Ocean Engineering 35 (2008) 953–962

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Learning from data for wind–wave forecasting

Ahmadreza Zamani [a,*], Dimitri Solomatine [b], Ahmadreza Azimian [a], Arnold Heemink [c]

[a] Department Mechanical Engineering, Isfahan University of Technology, Isfahan 84156, Iran
[b] International Institute for Infrastructural, Hydraulic and Environmental Engineering, 2601 DA Delft, The Netherlands
[c] Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands

ARTICLE INFO

Article history:
Received 17 September 2007
Accepted 18 March 2008
Available online 26 March 2008

Keywords:
ANN
Data-driven models
Instance-based learning
Wind–wave

ABSTRACT

Along with existing numerical process models describing the wind–wave interaction, the relatively recent development in the area of machine learning make the so-called data-driven models more and more popular. This paper presents a number of data-driven models for wind–wave process 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) .To capture the wind–wave relationship at measurement sites, these models use the existing past time data describing the phenomenon in question. Three feed-forward ANN models have been built for time horizon of 1, 3 and 6 h with different inputs. The relevant inputs are selected by analyzing the average mutual information (AMI). The inputs consist of priori knowledge of wind and significant wave height. The other six models are based on IBL method for the same forecast horizons. Weighted *k*-nearest neighbors (*k*-NN) and locally weighted regression (LWR) with Gaussian kernel were used. In IBL-based models, forecast is made directly by combining instances from the training data that are close (in the input space) to the new incoming input vector. These methods are applied to two sets of data at the Caspian Sea. Experiments show that the ANNs yield slightly better agreement with the measured data than IBL. ANNs can also predict extreme wave conditions better than the other existing methods.

1. Introduction

Wave height forecasts are typically made based on the numerical wave models. There are several types of such models that can provide wave information for regional or global studies. In these models, the sea state can be described with a wave spectrum, which represents the wave energy density per frequency and direction. By means of energy-balance equation the evolution of wave spectrum can be computed in space and time and hence wave forecasting could be obtained in the region of study. Presently WAM (The WADMI group, 1988), Wave Watch III (Tolman, 1999) and SWAN (Booij et al., 1999) are well-known mathematical-based models which are used in the most meteorological centers. When such models are used, preparation of meteorological data and heavy computer processing is a challenging job.

Along with numerous existing physical models, the relatively recent development in the area of machine learning makes the so-called data-driven models more and more popular. These methods are based on the analysis of all data characterizing the system under study to find an unknown mapping or dependencies between the systems input and output from the available data. When the observed wave data are available, these methods can be applied with relatively simple set up. There are many applications of machine learning in water-related modeling, including some in wind–wave modeling. Solomatine (2005) and Solomatine and Ostfeld (2008) reviewed various aspects of data-driven modeling and computational intelligence methods in water-related issues. Jain and Deo (2006) reviewed the application of ANN in several disciplines in ocean engineering. Mynett (1999) presented some applications of ANN and self-organizing feature maps (SOFM). Zijderveld (2003) investigated some neural network applications for prediction and classification tasks in a hydro-informatics context. Solomatine et al. (2007) compared instance-based learning (IBL) to other data-driven models in hydrological forecasting. Bhattacharya et al. (2003) applied neural network in the reconstruction of missing wave data in sedimentation modeling. Puca et al. (2001) designed a neural network approach to the problem of recovering lost data in a network of marine buoys. The potential of ANN to provide accurate estimates of nonlinear interactions for wind–wave spectra by direct mapping is considered by Tolman et al. (2005). Kazeminezhad et al. (2005)

* Corresponding author. Tel.: +98 311 3912539, +98 311 52785813; fax: +98 311 3912518.
E-mail addresses: arzamani@cc.iut.ac.ir (A. Zamani), d.solomatine@unesco-ihe.org (D. Solomatine), azimian@cc.iut.ac.ir (A. Azimian), A.W.Heemink@ewi.tudelft.nl (A. Heemink).

0029-8018/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.oceaneng.2008.03.007

proposed a fuzzy logic methodology to determine the wave parameters from wind speed and fetch length. Also Ozger and Zekai (2007) investigated the relation between wind speed and wave characteristics by fuzzy logic approach. Solomatine et al. (2001) used chaos theory (nonlinear dynamics) and ANN in predicting surge water levels in the North Sea based on the 5 years of the collected data on water level, wind and air pressure at several off-shore stations. Vaziri (1997) predicted the Caspian sea surface water level by ANN and ARIMA models.

Regarding forecasting wave height at one location, some latest works in wave forecasting is pertained to Deo and Naidu (1999), Deo et al. (2001), Agrawal and Deo (2002), Zamani and Azimian (2004); Makarynskyy (2004), and Mandal and Prabaharan (2006). In the aforementioned works various type of neural networks such as feed-forward or recurrent networks are used. These works are used in different observation areas without considering the effect of wind direction in wind–wave forecasting. Comparison of results based on the work carried out by Mandal and Prabaharan (2006) showed the correlation coefficient of 0.95, 0.90 and 0.87 for 3 h, average 6 h and average 12 h wave forecasting, respectively. Also Ozger and Zekai (2007) reported values of 0.97, 0.96, 0.89 and 0.80 for 1, 3, 6 and 12 h forecasting based on fuzzy interference system.

In the present study, ANN and IBL methods are used to forecast significant wave heights for several hours ahead using buoy measurements in Caspian Sea. First, a brief explanation of ANN and IBL is given. Following these, some characteristics of measurements, their duration and locations are specified. Some considerations regarding selection of inputs of models are also addressed, together with the data transformations used to better account for the effect of wind speed and wind direction. Forecasting models are described, the modeling results are presented and discussed, and finally, the conclusions are drawn.

2. Study area

Two sets of meteorological and wave data have been used in this study. Wind and wave data was collected by a 3-m diameter discuss shape buoy. This buoy deployed by Khazar Exploration and Production COmpany (KEPCO) at two different locations (A) and (B) situated in the southern part of the Caspian Sea (Fig. 1). Location (A) is near the beach but location (B) is far from the coast. The water depth at these locations is 15 and 800 m, respectively. The period of data collection at location (A) is from November 20, 2005 to April 9, 2006 and at location (B) is from October 11, 2006 to May 2, 2007. Fortunately during these periods buoy collected data continuously and without any gaps in 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. Fig. 2 shows a sample time series plot of wave height, wind speed and wind direction at site A.

3. Machine learning

A machine learning method is an algorithm that enables the user to estimate an unknown mapping rule between a system's input and output data (Aha et al., 1991). By data we understand the known samples each being a combination of the input vector and the corresponding outputs. Once such a dependency is discovered it can be used to predict the future system output from the known input values. Suppose that K instances represented by $\langle X_i, Y_i \rangle$ where X_i and Y_i typically

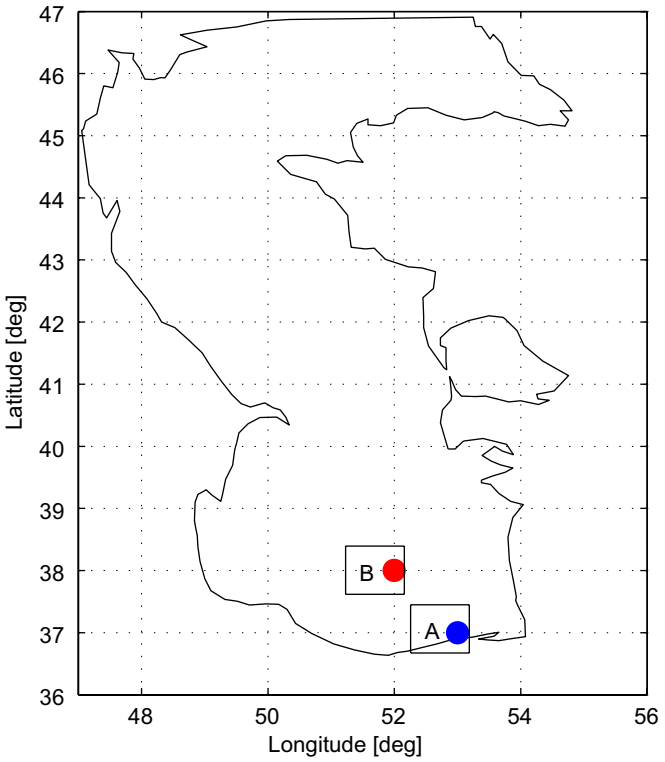

Fig. 1. Location of measurement sites at Caspian Sea: (A) shallow water, and (B) deep water.

contain multidimensional vectors in R^p and R^q, respectively (p is number of inputs and q is number of outputs). Now the objective is to build a function ('mapping' or 'model') similar to $Y = f(X)$ for finding functional dependency of inputs and outputs. In our study $q = 1$ and the model to build takes the form: $y = f(X)$.

Various kinds of machine learning methods can be used for building the models. Among them, in this study artificial neural network (ANN) and IBL including weighted k-nearest neighbors (k-NN) and locally weighted regression (LWR) with Gaussian kernel are used. Detailed information about above methods can be found in Haykin (1999) and Aha et al. (1991). In ANN the method is to learn the function f explicitly by training a network of neurons. ANN includes a set of weights, which should be determined by training optimization process aimed at minimizing the model error. After that, the model is used for prediction and there is no need to retain the training data. In contrast IBL is a memory-based (analog) method and it is a kind of non-parametric approach which retains the training data and uses it at each time which prediction needs to be made. In fact, in IBL the training data is stored in memory and when a new vector presented a set of similar related instances reviewed from memory and their corresponding outputs are used to predict the output for a new query vector. The similar related instances to new query can be classified by some distance metric such as Euclidean distance. Consider the k-nearest neighboring points of the query. It is possible to construct the local model (or approximation) to the modeled function that applies well near in the immediate neighborhood of the new query instance. It can be done by weighted k-NN or LWR algorithms. In weighted k-nearest point (X_i) according to their distance $d(X_q, X_i)$ to the query point (X_q) but in LWR the regression model is built on k-nearest instances, which are assigned weights according to their distance to the query instance. The predicted value for X_q is

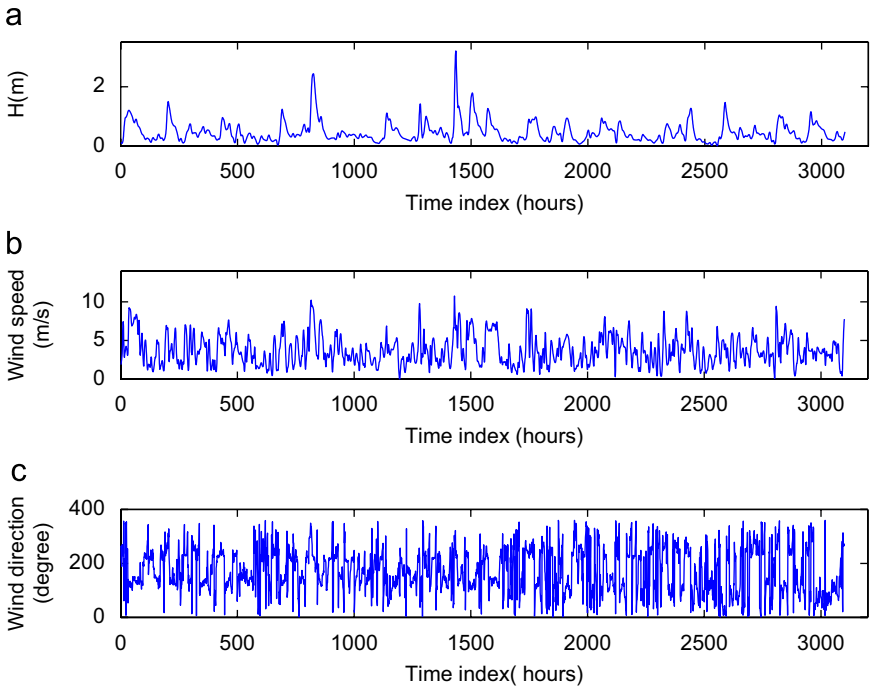

Fig. 2. Hourly time series plot of: (a) significant wave height, (b) wind speed, (c) wind direction, at location (A) for duration of 20 November 2005 till 9 April 2006.

calculated as follows:

$$f(X_q) = \frac{\Sigma_{i=1}^{k} w_i f(X_i)}{\Sigma_{i=1}^{k} w_i} \quad (1)$$

where w_i is a function of distance d with the following common weight functions:

$$w_i = 1 - d(X_q, X_i) \text{ (Linear)} \quad (2)$$

$$w_i = 1/d(X_q, X_i) \text{ (inverse)} \quad (3)$$

$$w_i = 1/d(X_q, X_i)^2 \text{ (Inverse-square)} \quad (4)$$

In LWR the weights are determined from the minimization of following distance-weighted square error:

$$E(X_q) = (1/2)\sum_{i=1}^{k}(y_i - f(X))(K(d)) \quad (5)$$

with y_i being the target output, $f(X)$ is regression function and K is Gaussian kernel function with the following definition:

$$K(d) = \exp(-d(X_q, X_i)^2) \quad (6)$$

Weight function should be maximum at zero distance and the function should decay smoothly as distance increase. Fedrov et al. (1993) addressed the issue of choosing weighting functions.

4. Determination of the inputs

One of the important steps in using models is determination of the adequate input and output variables. Usually, not all the input variables will be equally informative since some may be correlated, noisy or have no significant relationship with the output variables being modeled. Bowden et al. (2005) reviewed some methods of input determination for neural network models in water resource applications. For selecting appropriate inputs methods based on linear cross-correlation are often employed. The major disadvantage associated with using cross-correlation is that it is only able to detect linear dependence between two variables. Another method used in the present work is known as average mutual information (AMI) method. AMI measures the dependence between the two random variables. An AMI is calculated for two random variables A and B as follow (Abebe and Price, 2004):

$$\text{AMI}(A,B) = \sum_{i,j}^{N} P_{A,B}(a_i, b_j) * \log_2\left[\frac{P_{A,B}(a_i, b_j)}{P_A(a_i)P_B(b_j)}\right] \quad (7)$$

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. In order to calculate the mutual information between A and B, it is necessary to estimate joint probability density function and marginal density functions of A and B. Probabilities are estimated by the corresponding frequencies. Numerical axis for each variable is discretized and all values falling into each bin are given the same frequency value. Histogram and Kernel methods are widespread to estimate probability density functions (Scott, 1992).

Intuitively, mutual information measures the information that A and B share. It measures how much knowing about one of these variables reduces our uncertainty about the other variable. For two statistically independent random variables the AMI score in Eq. (7) is zero. Also if the random variables are strongly related, the AMI score would take a high value. The AMI score of wind–wave at measurement sites were computed for different lag hours indicated in Fig. 3. From this figure it is clear that the value of AMI at location (B) is higher than that at location (A). This means that more substantial information about the wave data is included in the wind field at location (B). On the other hand due to the boundary and sea bed effects at location (A) which is close to the coast, the wave system will be affected by other parameters which are not included in the wind information. The same can be inferred from Fig. 4 and AMI value of wave–wave at both

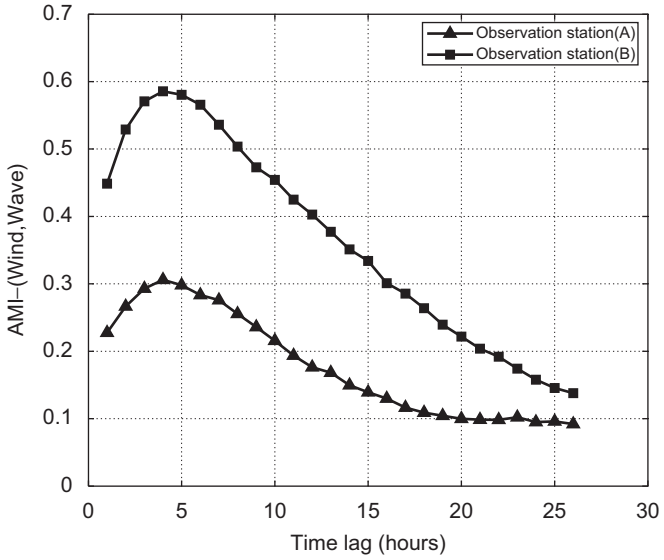

Fig. 3. Wind–wave average mutual information at locations (A) and (B).

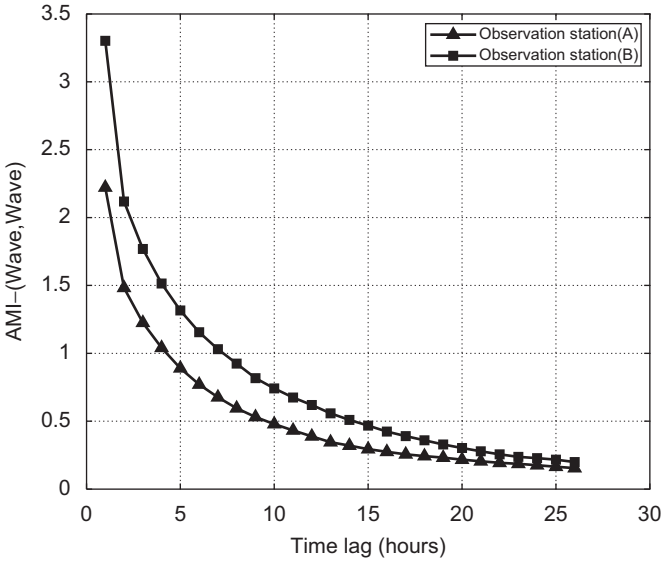

Fig. 4. Wave–wave average mutual information at locations (A) and (B).

locations. In addition, analysis of AMI has been conducted also for the following pairs of variables:

- wind speed difference and significant wave height
- wind speed difference and wave height difference
- wind speed and wave height difference

However, we found that AMI for the latter pairs was quite low and did not consider them as the inputs for the models. Another problem to address was to take into account the specifics of the method used to measure wind. The buoys measure wind speed at 3–5 m above the sea level. However, the standard 10 m reference level should be used. The following ($\frac{1}{7}$)th rule is used as an approximation for level adjustment (US Army, 2003):

$$U_{10} = U_z\left[\frac{10}{z}\right]^{1/7} \tag{8}$$

where U_{10} is wind velocity at 10 m height from the sea level and U_z is wind velocity at elevation z. In second stage in order to transfer the wind speed U_{10} to friction velocity U_* the transformation (9) has been used. It could be shown that by means of this transformation the range of friction velocity variations would be less than the wind speed variations. In actual computation, U_{10} is converted to the friction velocituy U_* with:

$$U_*^2 = C_D U_{10}^2 \tag{9}$$

where C_D is the wind drag coefficient. Later the effect of this transformation on the result will be shown. From WAM Cycle III and SWAN models, the value of C_D is determined with an expression due to Wu (1982):

$$C_D = \begin{cases} 1.2875 \times 10^{-3} & \text{if } U_{10} < 7.5\,\text{m/s} \\ (0.8 + 0.065 \times U_{10}) \times 10^{-3} & \text{if } U_{10} \geqslant 7.5\,\text{m/s} \end{cases} \tag{10}$$

Yet another consideration was taken into account. Wind is a vector and it has direction as well as magnitude. If measured in degrees, there would be a discontinuity in wind direction around the north direction. For example both 0° and 360° show north direction. Also directions of 3° and 357° are quite close to each other, but differ numerically. To account for wind direction the encoding method used for example in Bhattacharya et al. (2003) would be used as follows:

$$\Theta = \begin{cases} 1 - (\Psi/180) & \text{if } 0^\circ \leqslant \Psi \leqslant 180^\circ \\ (\Psi - 180)/180 & \text{if } 180^\circ < \Psi < 360^\circ \end{cases} \tag{11}$$

where Ψ is wind direction in degree and Θ is encoded wind direction between 0 and 1.

5. Models setup and training

5.1. Models structure

The problem associated with the wind–wave forecasting is to forecast wave height several hours ahead with respect to the previous information of the system. The available information can be wind speed vector and a priori knowledge of wave height. The AMI helps to find out how much information about the future wave is available from the past wind and wave data. The lag time corresponding to maximum AMI from Fig. 3 is about 4 h. Also the AMI between wind and wave is close to maximum value for lag times between 1 and 7 h. Although there is no maximum point for AMI of wave and wave (see Fig. 4) but due to high AMI values at time t and $t-1$, they can be used for model building. The AMI-based analysis does not directly lead to identifying the exact relationship between wind and waves but it helps to bring in the relevant physical variables, properly lagged, into DDM models. Using AMI analysis to assess dependencies between variables and the lag time, the following local models were built:

$$H_{ANN}(t+1) = f_1(U_t, U_{t-1}, U_{t-2}, U_{t-3}, U_{t-4}, U_{t-5}, U_{t-6}, U_{t-7}, \overline{\Theta}, H_t, H_{t-1}) \tag{12}$$

$$H_{ANN}(t+3) = f_2(U_t, U_{t-1}, U_{t-2}, U_{t-3}, U_{t-4}, U_{t-5}, \overline{\Theta}, H_t, H_{t-1}) \tag{13}$$

$$H_{ANN}(t+6) = f_3(U_t, U_{t-1}, U_{t-2}, \overline{\Theta}, H_t, H_{t-1}) \tag{14}$$

where t is discretized time, H is the wave height, f is the model's function, U is wind speed and $\overline{\Theta}$ is average wind direction for time span which has been included in the model. These models are global ANN. They are trained using the full set of input data. Eqs. (13) and (14) obtained from Eq. (12) with changing time index of wind speed from t to t+2 and t+5, respectively. Also time indexes greater than time t should be ignored at the right-hand side of the equations.

The following instance-based models were built as well:

$$H_{\mathrm{kNN}}(t+1) \Rightarrow f_4(\cdots) \tag{15}$$

$$H_{\mathrm{kNN}}(t+3) \Rightarrow f_5(\cdots) \tag{16}$$

$$H_{\mathrm{kNN}}(t+6) \Rightarrow f_6(\cdots) \tag{17}$$

$$H_{\mathrm{LWR}}(t+1) \Rightarrow f_7(\cdots) \tag{18}$$

$$H_{\mathrm{LWR}}(t+3) \Rightarrow f_8(\cdots) \tag{19}$$

$$H_{\mathrm{LWR}}(t+6) \Rightarrow f_9(\cdots) \tag{20}$$

In Eqs. (15)–(20) the notation $(\cdots)$ means the same inputs as ANN models for each forecast time horizon. The above notation has been chosen to distinguish between ANN and other models. These models do not construct an approximation designed to perform well over the entire instance space but use only the neighbors of the new input vector for forecasting and in this sense can be called "local" models.

It is possible to define models which map wind–wave data to higher forecast horizons. But due to the short memory of local wind–wave process as indicated by AMI analysis, up to 6 h forecast has been considered in this study.

5.2. Data division

The way the data is divided into training and test sets have to ensure their approximate statistical similarity. However, another criterion is to try to preserve the contiguousness of these sets. A compromise approach has been chosen: 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 latest 400 observation 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 for both locations. Still, this

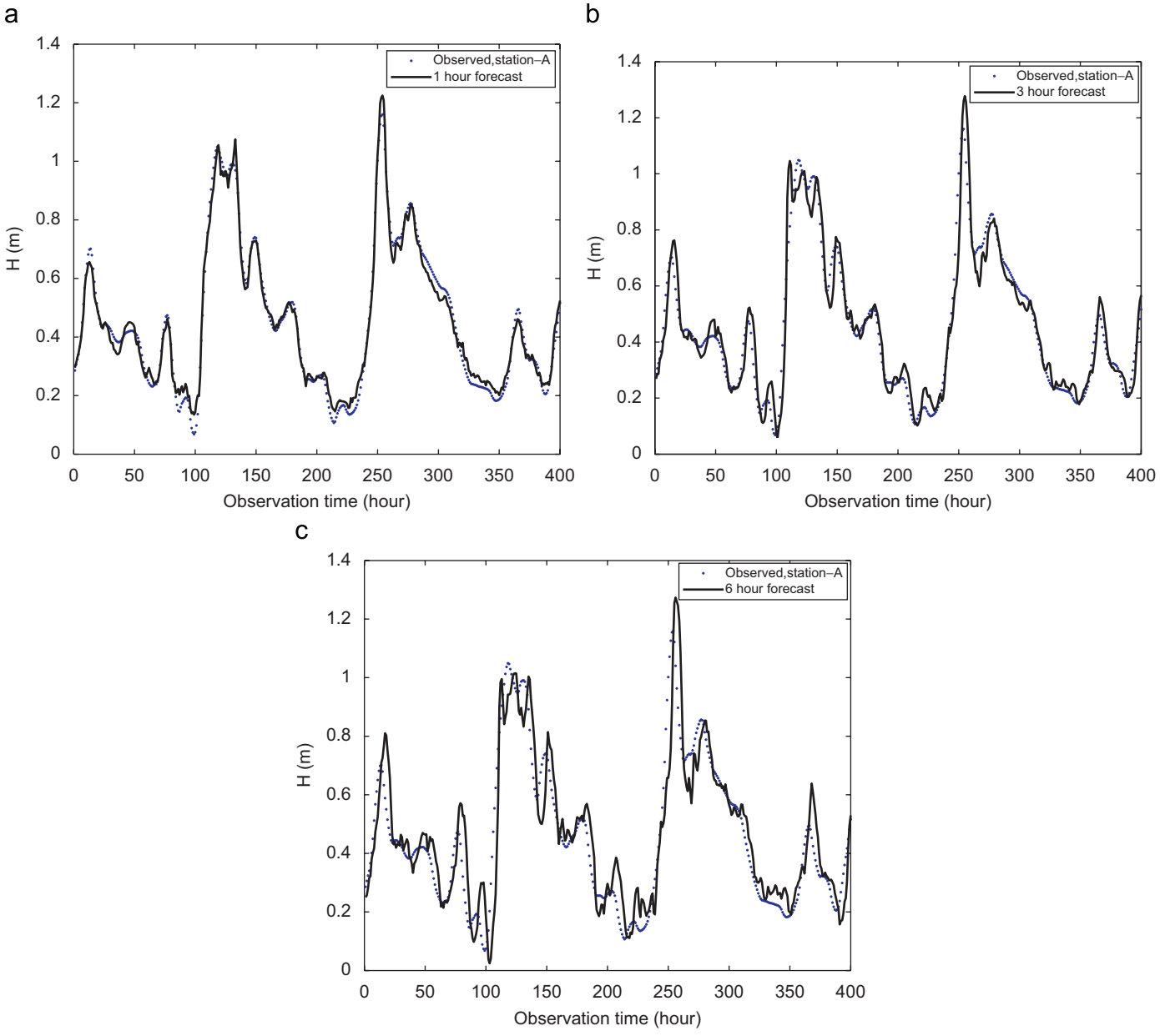

Fig. 5. (a) Forecast of model f_1 for location (A), (b) forecast of model f_2 for location (A), and (c) forecast of model f_3 for location (A).

has not ensured very similar statistical properties of the two-data subsets because of seasonal effect. For example most of training set in location (A) belongs to winter while the test period is in the spring (see Fig. 2). This situation has implications on the way models are trained.

5.3. Models training

For evaluation of models f_1–f_3 three-layer feed-forward ANN with sigmoid transfer function for hidden layer and linear transfer function for output layer have been selected. The network with five neurons in the hidden layer appeared to be the best model for both the locations. Since the statistics of training and test sets are different, the networks are trained by the so-called method of "n-folding". In this method n different sets of training-validation data should be prepared based on the original training data with the length of N_t. In each set (N_t/n) number of data vectors is used for validation and $(1-1/n)N_t$ of them for training. n models are built, and for each model training and validation sets are different. For model i the validation set comprises (N_t/n) vectors starting from vector$N_t*(i-1)/n+1$, and the rest of the data set is used for training. The stopping criteria were one of the following:

- minimum error in validation set
- mean square error in training reaching threshold of 0.0001
- or the number of epochs reaching 2500

6. Results

The performance of models can be measured using various metrics. In this study the correlation coefficient (r), the root mean square error (RMSE) and the scatter index (SI) are used:

$$r=\frac{\Sigma_{i=1}^{N}(o_i-\overline{o})(y_i-\overline{y})}{\left[\Sigma_{i=1}^{N}(o_i-\overline{o})^2\Sigma_{i=1}^{N}(y_i-\overline{y})^2\right]^{1/2}} \tag{21}$$

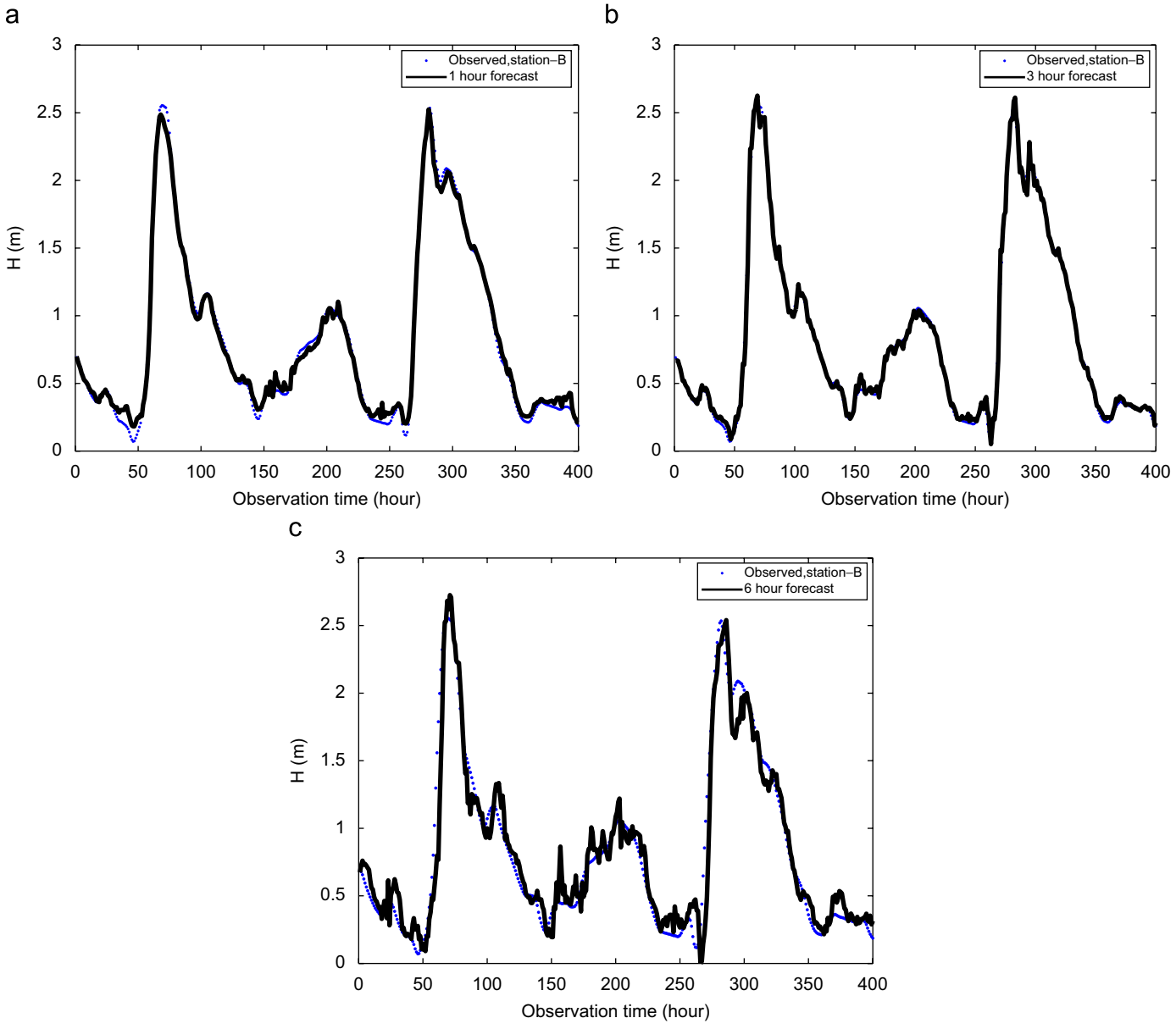

Fig. 6. (a) Forecast of model f_1 for location (B), (b) forecast of model f_2 for location (B), and (c) forecast of model f_3 for location (B).

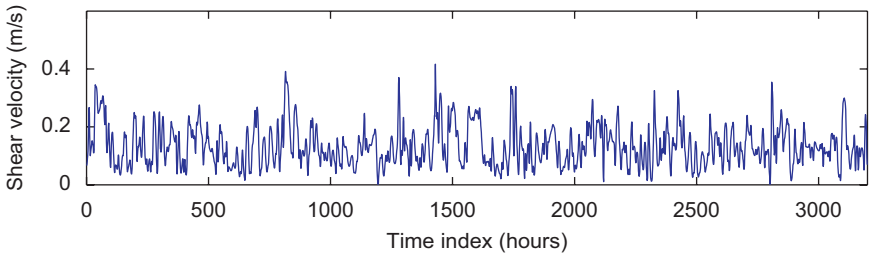

Fig. 7. Shear velocity for location (A).

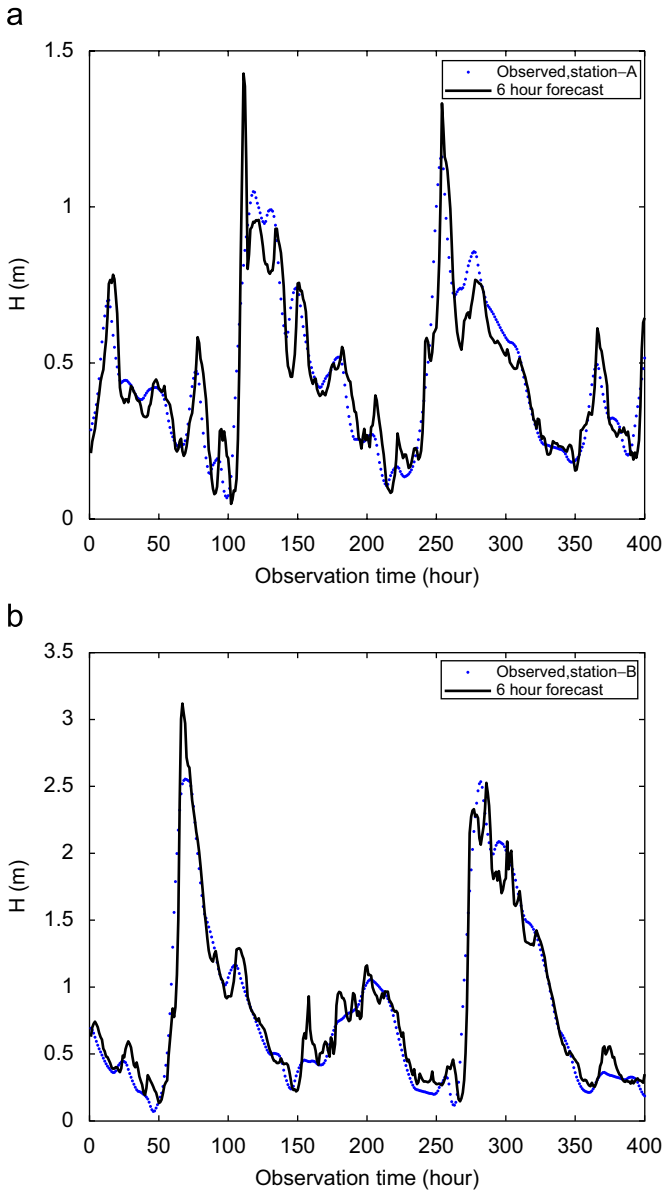

Fig. 8. (a) Forecast of model f_3 for location (A), model driven with wind speed instead of shear velocity, and (b) forecast of model f_3 for location (B), model driven with wind speed instead of shear velocity.

$$\mathrm{RMSE} = \sqrt{\frac{\Sigma_{i=1}^{N}(o_i - y_i)^2}{N}} \tag{22}$$

$$\mathrm{SI} = \frac{\mathrm{RMSE}}{\bar{o}} \tag{23}$$

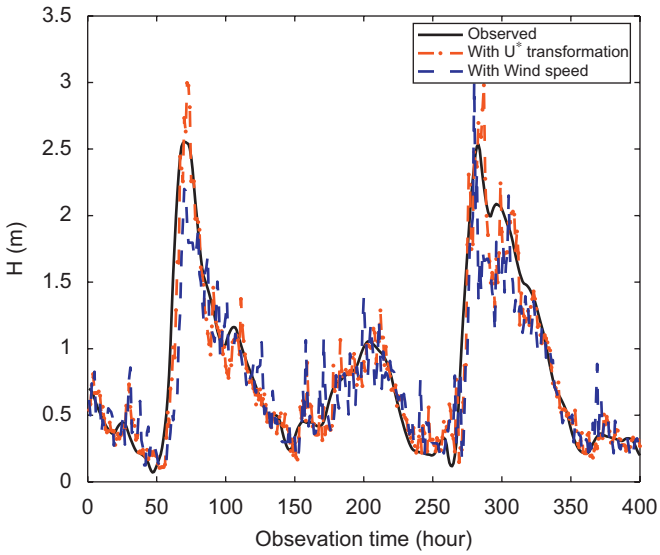

Fig. 9. Forecast of model f_6 for location (B) and for two different runs, First run with U input and second run with U_* input.

where o_i is an observed value at ith time step, y_i is a forecasted value at the same moment of time, N is the number of time steps, $\bar{o}$ is the mean value of observed data, and $\bar{y}$ is the mean value of forecasted values.

The outputs of models f_1–f_3 are shown in Figs. 5(a)–(c) and 6(a)–(c). Their error statistics are explained later. These figures show reasonable accuracy of ANN models in most cases except for the maximum value given by model f_3 in shallow water. Generally speaking the results obtained from ANN at site (B) agree well with the measured data and coincides with our expectation after AMI analysis.

For some of the models presented in Section 5 the effect of wind transformation from the common wind velocity to the shear wind velocity has been examined as well. The input variable U was treated in two different ways, as the actual wind velocity U and as the shear velocityU_*. This transformation simply plays the role of scaling. Scaled data covers the same range for all variables and therefore errors in each variable contribute in the same proportion to changes into the network weights. The variation of shear velocity is shown in Fig. 7 for location A.

Fig. 8 illustrates the output of model f_3 for the two locations where the input variable U used in the model formulation is the actual wind velocity. Comparison of model outputs for these runs show that the sensitivity of model to the actual wind speed is more than that when the shear velocity U_* is used. Also the model with U_* as input shows a better behavior in the region of extreme event. See Figs. 5c and 6c where the forecast horizon of 6 h is shown.

The effect of this transformation could be also seen when IBL method is used. Fig. 9 illustrates the output of model f_6 for the two

different options, one with the wind velocity and the other one with the shear velocity. Errors are obviously found in Fig. 9 at every peak. Generally speaking, this is due to the effect of number of neighboring points in IBL method and lack of similarity of samples in memory with the predicted points. In spite of errors in predictions, in this experiment, the statistics of IBL method with shear velocity is better than real wind according to Table 1. Simultaneous comparison of model outputs for these runs show that another feature. Usually model's responses to a sudden input change are slow, but in the second run the highest peak is forecasted well and the variation of wave height is closer to the real observed data.

Table 1
Two different runs of f_6

Wind input			Shear input		
r	RMSE (m)	SI	r	RMSE (m)	SI
0.901	0.316	0.307	0.951	0.201	0.242

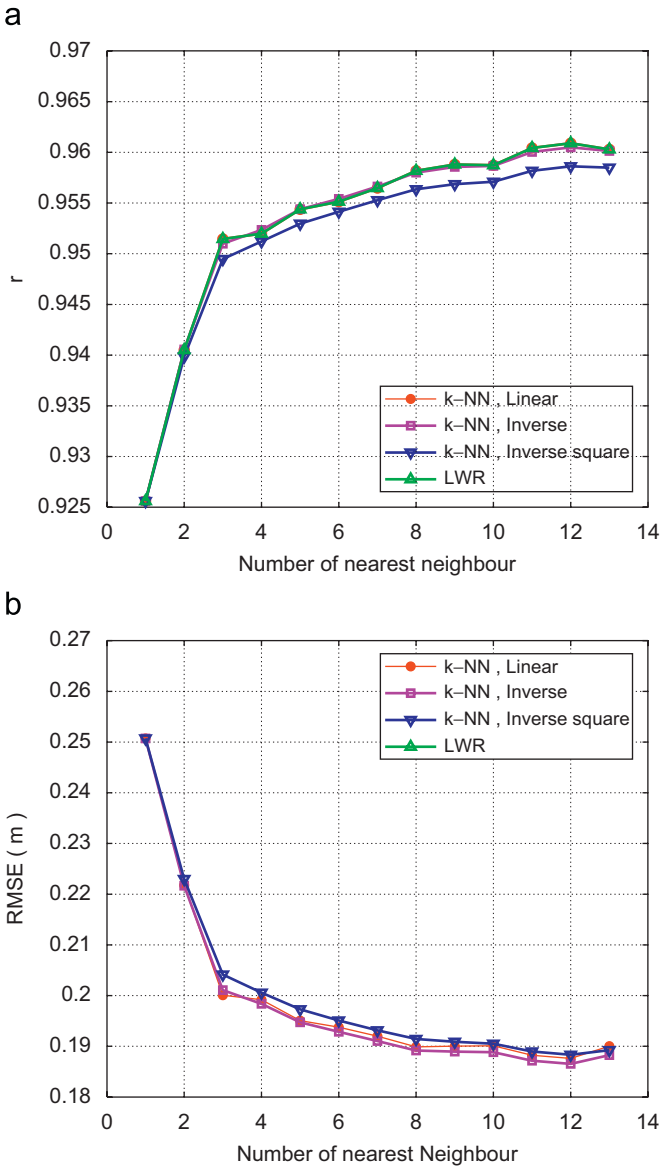

Fig. 10. Correlation coefficients and root mean square error for different weighing functions and different number of neighboring for model f_6 (k-NN) and model f_9 (LWR).

Most IBL models can be optimized with respect to their parameters and the weighting functions used. This can be done by running the models and comparing their performance on cross-validation set. Since the available data was limited, such optimization for the presented cases cannot be properly done, so the results obtained for the test set instead. Fig. 10 shows the correlation coefficient r and RMSE of the selected models in terms of the number of nearest neighboring points and different weighting functions. All the IBL methods will obviously show the same value of r and RMSE in case of using one nearest neighbor. Increasing number of neighbors gives the chance to the points located further away from the new data vector to influence the result. With the increase of the number of neighbors, the forecast will tend to be closer and closer to the global average, and RMSE will be decreasing so that the detailed information about the signals will be lost. For a small number of neighboring points the forecast will be too noisy. After experiments a compromise number of five neighbors was adopted to be used in all IBL models, and the obtained results are compared with those obtained with ANN. (More research however, would be desirable into the proper optimization of IBL models with respect to the number of neighbors.)

Verification statistics of ANN and IBL method indicated in Table 2. IBL method shows better performance for 1 h forecast. In IBL pervious training data are used to predict next step data in a recurrent way, these predicted data would achieve better than ANN when only trained weights and biases are used for future prediction. Mostly this is in agreement with statistics of 1 h forecast. In another way for one step ahead, the latest incoming data set may be considered as one of the high correlated neighboring points. For higher forecast horizon, there is no guarantee for latest data set to be considered as neighboring point. Also there is no guarantee to find suitable nearest points in the memory with limited number of similar events. For this

Table 2
Verification statistics of ANN and IBL method

Model	Location A			Location B		
	r	RMSE (m)	SI	r	RMSE (m)	SI
f_1	0.996	0.038	0.083	0.997	0.042	0.067
f_2	0.991	0.032	0.070	0.992	0.077	0.094
f_3	0.946	0.081	0.172	0.980	0.129	0.150
f_4	0.996	0.022	0.049	0.998	0.039	0.047
f_5	0.975	0.056	0.124	0.990	0.092	0.111
f_6	0.879	0.122	0.269	0.951	0.201	0.242
f_7	0.996	0.023	0.050	0.998	0.091	0.049
f_8	0.974	0.057	0.125	0.989	0.093	0.112
f_9	0.882	0.121	0.267	0.956	0.200	0.241

Table 3
Maximum errors between forecasted and observed wave heights

Model	Location A (m)	Location B (m)
$f1$	0.07	0.10
$f2$	0.15	0.22
$f3$	0.32	0.61
$f4$	0.07	0.10
$f5$	0.25	0.31
$f6$	0.64	0.64
$f7$	0.07	0.10
$f8$	0.24	0.31
$f9$	0.64	0.63

Table 4
Verification of wave height forecasts with other methods

Reference	+T	r	RMSE	SI	Method	Inputs
Ozger and Zekai (2007)	+1	0.974	0.282	0.110	Fuzzy logic approach	Wind speed and wave height
	+3	0.960	0.347	0.135		
	+6	0.899	0.541	0.211		
	+12	0.800	0.741	0.289		
Ozger and Zekai (2007)	+1	0.972	0.293	0.114	ARMAX	Wind speed and wave height
	+3	0.925	0.471	0.184		
	+6	0.842	0.666	0.260		
	+12	0.690	0.895	0.349		
Mandal and Prabaharan (2006)	+3	0.95	–	–	NARX	Wave height
	+6	0.90	–	–		
	+12	0.87	–	–		
	+24	0.73	–	–		
Zamani and Azimian (2004)	+3	0.911	0.16	–	MLP	Wave height
	+6	0.889	0.187	–		
	+12	0.585	0.311	–		
	+24	0.356	0.352	–		

reason IBL shows less performance for higher forecast horizons than ANN. Higher correlation of all models in deep water can be inferred from Table 2 which coincides with our expectation from AMI analysis. Also the lower scatter index in deep water can be seen form the same table. On the other hand, analyzing wave height time series of location B reveals that the average wave height for location B is more than the average wave height at location A (see the range of vertical axis in Fig. 8 for both locations). With the same number of test points (400) and lower scatter index, Eq. (22) results in a higher RMSE error for location B. Table 3 summarize the maximum errors between forecasted and observed wave heights during the test period. The increase of the forecasting time intervals results in an increase on the error in forecasting wave height. One of the ways of improving such predictions would be collecting more data at the distant locations that have high impact on the wave climate at the considered locations, and determining the proper lags for all observations used in the predictions.

Some other statistical methods were used by other researchers for various case studies. It is not possible to make a detailed comparison between different methods as used with other researchers. However, summary of statistics of various methods is presented in Table 4 for reference. Comparison of the results presented in Tables 2 and 4 reveals that both ANN and IBL models as defined in this work favorably compare to the other published results.

7. Conclusion

This presented results show how the data-driven (machine learning) models could be effectively used to perform the short-term wind–wave forecasting. A number of models were built and their inputs were selected by analyzing the AMI that gave some insight into the dependency of input and output parameters in the measurement locations. The effect of transforming the wind speed to shear velocity was investigated. Comparison of the statistical results obtained from the models using wind speed with the results of the models using the shear velocity showed that the performances of the latter models are better. The possibility of using a global model and also a local model was investigated too. The selected case studies at shallow water (location A) and deep water (location B) showed that all models perform much better in deep waters. Also comparison of statistics of IBL with ANN showed that IBL appears to be more accurate and robust for one step forecasting. For multi-step forecasting a global ANN model showed higher performance both in terms of the correlation coefficient and RMSE. A significant difference between the results of *k*-NN and LWR in this study was not observed.

The presented research was aimed at investigating the possibilities of the data-driven methods. The results obtained could have been better if we would have had more data for the considered locations, and especially if there would be data available for the other locations in the Caspian Sea. This would have allowed for building more accurate models with the higher forecasting horizon. Obviously, as the collected data set increases, the accuracy of the data-driven models will improve, and the intra- and inter-yearly variations could then be studied.

An interesting research direction is combination of machine learning methods like ANN and IBL with the methods developed in the framework of chaos theory (for example, Solomatine et al., 2001) and combining the data-driven methods with the physically based hydrodynamic models.

Acknowledgments

The authors thank Isfahan University of Technology (IUT) and Technical Delft University for their full support to this research. They are also thankful to the Khazar Exploration and Production COmpany (KEPCO) for providing the wave data. Some of the research findings used in this study was supported of the Delft Cluster Research program supported by the Dutch government (project "Morphodynamics of the North Sea").

References

Abebe, A.J., Price, R.K., 2004. Information theory and neural networks for managing model uncertainty in flood routing. Journal of Computing in Civil Engineering, ASCE 18 (4), 373–380.

Agrawal, J.D., Deo, M.C., 2002. Online wave prediction. Marine Structures 15 (1), 57–74.

Aha, D., Kibler, D., Albert, M., 1991. Instance-based learning algorithms. Machine Learning 6, 36–37.

Bhattacharya, B., Shrestha, D.L., Solomatine, D., 2003. Neural networks in reconstructing missing wave data in sedimentation modeling. In: Proceedings of the XXXth IAHR Congress, Thessaloniki, Greece.

Booij, N., Ris, R.C., Holthuijsen, L.H., 1999. A third generation wave model for coastal regions, Part I, model description and validation. Journal of Geophysical Research 104 (C4), 7649–7666.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources application. Part 1—background and methodology. Journal of Hydrology 301, 75–92.
Deo, M.C., Naidu, C.S., 1999. Real time wave forecasting using neural network. Ocean Engineering 26 (3), 191–203.
Deo, M.C., Jha, A., Chaphekar, A.S., et al., 2001. Wave prediction using neural networks. Ocean Engineering 28 (7), 889–898.
Fedrov, V.V., Hackel, P., Muller, W.G., 1993. Moving local regression: the weight function. Non-parametric Statistics 2 (4), 335–368.
Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, second ed. Prentice-Hall, Englewood Cliffs, NJ.
Jain, P., Deo, M.C., 2006. Neural networks in ocean engineering. SAOS 1 (1), 25–35.
Kazeminezhad, M.H., Etemad-Shahidi, A., Mousavi, S.J., 2005. Application of fuzzy interface system in prediction of wave parameters. Ocean Engineering 32 (14–15), 1709–1725.
Makarynskyy, O., 2004. Improving wave prediction with artificial neural network. Ocean Engineering 31, 709–724.
Mandal, S., Prabaharan, N., 2006. Ocean wave forecasting using recurrent neural networks. Ocean Engineering 33–10, 1401–1410.
Mynett, M., 1999. Hydroinformatics and its applications at Delft Hydraulics. Journal of Hydroinformatics 12, 83–102.
Ozger, M., Zekai, S., 2007. Prediction of wave parameters by fuzzy logic approach. Ocean Engineering 34, 460–469.
Puca, S., Tirozzi, B., Arena, G., et al., 2001. A neural network approach to the problem of recovering lost data in a network of marine buoys. In: Proceedings of the International Conference on Offshore and Polar Engineering, Scavenger, pp. 17–22.
Scott, D.W., 1992. Multivariable Density Estimation: Theory, Practice, and Visualization. Wiley, New York.
Solomatine, D., 2005. Data-Driven Modeling and Computational Intelligence Methods in Hydrology, Encyclopedia of Hydrological Science. Wiley, New York.
Solomatine, D.P., Ostfeld, A., 2008. Data-driven modeling: some past experiences and new approaches. Journal of Hydroinformatics 10 (1), 3–22.
Solomatine, D.P., Velickov, S., Wust, J.C., 2001. Predicting water levels and currents in the North Sea using chaos theory and neural networks. In: Proceedings of the 29th IAHR Congress, Beijing, China.
Solomatine, D., Maskey, M., Shrestha, D.L., 2007. Instance-based learning compared to other data-driven methods in hydrologic forecasting. Hydrological Processes 22 (2), 275–287.
The WAMDI group (13 authors), 1988. The WAM model-third generation ocean wave prediction model. Journal of Physical Oceanography 18 (12), 1775–1810.
Tolman, H.L., 1999. User Manual and System Documentation of WAVE-WATCH III, Version 1.18, NOAA/NWS/NCEP/OMB Technical note 166.
US Army, 2003. Coastal Engineering Manual Meteorology and Wave Climate. Engineer Manual 111021100. US Army Corps of Engineers, Washington, DC (Chapter II-2).
Tolman, H.L., Krasnopolsky, V.M., Chalikov, D.V., 2005. Neural network approximation for nonlinear interactions in wind wave spectra: direct mapping for wind seas in deep water. Ocean Modeling 8, 253–278.
Vaziri, M., 1997. Predicting Caspian sea surface water level by ANN and ARIMA models. Journal of Waterway, Port, Coastal and Ocean Engineering, 158–162.
Wu, J., 1982. Wind stress coefficients over sea surface from breeze to hurricane. Journal of Geophysical Research 87 (C12), 9704–9706.
Zamani, A., Azimian, A., 2004. On line wave prediction at Caspian Sea by using artificial neural network. In: Proceedings of the Ninth Fluid Dynamic Conference, Tehran, pp. 48–60.