Wave height prediction at the Caspian Sea using a data-driven model and ensemble-based data assimilation methods

Ahmadreza Zamani, Ahmadreza Azimian, Arnold Heemink
and Dimitri Solomatine

ABSTRACT

There are successful experiences with the application of ANN and ensemble-based data assimilation methods in the field of flood forecasting and estuary flow. In the present work, the combination of dynamic Artificial Neural Network and Ensemble Kalman Filter (EnKF) is applied on wind-wave data. ANN is used for the time propagation mechanism that governs the time evolution of the system state. The system state consists of the significant wave height that is affected by wind speed and wind direction. The relevant inputs are selected by analysing the Average Mutual Information. By help of the observations, the EnKF will correct the output of the ANN to find the best estimate of the wave height. A combination of ANN with EnKF acts as an output correction scheme. To deal with the time-delayed states, the extended state vector is taken and the dynamic equation of the extended state vector is used in EnKF. Application of the proposed scheme is examined by using five-month hourly buoy measurement at the Caspian Sea and several model runs with different assimilation–forecast cycles. The coefficient of performance and root mean square error are used to access performance of the method.

Key words | assimilation forecast cycle, data assimilation, dynamic neural network, ensemble Kalman filter, wind-waves

Ahmadreza Zamani (corresponding author)
Ahmadreza Azimian
Department of Mechanical Engineering,
Isfahan University of Technology,
Isfahan 84156,
Iran
Tel.: +98 311 3912539
E-mail: *arzamani@cc.iut.ac.ir*

Arnold Heemink
Delft Institute of Applied Mathematics,
Delft University of Technology,
P.O.Box 5031, 2600 GA, Delft,
The Netherlands

Dimitri Solomatine
UNESCO-IHE Institute for Water Education and
    Section of Water Management of Delft
    University of Technology,
2601 DA, Delft,
The Netherlands

INTRODUCTION

There are some parametric approaches to model the process based on data. These are data-driven models. These modeling techniques are mainly based on data, either gained by measuring processes or simulation using physically based models. In traditional approaches of data-driven modeling, relatively simple linear or nonlinear regression methods are used for assessment of the system relations. But in many disciplines in ocean engineering and hydrology, an increasing trend is to use Artificial Neural Network (ANN). These methods are based on the analysis of all data characterising the system under study to find unknown mapping or dependencies between the system's input and output from the available data. These methods can be used for engineering applications at low cost and simple set-up.

There are many applications of data-driven methods in water-related modeling. Solomatine (2005) and Solomatine & Ostfeld (2008) reviewed various aspects of data-driven modeling and computational intelligence methods in water-related modeling. Standard types of ANN such as multilayer percepteron and radial bias functions have been successfully used in ocean engineering. Jain & Deo (2006) reviewed the application of ANN in several disciplines in ocean engineering. They referred to a stock of research studies reported so far in this area. Mynett (1999) reviewed a number of issues in information technology such as ANN and self-organising feature maps (SOFM) at Delft Hydraulics. Zijderveld (2003) investigated a number of neural network applications for prediction and classification tasks in the hydroinformatics context.

Forecasting of wave parameters at the desired location directly from wave records was carried out by several authors. Some recent works related to wave predictions are from Deo *et al.* (2001), Agrawal & Deo (2002), Makarynskyy (2004), Mandal & Prabaharan (2006) and Zamani *et al.* (2008). These papers describe various types of neural networks such as feed-forward or recurrent networks in different observation areas.

Deo *et al.* (2001) demonstrated the use of neural networks for wave forecasting for three different sets of data. It shows that a proper trained network could yield good results in open wider areas, in deep water. Agrawal & Deo (2002) forecasted wave heights with a lead time of a few hours, days or weeks using autoregressive neural networks. They compared their results with the traditional time series schemes of AR, ARMA and ARIMA. It was found that, when small forecasting intervals were involved, the neural networks gave more accurate results. Makarynskyy (2004) predicted significant wave heights and periods with lead times of 1–24 h with the help of NN using some novel schemes like the merger of initial forecasts with measurements. Londhe & Panchang (2006) examined a modeling strategy that predicts wave heights up to 24 h on the basis of judiciously selected measurements over the previous 7 d. They used data from six National Data Buoy Center (NDBC) buoys with diverse geographical and statistical properties. They showed that 6 h forecasts can be obtained with a high level of fidelity, and forecasts up to 12 h showed a correlation of 0.67 or better relative to a full year of data. Mandal & Prabaharan (2006) forecast significant wave heights with the recurrent neural network. They used a three-layer feed-forward recurrent network with eight inputs and one output. They showed a correlation coefficient of 0.95, 0.90 and 0.87 for 3 h, average 6 h and average 12 h wave forecasting, respectively. Zamani *et al.* (2008) presented a number of data-driven models for wind-wave processes at the Caspian Sea. The problem associated with these models is to forecast significant wave heights for several hours ahead using buoy measurements. Models are based on Artificial Neural Network (ANN) and Instance-Based Learning (IBL). Three feed-forward ANN models have been built for time horizons of 1, 3 and 6 h with different inputs. The inputs consist of the wind characteristics and *a priori* knowledge of significant wave height.

The other models are based on the IBL method for the same forecast horizons. Weighted k-Nearest Neighbors (k-NN) and Locally Weighted Regression (LWR) with Gaussian kernel were used.

Altunkaynak & Ozger (2004) forecast significant wave heights through Kalman filtering where the parameters were obtained from percepteron. Percepterons are the simplest form of artificial neural network without any hidden layer and linear transfer function. The input of their network is wind speed and significant wave height at time t and the output of their network is the same parameters at the next time step. The authors have mentioned the restriction of uses of regression methods for predicting wind generated waves and have compared it with the perceptron Kalman filter. They used one-month hourly buoy data which is located in Coos Bay, Oregon, USA. The transition matrix elements in their work are taken as modes of weights that are calculated by the perceptron model during the 15 d training period.

In other research, an attractive combination of dynamic neural network with EnKF was proposed and carried out by Aguilar *et al.* (2006) and Aguilar (2006). Discharge and water level forecasting has been an essential part of flood forecasting systems and river basin engineering. In their research, an artificial neural network is used to replace the numerical hydraulic model and then the data assimilation technique is implemented to use the measurement available to improve the one hour forecast at the location of interest. Hourly discharge and water level at eight stations along the Rhine River were used for implementation of the combined method. Two different neural network architects with three layers were used for the functional relationship of discharge and water level. Considering the water level and discharge as system states, their network relates the states at time t to the time at the next time step.

The combination of ANN and sigma-point Kalman filter was investigated by Lu *et al.* (2007). In order to address the highly nonlinear dynamics in estuary flow, they proposed a data assimilation system based on components designed to accurately reflect nonlinear dynamics. The core of the system is a sigma-point Kalman filter coupled to a fast neural network emulator for the flow dynamics. In order to be computationally feasible, the entire system was operated on a low-dimensional subspace obtained by principal

component analysis. Experiments on a benchmark estuary problem showed that this method can significantly reduce prediction errors.

Although the neural network and other data-driven models are used for forecasting of wave parameters, there is no attempt at application of the combined method in the field of wind-wave modeling. In the present study a recurrent dynamic neural network with nonlinear transfer function and a nonlinear data assimilation scheme such as EnKF will be used. In this case the ANN emulates the physical relationship between wind and wave by recognition of patterns in the data presented to the network. This ANN is fast but may deteriorate with time. To correct the dynamic behavior of the system, ensemble-based data assimilation is applied. To achieve the required accuracy, often the ensemble size has to be large. Fortunately the ANN is fast and it can be used in combination with EnKF which needs many runs of the dynamic model during each time step of the assimilation.

The structure of the present paper is as follows. The ANN set-up and its training is given first. Some consideration regarding the selection of inputs for the model is also addressed. The concept of data assimilation and the EnKF algorithm are explained and the combined formulation of the problem will be presented. The assumption on system noise and its estimation is presented as well. In the next section, the type of observations, their duration and the locations of the measurements will be specified. The result of the combined method is presented for wind-wave forecasting. Finally, a conclusion is made at the end to point out the direction for future research.

ANN SET-UP AND TRAINING

A system of simple processing elements, neurons, that are connected into a network by a set of (synaptic) weights is called a neural network. An objective of ANN is to imitate some of the functions of the human brain. An ANN is normally used to map a random input vector with the corresponding output vector. The physics of the underlying system need not be known beforehand and, unlike the statistical methods, the network does not need mathematical assumptions *a priori*. This feature makes ANN a suitable tool to approximate nonlinear function relationships without a pre-existing model and without, or with only a little, knowledge about the physics of the system.

For data assimilation purposes, a model which evolves in time is required. Looking to the wind-wave process as a dynamic system, the state of the system (wave height) at time t depends on the system's state in one or more preceding time steps as well as the present or preceding forcing of the system (wind speed and direction). To account for such a temporal evolution, the dynamic model must be used in 'state-space' form. Such a state-space form can be achieved by the design of the recurrent network. This network makes it possible to include time-delay behavior into the model.

In building the network, time lags should be specified properly. Usually, not all the input variables will be equally informative. Often a method based on linear cross-correlation is employed for selecting appropriate inputs. The major disadvantage associated with using cross correlation is that it is only able to detect linear dependences between two variables. Another method is Average Mutual Information (AMI). This method was used in the present work for determination of the inputs of the recurrent network.

AMI measures the dependence between two random variables. The AMI function between two random variables A and B is given by

$$\text{AMI}(A,B) = \sum_{i,j}^{N} P_{A,B}(a_i,b_j)^* \log_2 \left[\frac{P_{A,B}(a_i,b_j)}{P_A(a_i)P_B(b_j)} \right] \qquad (1)$$

where a_i and b_j are the ith or jth bivariate sample pair in a sample size N and $P_A(a_i)$, $P_B(b_j)$ and $P_{A,B}(a_i, b_j)$ are the respective univariate and joint probability densities estimated at the sample data points. Mutual information measures the information that A and B share. It measures how much knowing about one of this variables reduces our uncertainty for other variables.

An AMI criterion is able to adequately capture all the linear and nonlinear dependency between the variables and avoids the need for making any major assumption to the underlying model structure. More information about input determination of networks could be found in Bowden *et al.* (2005). They reviewed several methods for input

determination of neural network models in water resource applications.

The ANN model used in this study is a three-layer feed-forward network with feedback from the output to input layer. Based on an AMI analysis which is similar to the work of Zamani *et al.* (2008), the following ANN is built:

$$h(t_{k+1}) = f\Big[h(t_k), h_k(t_{k-1}), U(t-1), U(t-2), \ldots,$$
$$U(t-7), \bar{\Theta}\Big] \tag{2}$$

where t is the discretized time, h the wave height, f the model's function, W the wind speed and $\bar{\Theta}$ is the average wind direction for a time span which has been already included in the model. The lag time corresponding to maximum AMI is about 4 h. Also the AMI between wind and wave is close to the maximum value for lag times between 1 and 7 h. Although there is no maximum point for AMI of wind and wave but, due to high AMI values at times t and $t-1$, they can be used for model building.

The ANN architecture with ten neurons in the input layer and one neuron at the output layer has been used throughout in this study. Also the optimal size of the hidden layer was found by systematically increasing the number of hidden neurons until the network performance on the test set did not improved significantly. In this simulation the hyperbolic tangent function and the identity function have been used as transfer functions in the hidden and output layers, respectively. This combination of activation functions is recommended for function approximation problems (Bishop 1995). Results showed that there is no longer any improvement in the performance of the model for more than five hidden neurons.

ENSEMBLE KALMAN FILTER

In some situations models may not perform perfectly because of unresolved physics in the model, errors in boundary conditions, errors in forcing and also errors in some model parameters. Due to these uncertainties in the model, there will be a difference between model outputs and actual observed data. Data assimilation schemes exploit such a difference to correct the model. The Kalman filter is the well-known data assimilation scheme which was originally introduced by Kalman (1960) for linear systems. The detailed theoretical background of Kalman filtering for linear systems, and its extension to nonlinear systems, is given by Jazwinski (1970).

This section will briefly review the mathematical formulation of EnKF for nonlinear systems. This scheme was introduced by Evensen (1994) and has been successfully implemented and used in different types of applications (Evensen 2003). The EnKF is a Monte Carlo approach based on the representation of the probability density of the state estimate by a finite number of randomly generated system states.

Consider the system of a general, nonlinear stochastic model M and the observations

$$x^f(t_k) = M[x^f(t_{k-1}), U(t_{k-1})] + w(t_{k-1}) \tag{3}$$

$$y^o(t_k) = H(t_k)x^f(t_k) + v(t_k) \tag{4}$$

where $x^f(t_k) \in \mathfrak{R}^n$ denotes the forecast of the system state at time t_k, $U(t_k)$ is the forcing of the system and M represents one time step of the model. A normal distributed system noise $w(t_k) \in \mathfrak{R}^n, w(t_k) \sim N(0, Q)$ is introduced to take the uncertainties of the model into account. The vector $y^o \in \mathfrak{R}^r$ represents the measurements, which are supposed to be a linear combination of the states represented by the operator H. The observation noise is represented by $v(t_k) \sim N(0, R)$. In order to obtain an optimal estimate it is necessary to combine the measurement taken from the actual system and modeled by Equation (3) with the information given by the system model (Equation (2)). The forecast state at time t_k, denoted by $x^f(t_k)$, is the forecast from observation time t_{k-1} to observation time t_k by the Equation

$$x^f(t_k) = M\Big[x^a(t_{k-1}), U(t_{k-1})\Big] \tag{5}$$

where $x^a(t_{k-1})$ is the analyzed system state. At time t_k, an observation $y^o(t_k)$ is available and the estimate is updated by the analysis step:

$$x^a(t_k) = x^f(t_k) + K\Big[y^o(t_k) - H(t_k)x^f(t_k)\Big] \tag{6}$$

where

$$K(t_k) = P^f(t_k)H(t_k)^T\Big[H(t_k)P^f(t_k)H(t_k)^T + R\Big]^{-1} \tag{7}$$

is the minimum variance gain and $P^f(t_k)$ is the forecast error covariance matrix. In the EnKF method, the covariance can be calculated by a finite number of randomly generated system states.

For the initial state estimate x_0, the uncertainty is expressed by an ensemble ξ_i^a, $i = 1, \ldots, N$ of randomly generated states. The ensemble members are propagated from one time step to another using the original model operator:

$$\xi_i^f(t_k) = M[\xi_i^a(t_{k-1}), U(t_{k-1})] + w_i(t_{k-1}) \tag{8}$$

with $w_i(t_k)$ realizations of the noise process $w(t_k)$. Noise is added to the most uncertain parts of the model to estimate the covariance between observations and the model variables. The ensemble mean

$$\bar{x}^f(t_k) = \frac{1}{N} \sum_{i=1}^{N} \xi_i^f(t_k) \tag{9}$$

represents the state estimate at time t_k. Using this estimate the error covariance can be computed as

$$E^{f,EN}(t_k) = \left[\xi_1^f(t_k) - \bar{x}^f(t_k), \ldots, \xi_N^f(t_k) - \bar{x}^f(t_k) \right] \tag{10}$$

$$P^f(t_k) = \frac{1}{N-1} E^{f,EN}(t_k) \left[E^{f,EN}(t_k) \right]^{\mathrm{T}} \tag{11}$$

With the error covariance P^f calculated, the Kalman gain $K(t_k)$ is obtained from Equation (7). The update equations for the analyzed ensembles are

$$\xi_i^a(t_k) = \xi_i^f(t_k) + K(t_k[y^o(t_k) - H(t_k)\xi_i^f(t_k) + v_i(t_k)] \tag{12}$$

where $v_i(t_k)$ represents realizations of the measurement noise $v(t_k)$.

COMBINATION OF ANN AND ENKF

In the combined method, the output of the ANN will be considered as state vectors. The trained network acts as a deterministic model. Process noise is added to this model to prepare the stochastic model:

$$h(t_k) = f\left[h(t_{k-1}), h(t_{k-2}), \; , U(t_k), U(t-1), \right] + w(t_k) \tag{13}$$

A successful state estimation requires a well-designed process noise model and a well-designed measurement noise model. We assume that errors of a trained network are time-independent, zero mean and Gaussian-distributed. The difference between outputs of ANN and targeted values are used as an indication of system error. The covariance matrix Q can be estimated from the residual of ANN predictions:

$$Q = \alpha \frac{1}{N_t} \sum_{k=1}^{N_t} \left[h(t_k) - \hat{h}(t_k) \right]^2 \tag{14}$$

where N_f is the number of time steps used for covariance matrix estimation and $\hat{h}(t_k)$ is the true wave height. Also α is a tuning factor which can be found by trial and error. Based on the buoy measurement techniques, the measurement error and its covariance matrix are taken. Here the wave height measurements are used to correct the states. By help of the observations, the EnKF will correct the output of the ANN to find the best estimate of the system (or analyzed state). The analyzed state will return to the inputs of the ANN for the next time step.

In the inputs of the ANN, there is a difference between the inputs of the network which comes from forcing or from the feedback loop. One should carefully note that the inputs of the EnKF are different from the inputs of the ANN. The above procedure can work without observation as well. In this case the recursive application of the method leads to prediction at higher forecast horizons.

To deal with the time-delayed states in Equation (13), the Kalman filter formulation proposed by Gibson *et al.* (1991) is employed and an extended vector is taken as follows:

$$x^f(t_k) = \left[h(t_k), h(t_{k-1}), \ldots \right]^{\mathrm{T}} \tag{15}$$

Using the extended vector allows us to write a dynamic state space model in the standard form. After each new measurement, the state vectors will be updated for all ensembles by using Equation (12), Kalman gain Equation (7) and error covariance matrix Equation (11).

OBSERVATION AND STUDY AREA

In the present study, input–output patterns for training of network are from a measurement site in the southern part of the Caspian Sea. One hourly wind speed, wind direction and significant wave height are deduced from a discus wave buoy. The depth of the measurement site was 800 m and the period of data collection was 5 months ranging from 11 October 2006 to 2 May 2007. Fortunately, during these periods the buoy collected data continuously and without any gaps in the transmitting of data. Wave data was collected for 20 min at 1 h intervals at a sampling frequency of 2 Hz. Wind data was also collected for 10 min at 1 h intervals at a sampling frequency of 2 Hz. Figure 1 shows the measured significant wave height and wind speed at the site of the study. The total number of datasets is 4,800.

RESULTS

It is difficult to know which training algorithm will be suitable for the given problem. Therefore it is necessary to examine several training methods. The supervised algorithms were tested in this research are the quasi-Newton(QN), back propagation with variable learning rate and momentum (BPvm), resilient back propagation (RBP), conjugate gradient (CG) and Levenberg–Marquardt (LM). Table 1 shows the one-step-ahead forecast performance in terms of Root Mean Square Error (RMSE) along with the

Table 1 │ Indication of predictive dynamic model using various training methods

Algorithm	RMSE (m)	Epochs
QN	0.098	769
BPvm	0.090	2,500
RBP	0.093	2,500
CG	0.091	400
LM	0.085	317

number of epochs. The limiting criterion was to reach the threshold of 0.01 m in mean square error of the cost function. Because of its lower RMSE and the moderate epochs of LM, this method was chosen for the reminder of the study.

The data was divided into two parts in such a way that both blocks of data would have extreme events of similar nature. For all the experiments the 400 latest observations were selected for testing the models, and the rest of the data was used for their training. This ensured both blocks have at least one extreme event.

The training interval includes 4,400 data values. In spite of the well-trained network, the ANN could not predict the wave height in the test period. To show this issue, it is supposed that we have an exact wind forecast during the test period. Results from running this dynamic model based on wind forcing and also using a one-step forecast are illustrated in Figure 2. In this run, no assimilation of data
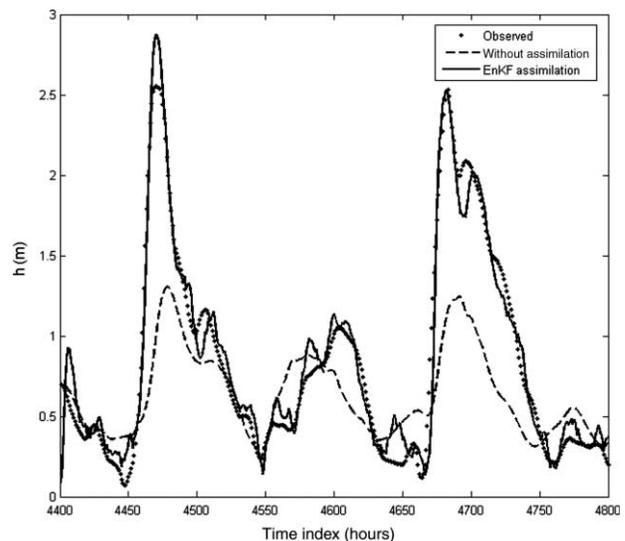

Figure 1 │ Time series plot for 1,500 h of data points. Upper panel: significant wave height (m), lower panel: wind speed (m/s).

Figure 2 │ Simulation without data assimilation and with data assimilation using EnKF and 10 ensembles.
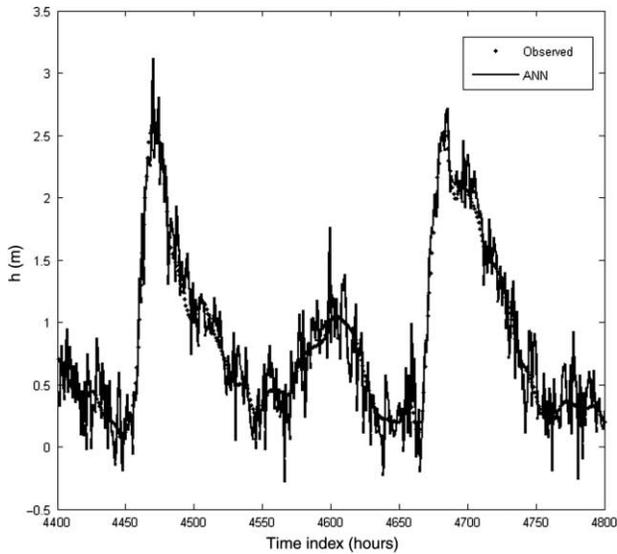
Figure 3 │ Direct assimilation of data in ANN input.

Figure 5 │ Innovations for time series plot of Figure 4.

has been used and the output of the network was fed into the input in a recursive manner.

In order to consider the effect of assimilation, the result of another run with EnKF assimilation is shown on the same figure. The new run is based on 10 ensemble members and an assimilation–forecast cycle for one hour. The effect of data assimilation can clearly be inferred from this figure. Although it is possible to use wave height measurements as the input to the network, it is not helpful in the case of noisy measurements. As a well-trained network will follow all

changes of input patterns, therefore, in the case of noisy input data, the network output would also be a noisy one. Using a few neurons in the hidden layer may tend to smooth out variations in the dependency structure between the input and output but this was not the suitable solution to this study. Forecast results using a direct assimilation of noisy measurements into the ANN is shown in Figure 3. In this experiment, at each time step the execution of the network is interrupted and the exact measurement of wave height is fed to the input of the network. These
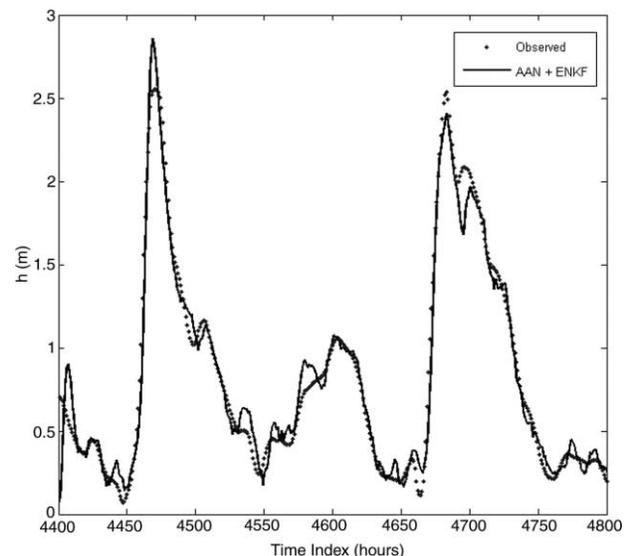

Figure 4 │ Wave height for 1 h forecast/assimilation cycle with 5 ensembles.

Figure 6 │ Wave height for 1 h forecast/assimilation cycle with 10 ensembles.

Figure 7 │ Innovations for time series plot of Figure 6.
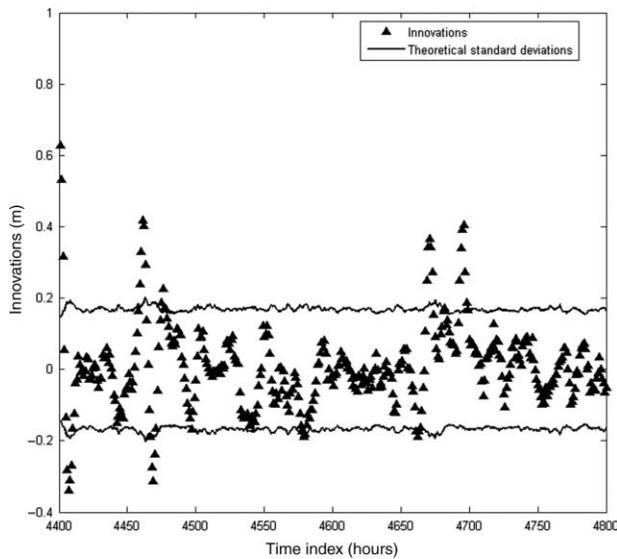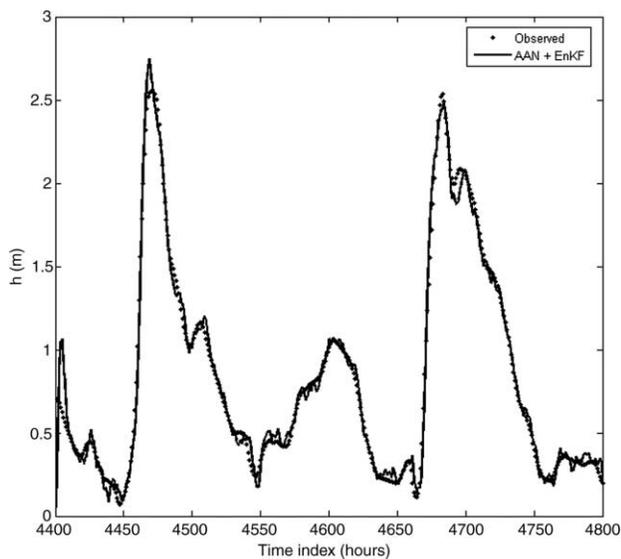

Figure 9 │ Innovations for time series plot of Figure 8.

measurements are noisy and they will cause the fluctuation in the prediction of wave height. The ANN mimics the behavior of noise while, on the other hand, the combined method is more robust.

To find the effect of ensemble members the dynamic model is run with different numbers of ensembles. Also it is possible to check filter performance by calculating the so-called innovation and plotting them along with their standard deviations.

Figures 4–9 indicate the role of increasing the number of ensembles and its effect on the innovation. In all experiments a 1 h forecast–assimilation cycle has been selected with 5, 10 and 50 numbers of ensembles. Figure 4 shows the 1 h forecast using 5 ensembles. In this figure, there is a large amount of error, especially in the region of extreme events. Figure 5 indicates the innovation based on the results of Figure 4. One can conclude that filter performance is not on a satisfactory level and innovations are not bound by their standard deviations with 5

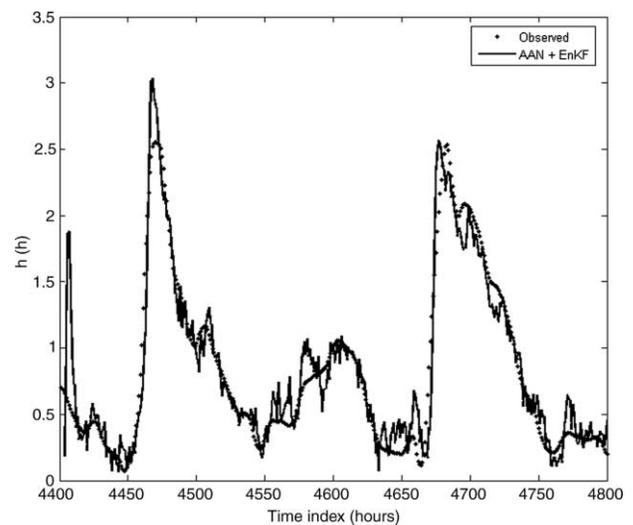**Figure 8** │ Wave height for 1 h forecast/assimilation cycle with 50 ensembles.

Figure 10 │ Four hour forecast of significant wave height with 1 h data assimilation and 50 ensembles.
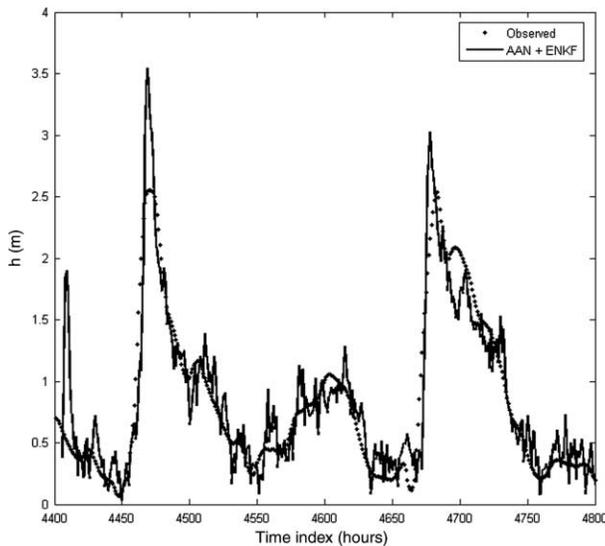
Figure 11 | Six hour forecast of significant wave height with 1 h data assimilation and 50 ensembles.

ensembles. Increasing the number of ensembles, as indicated in Figures 6 and 8, will improve the prediction results. As is shown in Figures 8 and 9, innovations are almost entirely bounded by their standard deviations for higher numbers of ensembles. Increasing the number of ensembles will increase computational cost but it will create some acceptable results.

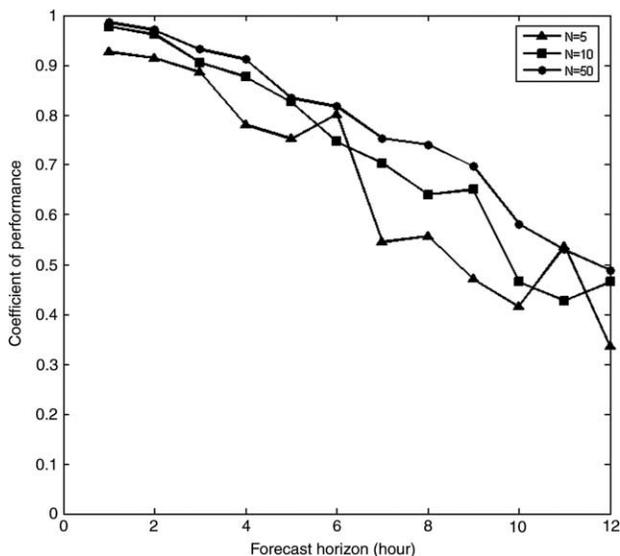Forecasts for higher horizon times can be achieved by implementing the procedure for different assimilation–

forecast cycles. The cycle has the following structure. At each time, a run is performed which consists of an analysis over the previous period in which the observation is assimilated followed by the forecast for the selected time horizon. The initial condition for each run is supplied by the previous analysis and for the beginning of the procedure the initial condition can be set to zero. Results of 4 h forecast and 1 h data assimilation is shown in Figure 10. Also the prediction of 6 h forecast and 1 h assimilation cycle is shown in Figure 11. Comparison of Figures 10 and 11 shows the increasing errors due to the higher forecast horizon.

The coefficient of performance can be used to measure the performance of the combined ANN and EnKF method:

$$Ce = 1 - \frac{\sum_{i=1}^{N} (h_i - \hat{h}_i)^2}{\sum_{i=1}^{N} (h_i - \bar{h})^2} \tag{16}$$

This coefficient is also called the Nash–Sutcliffe coefficient, which is used to assess the predictive power of the hydrologic models (Nash & Sutcliffe 1970). In Equation (16), Ce is the coefficient of performance, N is the number of samples in the test period and $\bar{h}$ is the average of the true wave heights.

Figure 12 shows the performance of the method for different forecast horizons and different numbers of ensembles in terms of the coefficient of performance. Also
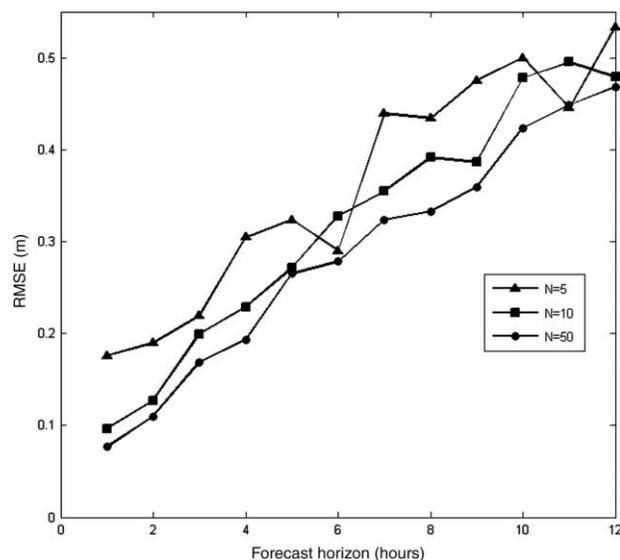

Figure 12 | Coefficient of efficiency for different forecast horizons and different numbers of ensembles.

Figure 13 | RMSE for different forecast horizons and different numbers of ensembles.

Figure 13 indicates the RMSE error for the same forecast horizons and number of ensembles. As can be seen from the figures, the efficiency and RMSE error of the model is reasonable for forecast horizons up to 6 h as, for higher forecast horizons, the RMSE grows rapidly.

CONCLUSION AND FUTURE RESEARCH

A wind-wave data assimilation method has been presented which is based on an Artificial Neural Network and an efficient low rank approximation of the Kalman filter. The dynamic equation for the relation between the state vectors and the forcing of the system is derived from data at the Caspian Sea. This ANN data-driven dynamic is updated using new measurements as soon as they are available. The analyzed or corrected states will then be used for short-term forecasts of the wind and wave height. The following conclusions can be deduced according to the experiments with the combined method:

- The framework of the combined method can be applied to estimate the state of a complex, nonlinear wind-wave model. Due to the predictor corrector behavior of the algorithm, the forecast is more accurate than with the use of the ANN alone.
- The combined method can be used with noisy data. Therefore, overfitting of an ANN to noisy patterns can be reduced.
- The combined scheme is fast and suitable for operational one-point wind-wave forecasting.
- The performance of the method for different forecast horizons and different numbers of ensembles indicates reasonable results for forecast horizons up to 6 h. For higher forecast horizons the RMSE grows rapidly.

The scheme is more general and extendable to other problems. For example, one can use the numerical wave model for preparing a lot of realizations of wind-wave patterns for the area under study and for preparing a new ANN which describes both the spatial and the temporal behavior of wind waves in this area. Then, by help of some observations, the state which consists of wind waves in various locations can be modified sequentially.

The combination method led us to build a portable data assimilation method for different types of large scale problems in wind-wave hindcast and forecast studies. The ANN has the ability to emulate a numerical model through the recognition of patterns in the data presented to the network and the combination with the EnKF will improve the ANN prediction. The combination will reduce the computational burden of applying the EnKF to the numerical model itself. A subject that will be addressed in a future paper is the use of simulation results of a full numerical wave model running at the Caspian Sea and the combined method as described in this paper. The full numerical model can be replaced with the ANN surrogate model. Surrogate models (or emulators or meta-models) are approximate models that mimic the behavior of large scale models as closely as possible while being computationally cheap. Also, in order to be computationally feasible, the entire system operates on a low-dimensional subspace obtained by principal component projection of wave and wind fields.

The ANN surrogate describes both the spatial and the temporal behavior of the wind waves in this area. The surrogate is updated using new observations as soon as they are available. The analyzed or corrected states will then be used for the next forecast of the wave field. In the context of data assimilation, accelerating the simulation of the dynamics of the system via the surrogate successfully reduces the time needed to propagate the ensemble forecasts.

REFERENCES

Agrawal, J. D. & Deo, M. C. 2002 Online wave prediction. *Mar. Struct.* **15** (1), 57–74.

Aguilar, S. G. 2006 *Neural Networks and Ensemble Kalman Filter Application for Hydrologic Models*. MSc Thesis,

Faculty of Electrical Engineering, Mathematics and Computer Science, Technical University of Delft, The Netherlands.

Aguilar, S. G., El Serafy, G. Y. & Mynett, A. E. 2006 Combined neural network and ensemble kalman filter application for discharge and water level forecasting in the river Rhine. In: *Proc. 7th International Conference on Hydroinformatics, HIC 2006, Nice, France* Vol. II, pp. 1447–1454.

Altunkaynak, A. & Ozger, M. 2004 Temporal significant wave height estimation from wind speed by perceptron Kalman filtering. *Ocean Eng.* **31**, 1245–1255.

Bishop, C. 1995 *Neural Network for Pattern Recognition*. Oxford University Press, Oxford, UK.

Bowden, G. J., Dandy, G. C. & Maier, H. R. 2005 Input determination for neural network models in water resources application. Part 1-background methodology. *J. Hydrol.* **301**, 75–92.

Deo, M. C., Jha, A. & Chaphekar, A. S. 2001 Wave prediction using neural networks. *Ocean Eng.* **28** (7), 889–898.

Evensen, G. 1994 Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast model statistics. *J. Geophys. Res.* **99** (C5), 10143–10162.

Evensen, G. 2003 The ensemble Kalman filter: theoretical formulation and practical implementation. *Mon. Weather Rev.* **123**, 1128–1145.

Gibson, J., Koo, B. & Gray, S. 1991 Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Signal Process.* **39** (8), 1732–1742.

Jain, P. & Deo, M. C. 2006 Neural networks in ocean engineering. *SAOS* **1** (1), 25–35.

Jazwinski, A. 1970 *Stochastic Process and Filtering Theory*. Academic, New York.

Kalman, R. E. 1960 A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.* **82**, 35–45.

Londhe, S. N. & Panchang, V. 2006 One day wave forecast based on artificial neural network. *J. Atmos. Oceanic Technol.* **23**, 1593–1603.

Lu, Z., Lean, T. K., VanderMerwe, R., Frolov, S. & Baptista, A. M. 2007 *Sequential Data Assimilation with Sigma-point Kalman Filter on Low dimensional Manifold*. Technical Report TR-07-001, NSF-STC for coastal margin observation and prediction. Available at: http://www.stccmop.org/files/CMOP-TR-07-001.Pdf

Makarynskyy, O. 2004 Improving wave prediction with artificial neural network. *Ocean Eng.* **31**, 709–724.

Mandal, S. & Prabaharan, N. 2006 Ocean wave forecasting using recurrent neural networks. *Ocean Eng.* **33** (10), 1401–1410.

Mynett, M. 1999 Hydroinformatics and its applications at Delft hydraulics. *J. Hydroinf.* **1** (2), 83–102.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I–a discussion of principles. *J. Hydrol.* **10** (3), 282–290.

Solomatine, D. 2005 Data-driven modeling and computational intelligence methods in hydrology. In *Encyclopedia of Hydrological Science* (ed. in M. Anderson). John Wiley and Sons, Ltd., Chichester.

Solomatine, D. P. & Ostfeld, A. 2008 Data-driven modelling: some past experiences and new approaches. *J. Hydroinf.* **10** (1), 3–21.

Zamani, A., Solomatine, D., Azimian, A. & Heemink, A. 2008 Learning from data for wind-wave forecasting. *Ocean Eng.* **35**, 953–962.

Zijderveld, A. 2003 *Neural Network Strategies and Modeling in Hydroinformatics*. PhD Thesis, Delft University of Technology, Delft, The Netherlands.