U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5830 University Research Court
College Park, MD 20740-3818

Office Note 490

http://doi.org/10.7289/V5/ON-NCEP-490

Improving NCEP's Probabilistic Wave Height Forecasts Using Neural Networks: A Pilot Study Using Buoy Data

Ricardo Martins Campos[1], Vladimir Krasnopolsky[2], Jose-Henrique Alves[3], Stephen Penny[1]

[1]Dept.of Atmospheric & Oceanic Science / University of Maryland
[2]EMC/NCEP / NOAA Center for Weather and Climate Prediction
[3]SRG/EMC/NCEP / NOAA Center for Weather and Climate Prediction

October 2017

MMAB Contribution No. 331

Email: ricardo.campos@noaa.gov; Phone: 1-(301)405-8022

Abstract

This technical note presents preliminary results (or a pilot study) of neural network models applied to produce non-linear ensemble averaging and bias correction of the Global Wave Ensemble System (GWES) of the US National Weather Service (NWS). Our work seeks to improve the skill of GWES products, including significant wave height (Hs), peak wave period (Tp), and 10-m wind speed from the Global Ensemble Forecast System (GEFS). We present an initial strategy, whereby one location in the Atlantic Ocean and one in the Pacific Ocean, both with reliable and quality controlled buoy data, are used to train and test our statistical models at single points. The GWES was evaluated against National Data Buoy Center (NDBC) measurements at these two points; this comparison indicated an increase of forecast errors and spread with time, as well as an increase of error as a function of percentile levels – indicating the value below which a given percentage of observations in a group of observations fall. Among several tested architectures, the best identified neural network model used two layers, each with 11 neurons at the intermediate layer, a hyperbolic tangent basis function, optimization using sequential training, and normalization applying the log function to time series of Hs. Many different random initializations, with different seeds, were found to have a significant impact on the results. An approach based on an ensemble of neural networks was successfully applied, providing an improvement on the 5-day forecast of 64% in the bias, 29% in the RMSE and scatter index, and 11% in the correlation coefficient. A final neural network model was trained to predict the difference of observations minus the ensemble mean, i.e., the ¨error¨ (called residue) of current ensemble average. This approach ensures that no range of values (from calm to extreme events) is deteriorated by the statistical model, and as a consequence expanded the improvement of the neural network post processing to higher percentiles, associated with waves above 2.5 meters.

1. Introduction

The Global Wave Ensemble System (GWES) was implemented in 2005 (Chen, 2006) and initially validated by Cao et al. (2007). It is run operationally at the US National Centers for Environmental Prediction (NCEP) within four daily cycles, with a forecast range of 10 days. The GWES is composed of 20 ensemble members, plus a control member (deterministic). The members are forced by Global Ensemble Forecast System (GEFS) winds on WAVEWATCH III model (Tolman, 2016). The GWES runs on a global domain with 0.5^{o} spatial resolution, with one-hour internal time step, generating point outputs every hour, and gridded fields every three hours. A more complete description of the system and additional validation is provided in Alves et al. (2013). The current operational configuration of the GWES, which provides data used in the present study, includes parameterizations for wind input and wave dissipation using the source terms developed by Ardhuin et al. (2010); nonlinear wave–wave interactions are calculated using the discrete interactions approximations (DIA) of Hasselmann and Hasselmann (1985); and propagation is computed using a third-order accurate scheme (Leonard 1991).

A previous implementation of the GWES, using different wind input and wave dissipation source terms following Tolman and Chalikov (1996), and a global grid with spatial resolution at 1^{o}, was validated by Alves et al. (2013). Model data was compared to a 2-yr-long database (04/2010 to 03/2012) of along-track altimeter measurements of significant wave height (Hs) made by Jason-1, Jason-2, and Envisat. Their results showed that although the general bias of the ensemble system does not show significant improvement over the deterministic global wave, after the fifth forecast day, root mean square errors from the GWES become smaller than the deterministic run. Furthermore, the GWES continuous ranked probability scores (CRPS) systematically outperforms the corresponding deterministic model's mean absolute error (MAE) at all forecast times.

Despite those advances, the GWES still suffers from shortcomings that limit its skill. These shortcomings are evaluated in the present manuscript considering the period from 02/2015 to 02/2017. The typical arithmetic mean is currently used to calculate the ensemble mean (EM) in the GWES. The arithmetic EM for a variable p is currently calculated as,

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

where n is the number of ensemble members and p_i is the i-th ensemble member. Using the arithmetic EM as a 'best representative' of the ensemble assumes that a linear relationship between the EM and ensemble members is optimal. However, in our application this relationship may be strongly nonlinear. An optimized neural network (NN) model can be developed to calculate nonlinear ensemble averages as well as to reduce the GWES error by training the statistical model using reliable measurements. The post-processing methodology proposed here applies the nonlinear statistical model:

$$NEM = NN(p_1, p_2, \cdots, p_n) \tag{2}$$

Further tests also considered the combination of both methodologies, using the NN model to simulate the nonlinear part of the signal together with the model error, appended to the arithmetic EM; i.e., the differences from the current ensemble average (equation 1) from the measurements. In this case, the target variable to simulate is the error signal, or 'residue', of the arithmetic EM compared to observations, as presented by equation (3) and illustrated in Figure 1.

$$NEM = EM + NN_r(p_1, p_2, \cdots, p_n) \tag{3}$$

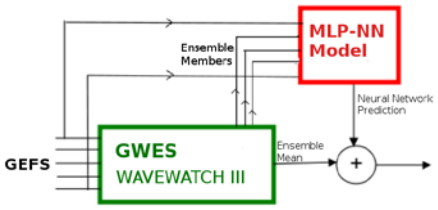

Figure 1 – Scheme of MLP-NN predicting the residue, attached to the GWES model, as represented by equation (3).

The NN models have been constructed based on the theory of Haykin (1999), Krasnopolsky (2013) and Krasnopolsky and Lin (2012). Additional references used are Deo et al. (2001), Deo and Naidu (1999), and Mandal and Prabaharan (2006), who developed network systems to predict Hs in India, and Tsai et al. (2002) in Taiwan. Campos and Guedes Soares (2016) applied the same equation (3), predicting the residue, to simulate Hs in Brazil.

A multilayer perceptron model (MLP-NN) with hyperbolic tangent as the activation function is considered (Krasnopolsky, 2013):

$$NN(p_1, p_2, \cdots, p_n; a, b) = y_q = a_{q0} + \sum_{j=1}^{k} a_{qj} \, . \, tanh\left(b_{j0} + \sum_{i=1}^{n} b_{ji} . p_i \right); q = 1,2, \ldots, m \qquad (4)$$

Here, p_i is the input, y_q is the output, a and b are the NN weights, n and m are the numbers of inputs and outputs respectively, and k is the number of nonlinear activation functions (hyperbolic tangents, or ¨neurons¨). The first summation (sigma) in the right hand side (RHS) of equation (4) represents a linear expansion (a linear combination of hyperbolic tangents), while the second summation (sigma) is the weighted sum of input variables. The combination of both composes a very flexible set of non-orthogonal basis functions that have great potential to adjust to the functional complexity of the mapping to be approximated (Krasnopolsky, 2013). The target variables are significant wave height (Hs), peak wave period (Tp), and 10-meter wind speed (U10); they will be evaluated against buoy measurements during the training process. Hs and Tp are diagnostic variables in GWES, but constitute the main operational guidance products used by forecasters. Thus, the primary goal of this project is better estimation of these quantities. The U10 is provided by the GEFS. Although U10 is a product generated from a separate ensemble system, the introduction of this variable improves the prediction of Hs, due to the high correlation of Hs with U10. The input variables consist of the 21 ensemble members (20 plus the control member), associated with these three variables, as well as the sine and cosine of time (Julian days) to properly include information about the seasonality of the signal. Thus, a total of 21*3+2=65 variables compose the inputs for the NN model. A single forecast time is initially fixed during the first tests, equal to the fifth day, which is approximately the lead time at which ensemble forecasts start to have a better performance than deterministic forecasts (Alves et al., 2013). A range of forecast days from 0-21 days is planned to be considered in future analyses.

2. Input Data and Assessment

The observations used for the present analysis consist of global quality-controlled buoy and altimeter measurements. The Centre ERS d'Archivage et de Traitement (CERSAT) of the French Research Institute for Exploitation of the Sea (IFREMER) is a center that continuously organizes, evaluates, quality controls, and calibrates all publicly available altimeter data, providing standardized netcdf output files. Queffeulou and Croizé-Fillon (2017) describe the data and the methodology applied in the CERSAT/IFREMER satellite wave data processing and quality control. Because the Operational GWES historical data is available from March/2015 to present, we downloaded altimeter data from the CERSAT/IFREMER ftp corresponding to the period from March/2015 until February/2017. However, these data are not available for operational use since there is a delay due to quality control and calibration. The matching period of GWES and altimeter data includes three altimeter missions: JASON2, CRYOSAT and SARAL. Comparisons with buoy data (Queffeulou 2003, 2004) show that the altimeter estimate of Hs is, in general, in agreement with in situ data, showing standard deviations of differences of the order of 0.30 meters. The along-track altimeter data was organized and collocated onto a regular grid (same grid and resolution of GWES) using a kd-tree, following Sepulveda et al. (2015). A maximum distance of 50 km and time lag of 30 minutes to the nearest GWES model grid point is defined, and all altimeter data within this space-time range was averaged. The global concentration of altimeter data during the two year period used for the experiments, from 03/2015 to 02/2017, is presented in Figure 2.

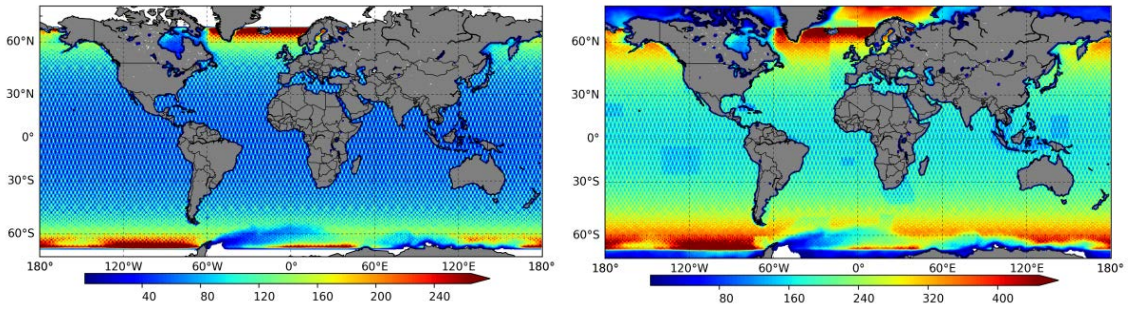

Figure 2 – Quantity of altimeter measurements at each grid point for wind speed (left) and Hs (right). Color bars indicate the total amount of satellite measurements averaged and allocated to the GWES regular grid within the 2-yr period considered.

Buoy measurements were obtained from two sources, the National Data Buoy Center (NDBC, in blued at Figure 3) and the CERSAT/IFREMER database previously described (in red at Figure 3), which has been working in partnership with the Copernicus Marine Environment Monitoring Service, MyOcean, and the World Meteorological Organization (WMO). Both are systematically quality-controlled data. A total of 154 buoys were selected, shown in Figure 3, mostly concentrated in the North America and Europe. The observed quantities obtained from the buoys are Hs, Tp, and wind speed (converted to the 10-m level).

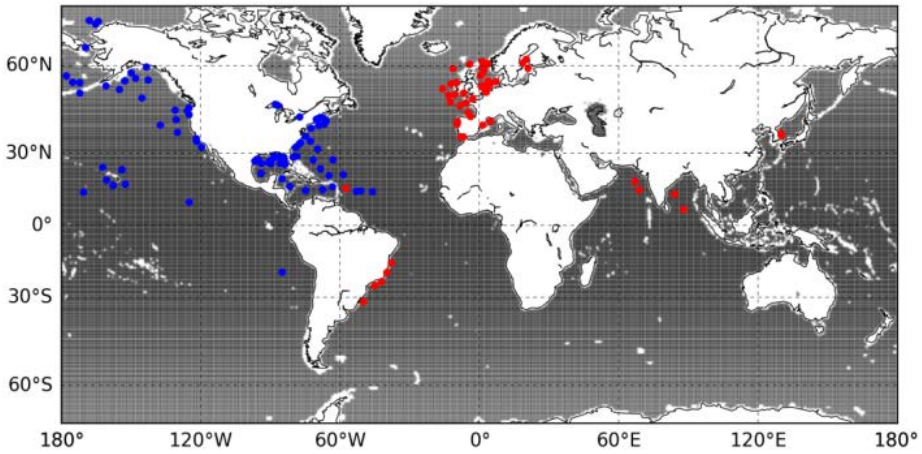

Figure 3 – Buoy positions involving all data obtained from NDBC (in blue) and IFREMER (in red) for the period from March/2015 until April/2017. Coastal and shallow water buoys were excluded from the database. Black points are the GWES grid points and the white areas represent the mask used to exclude land and coastal points.

Observations close to the coast in shallow waters were excluded because of two concerns: (1) Altimeter data present increasing error close to the coast, as explained by Sepulveda et al. (2015), Queffeulou and Croizé-Fillon (2017), and Shanas et al. (2014); and (2) the probability distributions and wave climate change rapidly due to the effect of bathymetry and coastline, which would require a different NN training strategy in these regions. A mask was applied with the exact latitude and longitude of the valid grid points used in the numerical wave model (WAVEWATCH III), excluding coastal and shallow areas to avoid such difficulties. Two sources of global data were used to build the mask: ETOPO1 NOAA´s bathymetry (Amante and Eakins, 2009 - National Geophysical Data Center/Geodas Databases NGDA/GEODAS/NOAA) with one arc-minute of resolution; and distance from the coast database from NASA's Goddard

Space Flight Center, with 0.04 degrees of resolution. Initially, minimum water depths of 10 meters and minimum distance from the coast of 50 km were imposed as constraints.

After compiling the global buoy and altimeter data, we began initial tests and development of MLP-NN models using a single point analysis. In our first set of tests, we first use one year, and later expand to two years of buoy measurements. For the first batch of tests, in order to have an independent validation, another nearby buoy was selected to provide data for the validation set; i.e., optimal MLP-NN parameters are obtained and tested on the principal buoy and then applied and validated using another neighboring buoy. A limitation with this method is the distance between buoys (Table 1) that leads to small differences in the wave climate and can induce NN prediction errors. Therefore, in the second batch of tests, MLP-NN was trained and tested using the first year of measurements and validated using the second year of measurements.

Table 1 – Information about the NDBC buoys selected. In bold are the buoys used for training the NN corresponding to the blue stars in Figure 4.

NDBC buoy	**Ocean**	**Latitude**	**Longitude**	**Depth (m)**	**Data Availability (%)**	**Distance from model grid point (km)**	**Distance between buoys (km)**	**Corr-Coeff of Hs between buoys**
41004	Atlantic	32.501	-79.099	37	99.5	9.3	164	0.89
41013		33.436	-77.743	28	99.7	23.6		
46047	Pacific	32.398	-119.498	1488	98.7	11.3	428	0.81
46028		35.712	-121.858	1048	82.1	26.8		

Buoys with almost no gaps in data and high quality of measurements were selected offshore South and North Carolina / USA. Another pair of buoys was selected in the Pacific Ocean, providing an independent set relative to the Atlantic analyses. These two additional buoys are moored offshore California / USA, and have data at the same high quality levels as the Atlantic buoys. Figure 4 illustrates selected buoy locations, and Table 1 provides details of each point/buoy. The blue stars show the position of the main buoys where the MLP-NN is trained and tested and the cyan dots are the buoys used for validation. The small black dots are the regular grid points of GWES selected after applying the mask previously described. Figure 4 (right) presents the histogram of Hs for buoys in the Atlantic and Pacific Oceans.

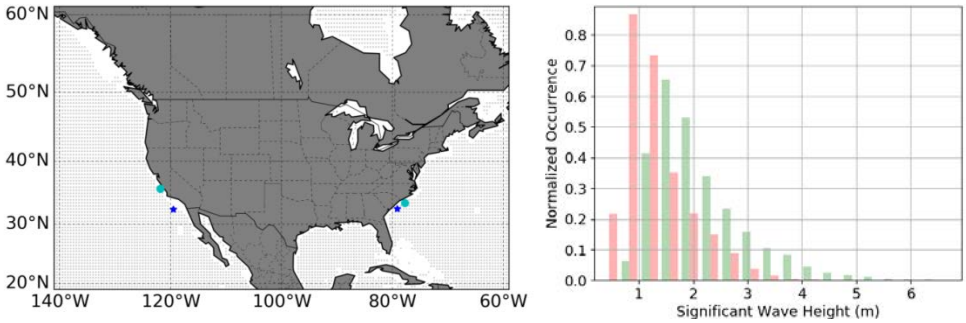

Figure 4 – On the left: Pairs of buoys selected for training, test, and validation sets in the NN model. The blue stars represent buoys selected for training. In the first batch of tests the validation was performed using buoys represented by cyan dots, while in the second batch of tests the validation was performed using the same buoys represented by the blue stars, but in a different year. On the right: Histogram of Hs (meters) for the NDBC buoys 41004 (red) in the Atlantic Ocean and 46047 (green) in the Pacific Ocean.

An assessment of the GWES error is presented in Figure 5. The right figure panels show the increase of error (scatter indexes) as a function of forecast lead time for all ensemble members. Results for the buoy in the Pacific Ocean (top-right) produce errors from 10% to 30% within the 10-day forecast range. The results for the buoy in the Atlantic Ocean (bottom-right) produce errors from 15% to 50%. The left panels show all cycles of 10-day forecasts and their evolution in time. For the same event, we visualize how it is simulated in the 10th forecast day and its improvement and modification when it approaches the analysis (t0) as time goes by. There are specific events that have large biases that are more significant than the increase of bias with forecast time. Therefore, the misrepresentation of certain events (storms) may result in more impactful errors than the expected deterioration of forecast with time. The control member outperformed all GWES ensemble members for all error metrics and variables analyzed (Hs, Tp and U10). The better results of the deterministic model compared to the ensemble simulation, for the same model resolution and set-up, are unexpected and will be further investigated.

One metric to evaluate the GWES and future NN simulations is the error as a function of percentiles and quantiles (Figure 7). This metric is constructed by re-sampling the data moving a minimum percentile level from 0 to 99.9 with several iterations to generate error metrics and curves. In other words, each percentile defines a quantile used as the minimum threshold, selecting all the values above it. Percentile is the value below which a percentage of data falls, from 0 to 99%, while the quantile is the level (in our case, values of Hs, U10 and Tp) associated

with each percentile. In Figure 7 the top X-axis presents the percentiles while the bottom X-axis shows the quantiles associated to the array of percentiles. When data is re-sampled, the length of time-series and thus the statistical confidence is decreased. Therefore, in Figure 7, more data is used to calculate the error metrics for the lower percentiles than the higher percentiles. This reduction of data with increasing percentile/quantile is shown in Figure 6, and must be considered when analyzing the following results in Figure 7. Figure 7 shows the agreement of ensemble members in the first percentiles (calm events) that rapidly diverge and spread above the 50th percentile (during more severe events, such as tropical storms, hurricanes and extra-tropical cyclones). The RMSE systematically increases with the percentiles, for both Hs and U10. The goal of the MLP-NN model is to improve the ensemble averaging without compromising that accuracy at any percentile level.

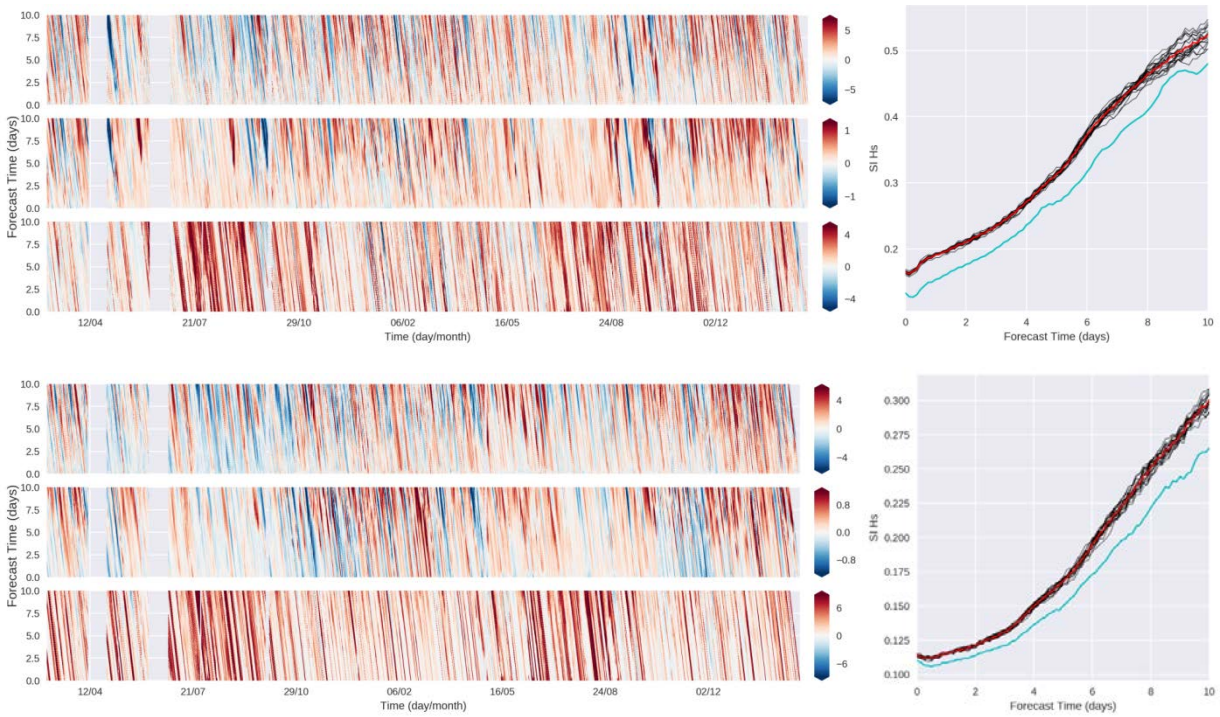

Figure 5 – Evolution of the GWES error with forecast time (up to 10 days); the top sub-figures related to buoy 41004 (Atlantic) while the bottom sub-figures are related to buoy 46047 (Pacific). The left plots: two years (x-axis) of GWES bias containing the whole forecast range (y-axis) and showing the average of the 21 ensembles. Among the three lines of subplots, top: U10 (m/s), middle: Hs (meters), bottom: Tp (seconds). The right plots: Scatter Index (SI) of Hs; in black the 20 ensemble members, in cyan the control member (deterministic) and in red the 20 ensemble average. Small values of SI indicate better results.

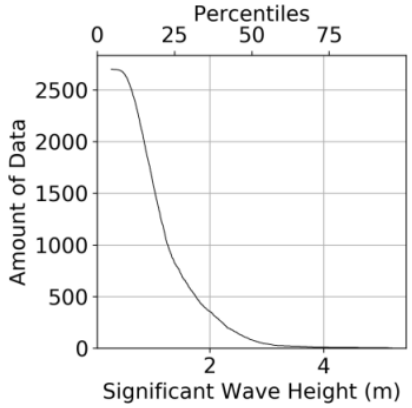

Figure 6 - Amount of data above each percentile and quantile of significant wave height.

The error metrics adopted in this technical note are calculated following the equations:

$$Bias = \frac{\sum_{i=1}^{n}(R_i - S_i)}{n} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(R_i - S_i)^2}{n}} \tag{2}$$

$$SI = \frac{RMSE}{\bar{S}} \tag{3}$$

$$r = \frac{\sum_{i=1}^{n}(S_i - \bar{S})(R_i - \bar{R})}{(\sum_{i=1}^{n}(S_i - \bar{S})^2 \sum_{i=1}^{n}(R_i - \bar{R})^2)^{1/2}} \tag{4}$$

where S are the measurements (buoy data) and R the simulation results (from GWES or from a NN). The overbar indicates mean values through time, and n is the number of observations.

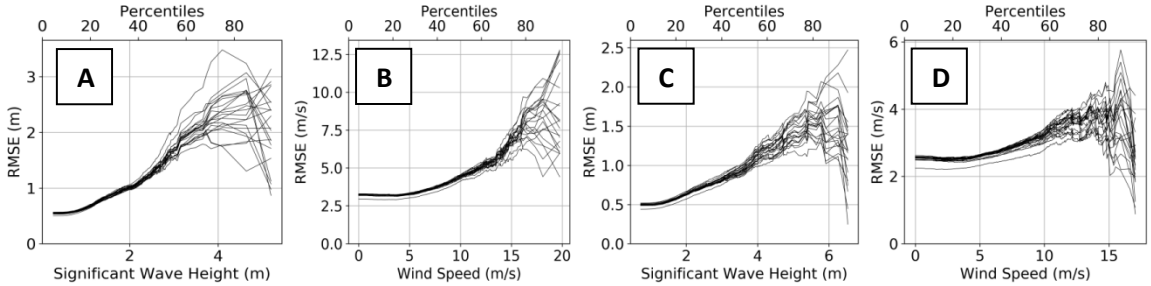

Figure 7 - Error as a function of percentiles and quantiles. Atlantic Buoy 41004 (A,B); Pacific Buoy 46047 (C,D).

3. Neural Network Testing and Evaluation

Several MLP-NN models were constructed and compared, testing different architectures and pre-processing. As a regression model, the NNs contain the nonlinear basis function ($tanh$) in the intermediate layer and linear basis function for the output layer; with a structure of two layers which is sufficient for the complexity of the problem. A back-propagation training algorithm (gradient descent) was applied using the 65 input variables previously explained (e.g., ensembles of Hs, Tp and U10 plus the time variables) and 3 variables (e.g., Hs, Tp and U10) from the buoys, considered ¨ground truth¨ and targeted by the model. The derivative of the non-linear neurons used for training is $1 - tanh^2$ and a learning rate of 0.001 was set – which is modified inside the NN code using an adaptive rate to better converge to the local minima. Sequential training was shown in previous MLP-NN tests to give better results than the Batch Training (Hsieh 2009), therefore sequential training was applied. For sequential training, weights and biases are sequentially updated for every new index of the time series. Approximately 1,000 epochs, each considering the entire 2-yr period, were used in the training process.

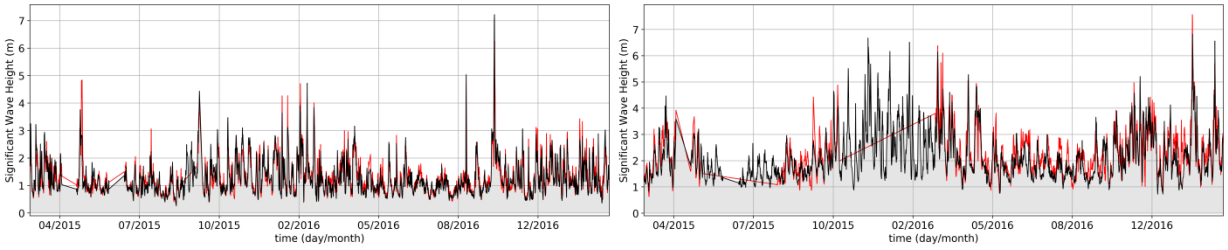

Figure 8 – Time series of Hs (meters) for the two pairs of buoys. Left: Buoy NDBC 41004 (black) and 41013 (red). Right: Buoy NDBC 46047 (black) and 46028 (red).

All input and output variables were normalized to the interval between 0 to 1 ($x_i^{[0,1]}$ of equation 5) when working with Hs, Tp and U10, and to the interval between -1 to 1 ($x_i^{[-1,1]}$ of equation 5) when working with the residue of Hs, Tp and U10 (e.g., the difference from buoy measurements to the arithmetic EM). It has been confirmed that applying the log function to Hs leads to better results due to the more homogeneous distribution of values. The signal is de-normalized at the end of the program and metrics are calculated against observations (Figure 8).

$$x_i^{[0,1]} = \frac{\left(x_i - x_i^{min}\right)}{\left(x_i^{max} - x_i^{min}\right)} \qquad x_i^{[-1,1]} = \frac{2.\left(x_i - x_i^{min}\right)}{\left(x_i^{max} - x_i^{min}\right)} - 1 \tag{5}$$

3.1. Nonlinear Means

Convergence and sensitivity tests were conducted, changing the number of neurons in the intermediate layers and excluding/including variables (Hs, Tp, and U10). Data are divided into training, test, and validation sets. The weights and biases of MLP-NN models are obtained using the training set only. When only one year of data is used, the test set is taken as a sub-sample of the total data set (and not used for training), but is still related to the same buoy and time range. This means the training and test sets have very similar probabilistic moments and distributions. Two thirds of the data are taken for the training set and one third for the test set. The number of neurons was changed from 1 to 50 and the bias, root mean square error (RMSE), scatter index (SI), correlation coefficient (CC), scatter plots and graphics of error in function of percentiles were used to evaluate the results.

Table 2 – Results of Hs (meters) assessment for the MLP-NN in the Atlantic Ocean using buoy 41013 (validation set), comparing number of neurons. EM is the arithmetic mean (equation 1) of the ensemble members. NN-validation is the NN model trained using data from buoy 41004 and run and validated against buoy 41013. The last two lines are related to the assessment in the Pacific Ocean, where the MLP-NN was trained using buoy 46047 and validated against buoy 46028.

	neurons	Set	bias	RMSE	SI	CC
Atlantic		EM GWES (41013)	0.026	0.445	0.334	0.759
	1	NN-validation (41013)	0.158	0.478	0.359	0.722
	11	NN-validation (41013)	-0.011	0.420	0.315	0.765
	20	NN-validation (41013)	0.005	0.434	0.326	0.747
Pacific		EM GWES (46028)	0.227	0.550	0.237	0.860
	11	NN-validation (46028)	-0.018	0.481	0.207	0.846

Table 2 shows the results of the arithmetic EM, as this quantity has been computed operationally, and the results from the NN nonlinear averaging using the scheme represented by equation 2. The errors are presented for the validation set, independent from the training set. Table 2 and Figure 9 (first line) indicate that one or few neurons are not sufficient. All error metrics are worse than the simple arithmetic EM. By increasing the number of neurons, results begin to improve. However, the model deteriorates as shown by Table 2 as the number of neurons is increased to 20. It was found after experimentation that the best results for this NN configuration are given with 11 neurons, at which point all error metrics of Table 2 are improved

compared to the arithmetic EM. The bias of 1 cm, RMSE of 42 cm, percentage error of 31%, and correlation of 0.76 are slightly better than the arithmetic EM. Figure 9 panels A, B and C point to a larger disagreement between measurements and simulation, especially in the validation set; whereas, in Figure 9 panels E, F, and G the cyan dots (NN-validation set) are more confined to the principal diagonal (dashed black, representing the perfect agreement) and contained within the cloud of black points (arithmetic EM). The better results of the training set (blue dots) relative to the validation set (cyan dots) in Figure 9 panels A, B and C, indicate another problem of using one neuron, associated with the low robustness – a fundamental challenge for NN model generalization.

Additional tests applied the MLP-NN with 11 neurons tested the importance of each input/output variable to the training process, keeping Hs as the main variable. By removing the wind information (U10) the results degrade, which reflects the relatively high correlation of Hs with U10, equal to 0.71. When the information of Tp is removed, the results are relatively unchanged, with a difference of only 0.2% in the RMSE. This reflects the low correlation between Hs and Tp for the buoy´s position, equal to 0.07. This is relevant for the next steps of the project, when satellite data, which does not provide Tp, will be included. Considering the single point analyses described above, the NN model with architecture using 11 neurons, Hs/U10 or Hs/Tp/U10 as input, normalization applying log function to Hs values, and using sequential training, has been found the most accurate configuration. This model was applied to the Pacific Ocean, where the NN weights and biases were re-trained using buoy 46047, and validated against buoy 46028. Results are presented in the last two lines of Table 2, showing the improvement of all error metrics apart from the correlation coefficient. Decreased values of bias, RMSE, and SI, combined with increased correlation coefficients indicate improved results.

In Table 2, we see that the MLP-NN model reduces the bias and results in a small improvement of 5% to 12% of RMSE for the NN model compared to the arithmetic EM, though there is deterioration of the correlation coefficient at buoy 46028. Figure 9C shows that the NN model is sensitive to sampling errors. The NN benefits the ensemble averaging only in the interval between 0 to 2 meters of wave heights, which was expected since it is an interval with a large number of observations (see Figure 4 and Figure 6). Extreme events, by nature, are rare, and provide a minor contribution during the NN training. The result is a major deterioration of the error metrics for wave heights above 3 meters, as shown by Figure 9C. We expect

improvements when moving to a spatial analysis in regions with a high density of buoys and with the inclusion of satellite data, as these new data will increase the amount of observations associated with extreme events.

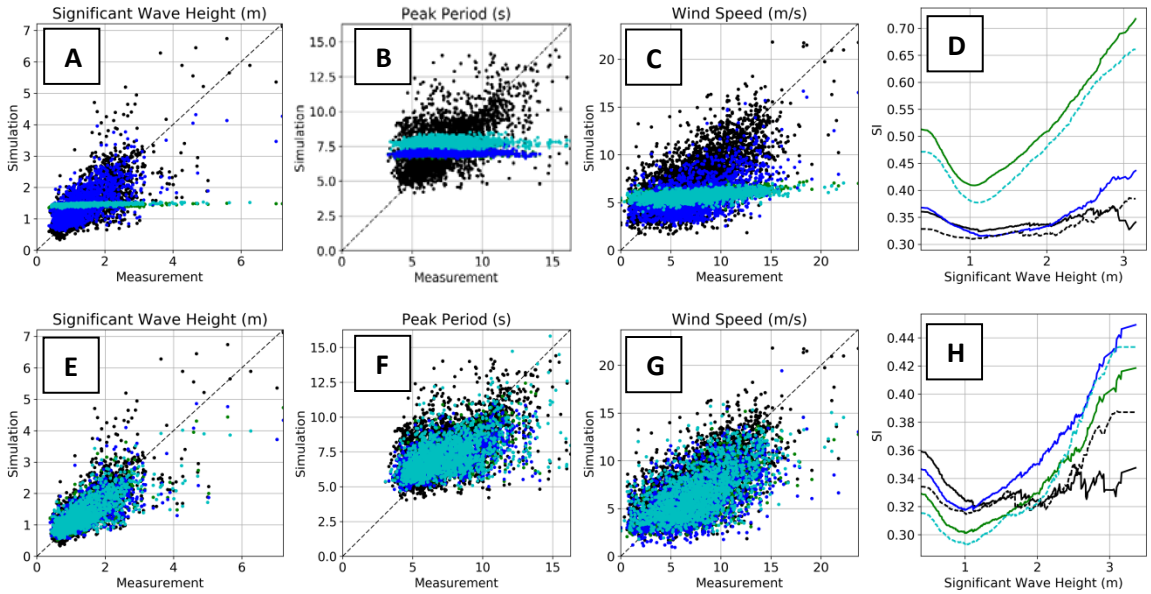

Figure 9 – Comparison of MLP-NN models in the Atlantic Ocean using 1 neuron (first line) and 11 neurons (second line) at the intermediate layer. The first three columns are scatter plots of Hs (meters), Tp (seconds) and U10 (m/s). The last column (D and H) are the curves of scatter indexes as function of quantiles. The black curves are the current arithmetic ensemble mean (EM), considered the control, that is intended to be improved by the NN models. In blue (dots and curves): NN-training set. In green: NN-test set. In cyan (dots and curves): NN-validation set. Solid lines indicate buoy 41004 (training and test sets) and dashed lines buoy 41013 (independent validation set). The target outcome in plots D and H is to have small values of SI of curves green and cyan, located bellow the black lines related to the arithmetic ensemble mean. This happens in plot H in the range between 0 to 2 meters. Above 2 meters, in this experiment, the NN performs worse than the arithmetic EM.

Additional methodologies were tested to improve the results and to ensure that, even when NN model does not improve the skill of the GWES, it at least preserves the skill of the simple arithmetic EM (equation 1). Two new approaches are applied, one using a set of NN ensembles, following the methodology of Krasnopolsky and Lin (2012), and another using the residue signal (equation 3; the NN is trained to predict the difference of GWES arithmetic mean from the measurements) as explained before.

3.3. Residue Signal

The purpose of the NN applied to the residue is to use the arithmetic EM (equation 1) of the GWES to maintain the linear processes that are sufficiently represented by a simple arithmetic mean of the ensemble members, while using the NN to simulate the deviations from this average due to nonlinear processes as observed via the measurements (Figure 1). Thus, the NN model is trained to predict the signal of the error of the current GWES ensemble mean. It is based on the assumption that GWES mean has a reasonably good skill and that the information of the error (residue) is partially contained in the ensemble members through a nonlinear relationship; i.e., the ensemble members are used to predict the error of their own arithmetic EM. Equation 3 represents the current approach.

A problem initially found is that the difference of measurements minus GWES mean is very noisy, with several peaks in the time series. This is a major difficulty for the optimization procedure used by the NN. A spectral analysis applied to the time series of Hs as well as the time series of the residue pointed to a small variance for periods below 12 hours, with the most important part of the signal associated with peaks around 24 to 48 hours, and other important peaks from 4 to 8 days. Therefore, a moving average filter of 24 hours was applied to the signal of the residue of Hs and U10 before the NN training process.

The same method and NN architecture was applied as in the previous experiments, and the results are presented in Table 3 and Figure 10. From Table 3 the bulk statistics do not show significant improvement compared to the last approach, based on equation 2 and using the NN to calculate directly the ensemble mean. The only error metric improved was the correlation coefficient. However, Figure 10 presents the error in function of the sea severity and shows the real benefits of using the NN applied to residue. The black line of Figure 10A presents the filtered signal of the error of the GWES mean compared to buoy 41013 in the Atlantic Ocean, which is the new target variable to predict. The cyan line of Figure 10A is the ¨prediction of the error¨ provided by the NN model, which is later added to the GWES arithmetic mean to provide the final forecast of Hs and U10. Figure 10B and C show a great reduction of the error with increasing quantiles, while cyan and blue lines (NN results) are below the black lines (GWES mean), indicating smaller scatter indexes and better result of NN compared to the arithmetic ensemble mean.

Comparing Figure 10B with Figure 9H we see better performance at the higher ranges of Hs, which are associated with severe sea states. When the NN is trained to predict the residue, it is weighted towards correcting cases in which the error is high, usually related to stormy conditions. When the GWES arithmetic mean is accurate, the residue is near zero and the influence on the NN training is small. Thus, the NN-residue has improved upon the poorly sampled cases from the NN experiments previously described.

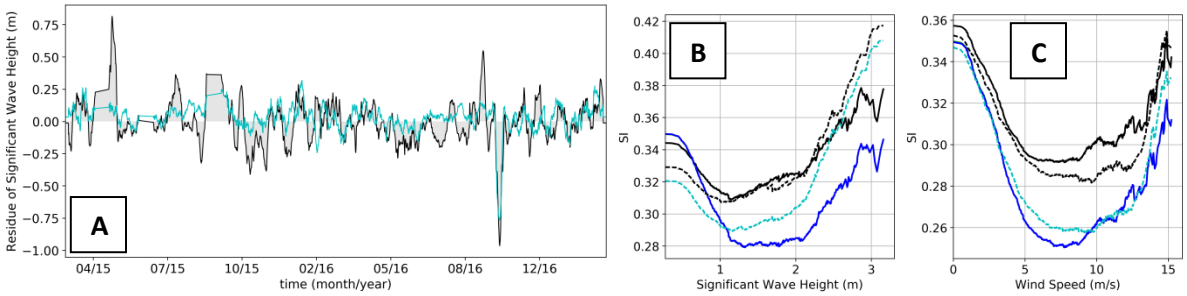

Figure 10 – Results of the NN simulation at the two Atlantic Ocean buoys. A: time series of filtered residue of Hs (meters) in black (buoy measurement minus the GWES arithmetic EM) and the predicted residue in cyan, for the independent validation set at buoy 41013. B and C: curves of scatter indexes in function of quantiles; in black: arithmetic mean of ensembles, in blue: NN-training set (buoy 41004), in cyan: NN-validation set (buoy 41013) - solid lines indicate buoy 41004 (training and test sets), and dashed lines buoy 41013 (independent validation set). The goal of plots B and C is to have small SI values of curves cyan and blue, located bellow the black curves.

Table 3 – Results of Hs (meters) assessment for the MLP-NN in the Atlantic Ocean at buoy 41013 (validation set), using the residues approach. First line is the control simulation with the current arithmetic mean of the members. The second line is the NN applied in the last item, using equation 2. The third line is the results with NN-Residue based on equation 3.

Set	bias	RMSE	SI	CC
EM GWES	0.026	0.445	0.334	0.759
NN (equation 2)	-0.011	0.420	0.315	0.765
NN-Residue (equation 3)	0.126	0.440	0.330	0.771

3.2. NN Ensembles

Table 4 shows the assessment for buoy 41004 considering four different results. As a baseline, we present the individual GWES ensemble member that produced the best error metrics compared to all other ensemble members. The ¨best member¨ in this case was found to be the deterministic control member, which passes through a bias correction algorithm. A linear

regression model was fitted to the same inputs of the NN model as a control experiment in order to compare the methodologies. Results from the linear regression are better than the arithmetic EM but still worse than the NN simulations, indicating the presence of nonlinearities in the signal and supporting the use of the NN approach. Finally, the NN ensemble that considers the five best individual NN models provides the best results, significantly improving upon than the arithmetic EM of the GWES (second line). For the 5-day forecast, there is a 64% reduction in bias, 29% reduction in RMSE and SI, and an 11% increase in the CC. Figure 11 further indicates that the NN ensemble (red dots) is in closer agreement with the diagonal of perfect agreement.

Table 4 – Comparison of different simulations against buoy 41004, for Hs (meters). The ¨Best Member GWES¨ is provided as a baseline. The EM is the arithmetic ensemble mean of the GWES members, without any correction. A linear regression statistical model is tested and presented. The last line shows the results for a 5-member NN ensemble model.

41004	bias	RMSE	SI	CC
Best Member GWES	-0.101	0.526	0.427	0.724
EM GWES	-0.115	0.457	0.371	0.755
Linear Regression model	0.094	0.433	0.352	0.739
NN-ensemble (5 members)	0.041	0.373	0.303	0.807

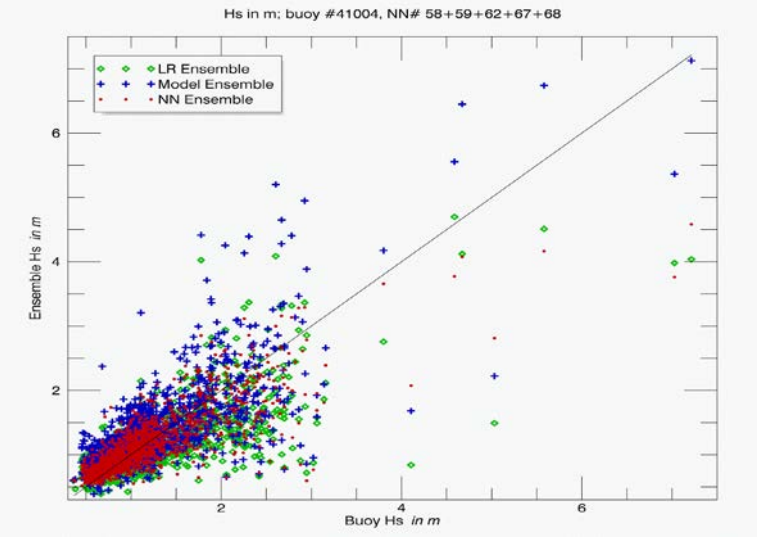

Figure 11 – Scatter plot of simulation results against buoy 41004 measurements, for Hs (meters). In green: Linear regression model. In blue: arithmetic mean of GWES members. In red: NN ensemble.

4. Final Discussion

This technical note presents developments and progress for the first 8 months of the NOAA Next Generation Global Prediction System (NGGPS) project ¨Improving Global Wind-Wave Probabilistic Forecasts and Products Beyond Week 2¨, (NA16NWS4680011), as follows:

1. We conducted initial data mining and organizing of all buoy and satellite data for the period from 03/2015 to 02/2017. The final database consists of reliable and QC'ed observations of Hs, Tp and U10.
2. We have implemented a strategy for the initial tests with MLP-NN models, considering two buoys in the Atlantic and two in the Pacific Ocean, where GWES was also evaluated. Results show an increase of GWES errors and ensemble spread with forecast time, as expected, while the control member produces better results than the ensemble members. This behavior has been discussed with scientist of EMC/NCEP but must be further investigated. The assessment results indicate that errors associated with stormy events tend to be higher than the usual increase of error with forecast time, which led to a different strategy for the NN training, instead using the residue of the variables versus the arithmetic EM.
3. The theory and architecture of the NNs were investigated. The best NN model found has used two layers with 11 neurons at the intermediate layer, a hyperbolic tangent basis function, optimization using sequential training, and normalization using equation (5), and applying the log function to the Hs time series. This solution proved to be successful in both the Atlantic and Pacific basins, off the east and west coasts of the USA. Moreover, it was found that the information of wave period (Tp) does not improve nor deteriorate results; therefore, future analyses will consider only Hs and U10 for NN output variables.
4. Three different types of MLP-NN models were developed and evaluated:
    - The first was nonlinear ensemble averaging (equation 2), which reduced by 5% to 12% the RMSE of the fifth day forecast of Hs but has the improvement confined to waves only from 0 to 2.5 meters, as a consequence of small amount of data at larger wave heights.
    - The second approach considered the NN trained to predict the difference of observations minus ensemble mean (equation 3 and Figure 1). The benefit of this

model was to reduce the impacts of sampling error at high wave heights, so that the statistical model improves the skill in all wind-speed and wave-height ranges, from calm to extreme events, even when the amount of data at a certain interval is small. Figure 10B shows that for all ranges of Hs the NN improved the skill of the model, with a benefit of 10% to 15% for most values. Additionally, the NN model improved the results for U10 as indicated by Figure 10C. Since bulk statistics were not generally deteriorated relative to other NN approaches, and a significant improvement was achieved for larger waves, this is a promising strategy that will be evaluated further in future work.

- The third approach, which produced the best results in terms of bulk statistics, considered an ensemble of the best five individual NN models. When compared to the GWES ensemble mean of Hs, the NN-ensemble reduced in the bias by 64%, RMSE and SI by 29%, and increased the CC by 11%. More than 40 random initializations of the NN model were tested, with different random number seeds, leading to differences in performance and convergence of the statistical model. The selection of the best NN models and their averaging represents an important step, as discussed by Krasnopolsky and Lin (2012) regarding the benefits of using ensembles of NN models. Therefore, the ensemble of NNs outperforms the GWES for all variables analyzed (Hs, Tp, and U10). A Linear Regression model was also tested, giving a small improvement to Hs but not to Tp or U10.

5. Conclusions and Next Steps

MLP-NN models with optimized architecture applied to single points with reliable buoy measurements have been validated in both the Atlantic and Pacific Oceans, using two years of NDBC data. Codes were developed in FORTRAN and python languages, giving very similar results, and providing the option to take advantage of sophisticated NN libraries available in python. For example, the same python code has been recently modified to use the scikit-learn module (python), with very similar results again but running much faster. It provides the support and confidence to move on to the next step of the research, when a large geographic area will be defined and the NN trained using buoy and satellite data; once again selecting two basins in the

Atlantic and Pacific Oceans. The methodology of the MLP-NN models addressed in this report is simple, and the software developed is also simple and straightforward, which should help to facilitate operational implementation. The output of optimal NN training consists of two matrices of weights and two vectors of biases (equation 4) and the NN can easily be re-trained when necessary, given any future modifications of the GEFS or GWES. Thus, we anticipate the results and software will maintain their relevance as NCEP transitions to the Unified Global Coupled System (UGCS) for many of its prediction products at multiple timescales.

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.

Campos, R.M., Guedes Soares, C., 2016. An hybrid model to forecast Hs. In: Guedes Soares, C., Garbatov, Y., Sutulo, S., Santos, T.A. (Eds.), Maritime Technology and Engineering. Taylor and Francis Group, CRC Press, London, pp. 473–479, ISBN 978-1-138-03000-8, DOI: 10.1201/b21890-138. http://www.crcnetbase.com/doi/pdfplus/10.1201/b21890-138

Cao, D., Chen, H.S., Tolman, H.L. 2007. Verification of ocean wave ensemble forecasts at NCEP. Proc. 10th Int. Workshop on Wave Hindcasting and Forecasting and First Coastal Hazards Symp., Oahu, Hawaii, Environment Canada, G1.

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

Deo, M.C., Sridhar Naidu, C., 1999. Real time wave forecasting using neural networks. Ocean Engineering, 26, 191–203.

Deo, M.C., Jha, A., Chaphekar, A.S., Ravikant, K., 2001. Neural networks for wave forecasting. Ocean Engineering, 28, 889–898.

Hasselmann, S., Hasselmann, K. 1985. Computations and parameterizations of the nonlinear energy transfer in a gravity-wave spectrum, Part I: A new method for efficient computations of the exact nonlinear transfer integral. J. Phys. Oceanogr., 15, 1369–1377.

Haykin, S., 1999. Neural Networks, A Comprehensive Foundation. 2nd Edition, Prentice Hall. ISBN-9780132733502.

Hsieh, W.W. 2009. “Machine learning methods in environmental sciences”, Cambridge University Press, Cambridge.

Krasnopolsky, V., 2013. ¨The Application of Neural Networks in the Earth System Sciences. Neural Network Emulations for Complex Multidimensional Mappings¨, Springer, 200p.

Krasnopolsky, V., Lin, Y. 2012. A Neural Network Nonlinear Multimodel Ensemble to Improve Precipitation Forecasts over Continental US. Advances in Meteorology, Volume 2012, Article ID 649450, 11 pages, doi:10.1155/2012/649450

Leonard, B. P., 1991: The ULTIMATE conservative difference scheme applied to unsteady one-dimensional advection. Comput. Methods Appl. Mech. Eng., 88, 17–74.

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

Queffeulou P., 2003. Long-term quality status of wave height and wind speed measurements from satellite altimeters. Proceedings of the ISOPE conference, Honolulu, Hawaii, USA, May 25-30.

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

Queffeulou, P., 2009. Altimeter wave height measurements - validation of long time series. Poster, OSTST meeting, Seattle, June 2009. Report available at http://www.aviso.altimetry.fr/en/user-corner/science-teams/ostst-swt-science-team/ostst-2009-seattle/posters.html

Queffeulou, P., 2013a. Cryosat-2 IGDR SWH assessment update – May, 2013. Report Cryosat2_igdr_swh_assessment_update.pdf available at ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/documentation/

Queffeulou P., 2013b. Merged altimeter wave height data base. An update. Proceedings of ESA Living Planet Symposium, 9-13 September 2013, Edinburgh, UK, ESA SP-722 December 2013, ESA Communications, ESTEC, PO Box 299, 2200 AG Noordwijk, The Netherlands. Report available at ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/documentation/publications/ESA_Living_Planet_Symposium_2013.pdf

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

Sepulveda, H.H., Queffeulou, P., Ardhuin, F., 2015. Assessment of SARAL AltiKa wave height measurements relative to buoy, Jason-2 and Cryosat-2 data. Marine Geodesy, 38 (S1),449-465, doi: 10.1080/01490419.2014.1000470. Report available at ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/documentation/publications/Sepulveda_etal_2015.pdf

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

Tolman, H.L, Chalikov, D.V., 1996. Source terms in a third-generation wind-wave model. J. Phys. ceanogr., 26, 2497–2518.