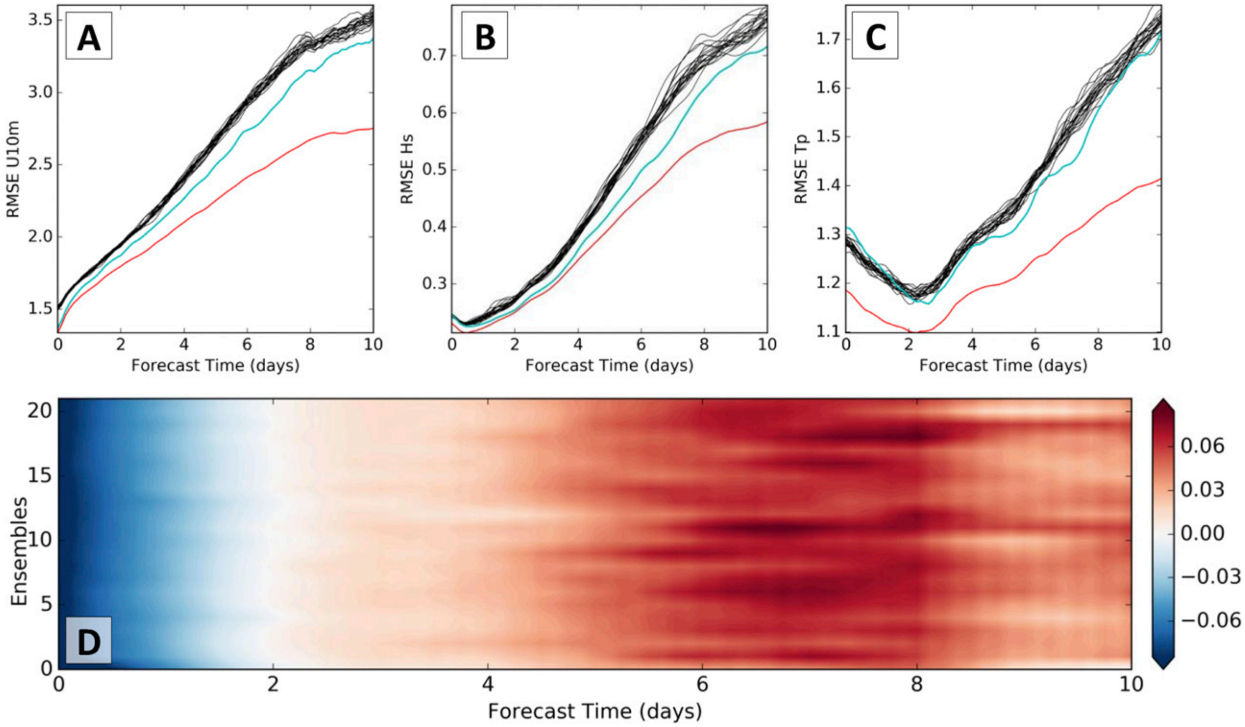

FIG. 4. Assessment of GWES in the Gulf of Mexico using six NDBC buoys in deep water. RMSE of (a) U10 (m s^{-1}), (b) Hs (m), and (c) Tp (s). Black curves show the ensemble members, cyan is the deterministic run, and red is the arithmetic mean of the ensemble members. (d) Normalized bias of Hs (m) illustrating the 20 ensemble members (*y* axis) through the 10 days of forecast (*x* axis), where reds indicate overestimation and blues represent underestimation of GWES.

the improvement of U10 at the eighth day. The higher surface wind skill is extended to the wave heights, where the ensemble means of Hs and Tp are also better than the deterministic runs. However, Fig. 4d presents the normalized bias of the ensemble members of Hs, where a heterogeneous distribution of the systematic error across the forecast days and ensemble members can be observed. The nowcast and first forecast day are underestimated, while after the third forecast day GWES tends to overestimate the measurements in the Gulf of Mexico. The ensemble members agree with each other during the first days, and they diverge at longer forecast ranges, increasing the spread confirmed by Figs. 4a–4c. Therefore, despite the improvements of GWES (Figs. 4a–c), a bias correction postprocessing algorithm should be implemented in order to reduce the severe systematic errors illustrated by Fig. 4d. The complexity of the bias distribution requires a robust and powerful approximator able to map the features and degrees of freedom of the GWES error signal, which will be handled by the ANNs in the next section.

The assessment of ANN results in the next section is analyzed with much more detail. The evaluation is based on the error metrics suggested by Mentaschi et al. (2013), who discussed the advantages of interpreting the systematic and scatter components of the error separately. Therefore, a total of seven metrics [Eqs. (3)–(9)] were introduced to evaluate the results, following the description of Mentaschi et al. (2013), where *x* is the buoy data and *y* is the forecast, and the overbar indicates the arithmetic mean:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i), \tag{3}$$

$$\text{NBias} = \frac{\sum_{i=1}^{n}(y_i - x_i)}{\sum_{i=1}^{n}x_i}, \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}, \tag{5}$$

$$\text{NRMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}x_i^2}}, \tag{6}$$

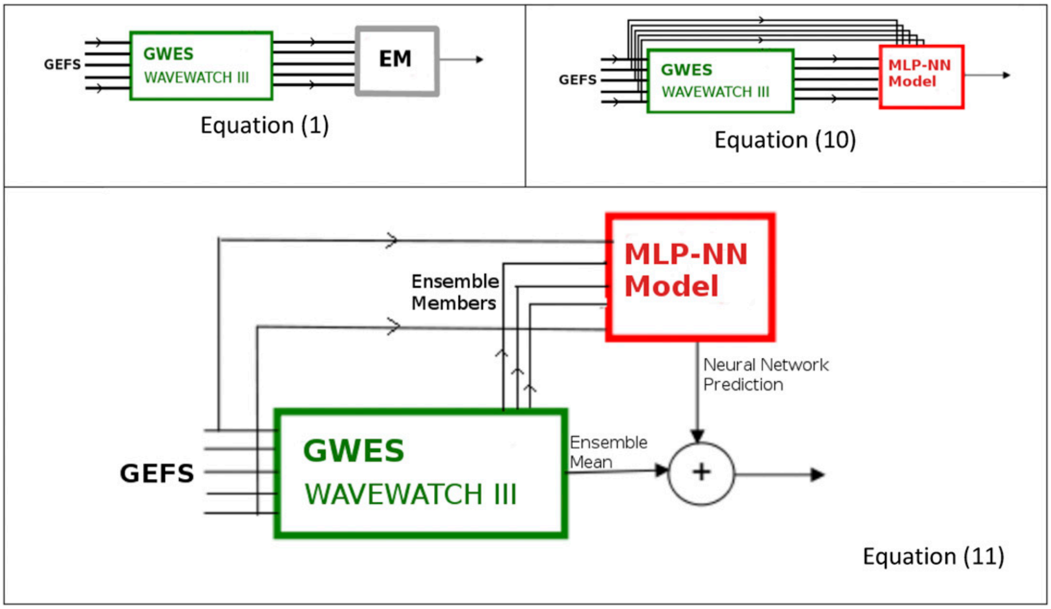

FIG. 5. Illustration of Eqs. (1), (10), and (11).

$$\mathrm{SCrmse} = \sqrt{\frac{\sum_{i=1}^{n}\left[(y_i-\overline{y})-(x_i-\overline{x})\right]^2}{n}} = \sqrt{\mathrm{RMSE}^2-\mathrm{Bias}^2}\,, \tag{7}$$

$$\mathrm{SI} = \frac{\sum_{i=1}^{n}\left[(y_i-\overline{y})-(x_i-\overline{x})\right]^2}{\sum_{i=1}^{n}x_i^2}\,, \quad \text{and} \tag{8}$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2\sum_{i=1}^{n}(x_i-\overline{x})^2}}\,. \tag{9}$$

The metric units of Bias, RMSE, and SCrmse are the same as the selected variables—that is, U10 in meters per second, Hs in meters, and Tp in seconds—while NBias, NRMSE, SI, and CC are nondimensional. It is important to note that these metrics are used for the assessment of the results, comparing the EM with the nonlinear ensemble averaging. The ANN algorithm minimizes only the square of the error in the training set.

4. Neural network models in the Gulf of Mexico

a. *ANN architecture and training*

The initial architecture of the ANNs is shown by Eq. (10), where the MLP-NN directly calculates the nonlinear ensemble average using Hs and Tp from GWES and U10 from GEFS as input:

$$\mathrm{NEM} = \mathrm{NN}(p_1, p_2, \ldots, p_n)\,. \tag{10}$$

The drawback of such an approach is that it uses the ANN model to calculate both linear and nonlinear components of the signal, when the benefits of ANN models are primarily found when applied to nonlinear problems (Krasnopolsky 2013).

To solve this problem, Eq. (11) brought a simple solution by taking the arithmetic mean and applying ANN to simulate the residue (difference between the target value and the EM):

$$\mathrm{NEM} = \mathrm{EM} + \mathrm{NN}_r(p_1, p_2, \ldots, p_n)\,. \tag{11}$$

Hence, the ANN model is dedicated exclusively to the nonlinear component, preserving the results of the EM on the linear part. Such an approach builds a more robust model that provides reliable ensemble averages at different metocean conditions and sea severities, as concluded by Campos et al. (2017). An illustration of Eqs. (1), (10), and (11) is provided by Fig. 5, where it is possible to compare the different ANN strategies described. The occasional problem associated with Eq. (11) is the excess of noise in the residue, which can increase the risk of overfitting (Krasnopolsky 2014). Figure 6 shows an example of the residue of Hs, where the level of noise and high-frequency changes can be visualized. The problem of noise and overfitting can be reduced by properly filtering the time series.

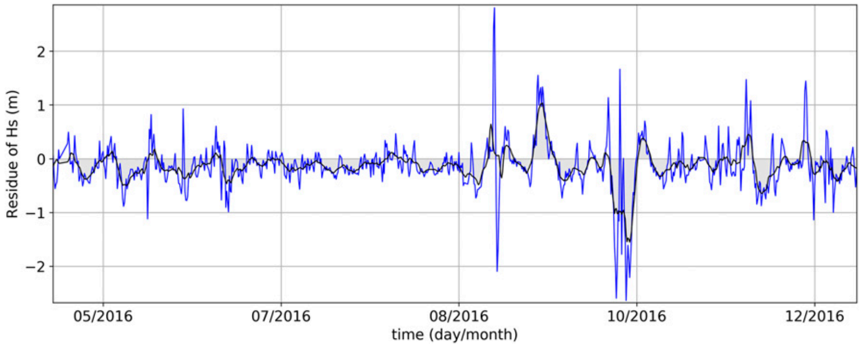

FIG. 6. Time series of the residue of Hs (blue) and the filtered signal using a moving average of 120 h (black).

Moving from a single spot analysis to a spatial approach in the Gulf of Mexico required a small modification in the ANN configuration. As ANN models converge to optimum weights and biases that cannot be directly interpolated or extrapolated in space, the strategy of introducing space in ANN models is to include the position (latitude and longitude) as a new degree of freedom (new ANN inputs). The ANN and tests were constructed for each forecast day, independently. This gives a total of 67 inputs, related to three variables (U10, Hs, and Tp) with 21 ensemble members, plus the sine and cosine of time [i.e., $\sin(2\pi t/T)$ and $\cos(2\pi t/T)$, respectively, where T is one year], and latitude and longitude. The three outputs are the residue of U10, Hs, and Tp.

b. Signal preprocessing

Figure 7 illustrates the randomness and spread of the residue against the buoy observations. The challenge of ANN models mapping the residue is equivalent to predicting the error of the EM of U10, Hs, and Tp using the ensemble members as input. Therefore, if there is any trend, pattern, or correlation between the error of the EM with the variables U10, Hs, and Tp itself, the ANN

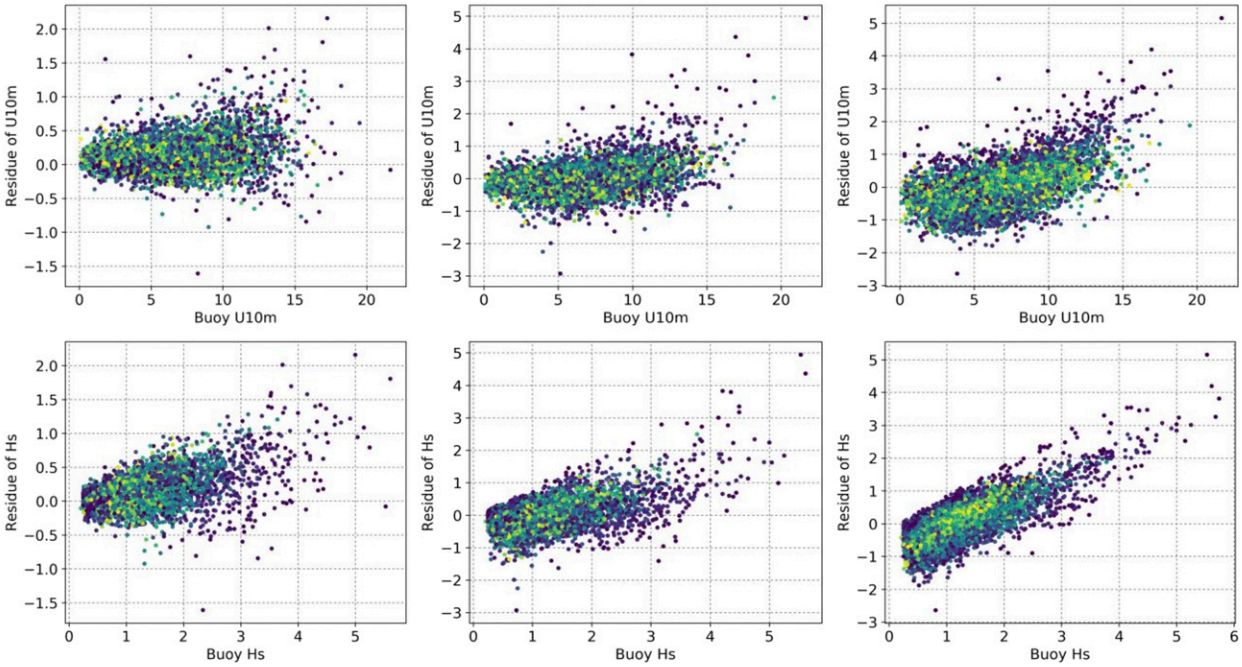

FIG. 7. Residue of (top) U10 ($\mathrm{m\,s^{-1}}$) and (bottom) Hs (m) related to the difference between EM and buoy measurements in the Gulf of Mexico. The y axis shows the residues, while the x axis shows the observations related to each of the 7913 inputs. (left) The nowcast (forecast day 0), (center) forecast day 5, and (right) forecast day 10.

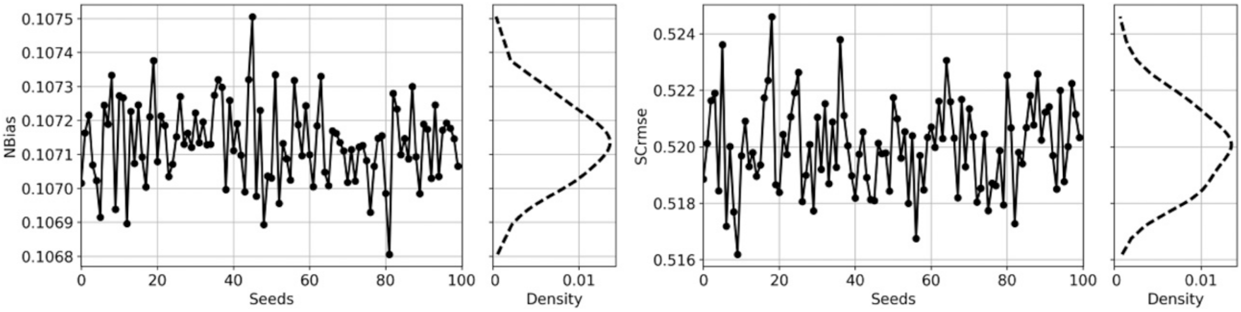

FIG. 8. NBias and SCrmse according to Eqs. (4) and (7), respectively, resulting from the ANN training tests on forecast day 10 in the Gulf of Mexico. Results involving different numbers of neurons and filtering windows were averaged to analyze the sensitivity to the initialization, using 100 different seeds. Each of the two sets shows (left) the main plot and (right) the empirical probability density function of the left panel, transposed, and sharing the same y axis.

can extract these dependencies from the data. Figure 7 shows a great spread at all plots, which is expected once they illustrate the difference between EM and the buoy measurements. However, there is a positive trend between the residue and the variables that suggests larger errors at higher values of U10 and Hs. This positive correlation is very small at day 0 plots, but it increases throughout the forecast range. Forecast day 10, for Hs, shows a strong linear relation between the residue and Hs. It gives an idea that the ANN will behave differently at distinct forecast instants and suggests better results for longer forecast ranges.

The spectra of the residues of U10, Hs, and Tp were also calculated for each buoy, showing three main peaks at 40, 70, and 200 h; that is, the errors of the EM are mainly concentrated at these three periods. The high level of noise in the signal of the residue, exemplified by Fig. 6, and the spectral analysis suggest that filtering techniques can significantly improve the ANN training. A simple low-pass moving average filter can remove the high-frequency oscillation (black curve of Fig. 6) mostly associated with random noise that is verified to compromise the ANN training. However, it is not known a priori the precise cutoff frequency that better optimizes the filtering of the residue signal, so several tests are performed with different filtering windows to investigate the level of smoothness on the validation metrics. After finding the optimum filtering strategy, the new filtered signal of the residue becomes the target value of the ANN, following Eq. (11).

c. *Sensitivity tests*

With the introduction of multiple buoy approaches, the complexity of the ANN and the training set increases compared to the single-location case. The performance in this case is highly dependent on the number of hidden neurons and the filtering of variables, which can be optimized by running multiple tests. Therefore, instead of working with one ANN model, a batch of several runs was constructed. Because of the partially random nature of the signal associated with the target variables (ANN outputs) and the MLP-NN training method, the initialization of weights also has some impact on the final skill of the model (Fig. 8). This means different seeds can lead to slightly distinct performances, so it was also considered in the ANN runs.

A total of 12 different combinations of neurons were tested, eight different filtering windows, and 100 seeds for the random initialization. It was constructed using separate ANNs for specific forecast lead times, from day 0 to day 10 forecasts. Therefore, taking the 11 forecast ranges selected, plus the tests with neurons, filtering, and initialization, a total of 105 600 ANNs were built and trained. ANNs with the number of hidden neurons $N = 2$, 5, 10, 15, 20, 25, 30, 35, 40, 50, 80, and 200, and filtering windows of magnitude 0, 24, 48, 96, 144, 192, 288, and 480 h were tested, using the moving average method.

For the ANN training, two-thirds of the records were selected for training and one-third for the test set, using a cross-validation scheme with three cycles, alternating the indices defined for training and testing. This approach was intended to ensure a fair evaluation of the training and to reduce the misinterpretation associated with overtraining of a specific set. The seven metrics from Eqs. (3)–(9) are calculated with the results reported in arrays with six dimensions for each variable (U10, Hs, Tp): forecast day; filtering window; number of neurons, seeds, set (EM, training, and testing); and error metrics. The results are analyzed for each variable, forecast day, and error metric.

Figure 8 shows the NBias and SCrmse using 100 different seeds for initialization, focusing on forecast day 10. The results indicate that the ANN is sensitive to the random initialization and show the spread associated with different ANN optimizations. It was observed that the scatter of error metrics is higher at longer forecasts.

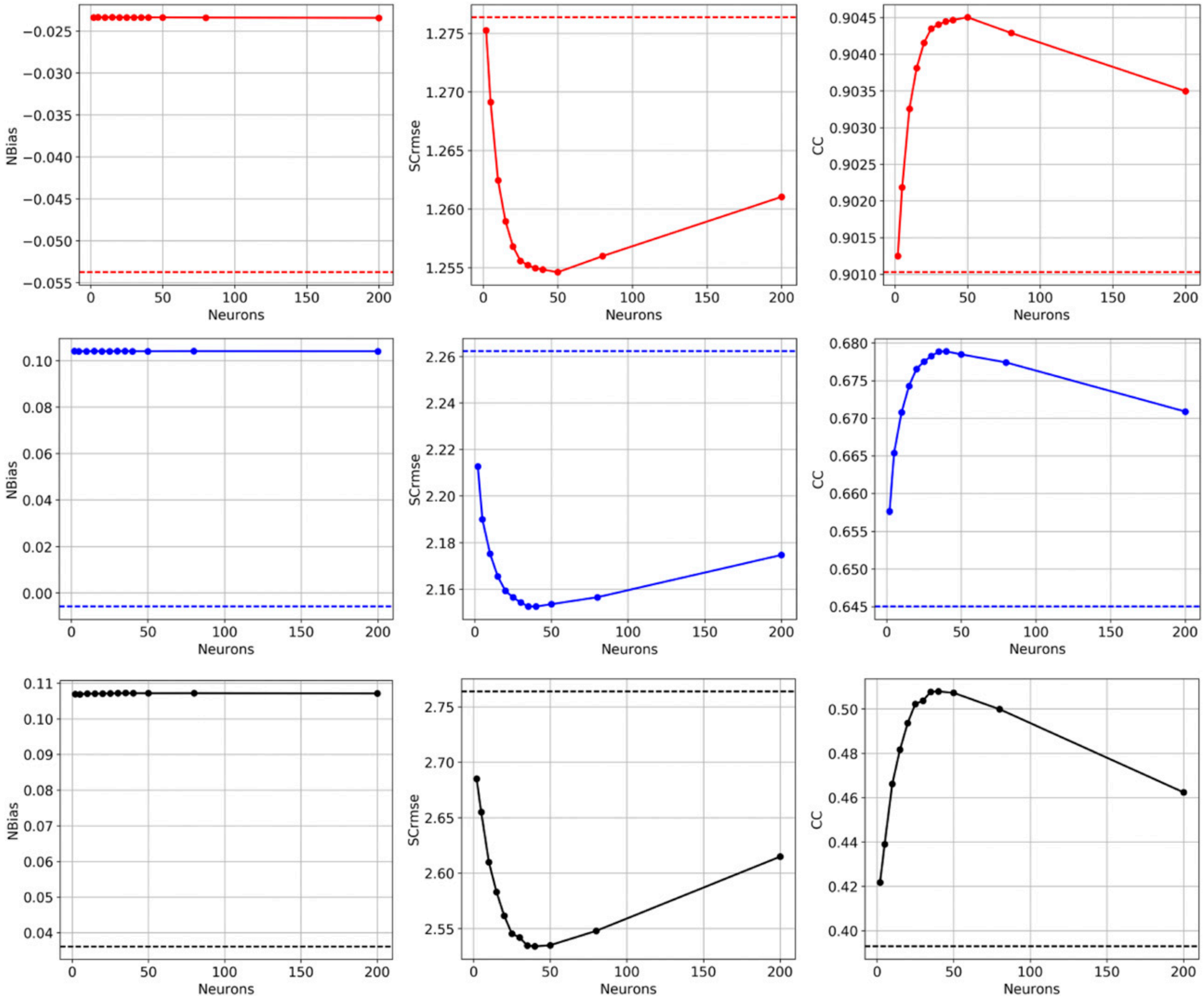

FIG. 9. NBias, SCrmse, and CC according to Eqs. (4), (7), and (9), respectively, resulting from the ANN training tests on forecast (top) day 0 (red), (middle) day 5 (blue), and (bottom) day 10 (black) for U10. Results involving different initializations and filtering windows were averaged to analyze the sensitivity to the number of neurons only. The solid line is the ANN model result, while the dashed line is the arithmetic EM, in order to compare their performances. Points on the plots represent the number of neurons equal to 2, 5, 10, 15, 20, 25, 30, 35, 40, 50, 80, and 200, respectively.

The objective is to look for the initialization (seed) that provides the best model performance.

Figures 9 and 10 present the results for different numbers of neurons at the hidden layer, for U10 and Hs, respectively. Different seeds and filtering windows were averaged to allow the analysis to focus on the number of neurons, from 2 to 200. It is easy to see that NBias is not very sensitive to the number of neurons, so a few neurons are sufficient to optimize the mapping compared to other error metrics, which is valid for all the variables, that is, U10, Hs, and Tp.

The scatter error is highly dependent on the number of neurons. The SCrmse and CC are continuously improved by a higher number of neurons until reaching an optimum around 40–50 neurons. At that point, the metrics begin to deteriorate with more neurons. Once again, the impact of different ANN models is higher on longer forecasts, which can be visualized, for example, comparing the range of values in the CC plots from day 0 with day 10. In Figs. 9 and 10, all the ANN architectures tested resulted in better values of SCrmse and CC than the traditional EM.

Figure 11 presents the error metrics as a function of the size of the filtering window, from 0 to 480 h (20 days), for Hs. Looking at NBias only and the first forecast days, the ANN models do not benefit from the filtering compared to the EM that already has small normalized biases around 0.05–0.10. The SCrmse and CC, instead,

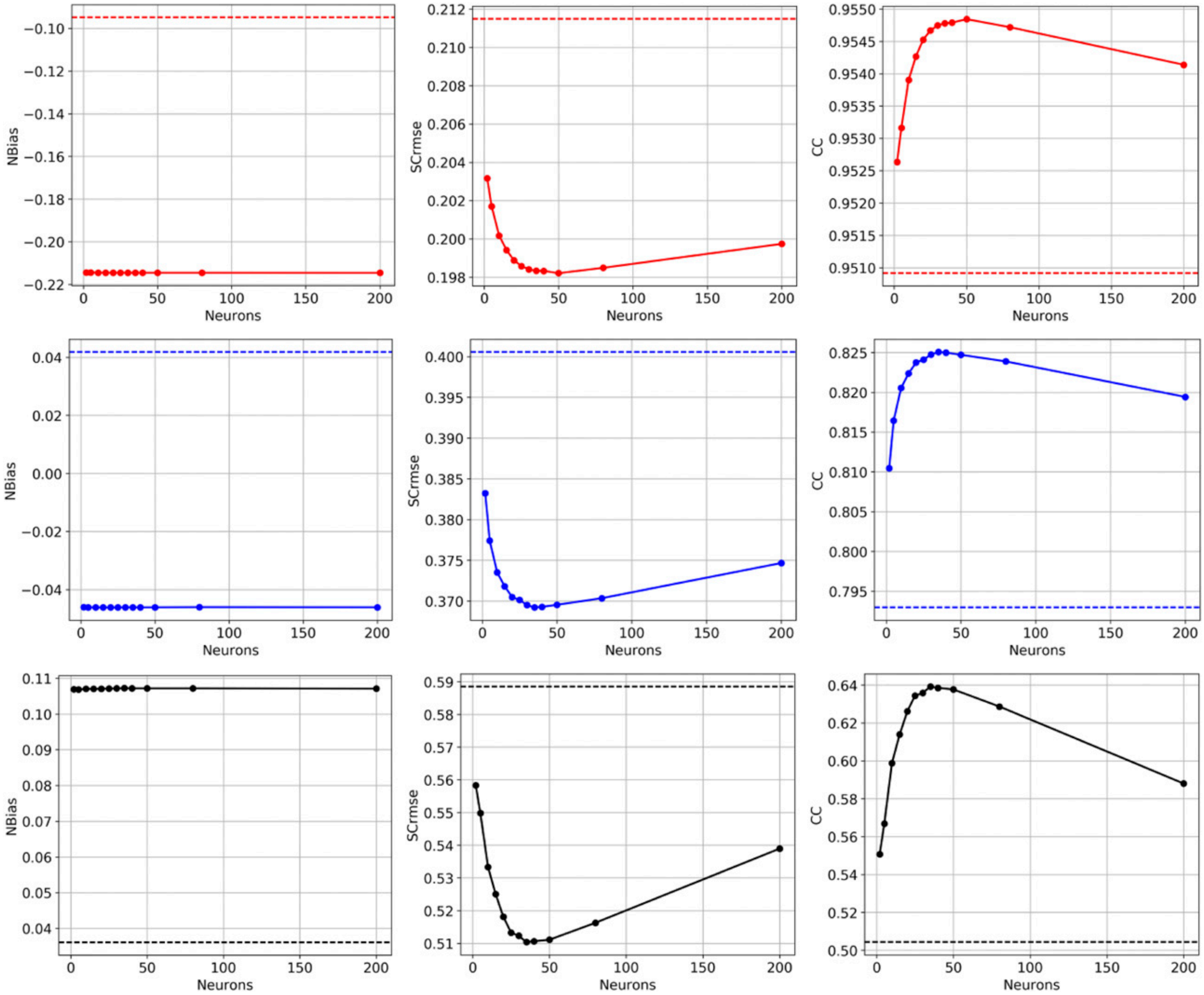

FIG. 10. NBias, SCrmse, and CC according to Eqs. (4), (7), and (9), respectively, resulting from the ANN training tests on forecast (top) day 0 (red), (middle) day 5 (blue), and (bottom) day 10 (black) for Hs. Results involving different initializations and filtering windows were averaged to analyze the sensitivity to the number of neurons only. The solid line is the ANN model result, while the dashed line is the result for the EM, in order to compare their performances. Points on the plots represent the number of neurons equal to 2, 5, 10, 15, 20, 25, 30, 35, 40, 50, 80, and 200, respectively.

show a significant improvement with window sizes between 48 and 192 h, which occurs mainly because the moving average filtering removes the high-frequency randomness from the signal that helps the ANN to minimize the scatter error. Recall the spectral analysis that identified three energetic peaks of the residues, at 40, 70, and 200 h. This behavior is confirmed in Fig. 11, where the ANNs could better improve the SCrmse and CC for filtering windows from 48 to 192 h. Again, the relative improvement is more evident at longer forecast ranges.

Regarding the bias, Figs. 9 and 10 present worse results for the ANNs than the conservative EM. This is initially because the systematic error does not have a strong dependence on the number of neurons, and the results of Figs. 9 and 10 have averaged different seeds and filtering windows, which have a greater importance for this type of error. When the bias is analyzed as a function of the size of the filtering window, in Fig. 11, it is possible to see the improvement of the bias of the ANNs, especially on day 10, and related to filtering windows around 144–192 h. Therefore, the bottom-left plot of Fig. 11 better exemplifies the benefit of the ANNs for bias correction of the GWES.

5. Results and discussion

We analyzed the errors of wind and wave parameters from the GWES, examined how the signal of the residue

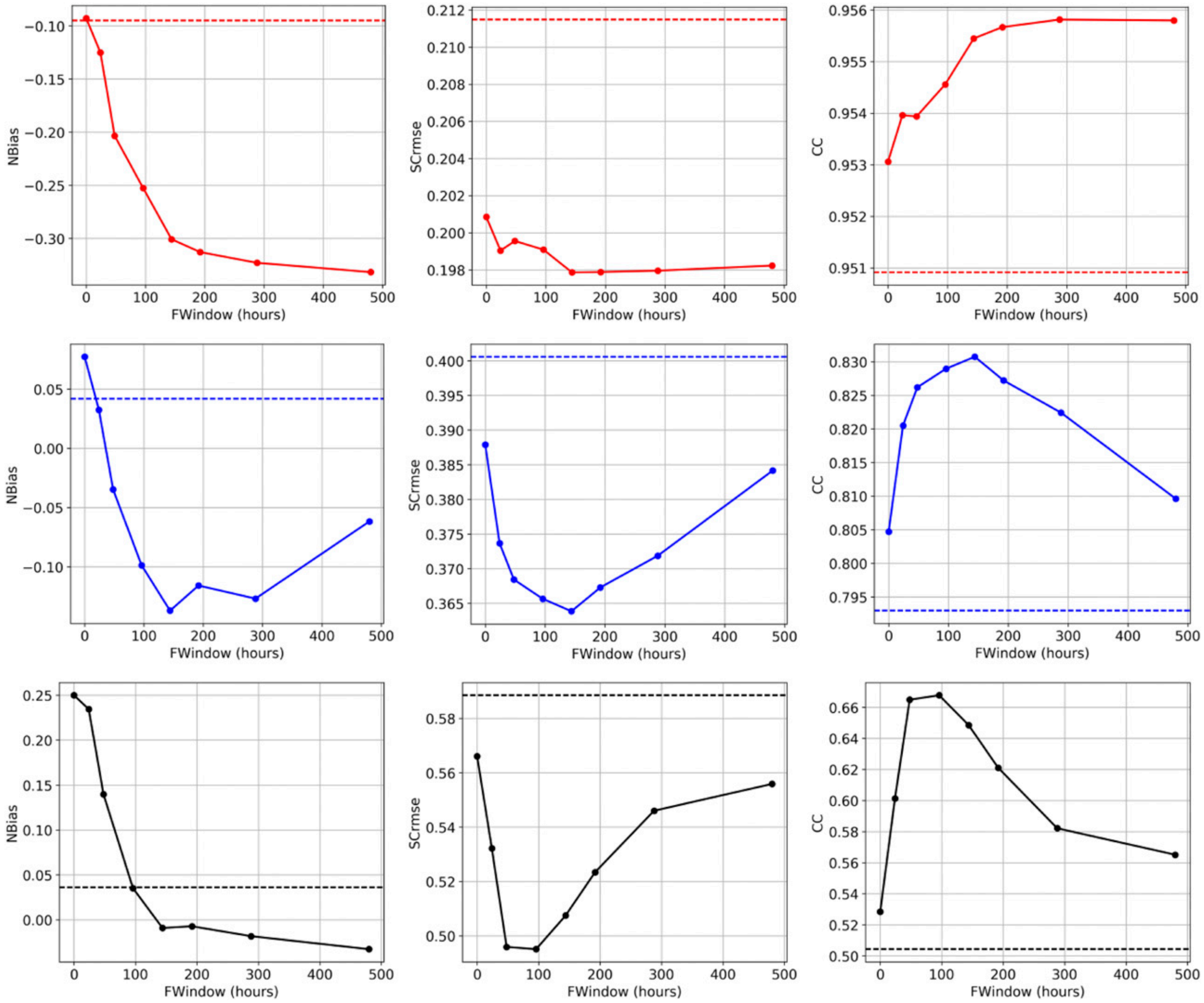

FIG. 11. NBias, SCrmse, and CC according to Eqs. (4), (7), and (9), respectively, resulting from the ANN training tests on forecast (top) day 0 (red), (middle) day 5 (blue), and (bottom) day 10 (black) for Hs. Results involving different initializations and the number of neurons were averaged to analyze the sensitivity to filtering windows only. The solid line is the ANN model result, while the dashed line is the result for the EM, in order to compare their performances. Points on the plots represent the site of the window (h) equal to 0, 24, 48, 96, 144, 192, 288, and 480, respectively.

is distributed, and explored the appropriate ANN architectures to best implement nonlinear ensemble averaging for simulations in the Gulf of Mexico. Figures 9–11 present the number of neurons and filtering windows that best improved certain error metrics for different forecast days and variables, using a large sensitivity test involving 105 600 ANN simulations. Figure 12 shows the results of the best independent ANNs, which are compared to the original GWES values and the arithmetic EM.

In terms of the brief assessment of the current NCEP ensemble forecast, the NBias of U10 in Fig. 12 indicates the nonhomogeneity of GEFS surface wind accuracy, with strong negative bias of the nowcast that is improved with forecast time. A direct impact is observed on the wave bias, which is consistent with the wind bias trend. Therefore, the ensemble approach of Zhou et al. (2017) implemented in the GEFS does not improve the systematic bias, as expected, confirmed by the red and cyan curves with similar values of NBias. The nonlinear ensemble averaging using ANNs, instead, could remove this trend of NBias on GEFS winds and reduce the systematic errors for all variables. The SCrmse plots of Fig. 12 confirm the success of the ensemble approach in reducing the scatter errors, when compared to the deterministic run.

Moving to the ANN nonlinear ensemble averaging results, Fig. 12 shows further improvement of the scatter

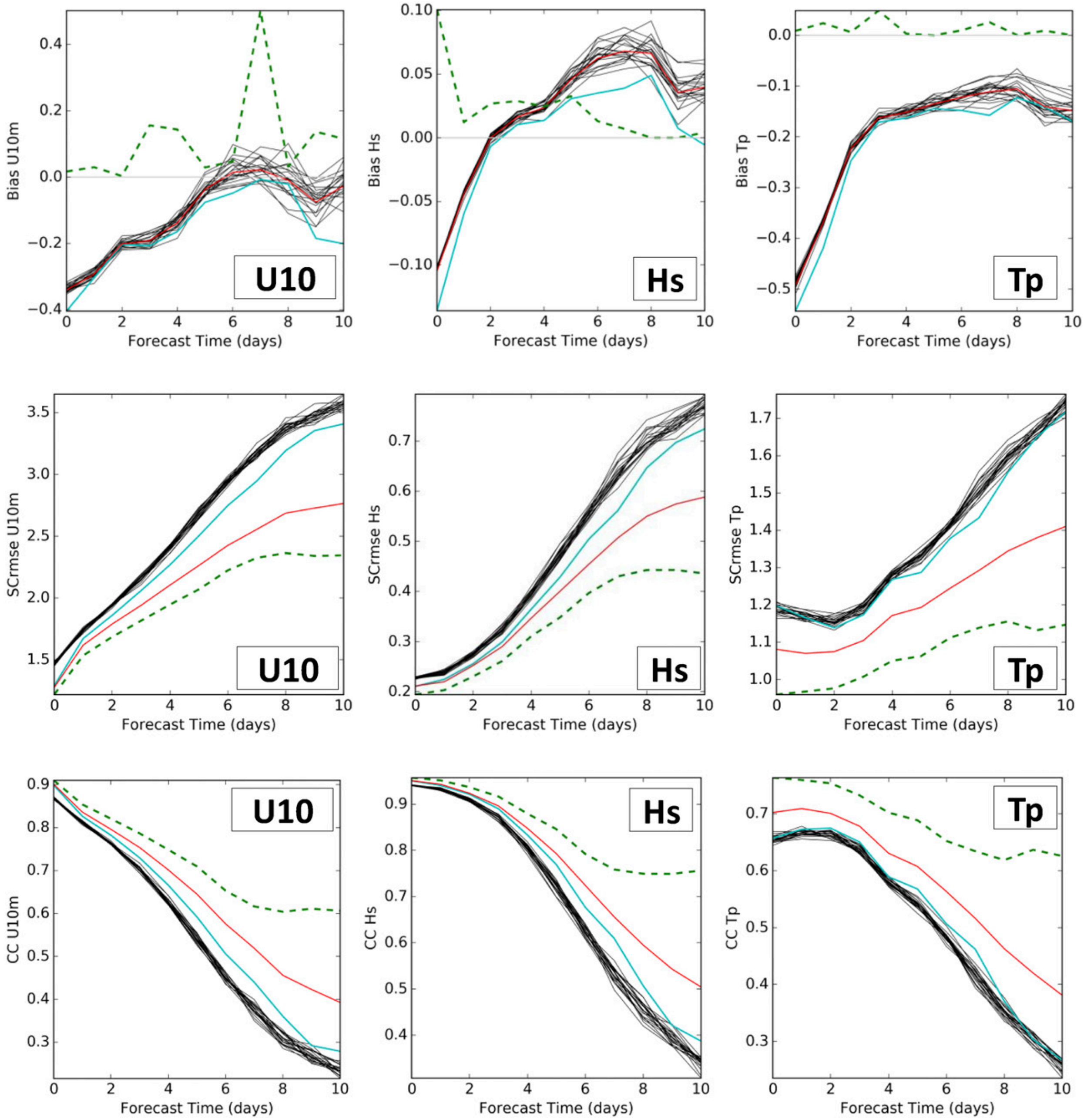

FIG. 12. Assessment of the results of the ANNs compared to the GWES and EM. Black curves show the ensemble members, cyan is the deterministic run, red is the arithmetic mean of the ensemble members, and dashed green is the nonlinear ensemble averages using ANNs. (top) Bias, (middle) SCrmse, and (bottom) CC according to Eqs. (3), (7), and (9), respectively.

errors, with smaller values of SCrmse. This shows that the ANNs are useful not only for bias correction but also for reducing scatter errors still present in the ensemble forecasts. The results for Tp have a slightly different evolution, with the reduction of SCrmse even for the nowcast and first forecast days, not seen in the plots of U10 and Hs. The CC of Fig. 12 confirms the success of the ANN method, where forecast day 10 has a skill similar to that of day 5 associated with the EM, a gain of five forecast days using the nonlinear ensemble average as described.

Apart from Tp, Figs. 9–12 show that the main advantage of using ANN models is not on the nowcast and short-range forecasts but on the longer-term forecasts. This is especially true for the reduction of the scatter errors, being related to the strong random component of

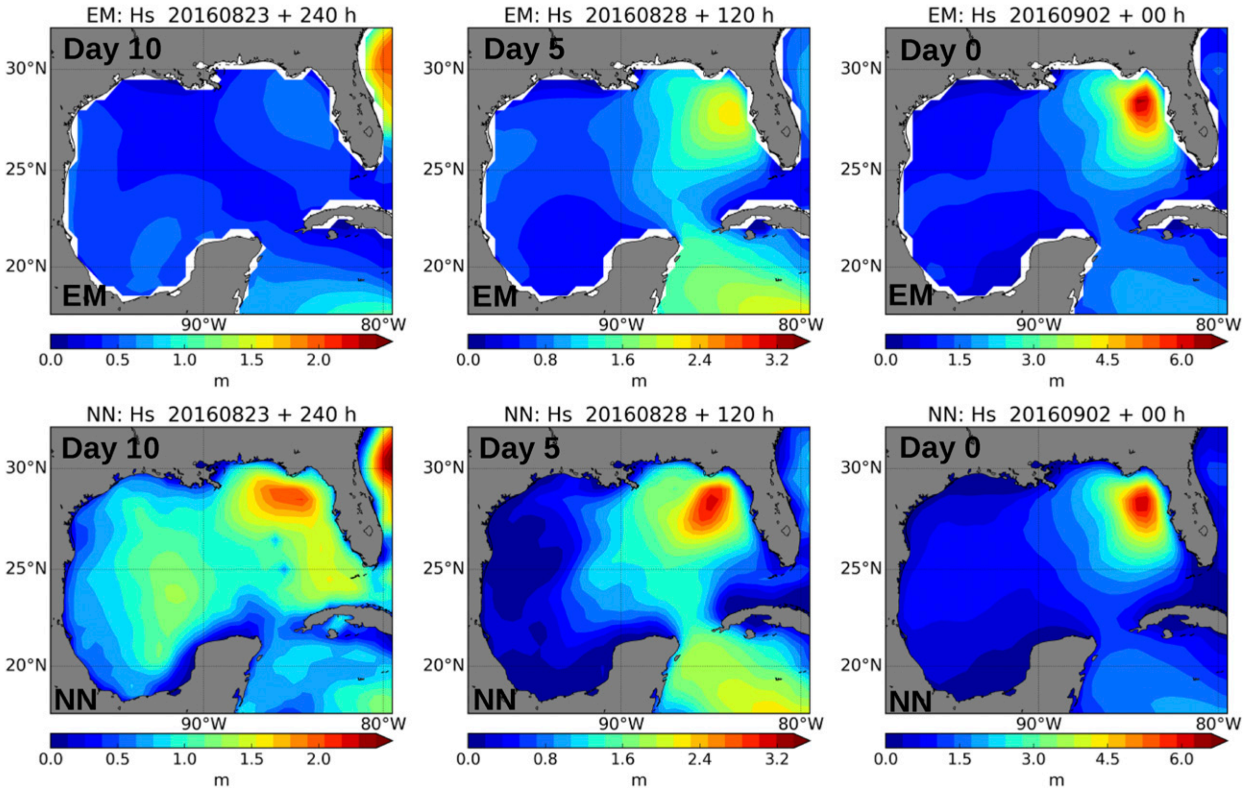

FIG. 13. Comparison between the arithmetic EM and the nonlinear ensemble average using ANNs for Hs of Hurricane Hermine in the Gulf of Mexico in 2016. All plots represent the same instant, 2 Sep 2016, but with three different forecast lead times. (right) The analysis (working as a benchmark), (center) forecast day 5, and (left) forecast day 10.

the residue signal of forecast day 0, at Fig. 7. This feature highlights the importance of filtering for the improvement of the scatter components of the error. However, even with various filtering windows, it is difficult for the ANNs to improve the scores provided by the EM because of the strong random component of day 0. The ANN architectures linked to the best scores provide important information to study the complexity and optimization of the nonlinear approximation.

Now that we have defined the best ANN architecture and calculated the preferred weights and biases, it is possible to implement the methodology for operational applications of the nonlinear ensemble averages of GWES data. To evaluate the ability of the ANNs, we apply it over the GWES grid covering the Gulf of Mexico and compare obtained fields with the conservative GWES ensemble average that is produced operationally. Figure 13 shows an example of an application relative to Hurricane Hermine (Berg 2017). All plots represent the same instant but show three different forecast lead times: nowcast, day 5, and day 10 forecasts. Hurricane Hermine had extreme winds up to 35 m s^{-1} and waves up to 6 m high, close to the coast of Florida. As mentioned before, the nowcast generally has smaller errors and the plots of EM and ANNs are very similar. Moving five days back in time, the day 5 forecast has significantly smoothed the sharp peak of the storm. The extreme winds of the EM dropped from 25 to 10 m s^{-1}, a severe underestimation. However, the ANNs better capture the peak of the storms with winds of 13 m s^{-1}.

The improved representation of the hurricane winds is propagated to the wave fields. The conservative EM of the day 5 forecast reduced Hs from 6.0 to 2.2 m in Fig. 13, while with the ANN it is still underestimated but with Hs of 3 m—confirming the better representation of the peak of the storm. Looking at the day 10 forecast, a very challenging forecast horizon, the EM does not show any signal of the hurricane. Meanwhile, a few GWES members were pointing to a tropical depression that was captured by the ANNs. The results from the nonlinear ensemble averaging are still underestimated but with a better representation of the peak of the storm than the EM. *Cryosat-2* altimeter measurements obtained from the satellite database of the National Environmental Satellite, Data, and Information Service (NESDIS) have covered the region close to Florida at 0400 UTC 2 September 2016.

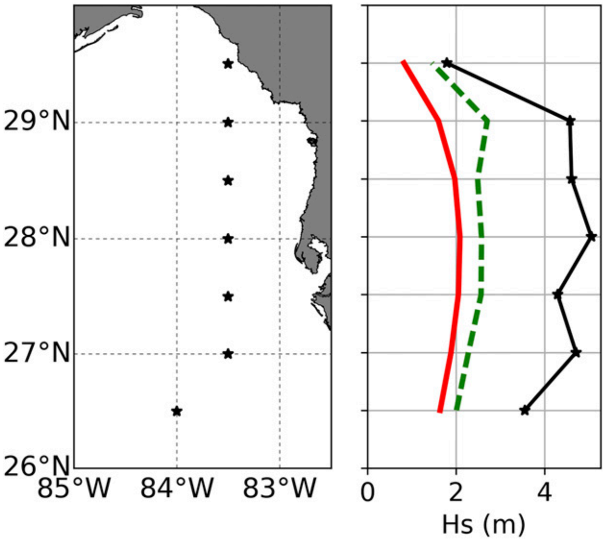

FIG. 14. Assessment of the results of Hurricane Hermine using altimeter data, at 0400 UTC 2 Sep 2016, close to the coast of Florida. (left) The seven points show the Cryosat altimeter data allocated at GWES grid points. (right) The plot of Hs associated with each grid point on the left, related to the fifth forecast day. Both figures share the same *y* axis (latitudes). Black is Hs from the altimeter, red is the EM, and dashed green is the nonlinear ensemble average using ANN. Unfortunately, the Cryosat tracks did not pass over the peak of the storm, both in time and position, but it is still useful to provide absolute values for the comparison between the EM and the ANN methods.

Comparative results for the forecast day 5 are presented by Fig. 14, which corroborates with the discussion, showing the better representation of the nonlinear ensemble average using ANNs than the EM.

6. Conclusions

The initial assessment of GWES presented in Fig. 4 pointed to a complex behavior of the systematic bias that cannot be addressed by simplistic bias correction algorithms. Further evidence illustrates the evolution of GWES errors, with systematic biases that are common to deterministic and ensemble forecasts (Fig. 8). When we considered this problem, as well as the limitations of the conservative arithmetic EM, we concluded that a proper nonlinear approximator could significantly improve the GWES forecasts.

A large experiment with 105 600 ANNs varying the random initialization, and the number of neurons and filtering windows, following the suggestions of Krasnopolsky and Lin (2012), was undertaken to study the best architecture and the complexity of ANNs to optimize the nonlinear ensemble averaging of GWES in the Gulf of Mexico. It was found that a small number of neurons is sufficient to reduce the bias, while 35–50 neurons are best suited to reduce both the scatter and systematic errors. The strong scatter component of the nowcast makes it difficult for the ANN to improve scores produced by the EM. The error metrics confirm that the main advantage of the methodology using ANNs is not on the nowcasts or short-range forecasts but primarily for longer-range forecasts. The correlation coefficient for forecast day 10, for example, was increased from 0.39 to 0.61 for U10, from 0.50 to 0.76 for Hs, and from 0.38 to 0.63 for Tp, representing a gain in skill of five forecast days using the nonlinear ensemble average.

Our study addressed the relatively small basin of the Gulf of Mexico using only buoy data for ANN training, because of the high temporal sampling of buoys that better capture the peak of the storms, as discussed by Alves and Young (2003). The ANNs were trained using only six deep water buoys and then was applied to the entire Gulf of Mexico basin. The comparison presented in Fig. 13 illustrates a very good generalization ability of the developed ANNs and strongly supports the validity of the presented approach. The next step of our study is to expand this approach to the whole globe, including altimeter data that will be crucial to support ANNs to proper simulate the spatial distribution of GWES errors. Our study also suggests that ANNs may be used effectively in an operational wave guidance context to produce bias-corrected data with improved skill, a path that will be pursued in more detail in a forthcoming study.

Acknowledgments. This study has been developed at the University of Maryland, College Park, and at the Environmental Modeling Center of NCEP, with funding from the National Weather Service Office of Science and Technology (NWS/OST), Award NA16NWS4680011. The authors would like to acknowledge Dr. Todd Spindler for providing support with coding and parallelization, the atmospheric ensemble team at NCEP, and the observations provided by NDBC and NESDIS.

REFERENCES

Alves, J.-H. G. M., and I. R. Young, 2003: On estimating extreme wave heights using combined Geosat, Topex/Poseidon and ERS-1 altimeter data. *App. Ocean Res.*, **25**, 167–186, https://doi.org/10.1016/j.apor.2004.01.002.

——, and Coauthors, 2013: The NCEP–FNMOC combined wave ensemble product: Expanding benefits of interagency probabilistic forecasts to the oceanic environment. *Bull. Amer. Meteor. Soc.*, **94**, 1893–1905, https://doi.org/10.1175/BAMS-D-12-00032.1.

Berbić, J., E. Ocvirk, D. Carević, and G. Lončar, 2017: Application of neural networks and support vector machine for significant wave height prediction. *Oceanologia*, **59**, 331–349, https://doi.org/10.1016/j.oceano.2017.03.007.

Berg, R., 2017: Hurricane Hermine (AL092016): 28 August–3 September 2016. National Hurricane Center Tropical Cyclone Rep., 63 pp., https://www.nhc.noaa.gov/data/tcr/AL092016_Hermine.pdf.

Campos, R. M., and C. Guedes Soares, 2016: Hybrid model to forecast significant wave heights. *Maritime Technology and Engineering 3*, C. Guedes Soares and T. A. Santos, Eds., CRC Press, 1027–1036.

——, V. Krasnopolsky, J.-H. G. M. Alves, and S. Penny, 2017: Improving NCEP's probabilistic wave height forecasts using neural networks: A pilot study using buoy data. NOAA Office Note 490, 23 pp., https://doi.org/10.7289/V5/ON-NCEP-490.

Cao, D., H. S. Chen, and H. Tolman, 2007: Verification of Ocean Wave Ensemble Forecast at NCEP. NOAA Tech. Note 261, http://polar.ncep.noaa.gov/mmab/papers/tn261.

Cavaleri, L., and Coauthors, 2007: Wave modelling—The state of the art. *Prog. Oceanogr.*, **75**, 603–674, https://doi.org/10.1016/j.pocean.2007.05.005.

Chen, H. S., 2006: Ensemble prediction of ocean waves at NCEP. *Proc. 28th Ocean Engineering Conf.*, Taipei, Taiwan, National Sun Yat-Sen University, 25–37.

Deo, M. C., and C. Sridhar Naidu, 1998: Real time wave forecasting using neural networks. *Ocean Eng.*, **26**, 191–203, https://doi.org/10.1016/S0029-8018(97)10025-7.

——, A. Jha, A. S. Chaphekar, and K. Ravikant, 2001: Neural networks for wave forecasting. *Ocean Eng.*, **28**, 889–898, https://doi.org/10.1016/S0029-8018(00)00027-5.

Det Norske Veritas, 2007: Environmental conditions and environmental loads. Recommended Practice DNV-RP-C205, 124 pp., https://rules.dnvgl.com/docs/pdf/dnv/codes/docs/2010-10/rp-c205.pdf.

Dixit, P., and S. Londhe, 2016: Prediction of extreme wave heights using neuro wavelet technique. *Appl. Ocean Res.*, **58**, 241–252, https://doi.org/10.1016/j.apor.2016.04.011.

Haykin, S., 1999: *Neural Networks: A Comprehensive Foundation*. 2nd ed. Prentice Hall, 842 pp.

Hsu, S. A., E. A. Meindl, and D. B. Gilhousen, 1994: Determining the power-law wind-profile exponent under near-neutral stability conditions at sea. *J. Appl. Meteor.*, **33**, 757–765, https://doi.org/10.1175/1520-0450(1994)033<0757:DTPLWP>2.0.CO;2.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Krasnopolsky, V. M., 2013: *The Application of Neural Networks in the Earth System Sciences: Neural Network Emulations for Complex Multidimensional Mappings*. Atmospheric and Oceanographic Sciences Library, Vol. 46, Springer, 189 pp., https://doi.org/10.1007/978-94-007-6073-8.

——, 2014: NN-TVS: NCEP Neural Network Training and Validation System. NOAA/NWS/NCEP/EMC Office Note 478, 60 pp.

——, and Y. Lin, 2012: A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental US. *Adv. Meteor.*, **2012**, 649450, https://doi.org/10.1155/2012/649450.

Lo, D.-C., C.-C. Wei, and E.-P. Tsai, 2015: Parameter automatic calibration approach for neural-network-based cyclonic precipitation forecast models. *Water*, **7**, 3963–3977, https://doi.org/10.3390/w7073963.

Mandal, S., and N. Prabaharan, 2006: Ocean wave forecasting using recurrent neural networks. *Ocean Eng.*, **33**, 1401–1410, https://doi.org/10.1016/j.oceaneng.2005.08.007.

Mentaschi, L., G. Besio, F. Cassola, and A. Mazzino, 2013: Problems in RMSE-based wave model validations. *Ocean Modell.*, **72**, 53–58, https://doi.org/10.1016/j.ocemod.2013.08.003.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1986: Learning internal representations by error propagation. *Foundations*, Vol. 1, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, 318–362.

Sánchez, A. S., D. A. Rodrigues, R. M. Fontes, M. F. Martins, R. A. Kalid, and E. A. Torres, 2018: Wave resource characterization through in-situ measurement followed by artificial neural networks' modeling. *Renewable Energy*, **115**, 1055–1066, https://doi.org/10.1016/j.renene.2017.09.032.

Tolman, H. L., 2016: User manual and system documentation of WAVEWATCH III version 5.16. NOAA/NWS/NCEP MMAB Tech. Note 329, 326 pp.

Tsai, C.-P., C. Lin, and J.-N. Shen, 2002: Neural network for wave forecasting among multi-stations. *Ocean Eng.*, **29**, 1683–1695, https://doi.org/10.1016/S0029-8018(01)00112-3.

Weigend, A. S., and N. A. Gershenfeld, 1994: The future of time series: Learning and understanding. *Time Series Prediction: Forecasting the Future and Understanding the Past*, A. S. Weigend and N. A. Gershenfeld, Eds., Santa Fe Institute Studies in the Sciences of Complexity, Vol. 15, Addison-Wesley Publishing Company, 1–70.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.