Improving NCEP's Global-Scale Wave Ensemble Averages Using Neural Networks

Ricardo Martins Campos[1*], Vladimir Krasnopolsky[2], Jose-Henrique Alves[3], Stephen G. Penny[4,5]

[1]Centre for Marine Technology and Ocean Engineering (CENTEC), Instituto Superior Técnico, University of Lisbon
[2]EMC/NCEP/NOAA Center for Weather and Climate Prediction
[3]SRG/EMC/NCEP/NOAA Center for Weather and Climate Prediction
[4]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder
[5]Physical Sciences Division, NOAA Earth System Research Laboratory

*Corresponding Author, e-mail address: riwave@gmail.com

ABSTRACT

The quality of metocean forecasts at longer forecast ranges has a significant impact on maritime safety and offshore operations. A nonlinear ensemble averaging technique is demonstrated using neural networks applied to one year (2017) of Global ocean Wave Ensemble forecast System (GWES) data provided by NCEP. Post-processing algorithms are developed based on multilayer perceptron neural networks (NN) trained with altimeter data to improve the global forecast skill, from nowcast to forecast ranges up to 10 days, including significant wave height (Hs) and wind speed (U10). NNs are applied as an alternative to the typical use of the arithmetic ensemble mean (EM). NN models are constructed using six variables sourced from 21 ensemble members, plus latitude, sin/cos of longitude, sin/cos of time, forecast lead time, and GWES cycle. The NN outputs are the residues of Hs and U10, i.e, the difference from the EM to the observations. One hidden (intermediate) layer is evaluated in terms of the optimum number of neurons (complexity) to map the given problem. The sensitivity test considered 26 different numbers of neurons, 10 seeds for initial conditions, and 3 equally-divided datasets; for a total of 780 NN experiments. Assessments using 2,507,099 paired satellite/GWES fields show that a simple NN model with few neurons is able to reduce the systematic errors for short-range forecasts, while a NN with more neurons is required to minimize the scatter error at longer forecast ranges. The novel method shows that one single NN model with 140 neurons is able to improve the error metrics for the whole globe while covering all forecast ranges analyzed. The bias of the widely used EM of GWES that varies from -10% to 10% for Hs compared to altimeters can be reduced to values within 5%. The RMSE of day-10 forecasts from the NN simulations indicated a gain of two days in predictability when compared to the EM, using a reasonably simple post-processing model with low computational cost.

Keywords: neural networks; ensemble forecast; non-linear ensemble averaging; wave modeling; altimeter data.

1. Introduction

Accurate forecasts of surface winds and waves are essential for activities such as ship routing, high-risk maritime operations, coastal management, and alerts of extreme events. Extending wave forecast skill throughout longer forecast horizons requires multiple research initiatives, from improved modeling, for example incorporating atmosphere-ocean coupling (e.g. Janssen et al., 2002), to improved data assimilation methods, such as coupled data assimilation (Penny et al., 2017). While these improvements represent important benefits to the predictability of metocean variables, deterministic forecasts are still limited in their usefulness for outlooks beyond one week due to the chaotic behavior of the atmosphere-ocean-wave coupled system, for example as pointed out by Lorenz (1963) using a simple model of the atmosphere. Using an ensemble of multiple forecasts can extend the range of skillful predictions often out to 10 days, with the additional benefit of providing a measure of the uncertainty via the spread of predictions (ensemble members). The arithmetic ensemble mean (EM) typically yields smaller forecast errors compared to the mean error of each individual member (Murphy 1988), which has been confirmed for both atmospheric ensemble forecasts (Zhou et al., 2017) and wave ensemble forecasts (Cao et al., 2009; Alves et al., 2013).

Since 1992, the European Centre for Medium-Range Weather Forecasts (ECMWF) and the U.S. National Centers for Environmental Prediction (NCEP) have produced operational ensemble forecasts. Saetra and Bidlot (2004) investigated the quality of the ECMWF ensemble prediction system using buoy and satellite data. An interesting improvement to ship routing using the ECMWF wave ensemble system was analyzed by Hoffschildt et al. (1999). The NCEP atmospheric global ensemble forecast system (GEFS) was recently assessed by Zhou et al. (2017) and the NCEP global wave ensemble system (GWES) has been described by Chen (2006), Cao et al. (2009), and evaluated by Alves et al. (2013). They found that after the day-5 forecasts, the root-mean-square error of the ensemble mean becomes smaller than the control forecast – however, the general bias does not show any improvement, as expected. This feature has been confirmed by Campos et al. (2018a), who calculated the systematic and scatter errors of 10-m wind speed (U10) and significant wave height (Hs) from NCEP ensemble forecast using buoy measurements. At longer forecast ranges, beyond one week, Campos et al. (2018a) found an improvement of 20% on the scatter index of the EM compared to the control run, and no significant improvement on the systematic error. Nevertheless, even with the benefits of the ensemble approach, large forecast errors are still present beyond the day-7 forecasts, demanding further post-processing techniques.

Zieger et al. (2018) implemented a regional wave ensemble forecast system and developed a technique to bias-correct the mean value using multivariate linear regression based on Glahn and Lowry (1972). Durrant et. al. (2009), based on Woodcock and Greenslade (2007), developed an operational consensus forecast scheme that uses past model performance to bias-correct and combine forecasts to produce an improved product at locations where recent observations are available. For 24-hour forecasts, their methodology produced improvements of 36% and 31% in RMSE of Hs and U10 compared to the mean raw model components. Following a similar idea, Harpham et al. (2016) developed a Bayesian statistical method that modifies the probabilities of ensemble forecasts based on recent performance of individual members against a set of observations. These works are examples of bias correction methods, which are mostly based on multivariate linear regression and estimation of dynamic weights applied to ensemble members. Moving to a nonlinear mapping, our goal is to develop post-processing algorithms

based on neural networks (NN) trained with altimeter data to improve the NCEP's GWES. This approach enhances the traditional EM to a nonlinear ensemble average that aims to reduce both the systematic and scatter errors of U10 and Hs. In our previous works (Campos et al. 2017, 2019a) we introduced a NN technique to perform a regional nonlinear ensemble averaging based on buoy data. In this work we generalize the previously developed technique to global scale forecasts, using altimetry data.

We describe the neural network model in Section 2. The global ensemble and observations, as well as the data organizing and pairing, are described in Section 3. Section 4 is dedicated to sensitivity tests and construction of neural network models, investigating the complexity necessary to address the global mapping. Section 5 shows the results and provide a discussion about the benefits and shortcomings of the method, and Section 6 presents the final conclusions, challenges, and suggestions of next steps.

2. Nonlinear Mapping using Neural Networks

The assessments of the NCEP ensemble forecast system performed by Campos et al. (2018a) and Campos et al. (2017), based on Mentaschi et al. (2013), draw attention to the multivariate and nonlinear aspects of the forecast error. Typically, the interpretation of ensemble outputs is mostly based on the mean and standard deviation (or spread) of the ensemble members. However, use of the EM assumes that a linear relationship between ensemble members is optimal, while this relationship may in fact be strongly nonlinear, particularly at longer lead times. In order to address these nonlinear relationships, we propose using feedforward NN models to produce an ensemble average as a post-processing alternative, trained with quality-controlled observations.

A multilayer perceptron NN model (MLP-NN, Rumelhart et al. 1986) has been selected due to its previous successes being a powerful universal mapping approximator (Hornik, 1991), while being flexible and easy to implement on regression problems. The MLP-NN is a feed-forward artificial NN that uses supervised learning, and consists of three or more layers: one input layer, one or more hidden layers, and one output layer. In this study, only one hidden layer is used, though we vary the number of nodes to properly identify the minimum complexity and avoid over-fitting. The MLP-NN implemented is based on the theory of Haykin (1999) and implementation support of Krasnopolsky (2013). Equation (1) presents the MLP-NN model, which is built with hyperbolic tangent as the activation function.

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

x_i are the inputs, y_q the outputs, a and b are the weights, n and m are the numbers of inputs and outputs respectively. The number of nodes (neurons), or hyperbolic tangents, is given by k. The optimization of parameters a and b is based on backpropagation training using gradient decent. At each iteration, the Loss function is calculated as the square of the error obtained from the forward propagation of the inputs minus the observations, which is then propagated backwards using the derivative of the activation function, $1 - \tanh(x)^2$ in order to correct the weights. It has been verified that efficient optimization is obtained with the stochastic gradient decent described by Kingma and Ba (2014), chosen for the NN training. The

selection of a large and reliable set of measurements, the number of iterations, and the number of neurons are key aspects during the training process. As a pre-processing step, a quality control is applied to exclude outliers and then input and output variables are rescaled between 0 and 1 according to equation (2). This process is later inverted after NN simulations.

$$\tilde{x}_i^{[0,1]} = \frac{\left(x_i - x_i^{min}\right)}{\left(x_i^{max} - x_i^{min}\right)} \tag{2}$$

Such NN models have gained increasing use in environmental problems. Examples for wave forecasting problems include Sánchez et al. (2018) concerning wave energy potential, Mandal and Prabaharan (2006) for forecast of Hs in India, Dixit and Londhe (2016) for extreme Hs simulations from hurricanes using a neuro-wavelet technique, Berbić et al. (2017) for short-term predictions of Hs, among other applications described by Krasnopolsky (2013).

The first step towards the nonlinear ensemble averaging using NN was taken by Campos et al. (2017), who developed MLP-NN models for two point-wise locations, on the east and west coasts of the United States, trained with NDBC buoy data. Despite initial problems with excess noise and risk of over-fitting, a simple NN model with 11 nodes (neurons) and one hidden layer was effective in reducing the 5-day forecast errors of Hs by 64% for bias, and 29% for RMSE. Rasp and Lerch (2018) applied a similar neural network model for postprocessing ensemble weather forecasts of 2-m temperature at surface stations in Germany – being able to outperform benchmark postprocessing methods with low computational cost. Later developments by Campos et al. (2019a) expanded the single-point approach of Campos et al. (2017) to a regional modeling application, introducing the spatial dimension into the NN. Using six NDBC buoys in the Gulf of Mexico, 105,600 NNs were built with different architectures and initial conditions in order to investigate the ability of NNs to reduce scatter errors and systematic errors present in the GWES. The most effective NN models of Campos et al. (2019a) were found with 35 to 50 neurons in the hidden layer, which improved the correlation coefficient of day-10 forecasts from 0.39 to 0.61 for U10, and from 0.50 to 0.76 for Hs, when comparing to the EM. We note that both Campos et al. (2017) and Campos et al. (2019a) developed one independent NN per forecast time, from 0 (nowcast) to 10 days (upper limit of GWES); and the training process was 'static', based on one year of measurements, and not dynamic (or online) as some of the references described before. This means that once the model is trained, the post-processing algorithm and NN parameters are not modified even when recent observations become available, unless a re-training is applied.

As a follow-up of Campos et al. (2019a), our present study has two specific technical challenges: (1) to expand the domain from a small basin (Gulf of Mexico, of Campos et al. 2019a) to the whole globe; and (2) develop a single NN model that can minimize the error at all forecast horizons, from the nowcast out to 10 days and beyond. Figure 1 illustrates the first challenge, where different wind and wave climates can be visualized through the correlation coefficient (CC) of U10 and Hs. Locations in red indicate large CC where Hs is usually high when surface winds are intense. On the other hand, locations in blue, where CC is low, often have relatively large waves without strong winds – probably due to the passage of mature swells at trade winds zones. Figure 1 is a simple illustration of how the homogeneous climate within the Gulf of Mexico, explored by Campos et al. (2019a), compares to the whole globe explored in the present study. This indicates the need for proper spatial modeling in the NN simulations and sufficient

amount of observations during the training process, in order to build a single best NN model to cover all forecast leads in global simulations.

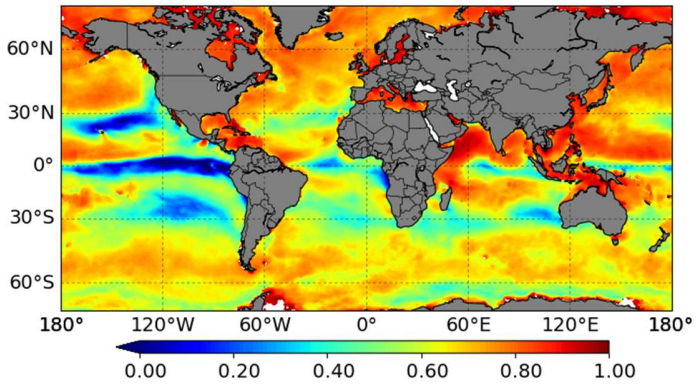

Figure 1 – Correlation coefficient map of Hs and U10, calculated using one year (2017) of the deterministic run (control) for the nowcast only. Red colors highlight areas strongly influenced by wind-sea.

We use observations of wind speed (U10) and wave height (Hs) covering the whole globe as provided by satellite altimeters, as described in Section 3, and consequently we use the same quantities as the output variables of the NNs. The inputs for the NN model include all of the variables that benefit the mapping, which consists of input information with high correlation with outputs and verified physical meaning, for each ensemble member, plus spatiotemporal parameters such as location, time, and forecast lead time – discussed in section 4. Inspired by the GWES forecast of Hurricane Mathew in the east coast of the United States (Figure 2), Campos et al. (2017) proposed a slightly different setup of the NN outputs. Figure 2 shows the ability of the EM of the day-5 GWES forecast in predicting an extreme event. The first part of the storm was very well simulated while the second peak was overestimated by the forecast. In this case, the traditional EM produces a skillful forecast for early part of the event, implying no need for post-processing intervention, while the later part of the storm has a significant drop in skill, indicated post-processing is required. As a result, the suggestion of Campos et al. (2017) was to use NNs to predict the anomaly (or residue) of the forecast, i.e., the difference between the measurement and the EM.

Predicting the residue (or residual) has an advantage during the training process, of not updating the NN parameters (Equation 1) during the backpropagation training when the EM is already relatively accurate, while reserving the largest updates to the weights a and b for the times when the EM severely deviates from the observations. This approach agrees with the paradigm that NN should be applied to nonlinear problems only (Krasnopolsky, 2013, Chpt.1 and 2), i.e., the linear part is adequately represented by the EM while the nonlinear component is simulated by the NN through the prediction of the residue. By using the residue predicted by the MLP-NN, combined with the EM, Equation 1 is now embedded in Equation 3 where the nonlinear ensemble averaging NEM is finally calculated. Campos et al. (2017) demonstrated the success of this approach for a range of percentiles, including extreme events with fewer samples in the database. The top percentiles are usually associated with larger errors (Campos et al., 2018a) that lead to larger updates of NN weights. Figure 3 illustrates the model, selected as the basis of our global NN simulations, and Table 1 shows the NN inputs, outputs, and NN experiments - described in detail in section 4.

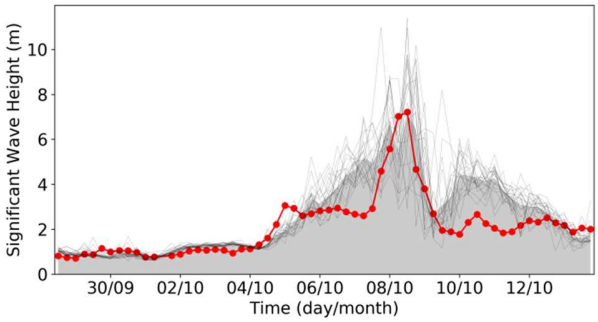

Figure 2 - GWES forecast (day-5 forecast) for a period in 2016 related to Hurricane Mathew. Black lines are the 20 NCEP ensemble members, shaded-grey is the arithmetic ensemble mean, and in red is the NDBC measurement for station 41004 at 32.501°N / 79.099°W.

$$NEM = EM + NN_r(p_1, p_2, \cdots, p_n) \quad (3)$$

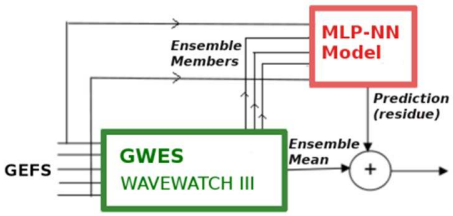

Figure 3 - Hybrid scheme proposed by Campos et al. (2017), where the NN model is dedicated to predict the residue that is added to the arithmetic ensemble mean to obtain the final value.

Table 1 – Summary of NN architecture and NN experiments, where U10 is 10-m wind intensity, Hs is significant wave height, Tp is peak wave period, Tm is mean wave period, WsH is significant wave height of wind-sea, and Tws is period of wind-sea.

<table>
<tr><th colspan="3">133 NN Inputs</th><th>2 NN Outputs</th><th>780 NNs</th></tr>
<tr><td rowspan="8">21
members</td><td rowspan="8">U10
Hs
Tp
Tm
WsH
Tws</td><td>Latitude</td><td rowspan="8">Residue U10
Residue Hs</td><td rowspan="2">10 seeds</td></tr>
<tr><td>Sine Longitude</td></tr>
<tr><td>Cosine Longitude</td><td rowspan="2">3 independent datasets</td></tr>
<tr><td>Sine Time</td></tr>
<tr><td>Cosine Time</td><td rowspan="3">26 different numbers of neurons</td></tr>
<tr><td>Forecast lead time</td></tr>
<tr><td>NCEP/GWES cycle</td></tr>
</table>

The hybrid system (Figure 3) uses the combination of the EM with the NN prediction of the residue to obtain final estimates of U10 and Hs, which are then assessed against altimeter observations. Willmott et al. (1985) provide a complete discussion about environmental model assessments, using metrics to analyze the accuracy and precision of model results. Mentaschi et al. (2013) present a recent valuable complement to this topic, with a discussion about limitations of RMSE and how it can be complemented by other metrics to have a reliable estimation of the systematic and scatter components of the error. Among the equations given by Mentaschi et al. (2013), we prefer to utilize normalized metrics since the model performance is assessed in a global domain, including wind and wave climates with different severities. Thus, three normalized error metrics are utilized to evaluate the results: normalized bias (NBias) that measures the systematic error; scatter index (SI) that measures the scatter error; and normalized RMSE (NRMSE) that combines the systematic and scatter components. Equations (4) to (6) describe the dimensionless metrics selected, where x is the altimeter data, y is the forecast, and σ_x is the standard deviation of x. The overbar indicates the arithmetic mean. By using these three normalized metrics without units, plots and tables of errors can be interpreted as ratios, or percentage errors divided by 100.

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

$$SI = \sqrt{\frac{\sum_{i=1}^{n}[(y_i - \bar{y}) - (x_i - \bar{x})]^2}{\sum_{i=1}^{n} {x_i}^2}} \tag{5}$$

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n} {x_i}^2}} = \sqrt{SI^2 + NBias^2\left(\frac{\bar{x}^2}{\bar{x}^2 + \sigma_x^2}\right)} \tag{6}$$

3. Input data and observations

We use one year of 2017 historical forecast data from the NCEP global ensemble system, and satellite observations selected for the same period. The GWES was implemented in 2005 (Chen, 2006; Cao et al., 2009), and is based on the third-generation wave model WAVEWATCH-III (Tolman, 2016). The current GWES version (Alves et al., 2013), used in the present paper, runs a 10-day forecast with four cycles per day, with a space-time resolution of 0.5° and 3 h, and produces ensemble forecasts using 20 GEFS-forced members plus a control member, described by Zhou et al. (2017). Winds and ice concentration are used as forcing fields from the GEFS, which was first implemented in 1992 (Toth and Kalnay, 1993). The GEFS initialization scheme was recently replaced (Zhou et al., 2017), from the breeding-based Ensemble Transformation with Rescaling (ETR) to the Ensemble Kalman Filter scheme (EnKF, Whitaker et al, 2008). The space-time resolution of surface winds from GEFS is the same as GWES, 0.5° and 3 h. Ice concentrations are obtained from NCEP's automated ice analysis (Wu and Grumbine, 2013).

Note that perturbations are solely added to the atmospheric ensemble in GEFS. Behrens (2015) and Farina (2002) argue that atmospheric forecast models represent highly nonlinear dynamic systems that

could generate chaotic forecasts due to small perturbations in the initial condition, while perturbations of the initial state in wave models have small effects on the results. Therefore, the wave ensemble integrates 21 independent simulations of the wave model that differ in the provided forcing conditions. GWES only propagates to the wave spectra the perturbations added to the atmospheric model.

The quality-controlled altimeter data used for training NN outputs were obtained from two sources: AVISO and NESDIS. The altimeter missions Jason2, Jason3 and Saral were downloaded from AVISO ftp area, while Cryosat2 was obtained from NESDIS. Complete assessments of altimeter data can be found at Queffeulou (2004), Queffeulou (2012), Queffeulou (2013), Sepulveda et al. (2015), and Queffeulou and Croizé-Fillon (2017). Comparisons with buoys show that the altimeter estimate of Hs is, in general, in agreement with in situ data, with the differences having a standard deviation around 0.3 m, depending on the satellite, with a small overestimation at low Hs and underestimation for high Hs. Taking into account that level of uncertainty is much smaller than forecast errors, altimeter data from the two sources above can be directly applied for the NN training, after a quick additional quality control.

The along-track altimeter data are organized and collocated into the regular grid of GWES, using the kd-tree algorithm and based on the methodology of Young and Holland (1996) and Sepulveda et al. (2015). Considering the high sampling rate of the satellite track, all measurements with a maximum space distance of 25 km and time distance of 0.5 hours are allocated to each grid point (Lat/Lon) at a specific time. The multiple altimeter records within this cube of Lat/Lon/Time are selected and a Gaussian function applied to weight values by distance to the center point, in order to give a single altimeter data matching the GWES grid-point. Figure 4 illustrates the collocated altimeter data over the globe for the duration of the NN experiment.

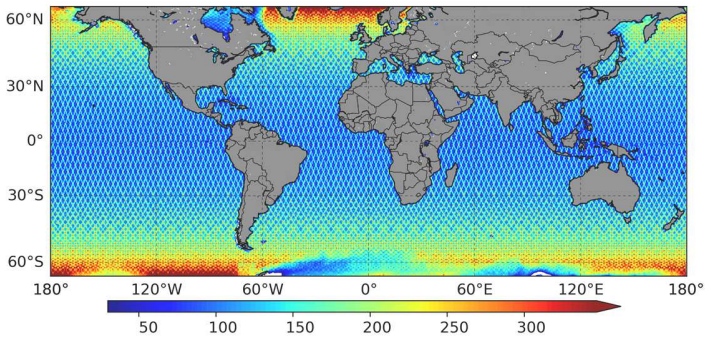

Figure 4 - Total count of altimeter observations per GWES grid point for 2017.

Furthermore, two additional criteria are imposed to exclude satellite/GWES matchups at shallow and intermediate waters or located close to the coast. These criteria avoid increasing errors of altimeter data due to footprint averaging size and restrict the NN emulation to deep waters. We use ETOPO1 bathymetry (Amante and Eakins, 2009) with 1 arc-minute of resolution, and a measure of distance from the coast with 0.04 degrees resolution from NASA's Goddard Space Flight Center database. We select a minimum water depth of 490 m and minimum distance from the coast of 100 km. Applying these criteria, a total count of 7,521,298 satellite/GWES matchups between 60°S and 60°N are allocated to GWES grid points at 3 hourly time resolution. For analysis of results, the largest oceans are delimited using the World Seas database (IHO, 1953), containing a demarcation of oceans and seas, giving: 817,516 matchups of

satellite/GWES in the North Atlantic, 1,280,934 in the South Atlantic, 1,601,452 in the North Pacific, 2,175,888 in the South Pacific, and 1,645,508 in the Indian Ocean.

For single hindcast simulations, the 7,521,298 matchups described before would consist of pairs of one vector of model variables (see Section 4) per altimeter value. However, there are two additional dimensions, ensemble members and forecast time, so each altimeter record is paired to a matrix exemplified by Figure 5. The inclusion of forecast time paired to altimeter data must be applied with caution, because the sequence of records at any specific location is made sparser by shifting satellite orbits. For each altimeter measurement at any Lat/Lon/Time, we move backwards in time and select GWES predictions all valid at the same location and time; for example, taking the 24-hour forecast step of the preceding 1-day GWES cycle, then the 48-hour forecast step of the preceding 2-day GWES cycle etc.

This process can be applied with the NCEP cycle resolution of 6 hours, giving 41 sets of forecast leads within the time horizon of 10 days. The matrix of 21 ensemble members per 41 forecast instants provides 861 model values that with an accurate forecast should be similar to the single satellite observation. Figure 5 exemplifies this matrix where, on the nowcast GWES is performing very well, for the first four forecast days GWES slightly overestimates the observations, and beyond the 5th day there is a severe underestimation of the forecast that should be attenuated by the NN post-processing model.

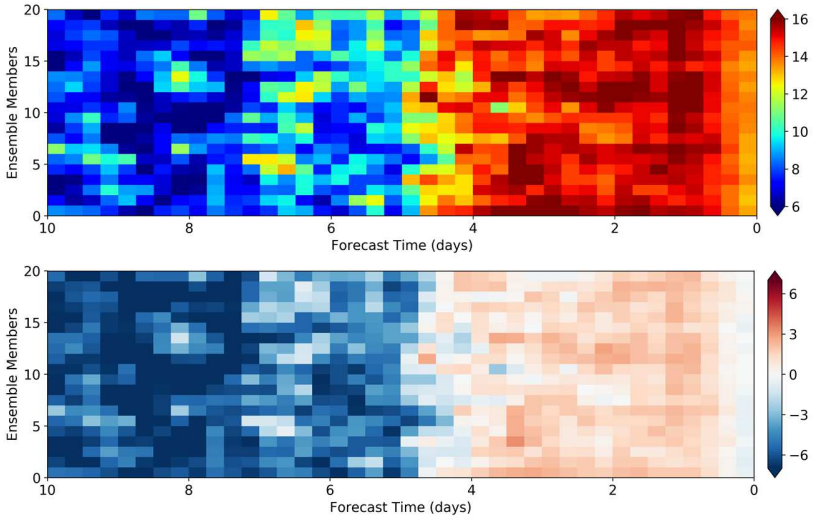

Figure 5 - Matrix representation of forecasts produced by GWES at 54.4°S / 74.5°W, related to the nowcast on 2017/06/10, 12Z, and up to 10-day forecasts (41 cycles) prior to the event. The top plot shows the significant wave height (Hs, meters) of GWES. The corresponding observation of Hs derived from the altimeter is 13.8 meters, on 2017/06/10, 12Z at the same position. The bottom plot shows the difference of observation minus GWES, in meters, where blue indicates underestimation of forecasts, red indicates overestimation, and the white color is the perfect agreement between model and observation.

4. Neural Network Architecture and Sensitivity Tests

The MLP-NN models were constructed based on equations (1) and (3). NN inputs include six variables: 10-m wind intensity (U10), significant wave height (Hs), peak wave period (Tp), mean wave period (Tm), significant wave height of wind-sea (WsH), and period of wind-sea (Tws), for each of 20 ensemble members plus the control member. Initial tests included only Hs, U10, and Tp as NN inputs, however after expanding the simulations to the whole globe, the addition of Tm, WsH, and Tsw, were

found to provide valuable information about the wave spectra needed to differentiate mature swell from young wind-sea, wave generation from propagation zones, and different wind and wave climates (Figure 1). Zieger et al. (2018), in their study of ensemble forecasts in Australia, confirmed the benefit of including wind-sea features, such as significant wave height and period of wind-waves, that improve the overall spectral information and the ensemble prediction.

The geographical space is represented with three additional NN inputs: latitude, and sine and cosine of longitude (equation 7). The sine and cosine of time (days) are added as inputs in order to include an annual cycle and seasonal effects to the mapping, presented in equation 8. Furthermore, forecast lead-time is included, varying from 0 to 10 days, as well as GWES forecast cycle (0,6,12,18). This results in a total of 133 variables as the NN inputs: 126 GWES variables, 3 variables for location, 2 for time, and 2 variables for forecast lead-time and cycle.

$$lonsin = \sin\left(\frac{2\pi lon}{360}\right), \qquad loncos = \cos\left(\frac{2\pi lon}{360}\right) \tag{7}$$

$$tsin = \sin\left(\frac{2\pi time}{365}\right), \qquad tcos(t) = \cos\left(\frac{2\pi time}{365}\right) \tag{8}$$

NNs do not automatically understand periodic and cyclic variables if not stated, for example "time" (where month 1 comes after month 12, and hour 0 comes after 23), "longitude" (-180 comes after 179) etc. Therefore, sine and cosine had to be applied to time and longitude, as presented by equations above, increasing the number of variables.

The NN outputs are composed of two variables only: the residues of U10 and Hs, presented by Figure 3. The hidden (or intermediate) layer, containing the hyperbolic tangents, controls the complexity of the mapping. The computational and functional complexity of the NN mapping (N_c) of MLP-NN (1) can be defined by equation (9), following Krasnopolsky (2013). As in equation (1), n and m are the total numbers of inputs and outputs, and k is the number of nodes (neurons) in one hidden layer. Once the NN inputs and outputs are defined and fixed, the complexity is controlled by k. We focus on identifying the most effective value for k, which is problem and domain dependent. The optimal N_c for the NN global modeling is unknown, so we conduct an experiment with several NN simulations with different number of neurons k. The test aims to find a single NN model with the best configuration of the hidden layer and optimized parameters a and b.

$$N_c = k \cdot (n + m + 1) + m \tag{9}$$

A total of 26 different numbers of neurons are tested through independent NN simulations: 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 180, 200, 250, 300, 350, 400, 450, and 500. Ten different seeds, for random initialization of a and b of equation (1), are used to estimate the sensitivity of the backpropagation training algorithm to the initial weights and to find better initial weights. The dataset of 7,521,298 matchups of satellite/GWES is randomly divided into three datasets where the NNs are trained and assessed independently. This allows a further analysis of the robustness of the NN model and reduces memory load during the computational-costly training step. The entire sensitivity test considers 26 different numbers of neurons, 10 seeds, and 3 datasets, giving a total of 780 independent NNs. Table 1 summarizes the NN architecture and NN simulations performed. Besides the data division, a cross-validation scheme with three cycles was applied to each dataset (previously divided, and independent to each other), alternating indexes defined for training and testing by using the leave-

one-out method. In summary, cross-validation was applied three times, where each one selected 2/9 of the entire dataset for training and 1/9 for testing. This approach was selected to ensure that NN assessments are applied to records that were not used during training and, combined to a sufficiently small number of neurons and iterations, avoids over-fitting, over-training and ensures better generalization.

Results of sensitivity tests

Results involving NN simulation for the three different randomly selected datasets (see previous paragraph) are averaged, since errors are very similar among them, and then presented as a function of the number of neurons and plotted divided into training and test sets. The systematic error of the EM of GWES is around 3% (written on top of the left column plots in Figure 6) while the NNs errors are bounded within -1% and +1%; a very small error involving NN with both small and large number of neurons. Similar values of NBias are found for U10 and Hs, presented by Figure 6. The scatter indexes (SI) indicate better results for Hs than U10, where the EM has 24.5% of error for U10 and 21% for Hs. These values drop to 23% for U10 and 19% for Hs when using NNs, a relative improvement that is smaller than the improvement found for NBias. The evolution of the SI with the number of neurons shows a minimum value that corresponds to more neurons than the same for NBias, where a sharp decay is seen between 2 to 50 neurons. These differences, however, are only 0.38% in the scatter error among various numbers of neurons. The best results of NRMSE, which combines the scatter and systematic errors (equation 6, shown in the right column of plots of Figure 6), are found between 60 to 180 neurons but within a range of only 0.31%. Above 200 neurons, the SI and the NRMSE start to increase again.

The difference between training and test sets is small (Figure 6), suggesting that there is no overtraining during the backpropagation training step. However, there is variation due to the use of different seeds such that the results tend to diverge above 200 neurons with increasing spread. This indicates that the NN models might be over-fitted and implies that the complexity of NN (9) is greater than needed. For the three different metrics and two output variables, the NN models have smaller errors than both the EM and control member (top of each plot in Figure 6). Thus, from aforementioned analysis we conclude that the best NN models should have between 60 to 180 neurons at the intermediate layer. However, this result comes from the assessment integrated over the entire GWES forecast range of 10 days, while errors increase significantly with forecast horizon, which impacts the NN training. Figure 7 presents the same results as Figure 6 but for three different forecast lead times: day 0 (nowcast), day 5, and day 10 - where each point related to a specific number of neurons is an average of 30 NNs (10 seeds and 3 datasets). The NRMSE for Hs reaches minimum at 80 and 90 neurons for different forecast leads. For U10, on day 0, the sharp decay of the curve suggests the best results with 50 to 80 neurons, and values above 90 have larger NRMSE. For day 5, a second minimum is found around 160 to 180 neurons. The longest forecast range, day 10, shows the best results between 120 and 180 neurons, also indicating larger NRMSE for NN with neurons equal or less than 110. Therefore, the increasing scatter error of the surface winds at longer forecast ranges is the main feature that requires more complex NNs. Another characteristic to notice is the distance between the NN curves of training and test set. On day 0 they are very close to each other, while for day 5 and 10, the test set changes to larger NRMSE than the training set, indicating the greater difficulty of the NN in simulating longer forecast ranges.

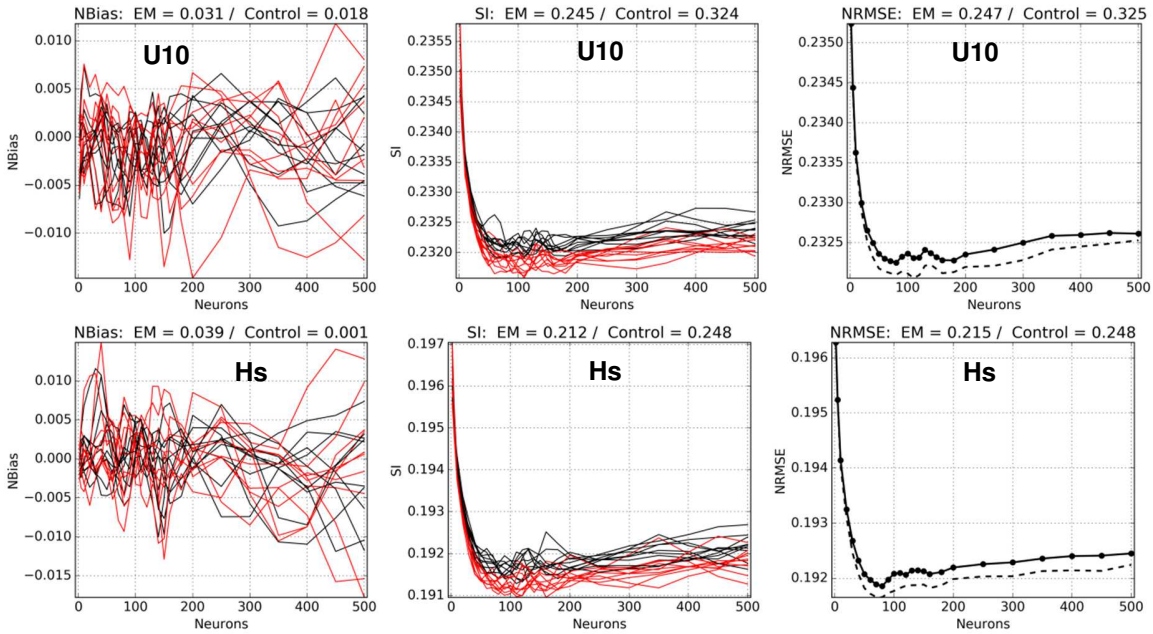

Figure 6 – Assessment of NNs performance statistics (vertical axes) as functions of the number of neurons at the hidden layer (from 2 to 500): Normalized Bias (left), Scatter Index (center), and Normalized RMSE (right). The red and black curves at the first two columns represent the training and test sets, respectively, showing the results for ten different seeds (initial conditions). The right column shows the dashed line that is the average (over different seeds) result of training set while solid line is the results for the test set. On top of each plot the same error metrics for the GWES control run and the arithmetic ensemble mean (EM) are presented to allow the comparison of results.

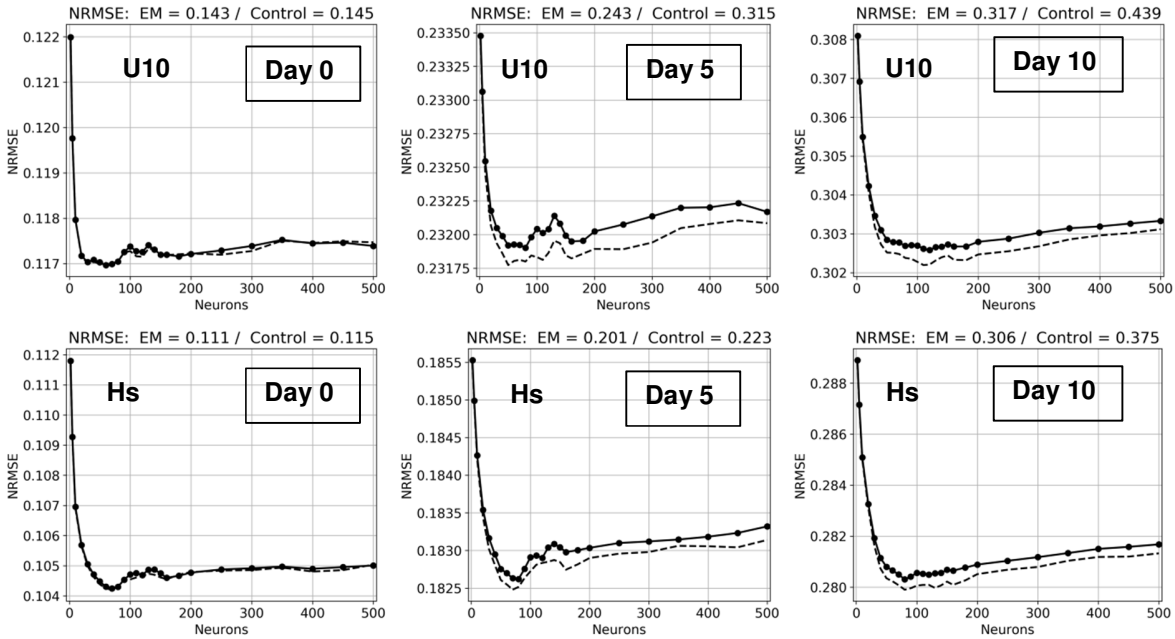

Figure 7 – NRMSE as functions of the number of neurons, for three different forecast horizons. The dashed line is the average (over different seeds) result for training set while solid line is the results for test set. On top of each plot the same error metrics for the GWES control run and the arithmetic ensemble mean (EM) are presented, to allow the comparison of results.

The assessment of 780 NNs averaged through the three different datasets (260 results presented) can be further visualized into the two-dimensional space of scatter and systematic error (Figure 8). All NNs, represented by the cloud of circles in Figure 8, performed better than the EM and control run (red and cyan squares, on the left plots), for both types of errors (systematic and scatter), and indicates the success of the approach. The systematic errors of Hs and U10 presented by NBias are especially small, between -2% and 2%, while the scatter errors are between 19% and 23%. The clouds of training and test points of Figure 8 are close to each other so NNs are not over-fitted and have generalization capability. The selection of the best NN model among the tests relies on a defined criterion assigning scores to each NN based on the error metrics. The right plots of Figure 8 provides a more detailed view of test set results, where the dot size is proportional to the variance of SI through the forecast range divided by the SI, i.e., NNs with good results for the whole range of forecast have small points, while NNs that improved some forecast ranges but not others are depicted by points that have large diameters. This last situation occurs mostly in NNs with few neurons and high values of SI, at the top of the cloud of points of Figure 8. Based on these plots, the best NNs are expected to have more than 30 neurons and can simulate very well a wide range of forecast leads.

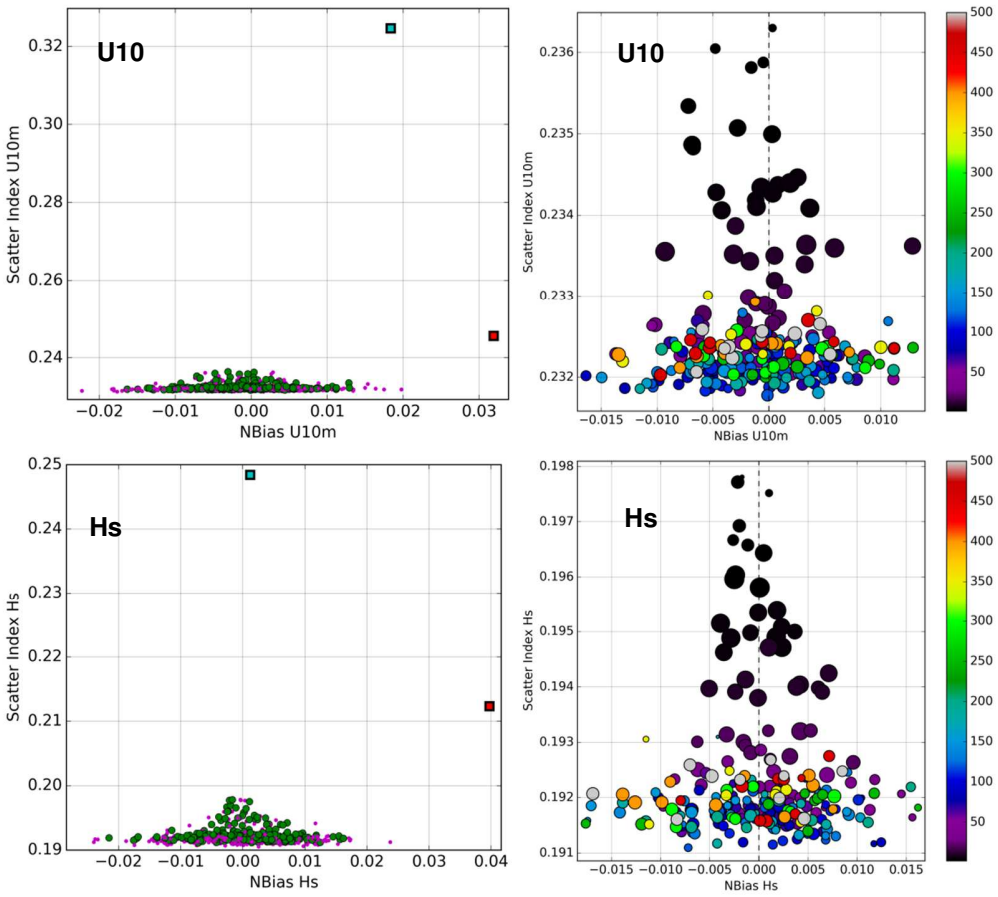

Figure 8 - Results of the neural network tests in terms of the scatter error (y-axis) and systematic error (x-axis). The left plots present the NN results (training set in magenta and test set in green) compared to the control run (cyan square, at the top) and the arithmetic EM (red square). The right plots are a magnification of the clouds of NN results on the test set, where the color indicates the number of neurons and the size of the dots indicates the normalized standard deviation of scatter error throughout different forecast ranges.

The decision about the best NN was based on three steps. The first one restricted the NN results within a maximum systematic error of 0.5% including Hs and U10, as NBias is very small for most of the NNs. A total of 326 NNs of the 780 have absolute NBias of U10 smaller than 0.005, while for Hs this amount is 294. The combined restriction for both variables leads to 144 NNs with extremely small and therefore acceptable biases. The second step sorted the arrays of NNs error metrics, building arrays in ascending order for each type of error, with IDs related to each NN model. The third step looked at the top values of the rank (best results) of correlation coefficient (CC) and SI for Hs and U10, searching for the best NN that minimize the scatter errors of both waves and winds. It has been verified that the top-ranking NN models that minimize certain scatter error metric such as SI, also maximize the correlation coefficient, which makes the final selection much easier. Three NNs were identified with very similar values, from which the best one was selected, containing 140 neurons at the hidden layer. Although the goal of the post-processing simulations is to prioritize Hs, the selection of optimum NNs that also minimize the error of U10 is important, since both output variables are correlated, and the wave generation process depends of the quality of surface winds.

5. Results of Global simulations and discussion

Once the best NN architecture and parameters have been determined, the performance of the selected NN was evaluated using an independent set of altimeter data that was not included in the backpropagation training. This includes 2,507,099 matchups of satellite/GWES distributed over the whole globe. Figure 9 and Figure 10 present global maps of systematic and scatter errors, comparing the EM with the NN nonlinear average. The matchups are grouped in bins (61 latitudes and 181 longitudes) within a radius of 2° to compute the error statistics for each location. NBias of Figure 9 shows a strong spatial dependence of GWES errors, reflecting areas of occurrence of tropical and extra-tropical storms where the atmospheric model data errors are expected to be larger than in other areas, and regions in the tropical ocean exposed to swell systems that may either propagate extratropical-storm wind-field errors or indicate intrinsic wave model source-term biases. In mid and high latitudes, the EM tends to overestimate observed values from altimeters. This is particularly evident in the Southern Hemisphere. In the tropics, the EM tends to underestimate the observed values, with an exception of area along the ITCZ in the Pacific Ocean. The systematic error of the EM varies from -10% to +10% at most locations. The nonlinear ensemble average using NN reduces this bias to values within 5%. The benefit is greater at mid-latitudes dominated by extratropical cyclones where the NBias of the EM can reach 12% for Hs. However, errors along the Equator in the eastern Pacific Ocean are not improved, possibly due to the small correlation of Hs and U10 as illustrated in Figure 1.

The global maps of SI (Figure 10), indicate significant errors for both U10 and Hs at extra-tropical latitudes, again reflecting areas where forcing errors are expected to be larger due to the occurrence of tropical and extra-tropical storms, or to the dominance of swells. The Hs maps present larger errors at western portions of the oceans and, concerning the Southern Hemisphere, the South Atlantic Ocean has larger errors than the Indian and South Pacific Oceans. SI in general reaches up to 40% for U10 and 30% for Hs. The NN provided additional skill that is not restricted to specific locations but distributed over the globe. Comparing Figure 9 and Figure 10 it is possible to conclude that the relative improvement due to the NN on the SI is smaller than the improvement found for NBias. In addition, for practical applications it is important to have the total RMSE with the same unit as the significant wave height (Hs, in meters)

combining both error components – presented by Figure 11. As a final global map, we include the deterministic forecast (control run) to provide an overview of the progress associated with each step. A significant improvement occurs moving from the control run to the ensemble mean (EM), where the RMSE at extra-tropical locations is reduced by approximately 30% and confirms the success of the ensemble methodology described by Zhou et al. (2017) and Alves et al. (2013). The NN post-processing simulation acts especially on the locations with large RMSE at mid-latitudes and provides an additional reduction of 20% at these locations so after the hybrid modeling (neural network attached to the ensemble forecast, illustrated by Figure 3 and equation 3) almost the entire globe has average RMSE of Hs bound to one meter.

In order to further contribute to the spatial discussion of results, Table 2 divides the assessment in five ocean basins using the World Seas database initially described. The differences in performance among the oceans are very small, and the NN is proven to be suitable for all parts of the globe. Table 2 indicates that the ensemble is adding bias to the control run, which is greatly reduced by the NN. In terms of scatter error, the EM significantly reduces the SI of the control run, by approximately 25%, and the NN provides an additional small reduction of 5% to 10% of SI values.

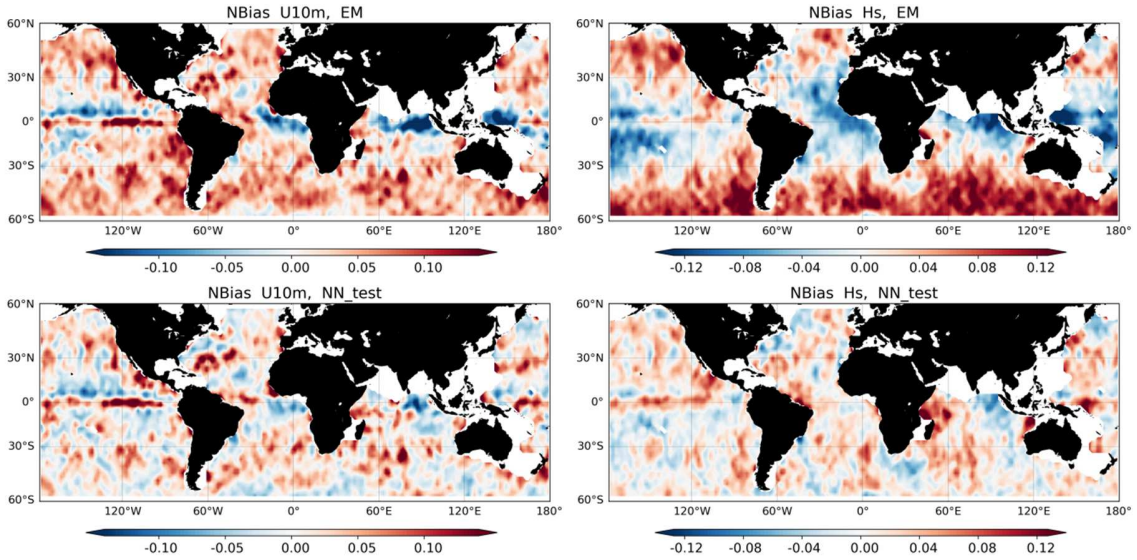

Figure 9 – Global assessments showing the normalized bias (NBias) for GWES ensemble mean (EM, top), and for NN ensemble mean (bottom) on an independent test set. The columns represent U10 (left) and Hs (right). Red indicates overestimation of the model compared to altimeter observations while blue indicates underestimation. Great part of large-scale biases in the mid- to high-latitudes has been eliminated by the NN ensemble mean.

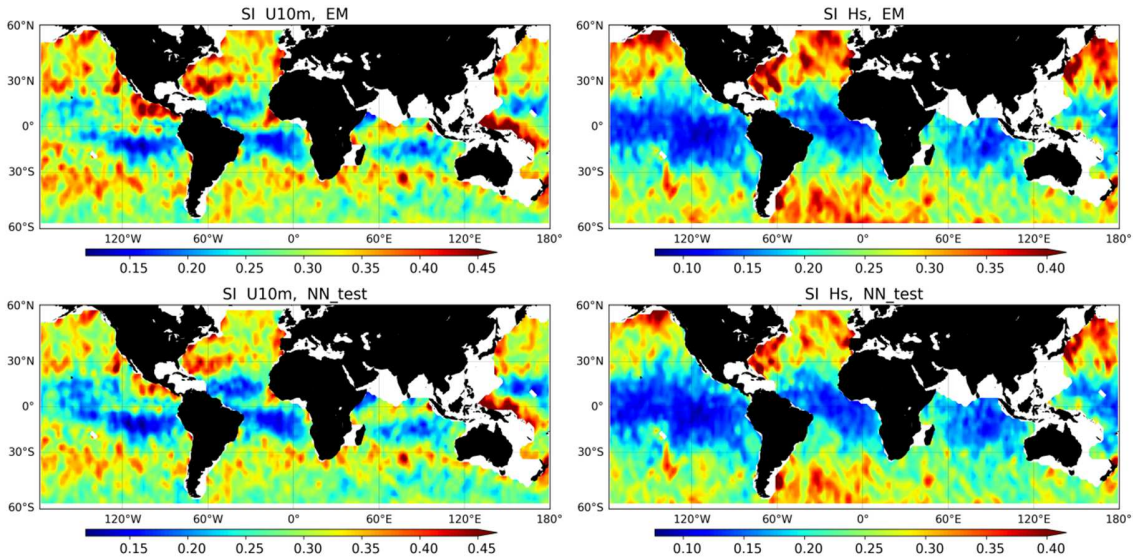

Figure 10 - Global assessments showing the scatter index (SI) for GWES ensemble mean (EM, top), and for NN ensemble mean on an independent test-set (bottom). The columns represent U10 (left) and Hs (right). A reduction of SI is seen in the NN results at some locations.

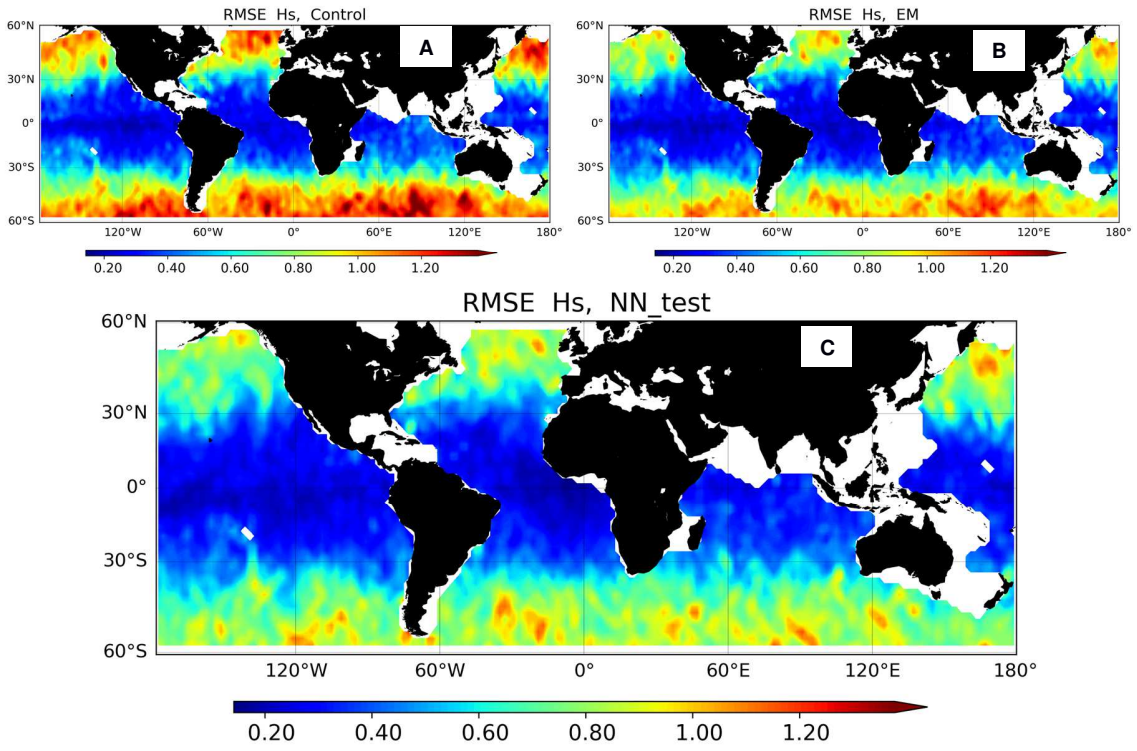

Figure 11 - Final Global assessment maps of Hs showing the RMSE (in meters) for the control run of GWES (A: top-left), the EM of GWES (B: top-right), and the NN post-processing result (C: bottom). It highlights the progressive improvement divided in two steps, first the arithmetic ensemble mean (EM) of the ensemble members compared to the deterministic single run (control), and the neural network post-processing compared to the arithmetic ensemble mean.

Table 2 – Systematic and scatter errors for each ocean, comparing the GWES control run with EM and NNs (test set).

		U10m					Hs				
		NAtlantic	SAtlantic	Indian	NPacific	SPacific	NAtlantic	SAtlantic	Indian	NPacific	SPacific
Nbias	Control	0.016	0.018	0.013	0.010	0.029	-0.031	0.002	0.022	-0.017	0.008
	EM	0.029	0.034	0.030	0.019	0.041	0.001	0.041	0.065	0.017	0.048
	NN-Test	0.007	0.008	0.005	0.006	0.009	0.006	0.007	0.006	0.012	0.005
SI	Control	0.338	0.329	0.314	0.335	0.320	0.265	0.269	0.243	0.248	0.237
	EM	0.258	0.244	0.235	0.259	0.241	0.223	0.229	0.206	0.214	0.202
	NN-Test	0.245	0.231	0.223	0.242	0.229	0.208	0.209	0.183	0.197	0.182

We do not divide the global assessment maps into several figures related to forecast lead days because it would reduce the total data volume of matchups at the bins over the globe. Therefore, Figures Figure 9 and Figure 10 as well as Table 2 integrate the results over the 10-days forecast range and inevitably insert more weight into the analyses and comparisons involving longer lead times associated with larger errors. The final Figure 12 shows the error metrics as a function of the forecast lead-time, providing a meaningful assessment of the nonlinear wave ensemble averaging using NN. A total of 61,149 matchups of satellite/GWES per forecast time is utilized to compose the plots. Figure 12 shows that NBias is reduced to values between 0 to 2% throughout the whole range. This improvement is especially important after the day-7 forecasts, when the control run tends to underestimate, and EM tends to overestimate the observations. The SI plot indicates a small reduction of the error by the NN, equally distributed over the lead times. Taking the right part of the SI plot, associated with the longest horizons, the results of the NN on day-10 has the same error of the EM on day-8, equal to 27%, which represents an extension of 2 days in terms of predictability if the NN averaging is used.

Equation (6) presents the combination of NBias and SI into the NRMSE, also included in Figure 12. The growth pattern is similar to the SI plot, which is expected after comparing the y-axis of NBias and SI plots that indicate much larger errors coming from the scatter component. The correlation coefficient (CC) is the most challenging metric to improve but the NN model was able to slightly improve the values compared with the EM, especially at longer forecast lead times. The comparison of plots in Figure 12 allows one to have a valuable overview of the benefits and shortcomings of the NN post-processing method.

The operational implementation of the post-processing algorithm is simple. Once the NN parameters (a, b of equation 1) and normalization parameters (equation 2) are obtained, the simulation is straightforward, following three steps. (1) The inputs must be downloaded (or linked) from the NCEP ensemble forecast system, which are then normalized and reshaped to build the input array for the NN program; (2) the NN simulation is run covering the latitude, longitude and forecast time, generating the global residue; and (3) outputs invert the initial normalization, the proper shape of the array is rebuilt, and the residues are added to the EM fields of Hs and U10 to construct the final output file (in our case, in netcdf or grib2 format). As described before, the training process is the step that requires more computational power. Daily runs following GWES cycles, however, can be performed by single-core processors (~2GHz) taking approximately five minutes, in Python language. We believe this might be further improved and time consumed can be reduced.

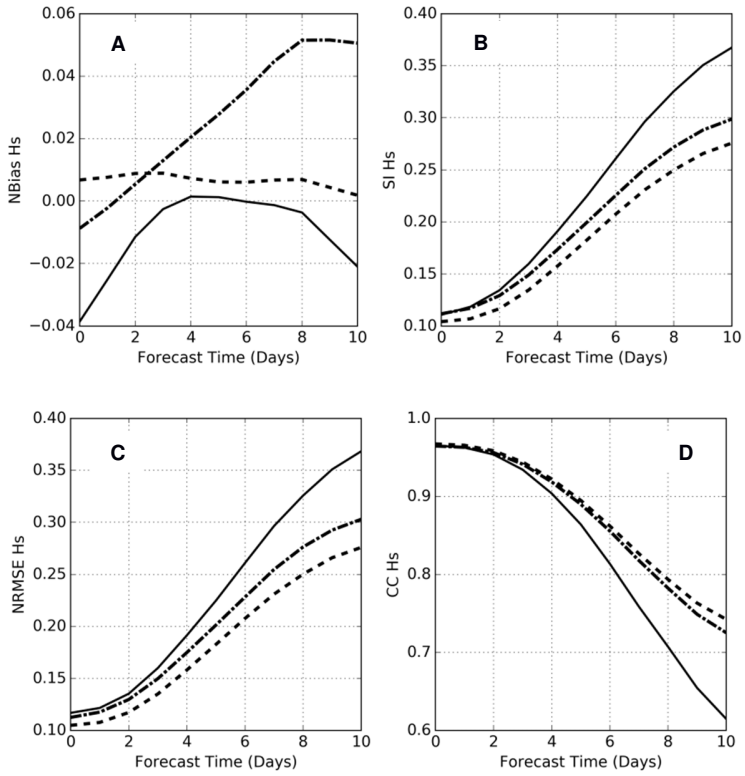

Figure 12 – NBias (A: top-left), SI (B: top-right), NRMSE (C: bottom-left), and correlation coefficient (D: bottom-right) of Hs versus forecast lead time, considering the independent assessment using 2,507,099 matchups of satellite/GWES (61,149 per forecast time). Solid curves show the deterministic run, dashed-point the EM, and dashed curves the nonlinear ensemble averaging using NN.

It is important to stress again that our methodology based on neural-networks does not intend to replace the numerical modeling of the physical processes or the ensemble approach. Instead, we have developed a framework maintaining the ensemble forecast, as provided by NCEP, and including a soft-computing neural network as a post-processing step - linking the traditional modeling with a machine learning algorithm that improves the ensemble mean, trained with a large amount of altimeter data. The post-processing algorithm using multilayer perceptron neural networks is simple enough to be used as a bias correction to deterministic forecasts (same methodology but with only "one member"). However, it is proved that ensemble forecasts significantly reduce the scatter errors at longer forecast ranges (Zhou et al., 2017; Campos et al., 2019b) so the best solution considering the hybrid modeling is to attach the neural network to outputs of ensemble forecast systems.

6. Conclusions

A large set of experiments was conducted to develop neural networks to post-process and bias-correct operational ensemble wave forecasts, where the target variable is primarily Hs followed by U10. The main goal was to build a NN model trained with altimeter data, able to calculate nonlinear ensemble averages that outperform the typical arithmetic ensemble mean, applied to the whole globe and covering a forecast range of 10 days. Simplicity of post-processing algorithms has been a priority during the project. An analysis of 780 NNs facilitated identifying an effective architecture and complexity of the problem, as well as testing the generalization and the distribution of the error with forecast lead times, latitudes, and longitudes.

A previous study by Campos et al. (2019a) focusing on the Gulf of Mexico found the best NN configurations with 35 to 50 neurons in the hidden layer. Expanding to our global simulation, it was found that 60 to 180 neurons produce the best results. The complexity of the NN models, described by equation (9), necessary to address global nonlinear ensemble averages, involves a deeper discussion that depends on the output variable of interest, forecast range, and the type of error. It was shown that minimizing Hs errors require fewer neurons (around 80) than U10 (more than 100). The same is valid for shorter and longer forecast ranges. Simpler NN models with 60 to 80 neurons produce the smallest errors in the nowcast, while 120 neurons are needed when considering the day-10 forecasts. Overall, simple NN models are able to reduce the systematic errors of Hs at short-range forecasts, while NN models with more neurons are necessary to minimize the scatter error of U10 at longer forecast ranges. After a limit around 200 neurons, increasing the complexity of NN models resulted in larger errors and loss of generalization.

The best NN overall configuration was found to have 140 neurons at the hidden layer. Taking the results from one year of simulations (2017), we found that the NNs are efficient in reducing global systematic errors. The average NBias was reduced from an average of 3.5% for the EM to less than 1% globally, which was further confirmed to be valid in all the five oceans analyzed separately. Scatter errors were more difficult to reduce; however, the NNs did provide a small improvement of SI, especially for Hs. The NRMSE combines the systematic and scatter components of error (equation 6), and confirms the effectiveness of the nonlinear ensemble average using global NN trained with altimeter data, which was able to improve the NRMSE throughout the whole range of forecasts. Using the NN-based nonlinear averaging, the day-10 forecasts have the same NRMSE as the day-8 forecasts for the arithmetic ensemble mean – a gain of two forecast days in predictability. We believe that the methodology described can be successfully extended to even longer forecast ranges, which requires a new setup of the operational wave ensemble forecast of NCEP/NOAA that nowadays is limited to 10 days. Further, we believe that integrating the NN methods with coupled atmosphere-ocean-wave forecasts and coupled data assimilation (Penny and Hamill et al., 2017; Penny et al., 2017) may further extend this prediction capability, as well as introducing wave parameters from spectral partitions into the NN inputs, which could benefit lower latitudes with multiple distant swells.

In terms of future developments, besides the extension of forecast horizon, the construction of neural network-based ensembles is a promising example of a growing trend to incorporate machine learning into weather forecasting (Boukabara et al., 2019). The criterion of selecting the best NN among the tests led to the choice of a single NN whereas Krasnopolsky and Lin (2012) showed that multiple NN simulations (developing an ensemble of NNs) produced successful results for precipitation forecasts in the United

States. Specific NN ensembles suitable for extreme wave conditions can also be developed in the future, following the track of the storms, as performed by Campos et al. (2018b), or even building NN members trained for tropical cyclones, which represent a unique family of events. Our study focused on large basins in deep water. Future NN developments are needed to cover coastal areas, lakes, small seas, and locations close to the Arctic influenced by sea ice. Our last suggestion and plan are to include multi-model ensembles in the NN post-processing algorithm, introducing more input variables into the NNs in addition to the NCEP ensemble members, for example: ECMWF, Canadian Meteorological Center (CMC), Fleet Numerical Meteorology and Oceanography Center (FNMOC), and Icosahedral Nonhydrostatic Model (ICON-DWD). We believe that this approach can expand the applicability of post-processing algorithms using neural networks and can significantly improve wind and wave forecast with relative low computational cost.

Acknowledgments

This study has been developed at the Department of Atmospheric and Oceanic Science of the University of Maryland, and at the Environmental Modeling Center of NCEP, funded by the National Weather Service Office of Science and Technology (NWS/OST), Award NA16NWS4680011, with further support in the last stage of development from Fundação para a Ciência e a Tecnologia (FCT – Portugal) under the project EXWAV (RD0504) number PTDC/EAM-OCE/31325/2017. The authors would like to acknowledge Dr. Todd Spindler for giving support with data management and coding, the atmospheric ensemble team at NCEP, and the satellite data provided by AVISO and NESDIS.

Data sources

NCEP's Global Wave Ensemble Forecast:

- ftp://ftpprd.ncep.noaa.gov/pub/data/nccf/com/wave/prod

Altimeters:

- ftp://avisoftp.cnes.fr/AVISO/pub/
- ftp://ftp.star.nesdis.noaa.gov/pub/sod/lsa/cs2igdr/

Distance to the nearest coastline:

- https://oceancolor.gsfc.nasa.gov/docs/distfromcoast/

World Seas database, IHO-Sea-Areas:

- http://www.marineregions.org/downloads.php#iho

References

Alves, J.H.G.M., Wittman, P., Sestak, M., Schauer,J., Stripling, S., Bernier, N.B., McLean, J., Chao, Y., Chawla, A., Tolman, H., Nelson, G., Klots, S., 2013. The NCEP–FNMOC combined wave ensemble product. Expanding Benefits of Interagency Probabilistic Forecasts to the Oceanic Environment. Bulletin of the American Meteorological Society, BAMS, December 2013.

Amante, C., Eakins, B.W., 2009. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA.

Behrens, A., 2015. Development of an ensemble prediction system for ocean surface waves in a coastal area. Ocean. Dyn. 65, 469-486.

Berbić, J., Ocvirk, E., Carević, D., Loncar, G., 2017. Application of neural networks and support vector machine for significant wave height prediction. Oceanologia, 59, 331–349.

Boukabara, S.-A., Krasnopolsky, V., Stewart, J.Q., Penny, S.G., Hoffman, R.N., Maddy, E., 2019. Artificial Intelligence May Be Key to Better Weather Forecasts. Earth & Space Science News: https://eos.org/opinions/artificial-intelligence-may-be-key-to-better-weather-forecasts

Campos, R.M., Krasnopolsky, V., Alves, J.H.G.M, Penny, S.G., 2019a. Nonlinear Wave Ensemble Averaging in the Gulf of Mexico using Neural Networks. Journal of Atmospheric and Oceanic Technology, 36, 113-127.

Campos, R.M., Alves, J.H.G.M., Penny, S.G., Krasnopolsky, V., 2019b. Global Assessments of the NCEP Ensemble Forecast System using Altimeter Data. Ocean Dynamics, ISSN 1616-7341. https://doi.org/10.1007/s10236-019-01329-4

Campos, R.M., Alves, J.H.G.M, Penny, S.G., Krasmopolsky, V., 2018a. Assessments of surface winds and waves from NCEP Ensemble Forecast System. Weather and Forecasting, 33, 1533-1546.

Campos, R.M., Alves, J.H.G.M, Guedes Soares, C., Guimaraes, L.G., Parente, C.E., 2018b. Extreme wind-wave modeling and analysis in the South Atlantic Ocean. Ocean Modelling, 124, 75–93.

Campos, R.M., Krasnopolsky, V., Alves, J.H., Penny, S.G., 2017. Improving NCEP's probabilistic wave height forecasts using neural networks: A pilot study using buoy data. NCEP Office Note 490,

23 pp., https://doi.org/10.7289/V5/ON-NCEP-490.

Cao, D., Tolamn, H., Chen, H.S., Chawla, A., Wittmann, P., 2009. Performance of the Ocean Wave Ensemble Forecast at NCEP. NOAA Marine Modelling and Analysis Branch (MMAB). Technical Note No.279.

Chen, H.S., 2006. Ensemble prediction of ocean waves at NCEP. Proc. 28th Ocean Engineering Conf., Taipei, Taiwan, NSYSU, 25–37.

Dixit, P., Londhe, S., 2016. Prediction of extreme wave heights using neuro wavelet technique. Appl. Ocean Res., 58, 241–252.

Durrant, T.H., Woodcock, F, Greenslade, D.J.M., 2009. Consensus forecasts of modelled wave parameters. Weather Forecast 24, 492–503.

Farina, L., 2002. On ensemble prediction of ocean waves. Tellus 54A, 148–158.

Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics (MOS) in objective weather forecasting. J Appl Meteor 11, 1203–1211.

Harpham, Q., Tozer, N., Cleverley, P., Wyncoll, D., Cresswell, D., 2016. A Bayesian method for improving probabilistic wave forecasts by weighting ensemble members. Environmental Modelling & Software 84, 482-493.

Hoffschildt, M, Bidlot, J.R., Hansen B, Janssen, P.A.E.M., 1999. Potential benefits of ensemble forecasting for ship routing. ECMWF, Technical Memorandum 287.

Hornik, K., 1991. Approximation Capabilities of Multilayer Feedforward Network. Neural Networks, 4, 251-257

International Hydrographic Organization, 1953. "Limits of Oceans and Seas", Special Publication N° .28, 3rd edition.

Janssen, P., Doyle, J.D., Bidlot, J., Hansen, B., Isaksen, L., Viterbo, P., 2002. Impact and feedback of ocean waves on the atmosphere. Atmosphere–Ocean Interactions, N. Perrie, Ed., Advances in Fluid Mechanics, Vol. I, WIT Press, 155–197.

Kalnay, E., 2003. Atmospheric modeling, data assimilation and predictability. Cambridge University Press, 341pp.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

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.

Krasnopolsky, V.M., Lin, Y., 2012. A neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental US. Adv. Meteor., 2012, 649450.

Lorenz, E.N., 1963. The predictability of hydrodynamic flow. Trans. NY Acad. Sci., Series II 25, 409-432.

Mandal, S., Prabaharan, N., 2006. Ocean wave forecasting using recurrent neural networks. Ocean Eng., 33, 1401–1410.

Mentaschi, L., Besio, G., Cassola, F., Mazzino, A., 2013. Problems in RMSE-based wave model validations. Ocean Modelling, 72, 53–58.

Murphy, J.M., 1988. The impact of ensemble forecasts on predictability. Q J R Meteorol Soc 114(480):463–493.

Penny, S.G., Hamill, T.M., 2017. Coupled Data Assimilation for Integrated Earth System Analysis and Prediction. Bulletin of the American Meteorological Society, doi: 10.1175/BAMS-D-17-0036.1.

Penny, S.G., et al., 2017. Coupled Data Assimilation for Integrated Earth System Analysis and Prediction: Goals, Challenges and Recommendations. World Meteorological Organization, WWRP-2017-3. https://www.wmo.int/pages/prog/arep/wwrp/new/documents/Final_WWRP_2017_3_27_July.pdf

Queffeulou P., 2004. Long-term validation of wave height measurements from altimeters, Marine Geodesy, 27, 495-510.

Queffeulou, P., 2012. Preliminary assessment of Jason-2 GDR version D for SWH and sigma0 data, September 2012. Laboratoire d'Océanographie Physique et Spatiale IFREMER. Report available at

ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/documentation/J2_versions_D_T.pdf

Queffeulou P., 2013. Cryosat-2 IGDR SWH assessment update – May, 2013. Laboratoire d'Océanographie Physique et Spatiale IFREMER. Report available at

ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/documentation/

Queffeulou, P., Croizé-Fillon, D., 2017. Global altimeter SWH data set. Laboratoire d'Océanographie Physique et Spatiale IFREMER. Report available at

ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/documentation/altimeter_wave_merge__11.4.pdf

Rasp, S., Lerch, S., 2018. Neural Networks for Postprocessing Ensemble Weather Forecasts. Monthly Weather Review, 146, 3885-3900.

781 Saetra, O., Bidlot, J.R., 2004. Potential Benefits of Using Probabilistic Forecasts for Waves and Marine
782 Winds Based on the ECMWF Ensemble Prediction System. Weather Forecast. 19, 673-689.
783
784 Sánchez, A.S., Rodrigues, D.A., Fontes, R.M., Martins, M.F., Kalid, R.A., Torres, E.A., 2018. Wave
785 resource characterization through in-situ measurement followed by artificial neural networks' modeling.
786 Renewable Energy, 115, 1055–1066.
787
788 Sepulveda, H.H., Queffeulou, P., Ardhuin, F., 2015. Assessment of SARAL AltiKa wave height
789 measurements relative to buoy, Jason-2 and Cryosat-2 data. Marine Geodesy, 38 (S1), 449-465.
790
791 Tolman, H. L., 2016. User manual and system documentation of WAVEWATCH III version 5.16.
792 NOAA/NWS/NCEP MMAB Tech. Note 329, 326 pp.
793
794 Toth, Z., Kalnay, E., 1993. Ensemble forecasting at NMC: The generation of perturbations. Bull. Amer.
795 Meteorol. Soc., 74, 2317-2330.
796
797 Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema,J.J., Klink, K.M., Legates,D.R., O'Donnel, J.,
798 Rowe, C., 1985. Statistics for the Evaluation and Comparison of Models. Journal of Geophysical
799 Research, 90, C5, pp 8995-9005.
800
801 Whitaker, J.S., Hamill, T.M., Wei, X., Song, Y., Toth, Z., 2008. Ensemble Data Assimilation with the
802 NCEP Global Forecast System. Mon. Wea. Rev., 136, 463–482.
803
804 Wu, X., Grumbine, R., 2013. Sea Ice in the NCEP Climate Forecast System Reanalysis. Science and
805 Technology Infusion Climate Bulletin. 38th NOAA Annual Climate Diagnostics and Prediction
806 Workshop.
807
808 Woodcock, F., Greenslade, D. J. M., 2007. Consensus of numerical model forecasts of significant wave
809 heights. Weather and Forecasting, 22, 792–803.
810
811 Young, I.R., Holland, G.J., 1996. Atlas of the oceans: Wind and Wave Climate. Pergamon Press, New
812 York, pp. 241.
813
814 Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J., Wobus, R., 2017. Performance of the new NCEP Global
815 Ensemble Forecast System in a parallel experiment. Wea. Forecasting, 32, 1989–2004.
816
817 Zieger, S., D., Greenslade, Kepert, J.D., 2018. Wave ensemble forecast system for tropical cyclones in the
818 Australian region. Ocean Dynamics, 68, 603–625.
819
820
821
822