Artificial Intelligence Tools to Forecast Ocean Waves in Real Time

Pooja Jain and M.C. Deo*

Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, 400076, India

Abstract: Prediction of wind generated ocean waves over short lead times of the order of some hours or days is helpful in carrying out any operation in the sea such as repairs of structures or laying of submarine pipelines. This paper discusses an application of different artificial intelligent tools for this purpose. The physical domain where the wave forecasting is made belongs to the western part of the Indian coastline in Arabian Sea. The tools used are artificial neural networks, genetic programming and model trees. Station specific forecasts are made at those locations where wave data are continuously observed. A time series forecasting scheme is employed. Based on a sequence of preceding observations forecasts are made over lead times of 3 hr to 72 hr. Large differences in the accuracy of the forecasts were not seen when alternative forecasting tools were employed and hence the user is free to use any one of them as per her convenience and confidence. A graphical user interface has been developed that operates on the received wave height data from the field and produces the forecasts and further makes them accessible to any user located anywhere in the world.

INTRODUCTION

Since recent past countries such as USA, Canada, Australia and India have implemented elaborate in-situ ocean data collection programs. Under these programs data of ocean related parameters including significant wave heights, zero cross wave periods and wind speed are measured at regular intervals – typically 1 or 3 hours - through instruments such as floating wave rider buoys and transmitted to the users through a web based data dissemination scheme. In India the National Institute of Ocean Technology (NIOT) located at Chennai practices such data collection in a big scale. The present work deals with wave measurements made by NIOT at three locations in the Arabian Sea area. (Fig. **1**). It attempts to provide real time forecasts of significant wave heights at these stations by considering this exercise as a time series forecasting problem and based on three alternative artificial intelligence techniques, namely, artificial neural network (ANN), genetic programming (GP) and model trees (MT). Different tools are employed to see if better results are possible by adoption of a different learning scheme.

The observations of wave heights considered in this work were made at the locations code named DS1 (15.326° N 69.371° E), SW2 (18.595° N 71.031° E), and, SW4 (12.932° N 74.716° E) off the western Indian coast shown in Fig. (**1**). Information on the underlying data collection program could be seen on the website [1]. The 3-hourly significant wave height data ranging from 3 to 7 years (1998 to 2004) were used. The location DS1 is in very deep water (3800 m) while the stations SW2 and SW4 are in 80 m and 24 m water depths, respectively.

Past works that incorporate the use of ANN to make online prediction of waves include [2, 3, and 4]. However effect of using the techniques of GP and MT for this task has not been reported so far. The need to explore usefulness of alternative approaches nonetheless exists in view of the highly random nature of wave occurrences and resulting inability to obtain entirely satisfactory predictions.

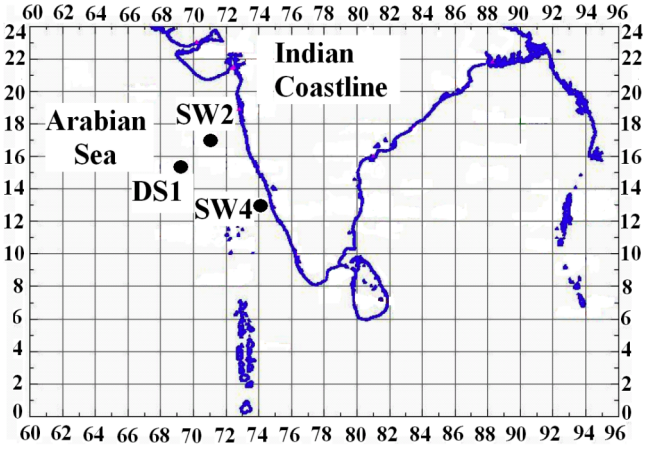

Fig. (1). Buoy locations in the Arabian Sea.

ANN MODELS

A typical artificial neural network (exemplified in Fig. (**2**)) consists of an interconnection of computational elements called neurons. Each neuron basically carries out the task of combining the input, determining its strength by comparing the combination with a bias (or alternatively passing it through a non-linear transfer function) and firing out the result in proportion to such a strength. Mathematically,

$$O = 1/(1 + e^{-S}) \quad (1)$$

$$\text{where, } S = (x_1 w_1 + x_2 w_2 + x_3 w_3 + \ldots\ldots) + \theta \quad (2)$$

In which, O = output from a neuron; x_1, x_2,….. = input values; w_1, w_2,…….= weights along the linkages connecting two neurons that indicate strength of the connections; θ = bias value. Equation (1) indicates a transfer function of sig-

*Address correspondence to this author at the Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, 400076, India; E-mail: mcdeo@civil.iitb.ac.in

1874-835X /08

moid nature, commonly used, although there are other forms available, like sinusoidal, Gaussian, hyperbolic tangent. Reference may be made to text books [5, 6 and 7] to understand the theoretical details of the working of a neural network, while a review of the ANN applications to ocean engineering in general can be seen [8]. A majority of the applications made in ocean engineering so far have involved a feed forward type of the network as against the feedback or recurrent one. A feed forward multi-layer network would consist of an input layer, one or more hidden layers and an output layer of neurons as shown in the Fig. (**2**).

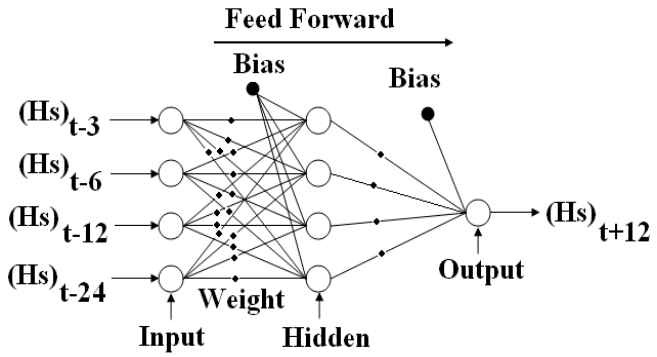

Fig. (2). A Feed forward network.

GENETIC PROGRAMMING

The GP is modeled out of the process of evolution occurring in nature, where the species survive following the principle of 'survival of the fittest'. It essentially transforms one population of individuals into another one in an iterative manner by following the natural genetic operations like reproduction, mutation and cross-over. Unlike the more widely known genetic algorithm (GA), its solution is a computer program or an equation as against a set of numbers in the GA. Koza [9] explains various concepts related to GP. In GP a random population of individuals - equations or computer programs - is created, the fitness of individuals is evaluated and then the 'parents' are selected out of these individuals. The parents are then made to yield 'offsprings' by following the process of reproduction, mutation and cross-over. The creation of offsprings continues in an iterative manner till a specified number of offsprings in a generation are produced and further till another specified number of generations is created. The most-fit offspring at the end of all this process is the solution of the problem.

Applications of GP in ocean engineering are difficult to find, although the same in civil engineering related to water flows started around 5 to 6 years ago. The tool of GP has been used for a variety of purposes like pattern recognition, classification and regression. Unlike the other soft computing tool of ANN, GP applications in water are restricted to relatively fewer areas and include rainfall-runoff modeling [10, 11], estimation of settling velocities of faecal pellet [12], modeling of risks in water supply [13], evaluation of ocean component concentration from sunlight reflectance or luminance values [14]. Most recently, usefulness of GP for infilling missing values in given wave height time series is reported in [15, 16].

MODEL TREES

In a model tree the computational process is represented by a tree structure consisting of a root node (decision box) branching out to numerous other nodes and leaves (Fig. **3**). The entire input or parameter domain is divided into subdomains and a multivariate linear regression model is developed for each sub-domain. It therefore considers a piecewise linear model to approximate a given nonlinear relationship between a dependent variable and corresponding independent variables. Depending on considerations like a domain splitting criterion there are alternative algorithms to build a model tree. Out of these the M5 algorithm of Quinlan [17] used in this work is popular [18]. It is as follows:

Let N = total number of training examples. Typically a set of all independent variable values together with corresponding dependent variable values would form a training example. Using some dividing or splitting criterion N is divided into many sub-domains. For every sub-domain an error measure is selected to decide if further division is necessary. In the M5 algorithm the error criterion is standard deviation (SD) of the class value reaching a node (decision box). The splitting stops when the class values of all collections in a sub-domain do not vary much (typically by 5 %), i.e., all samples have the same classification or when very few collections result. The splits are most of the times so large that an overfitting might happen, in which generalization does not take place and instead individual training patterns only are learnt. To overcome this, pruning of the structure is done by say merging together a few lower sub-regions producing similar models. There is also a possibility that the above pruned structure might have large discontinuities between neighboring models, especially when the training examples are less. Therefore a smoothing is done by revising nearby equations such that their output becomes closer to each other [19].

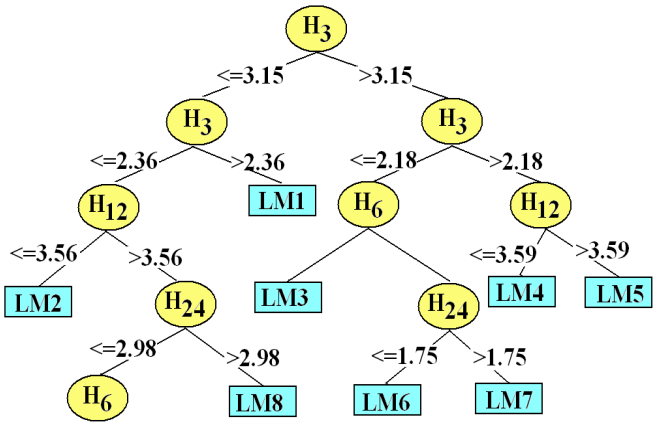

Fig. (3). A Model Tree (H_3, H_6, H_{12}, H_{24} : wave heights at preceding time t-3, t-6, t-12 and t-24 respectively; LM1,LM2…LM7 are different linear regression models. Typical linear regression models can be seen in Appendix II).

An advantage of MT over other data mining approaches like the ANN or program based GP is that its outcome is understandable and can be easily applied by another user. However it is to be noted that being a piecewise linear model it is not purely non-linear like ANN or GP.

APPLICATION OF AI MODELS

For developing the ANN usual 3-layer feed forward back-propagation type neural networks were considered. Various input – output combinations were tried to train the network. It was found that the input comprising of four unequally spaced preceding values of the significant wave height, Hs, (i.e., Hs, pertaining to 3rd, 6th, 12th and 24th hour backwards) and the output consisting of the forecasted Hs value for the lead time of 3, 6, 12, 24, 48 and 72 hr –one at a time – constituted the best network architecture [4]. The number of hidden neurons was of the order of 4 to 5 for various networks. The transfer function for the hidden layer was 'logsig' while the same for the output layer was 'purelin' for all the networks. The data were divided into two parts, with the first 60% portion used for training and the remaining one employed for testing the network. Out of a variety of training schemes attempted with a view to impart the best possible training, the algorithm of Levernberg- Marquardt gave most satisfactory results. Details on this training algorithm may be seen in [20].

The weight and bias matrices of the trained network were retained for testing the network. The prediction accuracy of the networks was judged by calculating the correlation coefficient, R, between the predicted and observed wave heights at these locations along with wave height time series plots and accompanying scatter plots. Additionally the error measures of root mean square error (RMSE) and mean absolute error (MAE) were also used to confirm the findings.

It was found that the measurements involved considerable amount of missing information. For the duration of 3 to 7 years considered for the analysis the gaps were 19% at DS1, 15% at SW2 and 2% at SW4 and ranged from a few hours to a few months at a time. Such lack of continuity in the observations had caused problems in network training due to which lower accuracies were realized. The gaps were filled up by a spatial correlation with adjacent buoys if they were longer and by temporal correlation with the preceding observations of the same time series if these were smaller in length. The details of the methods used to fill in the gaps in data can be seen in authors' previous work [21].

Figs. (**4**) and (**5**) show, for location SW2 examples of comparison between observed significant wave heights and their predictions over time step of 24 hr through the scatter and time series plot for the testing data. Table **1** shows a similar comparison through the error measure of correlation coefficients for all the locations for the testing data. Although the accuracy of predictions decreased with the forecast lead times the figures and this table show a satisfactory work done.

Forecasts over the lead time up to 24 hr were indeed very satisfactory in terms of the correlation coefficient while the same over the subsequent higher prediction interval of 48 hr and 72 hr were also acceptable as can be seen in the Table **1**.

Advanced artificial neural networks like radial basis function and ANFIS (Adaptive Neuro Fuzzy Inference System) were also employed for the forecasting purpose. But no improvement was seen in large lead time forecasting using these networks. The results were more or less similar to that of the feed forward network involved earlier.

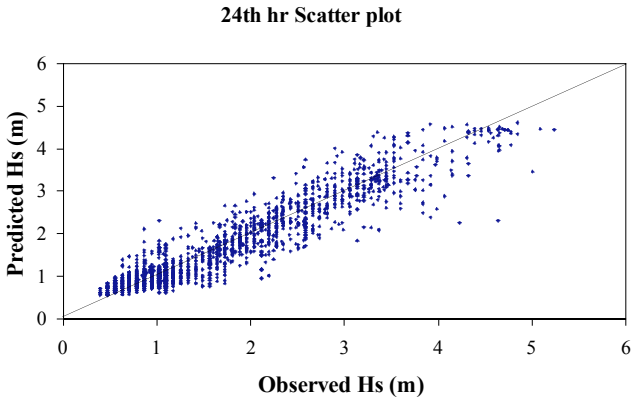

Fig. (4). Scatter Plot between observed and ANN-predicted wave heights for 24-hr lead time (Station SW2).

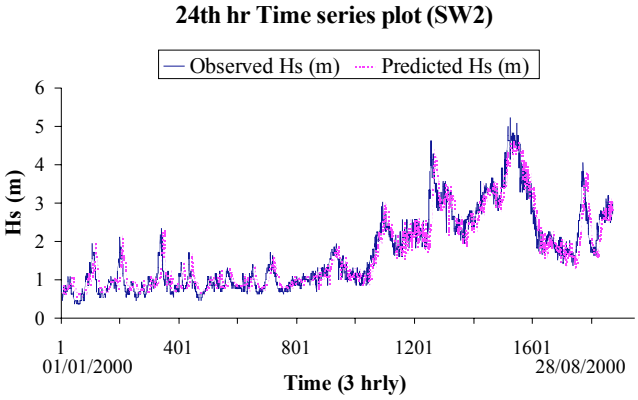

Fig. (5). Time series plot between observed and ANN-predicted wave heights for 24-hr lead time (Station SW2).

Table 1. ANN Testing Performance in Terms of Error Statistics

		3rd hr	6th hr	12th hr	24th hr	48th hr	72th hr
Station DS1	R	0.99	0.98	0.97	0.95	0.89	0.85
	RMSE(m)	0.19	0.22	0.28	0.36	0.50	0.60
	MAE(m)	0.12	0.14	0.18	0.23	0.31	0.37
Station SW4	R	0.98	0.97	0.96	0.94	0.91	0.88
	RMSE(m)	0.12	0.14	0.18	0.19	0.49	0.52
	MAE (m)	0.08	0.09	0.11	0.13	0.71	0.19
Station SW2	R	0.98	0.97	0.96	0.94	0.89	0.85
	RMSE(m)	0.21	0.23	0.26	0.33	0.46	0.52
	MAE(m)	0.15	0.17	0.19	0.24	0.33	0.38

The techniques of GP and MT were later employed to make predictions of significant wave heights in an exactly similar way to that of the earlier ANN. This was done to see how these alternatives, which are relatively new in ocean applications, work in comparison with the ANN. Tables **2**

and **3** shows the error measures for all the locations for the testing data for GP and MT respectively. Figs. (**6-9**) indicate scatter plot and time series plots of GP and MT models for SW2 location at lead time of 24 hr while Figs. (**10-12**) show an inter-comparison between the ANN, MT and GP models through the various error measures of R, RMSE and MAE for locations: DS1, SW4 and SW2 respectively. Appendix I and II give calibrated GP and MT programs as examples for a lead time of 24 hr at station SW4

Table 2. GP Testing Performance in Terms of Error Statistics

		3rd hr	6th hr	12th hr	24th hr	48th hr	72th hr
Station DS1	R	0.99	0.98	0.96	0.94	0.85	0.84
	RMSE(m)	0.17	0.22	0.31	0.39	0.57	0.61
	MAE(m)	0.12	0.14	0.17	0.23	0.34	0.39
Station SW4	R	0.98	0.97	0.96	0.94	0.90	0.87
	RMSE(m)	0.12	0.14	0.16	0.19	0.25	0.28
	MAE (m)	0.08	0.10	0.11	0.13	0.17	0.20
Station SW2	R	0.98	0.97	0.96	0.94	0.88	0.84
	RMSE(m)	0.20	0.23	0.28	0.24	0.48	0.55
	MAE(m)	0.15	0.17	0.20	0.16	0.34	0.40

Table 3. MT Testing Performance in Terms of Error Statistics

		3rd hr	6th hr	12th hr	24th hr	48th hr	72th hr
Station DS1	R	0.99	0.98	0.97	0.95	0.91	0.87
	RMSE(m)	0.19	0.21	0.25	0.35	0.45	0.54
	MAE(m)	0.11	0.13	0.16	0.21	0.29	0.35
Station SW4	R	0.98	0.97	0.96	0.95	0.90	0.88
	RMSE(m)	0.12	0.14	0.16	0.19	0.26	0.52
	MAE (m)	0.09	0.10	0.11	0.13	0.17	0.19
Station SW2	R	0.98	0.98	0.97	0.95	0.89	0.85
	RMSE(m)	0.18	0.21	0.25	0.33	0.47	0.52
	MAE(m)	0.11	0.14	0.16	0.21	0.30	0.38

These figures clearly indicate that the values of R are high enough and those of RMSE and MAE are sufficiently low to view the results as acceptable predictions. It is however recognized that these accuracies are possible in the present moderate environment where the target waves were less than around 6 m and 2.5 m for the offshore and coastal stations respectively. Most importantly a large difference in the results across the prediction tools employed was not seen and hence it is felt that the choice of one particular method might be guided by the user's convenience and confidence in the technique rather than its technological superiority.

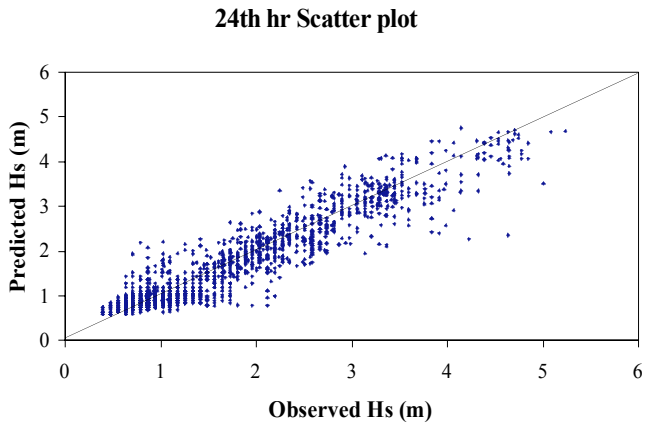

Fig. (6). Scatter Plot between observed and MT- predicted wave heights for 24-hr lead time (Station SW2).

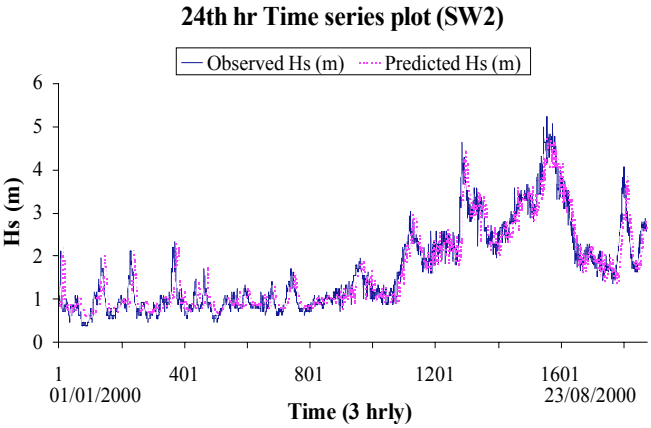

Fig. (7). Time series plot between observed and MT-predicted wave heights for 24-hr lead time (Station SW2).

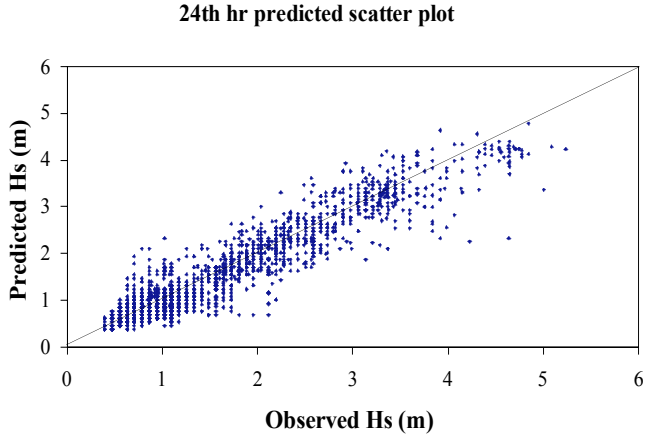

Fig. (8). Scatter Plot between observed and GP-predicted wave heights for 24-hr lead time (Station SW2).

GUI FOR REAL TIME WAVE FORECASTING

The techniques developed to carry out real time forecasting of significant wave heights as described in preceding sections can be put into practice for the benefit of users through a Graphical User Interface (GUI). This section describes the development of such GUI as well as its implementation over the observation stations. The buoy data in the

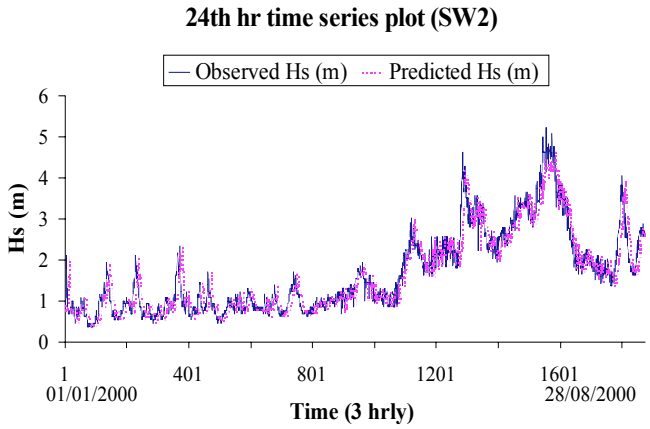

Fig. (9). Time series plot between observed and GP-predicted wave heights for 24-hr lead time (Station SW2).

form of 3-hourly values of Hs are routinely collected and transmitted to a web based server at NIOT, India through a satellite transceiver. These measurements form the input for operating on the GUI. The data are processed in the program itself and the model for the forecast is run. These models are based on the selected form of the ANN architecture for different time intervals. The GUI integrates all models for different stations and gives forecast for any of the selected location. A help file is also designed for the user to help in GUI implementation. The forecasts are displayed in the GUI window along with the plot of observed and predicted significant wave heights over the next 72 hrs. The GUI can also save and print the plots along with the numerical values. The development of the GUI including the data processing is done under the Matlab environment. The values of forecasted significant heights can be saved so that the trend can be reviewed from time to time. The plots and numerical values can be printed to maintain past records. The computational time to execute the GUI interface for a station is 19.129 sec. The GUI provides a logical and integrated interface to the user in an easy-to-use style. The operation of the GUI by any registered user will be as follows:

Open Matlab window and set the "current Directory" to the folder where GUI files are saved.

1. Type "INDIA" in the command window to run the GUI. (Fig. **13**)
2. Click on the selected station whose forecast you want to see. (Fig. **13**)
3. Load the data (XLS format) by clicking on the command "LOAD"(Fig. **14**). Thereafter run the program (click on "RUN") to get forecasts at the selected station over a period of the next 24 hrs.
4. Click on "Save PLOT" (Fig. **14**) to save the plot of the forecasted values over a period of subsequent 24-hr (as dotted line) as well as observed values over the preceding 24 hrs period (full line).
5. Click on command "Save"(Fig **14**). This will save the forecasts values in the XLS format which can be appended after every run. Also "PRINT" command is there to print the plot and values as seen in the GUI window.

The GUI includes intelligent software. It has the provision to take care of missing values in data. If a value at the

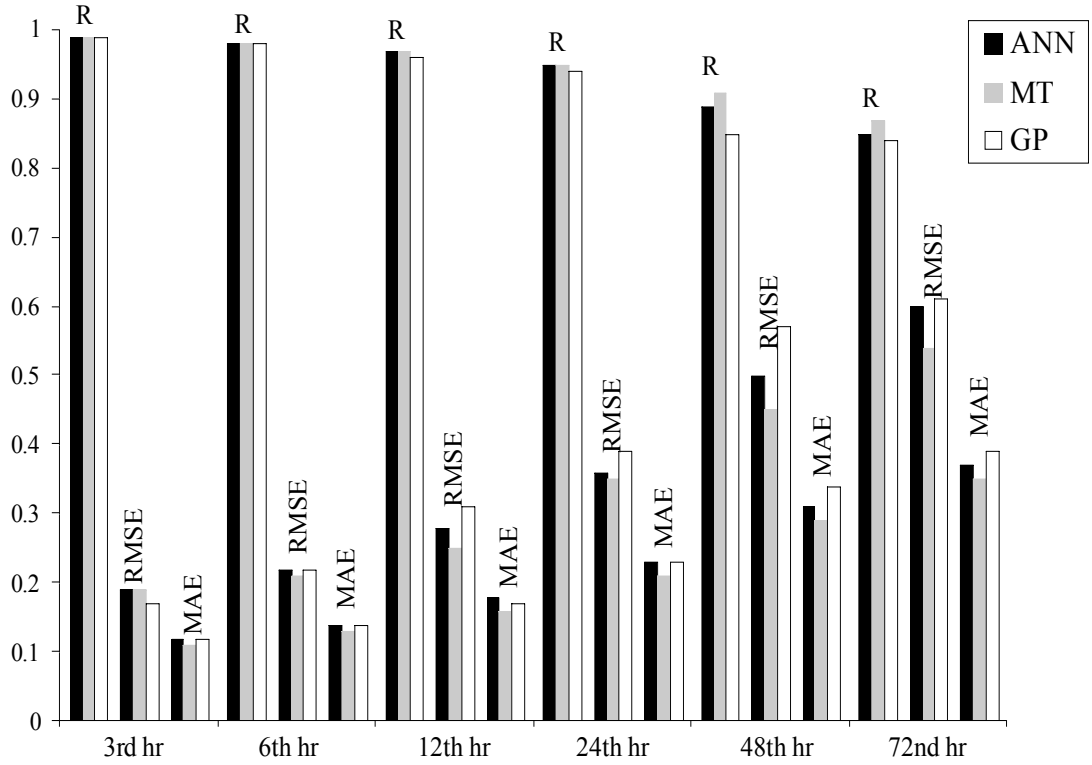

Fig. (10). Comparison between ANN, MT and GP error results for station DS1.

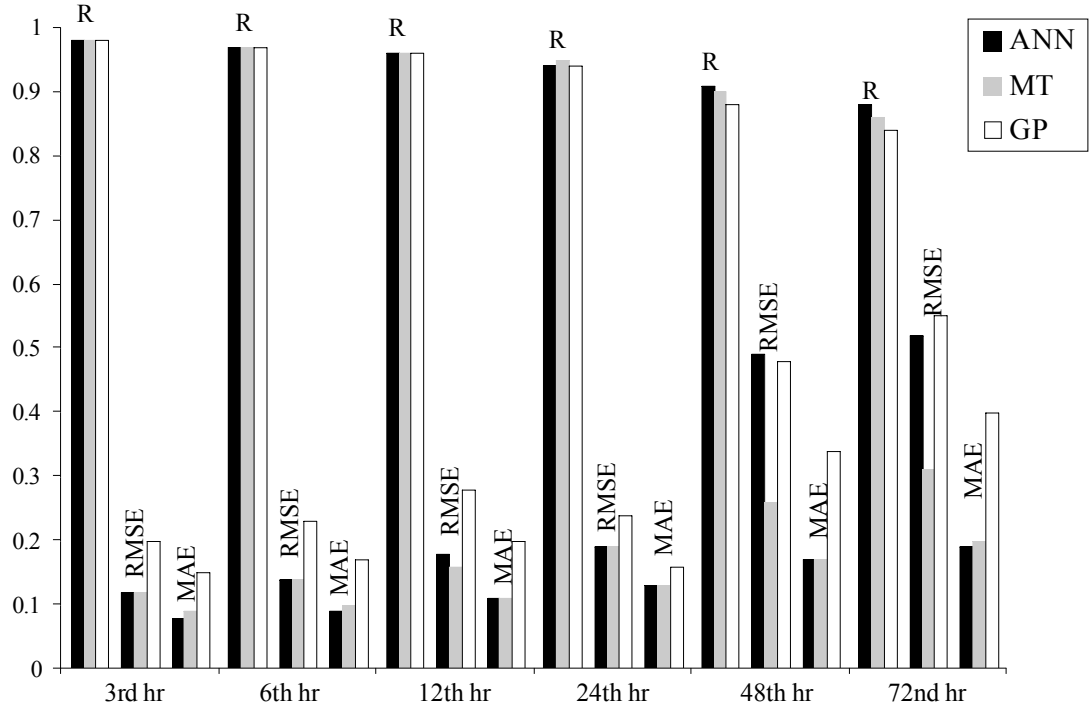

Fig. (11). Comparison between ANN, MT and GP error results for station SW4.

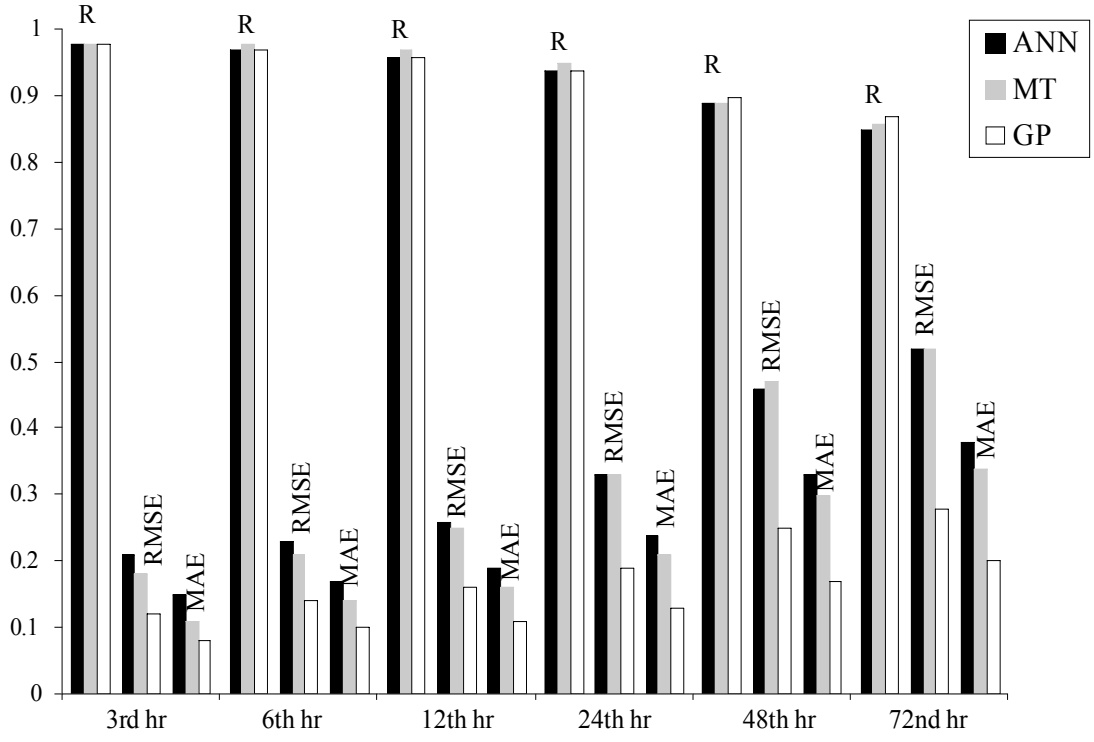

Fig. (12). Comparison between ANN, MT and GP error results for station SW2.

current time step is not recorded, such wave height is immediately evaluated using the temporal regression based on univariate time series analysis for each of the stations and thereafter used for the forecasting purpose. If the missing values exceed a period of two weeks then the same are calculated through a separate ANN that performs spatial mapping. Although at present the GUI shown in Fig (**13**) has been developed only for three stations, namely SW2, SW4 and DS1, it can be easily extended to cover any other buoy stations as desired.

CLOSURE

The present study involves confirmation of accuracy of the predictions through comparison with actual wave height

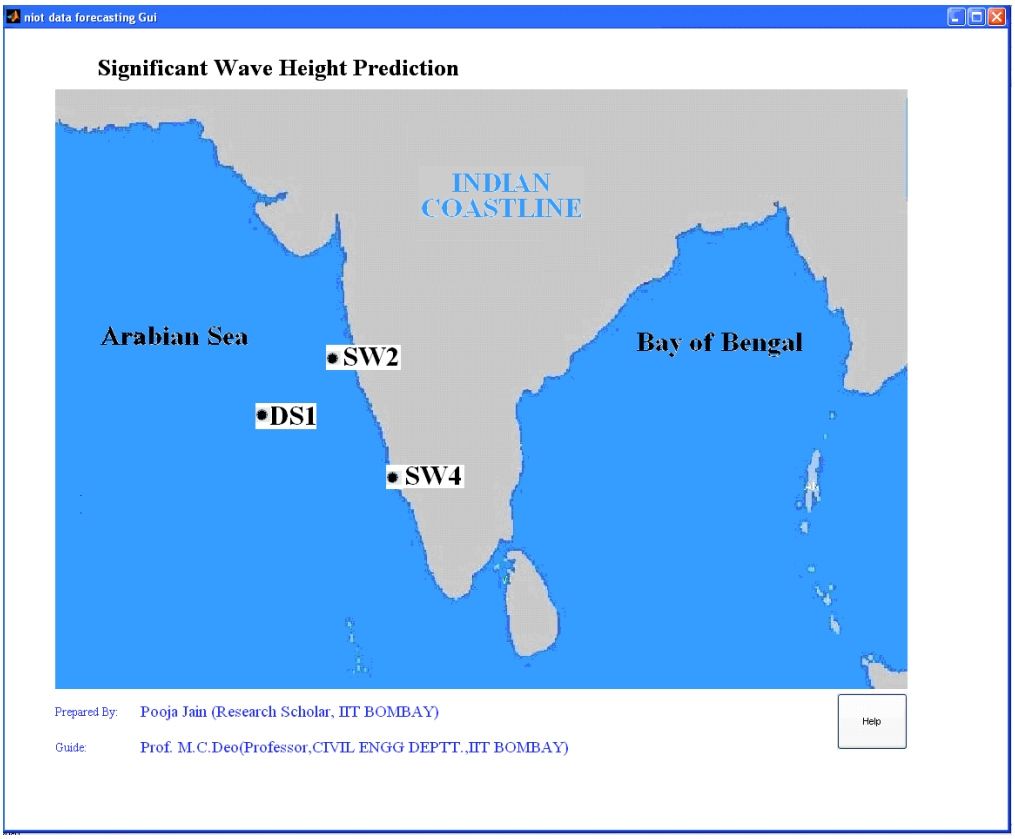

Fig. (13). Front page of the GUI displaying stations for online forecast.

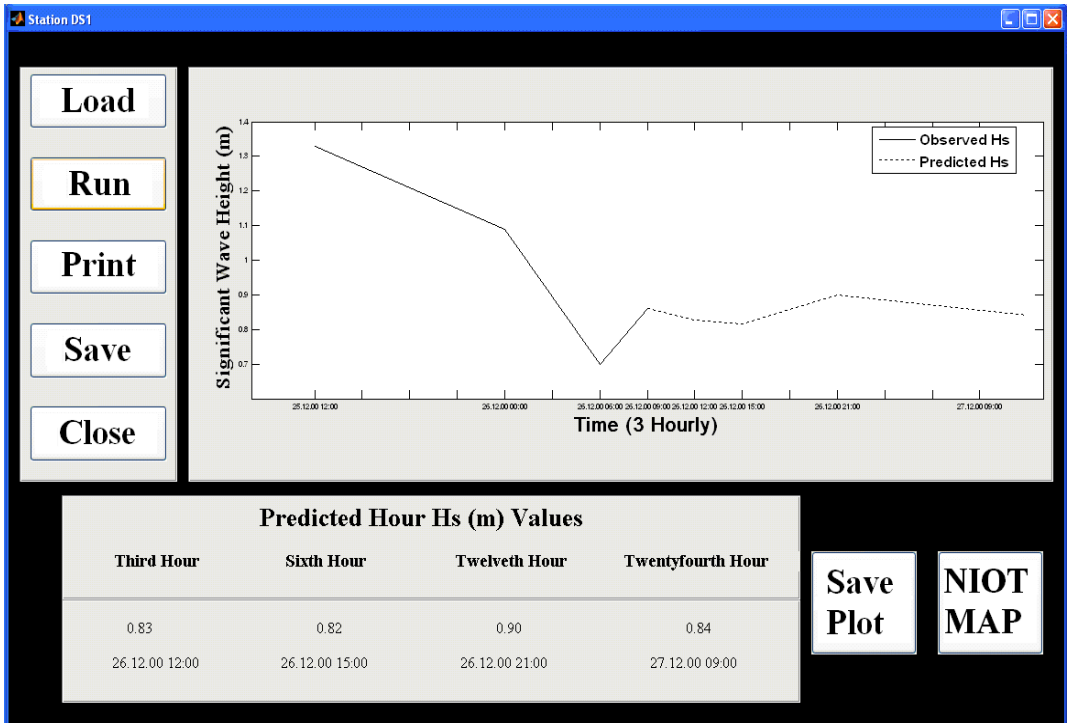

Fig. (14). GUI window for station DS1 after using the LOAD and RUN options.

observations. It would be of interest to know in future how these predictions compare with those of numerical models such as WAM and SWAN.

Similarly the current work does not include comparison with stochastic models such as AR, ARMA, ARIMA since past studies [22] have discussed this issue and have commented on usefulness of methods like ANN accordingly.

CONCLUSION

The preceding sections addressed the problem of real time forecasting of waves for warning times up to 72 hours at three locations along the Indian coastline using alternative techniques of ANN, GP and MT. The models have been calibrated after experimenting with innumerable input-output combinations and ANN training algorithms. A simple feed forward type of network trained using an appropriate training scheme was found to be sufficient for this work involving auto-regressive ANN's.

The missing values in the wave height time series were filled up using temporal as well as spatial correlation approaches.

Both MT and GP results were competitive with that of the ANN forecasts and hence the choice of a model should depend on the convenience of the user.

The selected tools were able to forecast satisfactorily even up to a high lead time of 72 hrs. It is however recognized that these accuracies are possible in the present moderate environment where the target waves were less than around 6 m and 2.5 m for the offshore and coastal stations respectively.

A graphical user interface was developed for the benefit of any user to obtain the forecasts at any of the locations considered through a web based operation. The software takes care of missing values in an intelligent manner. Although the current GUI caters to only three stations along the west coast of India, it can be easily extended to cover additional data buoy stations.

ACKNOWLEDGEMENTS

Authors thank Naval Research Board, India for providing financial support. Thanks are also due to Dr K Premkumar and Dr G Latha of NIOT, Chennai for sparing data and for fruitful technical discussions. The help of Dr S N Londhe was crucial in developing the GUI.

REFERENCES

[1] http://www.niot.res.in/ndbp/bm.pdf [March 1, 2007].

[2] M.C. Deo and C.S. Naidu, "Real time wave forecasting using neural networks", *Ocean Engineering*, vol. 26, pp. 191-203, 1999.

[3] O. Makarynskyy, "Improving wave predictions with artificial neural networks", *Ocean Engineering*, vol. 31, pp. 709-724, 2004.

[4] P. Jain and M.C. Deo, "Large lead time forecasts of significant waves on real time basis using ANN", *Coastal Engineering Journal*, World Scientific, 2008 (accepted CEJ-06010-21)

[5] B. Kosko, *Neural networks and fuzzy systems*. Prentice Hall, 1992.

[6] P.D. Wasserman, *Advanced methods in neural computing*. Van nostrand Reinhold, New York, 1993.

[7] K.K. Wu, *Neural networks and simulation method*s. Marcel Decker, New York,1994.

[8] P. Jain and M.C. Deo, "Neural networks in Ocean Engineering", *International Journal of Ships and Offshore Structures*, vol.1, pp. 25-35, 2006.

[9] Koza. Genetic Programming-On the programming of computers by means of natural selection. MIT Press, 1992.

[10] J.P. Drecourt, "Application of neural networks and genetic programming to rainfall runoff modeling", Danish Hydraulic Institute (Hydro-Informatics Technologies), HIT, 1999, June, D2K-0699-1.

[11] P.A. Whigham and P.F. Crapper, "Modeling Rainfall-Runoff using genetic programming", *Mathematical and Computer Modeling Canberra, Australia*, vol. 33, pp. 707-721, 2001.

[12] V. Babovic, R. Kanizares, H.R. Jenson and A. Klinting, "Neural networks as routine for error updating of numerical models", *Journal of Hydraulic Engineering, ASCE*, vol. 127, pp. 184-193, 2004.

[13] V. Babovic, J.P. Drecourt, M. Keijzer and P.F. Hansen, "A data mining approach to modeling of water supply assets", *Urban Water*, vol. 4, pp. 401-414, 2002.

[14] C. Fonlupt, "Solving the ocean color problem using genetic programming", *Journal of Applied Soft Computing*, vol.1, pp. 63-72, 2001.

[15] R. Kalra and M.C. Deo, "Genetic Programming to retrieve missing information in wave records along the west coast of India", *Applied Ocean Research, Elsevier*, in print, paper no. APOR-D-07-00029R2. 2008.

[16] K. Ustoorikar and M.C. Deo, "Filling up gaps in wave data with genetic programming", *Marine Structures,* Elsevier, 2008; doi:10.1016/j.marstruc.2007.12.001.

[17] J.R. Quinlan, "Learning with continuous classes", Proc. of Australian Joint Conf. on AI, World Scientific, Singapore, 1992, pp. 343-348.

[18] D.P. Solomatine and Y. Xue, "M5 model trees compared to neural networks, application of flood forecasting in the upper reach of the Huai River in China", *ASCE Journal of Hydrologic Engineering*, vol. 9, pp. 491-501, 2004.

[19] I.H. Witten and E. Frank, "*Data mining, practical machine learning tools and techniques with java implementations*", Morgan Kaufmann, Los Altos, CA, 2000.

[20] S. Haykin, *Neural networks: A comprehensive foundation*. Prentice-Hall, Englewood Cliffs, NJ, 1999.

[21] P. Jain and M.C. Deo, "Real time wave forecasts off western Indian coast" *Applied Ocean Research*, Elsevier, vol. 29, pp. 72-79, 2007.

[22] J.D. Agrawal and M.C. Deo, "Online wave prediction", *Marine Structures*, Elsevier, Oxford, UK, vol. 15, pp. 57-74, 2002.

Received: March 26, 2008 Revised: May 13, 2008 Accepted: July 29, 2008