Ocean Engineering 35 (2008) 1166–1172

Contents lists available at ScienceDirect

Ocean Engineering

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

Real-time wave forecasting using genetic programming

Surabhi Gaur, M.C. Deo *

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

ARTICLE INFO

Article history:
Received 4 March 2007
Accepted 17 April 2008
Available online 30 April 2008

Keywords:
Genetic programming
Wave forecasts
Wave heights
Real-time forecasting

ABSTRACT

The forecasting of ocean waves on real-time or online basis is necessary while carrying out any operational activity in the ocean. In order to obtain forecasts that are station-specific a time-series-based approach like stochastic modeling or artificial neural network was attempted by some investigators in the past. This paper presents an application of a relatively new soft computing tool called genetic programming for this purpose. Genetic programming is an extension of genetic algorithm and it is suited to explore dependency between input and output data sets. The wave rider buoy measurements available at two locations in the Gulf of Mexico are analyzed. The forecasts of significant wave heights are made over lead times of 3, 6, 12 and 24 h. The sample size belonged to a period of 15 years and it included an extensive testing period of 5 years. The forecasts made by the approach of genetic programming indicated that it can be regarded as a promising tool for future applications to ocean predictions.

1. Introduction

Real-time forecasting of waves over a time step of a few hours at a specified location is required in operational planning of any engineering activity in the ocean. This is traditionally done by converting wind-related information to waves. In 1960s and 1970s empirical models like SMB and Darbyshire (US Army, 1984) were practiced for this purpose, while numerical models such as WAM and SWAN solving the wave energy balance equation (WMO, 1988; Komen et al., 1994) became popular in 1980s and 1990s. The latter methods have the advantage of mathematical rigor as well as that of large temporal and spatial coverage associated with them. Since last 15 years or so collection of wave data through rider buoys moored at a large number of locations around the world became very common and this gave rise to demand and also availability of station specific forecasts on real-time or online basis (Deo and Naidu, 1999; Makarynskyy, 2004). For making such ‘point’ forecasts time-series-based models like auto-regressive (AR), auto-regressive moving average (ARMA), auto-regressive integrated moving average (ARIMA) and Kalman filter are suitable. While AR and ARMA models are applicable for stationary time series of a short term, ARIMA and Kalman filter models can cater to data non-stationarity and hence can model a long-term series. Apart from forecasting, such stochastic models are also popular in carrying out simulation of a wave time history. More information on such a task along with a good account of related past works can be seen in Cunha and Guedes Soares (1999). Artificial neural network (ANN) is another alternative in this category. While forecasting using a time history the model (autoregressive stochastic or ANN) is fed with an input consisting of a sequence of previous observations so that it recognizes a hidden pattern in such a sequence and accordingly forecasts the future value in continuation. It is presumed in this process that all causative factors are reflected in the very occurrence of historical values. The application of time-series-based methods to wave forecasting being relatively new compared to the numerical models it requires a continued research with innovative modeling methods.

Presumably, the first report on application of ANN to online wave forecasting is due to Deo and Naidu (1999), who used this soft computing tool to forecast ocean wave heights with varying warning times off the east coast of India. Later Agrawal and Deo (2002) compared the ANN results with those derived from stochastic models of ARMA and ARIMA and found that the ANN was more accurate than the latter for shorter intervals. However, they found that for longer intervals both the neural and stochastic techniques produced similar results. Makarynskyy (2004) attempted to improve long-range predictions of significant wave heights and zero crossing wave periods using ANN with the help of different correction approaches. Londhe and Panchang (2006) carried out real-time forecasts at five locations in the Gulf of Mexico and noticed anomalies in high interval forecasts of hourly wave heights. Their predictions were much satisfactory up to the interval of 6 h, following which inaccuracy (especially at the peaks) and lag started developing to a large extent and the coefficient of correlation between the predicted and observed values dropped to around 0.40 for a 24 h prediction. The authors

* Corresponding author. Tel.: +91 22 2576 7330; fax: +91 22 2576 7302.
E-mail address: mcdeo@civil.iitb.ac.in (M.C. Deo).

0029-8018/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.oceaneng.2008.04.007

further noticed that the low prediction accuracy at the peaks was due to the small period of time in which the peak forms compared with the prediction interval. Kanbua et al. (2005) provided online forecasting up to 24 h based on a cause-effect modeling in which the wind speed, wind fetch and duration, together with the water depth and the significant wave height at previous time steps were given as input. They reported that the root mean square error (RMSE) increased from 0.06 to 0.15 m when the forecasting interval changed from 3 to 24 h. The authors also found that the ANN performed better than the numerical model 'WAM' which underestimated waves by 20 percent.

From the above discussion it is clear that there is a scope to explore usefulness of alternative soft computing tools in the area of real-time wave forecasting. In this study the soft approach of genetic programming (GP) has been selected since like ANN it is also a purely non-linear technique of modeling.

2. Genetic programming

The concept of genetic programming is borrowed from the process of evolution occurring in nature, where the species survive according to the principle of 'survival of the fittest'. The GP is similar to genetic algorithms (GA) but unlike the latter its solution is a computer program or an equation as against a set of numbers in the GA. A good explanation of various concepts related to GP can be found in Koza (1992). Information on basic GP operations like reproduction, mutation, and crossover is given in Appendix A for the benefit of readers not familiar with this technique.

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 'offspring's' by following the process of reproduction, mutation and crossover. The creation of offspring's continues (in an iterative manner) till a specified number of offspring's in a generation are produced and further till another specified number of generations are created. The resulting offspring at the end of all this process (an equation or a computer program) is the solution of the problem. The GP thus transforms one population of individuals into another one in an iterative manner by following the natural genetic operations like reproduction, mutation and crossover.

The step-by-step procedure involved in the implementation of the GP is further explained below:

1. Create initial random population of individuals (equations or programs) of a certain size by randomly picking up the same from a set of terminals (consisting of input variables and constants) and functions (involving operators like, multiplication, addition, subtraction, division, square root, log, etc.).
2. Evaluate the fitness of each individual in a population through some criterion like the RMSE.
3. Select individuals or parents (normally probabilistically through a tournament involving comparing two parents at a time and thereafter short listing the winner for further competition).
4. Generate new offspring's (individuals) from these parents by a, b and c below:
   a. *Reproduction:* Copy the best program as it is as per the fitness criterion and include it in the new population. Increase individuals by 1.
   b. *Crossover:* Select two individuals as per the fitness. Perform crossover. Insert the two individuals into the new population. Increase individuals by 2.
   c. *Mutation:* Select one individual as per the fitness. Perform mutation. Insert the mutant into the new population. Increase individuals by 1.
5. If the number of individuals (offspring's) equals a maximum selected number, increase the number of generations by 1 and go to step 6; otherwise increase the individuals by repeating steps 2–5.
6. If the number of generations is equal to a certain maximum value, terminate the program; otherwise repeat steps 2–5.

2.1. Previous works on applications of GP

Applications of GP in coastal engineering are difficult to find, although the same in water-related engineering started around 6 years ago. Unlike the other soft computing tool of artificial neural networks, they are restricted to relatively fewer sub-areas and include rainfall-runoff modeling (Whigham and Crapper, 2001), modeling of risks in water supply (Babovic et al., 2002), modeling of waste water treatment plants (Hong and Rao, 2003) and river flow prediction (Drunpob et al., 2005). Applications of GP to solve problems in coastal engineering are conspicuous by their near absence. The current work therefore assumes significance from this consideration.

2.2. Implementing GP

In the current work software Discipulus (Francone, 1998) was used to generate the GP programs, while TurboC in the C++ environment was employed to run the evolved programs and to implement them over a new data set. As explained in the subsequent section about two thirds of the sample size had been used for calibration of the GP models while the remaining ones were employed to test or validate the same. A typical choice of the initial GP control parameters was as follows: The population size was 500 while the number of generations was 300. The mutation frequency was 90 percent while the crossover frequency was 50 percent. Values of these control parameters were selected initially and thereafter varied in trials till the best fitness measures were produced. The fitness criterion was the mean squared error between the actual and the predicted value of the significant wave height. The statistical error criteria of correlation coefficient (R), RMSE and mean absolute error (MAE), have been used in this study to compare the GP predictions with the actual observations and these were evaluated by using Matlab, which also facilitated generation of the scatter plots between the target output and the one obtained through GP.

3. Analysis and results

The present work aims at the use of GP to carry out the real-time forecasting of significant wave heights over time steps of a few hours or a day at a specified location.

The database used in the study pertains to two selected locations of the US coast and available for free download on the website of National Data Buoy Center (NDBC) (http://www.ndbc.noaa.gov). The choice of these two stations was governed by regularity and continuity in the reported values. One of the stations chosen (station 42001) is in deep water while the other one (station 42020) is near the coast of 88 m water depth (see Fig. 1). The choice of two stations enabled confirmation of the results obtained at one location with another site in a different depth regime. The measurements belonged to hourly values of the significant wave height over the duration of 1990–2004.

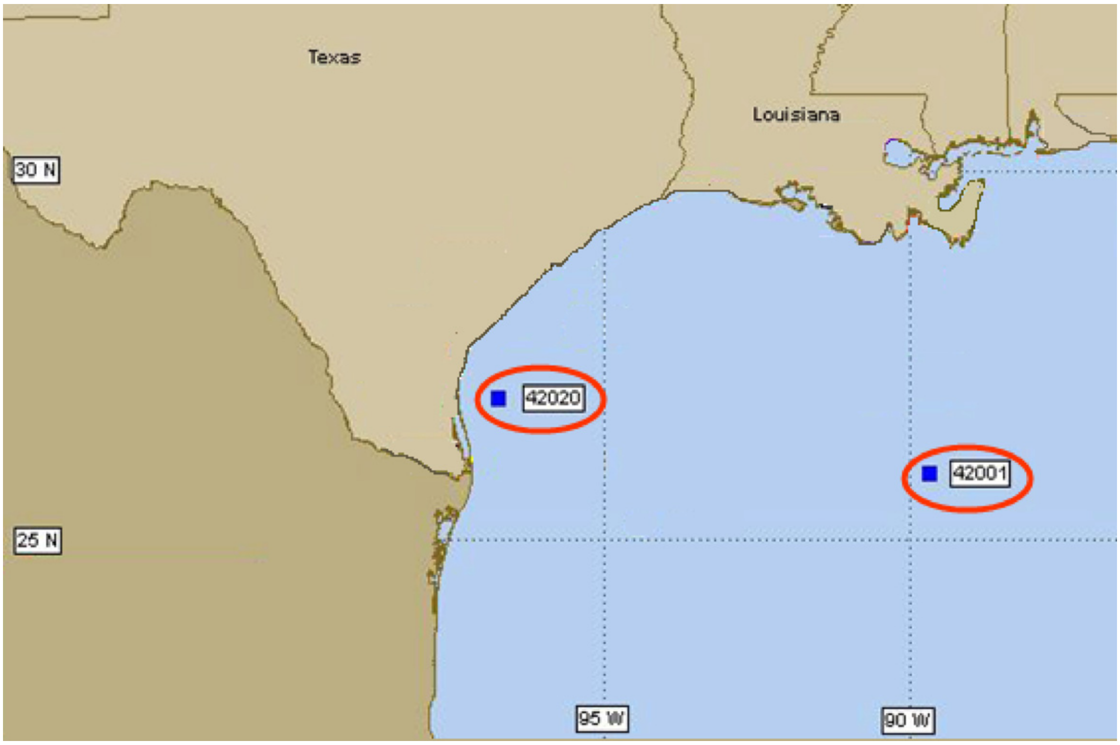

Fig. 1. The location map.

Considering the statistical homogeneity of the monthly measurements it was decided to develop separate GP models for every month. Hourly values of significant wave height, H_s, for each month from 1990 to 2004 were considered for this purpose. The training data belonged to the measurements of H_s for 5 years from 1990 to 1994, while the validation set consisted of hourly values of H_s for 5 years from 1995 to 1999. The testing or actual application was done by giving only the input set from 2000 to 2004. The output was then compared with actual measurements with appropriate lead times and the error plots and statistics were accordingly derived. Different GP models were developed to make forecasts for different lead times. The lead times were, 3, 6, 12 and 24 h. By trials it was observed that a sequence of three preceding observations was sufficient for the GP model to understand a hidden pattern in measured values and predict the subsequent value accordingly. The outcome from each model at the end of training and validation was in the form of the best computer program. This was further employed to make predictions of H_s with different lead times during the 5-year testing period of 2000–2004.

The closeness between the GP forecasts and their actual observations was qualitatively ascertained with the help of time history and scatter plots showing predicted values versus the observed ones along with their deviations from the ideal fit line and quantitatively checked using the three error criteria of correlation coefficient (R), RMSE, and MAE. The correlation coefficient provides the extent of a linear association between the measured and the predicted values, while the RMSE shows an overall prediction quality, especially when iterative procedures are involved. The MAE on the other hand shows a uniform goodness of fit without giving undue weightage to the higher observations unlike the correlation coefficient. Multiple criteria are necessary as a trade-off between the advantages and disadvantages of each one of them.

Figs. 2–5 show typical examples through scatter plots of predicted versus observed values along with corresponding time history comparisons for 3, 6, 12 and 24 h prediction cases, respectively, in the month of July at location 42020. It may be noticed from the figures that smaller prediction intervals are associated with better prediction accuracies. This can be expected considering higher correlations between the values separated by shorter periods. Some times during testing at longer prediction intervals the higher H_s values were found to be either over- or under-predicted as typically shown in the example of Figs. 4 and 5. A possible reason for this could be the lack of sufficient involvement of higher H_s values during the training imparted.

The match of the GP based forecasting with actual observations for the testing period of 5 years in terms of the three error criteria is shown in Figs. 6–8 for the location 42001 and in Table 1 for the site 42020. These figures and the table show how the three statistics change for each month from January to December over the two locations involved. For station 42001, the R values were noticed to be high, varying from 0.83 to 0.93 for 3 h predictions, indicating excellent predictions. Similarly the RMSE and MAE are low, ranging from 0.17 to 0.43 and 0.08 to 0.24 m, respectively. The corresponding 3 h predictions for station 42020 were marginally better than those of site 42001, where the R, RMSE and MAE changes from 0.89 to 0.97, 0.16 to 0.40 and 0.07 to 0.24 m, respectively, across all months. The better predictions at site 42020 could be due to relatively less gaps (and more consistency as a result) in the measured time series of H_s values. From the long-term average values of significant wave heights at these locations shown on the web site http://www.ndbc.noaa.gov/images/climplot/42001_jpg it appears that very high wave activity prevailed in the months of March, August and October. Such higher wave occurrences were however not found to have a clear link or one-to-one correspondence with the accuracy of predictions (higher R and lower RMSE and MAE) for different months.

When it comes to forecasting for a period of next 6 h, almost similar accuracy as the previous 3 h prediction was seen. Typically for station 42020 the R, RMSE, MAE vary from 0.82 to 0.95, 0.09 to 0.47 and 0.09 to 0.30 m, respectively.

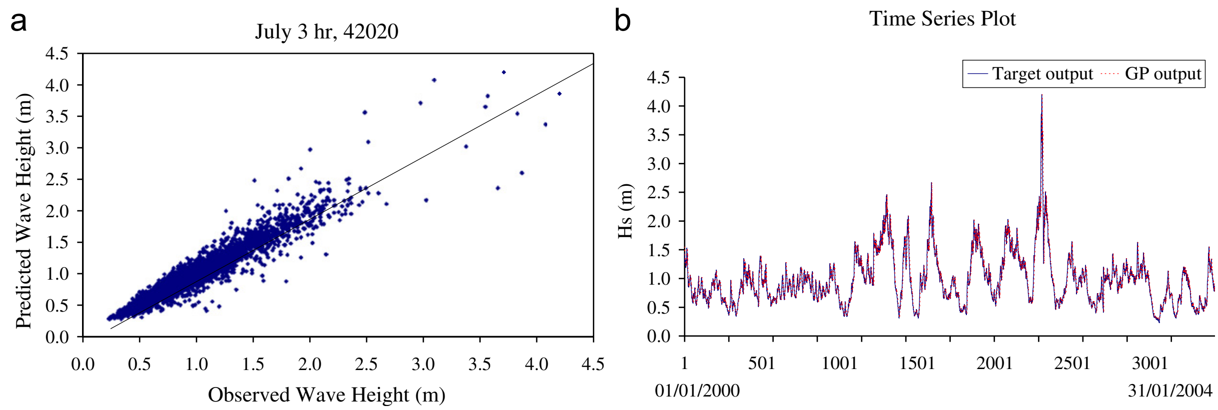

Fig. 2. Predicted versus observed *Hs* values (lead time: 3 h; month: July; Station:42020).

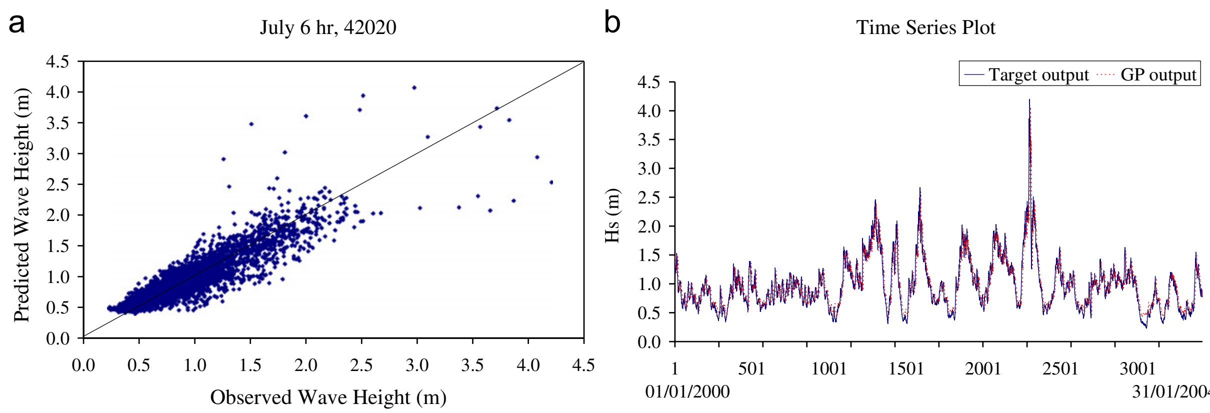

Fig. 3. Predicted versus observed *Hs* values (lead time: 6 h; month: July; Station:42020).

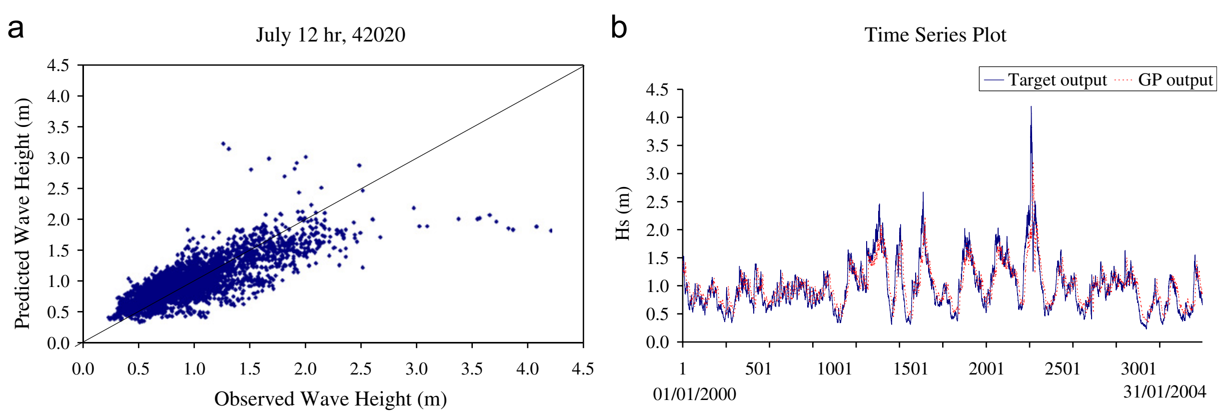

Fig. 4. Predicted versus observed *Hs* values (lead time: 12 h; month: July; Station:42020).

The accuracy levels were non-uniform, when it comes to 12 h predictions with R varying from 0.62 to 0.92 and RMSE and MAE changing from 0.25 to 0.59 and 0.12 to 0.44 m respectively at station 42020. The drop in the forecasting accuracy is expected in the light of the fact that values spaced widely away from each other would have smaller dependence and hence become difficult to model as a sequence.

For the lead time of 24 h, the prediction accuracy further dropped down. For station 42001 (Fig. 6) the R value with an exception of the month of June ranges from around 0.60 to 0.75 while the RMSE and MAE vary from 0.30 to 0.78 and 0.25 to 0.55 m, respectively. The error statistics at station 42020 generally agree with these values of station 42001 (or even look better in some cases) except the correlation coefficient, which is considerably lower for certain months. It is however mentioned that many previous works failed to achieve even this much accuracy level for the 24 h predictions. Deo and Naidu (1999) while forecasting at site Yanam along the east coast of India, observed that the correlation coefficient between the ANN predictions and actual observations significantly drops from 3 to 24 h prediction, so also Agrawal and Deo (2002) later, who reanalyzed the Yanam site data by using alternate training schemes. Both these works additionally had the shortcomings that they did not produce direct forecast of actual observations, but only provided an

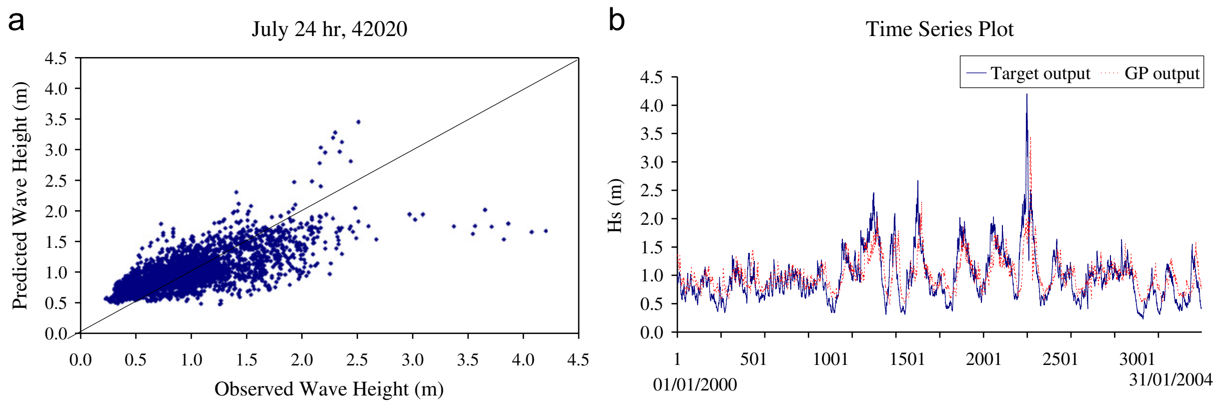

Fig. 5. Predicted versus observed *Hs* values (lead time: 24 h; month: July; Station:42020).

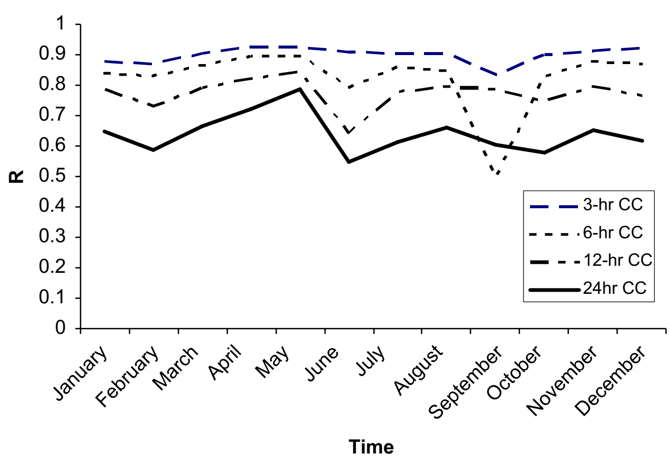

Fig. 6. Variation of correlation coefficients at station 42001.

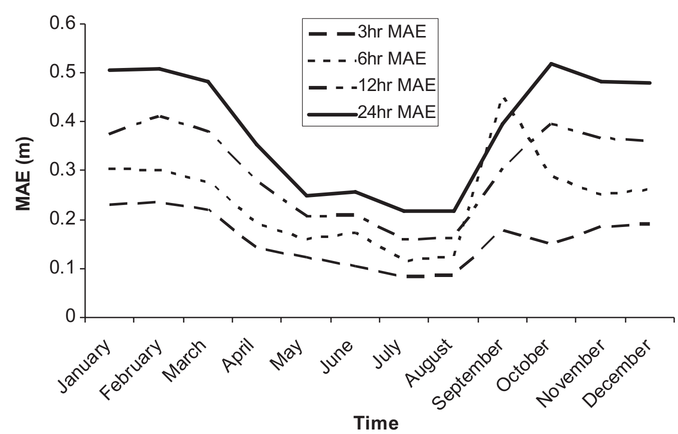

Fig. 8. Variation of MAE at station 42001.

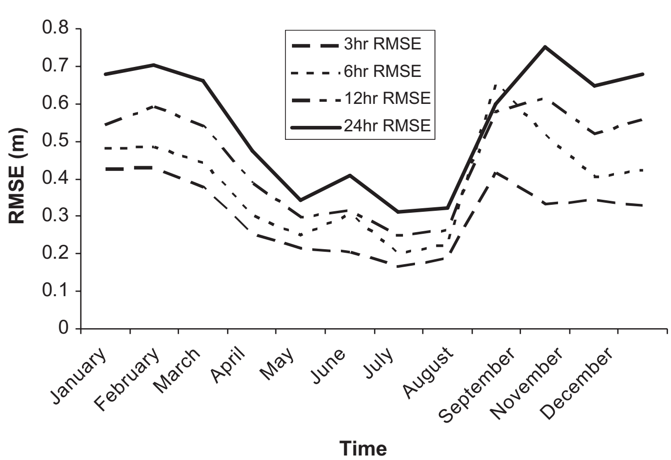

Fig. 7. Variation of RMSE at station 42001.

indirect estimation of average values. Further, these results involved a limited comparison based on low sample sizes unlike the present case and hence they may not be always generally realized at any other location. Makarynskyy (2004) during their ANN based analysis of Atlantic and Irish Sea measurements had similarly noted low correlation coefficients for the 24 h ahead predictions. While experimenting on coastal waters off Tasmania and Portugal with ANN, Makarynskyy (2005) and Makarynskyy et al. (2005) further confirmed high reductions in the correlation coefficients for the 24 h predictions. Londhe and Panchang (2006) have recently reported an exhaustive application of ANN to online forecasting at the US National Data Buoy Centre sites with a large sample size. The problem of handling large forecasting intervals still remained elusive, although they could achieve a correlation coefficient ranging from 0.40 to 0.63 in their works for the 24 h forecasts.

From this discussion, it is clear that the GP-based prediction reported herein may appear to be more attractive than many past works based on ANN especially when all the error statistics are viewed together. The testing sample size used in this study is one of the largest compared to the previous works (4 years) and hence these results should be regarded as relatively more reliable. One of the reasons why the GP worked better here can be the fact that as compared with the ANN the GP might manage a large number of training pairs (typically hourly values for a period of 10 years) more efficiently while the ANN may suffer from problems like over-fitting in such a case. The use of a large sample size as in the present work for training can be regarded as necessary for the problem under consideration in order to account for a large range of variations in long-term observations.

4. The ANN model

An attempt was also made to see if better predictions were possible for larger intervals with the help of an equivalent ANN. An ANN model of autoregressive type was developed. It was calibrated and tested using the same training as well as testing data employed for the previous GP model. The ANN used for this purpose was of feed forward type, which is most commonly used

Table 1
Error statistics for station 42020

	3rd hour	6th hour	12th hour	24th hour
January				
R	0.89	0.84	0.78	0.70
RMSE (m)	0.40	0.47	0.55	0.64
MAE (m)	0.24	0.28	0.40	0.44
February				
R	0.89	0.82	0.64	0.30
RMSE (m)	0.29	0.37	0.50	0.42
MAE (m)	0.17	0.27	0.39	0.50
March				
R	0.94	0.86	0.71	0.47
RMSE (m)	0.23	0.33	0.46	0.59
MAE (m)	0.15	0.24	0.34	0.47
April				
R	0.93	0.85	0.71	0.52
RMSE (m)	0.20	0.29	0.39	0.46
MAE (m)	0.13	0.20	0.28	0.34
May				
R	0.97	0.95	0.92	0.88
RMSE (m)	0.16	0.20	0.25	0.31
MAE (m)	0.07	0.09	0.12	0.16
June				
R	0.91	0.87	0.78	0.62
RMSE (m)	0.20	0.25	0.31	0.39
MAE (m)	0.12	0.17	0.23	0.31
July				
R	0.96	0.91	0.79	0.71
RMSE (m)	0.21	0.31	0.27	0.31
MAE (m)	0.11	0.17	0.20	0.24
August				
R	0.94	0.89	0.77	0.62
RMSE (m)	0.17	0.21	0.31	0.38
MAE (m)	0.09	0.13	0.20	0.24
September				
R	0.93	0.89	0.79	0.58
RMSE (m)	0.27	0.33	0.45	0.60
MAE (m)	0.15	0.20	0.29	0.40
October				
R	0.95	0.88	0.75	0.50
RMSE (m)	0.19	0.28	0.40	0.54
MAE (m)	0.13	0.20	0.30	0.42
November				
R	0.90	0.83	0.79	0.40
RMSE (m)	0.34	0.43	0.50	0.71
MAE (m)	0.19	0.30	0.38	0.57
December				
R	0.93	0.83	0.62	0.36
RMSE (m)	0.28	0.43	0.59	0.79
MAE (m)	0.19	0.30	0.44	0.64

and where the information flowed only in the forward direction, i.e. from the input layer to the output layer and through a hidden layer—all layers consisting of a set of neurons or computational elements. The number of neurons in the input layer was three and it belonged to three preceding hourly observations of the significant wave heights, like the earlier GP method while the number of output layer neurons was one corresponding to the predicted significant wave height value. The basic concepts and details of working of an ANN can be seen in text books such as Haykin (1999). A variety of learning algorithms were employed to impart training to the network and these ranged from ordinary error back propagation to advanced search-based techniques (Haykin, 1999). The Levernberg–Marquardt algorithm turned out to be the best method of training and hence the same was adopted for testing. In the end however the testing results of such an ANN model were not found to be encouraging for large interval predictions as compared to GP. Typically at station 42001, for the month of January, the testing over the period of 2000–2004 yielded 12 and 24 h wave forecasts with $R = 0.79$ and 0.65 based on GP while the same with ANN produced forecasts with $R = 0.71$ and 0.42, respectively. This difference in the results could be due to a more efficient handling by GP of a very large amount of calibration data involved in this study. However such difference between the GP and the ANN methods needs to be further confirmed with more detailed investigations.

5. Conclusions

The previous sections discussed an application of the soft computing tool of GP to the problem of making online wave forecasts over lead times of 3, 6, 12 and 24 h. The sample size belonged to a period of 15 years and this included an extensive testing period of 5 years. Wave rider buoy measurements available at two locations in the Gulf of Mexico were considered.

Like the recent works of past investigators on real-time wave forecasting based on ANN, the application of GP in this study also resulted in a situation where small-interval forecasts (3 or 6 h) were more accurate than the large-interval ones (12 or 24 h); however, the general level of prediction accuracy seen in the current study would indicate that the soft tool of GP held promise for future applications.

A comparison of results obtained at two different locations showed that a small number of gaps and resulting consistency in data might lead to relatively better forecasts.

A limited comparison of the GP-based predictions with corresponding ANN-based ones indicated the tendency of the GP to perform better at higher forecasting intervals and this could be due to its relatively efficient handling of a very large amount of calibration data involved in this study.

The present work showed that GP can be regarded as a promising tool for its future applications to problems of ocean predictions. It is worth exploring how it performs in carrying out a causal as well as a spatial mapping as against the present temporal one.

Appendix A. Examples of genetic operations

A.1. Generating population

A program $[-q+(\pi)^{1/2})/3\ p]$ is given in Fig. 9 in the form of a tree structure. A population of random trees representing the programs is initially constructed and genetic operations are performed on these trees to generate individuals with the help of two distinct sets; the terminal set T and the function set F. For the Fig. 9, $\{-,+,\surd,/\}\subseteq F$ and $\{\pi, 3, p, q\}\subseteq T$. In order to generate a random tree one has to pick randomly from $T\cup F$, until all branches end up in terminals.

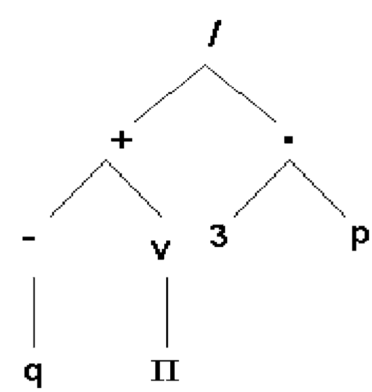

Fig. 9. Program $[-q+(\pi)^{1/2})/3\ p]$ in the form of a tree structure.

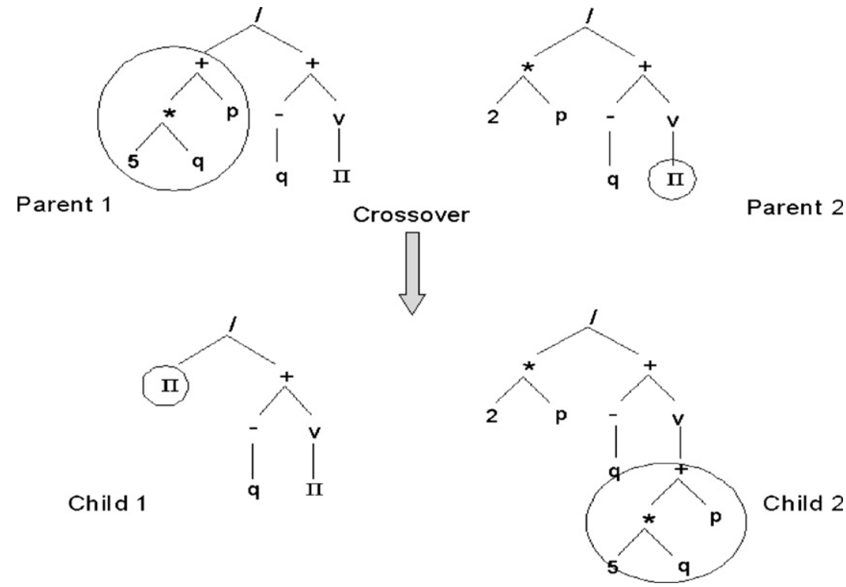

Fig. 10. Crossover.

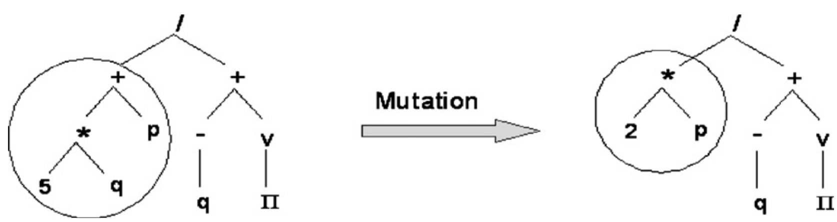

Fig. 11. Mutation.

A.2. Crossover

Two random nodes are selected from inside such program (parents) and thereafter the resultant sub-trees' are swapped, generating two new programs as in Fig. 10.

A.3. Mutation

A sub-tree is replaced by another one randomly (Fig. 11).

A.4. Reproduction

This means an exact duplication of the program if it is found to be acceptable by the fitness criteria.

References

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

Babovic, V., Drecourt, J.P., Keijzer, M., Hansen, P.F., 2002. A data mining approach to modeling of water supply assets. Urban Water 2002 (4), 401–414.

Cunha, C., Guedes Soares, C., 1999. On the choice of data transformation for modelling time series of significant wave height. Ocean Engineering 26 (1999), 489–506.

Deo, M.C., Naidu, S., 1999. Real-time wave forecasting using neural networks. Ocean Engineering 26 (3), 191–203.

Drunpob, A., Chang, N.B., Beaman, M., 2005. Stream flowrate prediction using genetic programming model in a semi-arid coastal watershed. In: Walton, R. (Ed.), Proceedings, World Water and Environmental Resources Congress-2005. ASCE, Anchorage, Alaska, pp. 15–19.

Francone, F.D., 1998. Discipulus Owner's Manual. Machine Learning Technologies, Inc., Littleton, Colorado.

Haykin, S., 1999. Neural Networks—A comprehensive Foundation, second ed. Prentice-Hall.

Hong, Y.S., Rao, B., 2003. Evolutionary Self-Organising Modeling of a Municipal Wastewater Treatment Plant, Water Research, vol. 37. Elsevier, pp. 1199–1212.

Kanbua, W., Supharatid, S., Tang, I.M., 2005. Short term predictions of ocean wave in the Gulf of Thailand: WAM and neural network approaches. Science Asia 31 (3), 243–250.

Komen, G.J., Cavelar, L., Donelan, M., Hasselmann, K., Hasselmann, S., Janssen, P.A.E.M., 1994. Dynamics and Modelling of Ocean Waves. Cambridge University Press.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Bradford Book.

Londhe, S.N., Panchang, V., 2006. One day wave forecasts using buoy data and artificial neural networks. Journal of Atmospheric and Ocean Technology, AMS 9, 1593–1603.

Makarynskyy, O., 2004. Improving wave predictions with artificial neural networks. Ocean Engineering 31 (5-6), 709–724.

Makarynskyy, O., 2005. Artificial neural networks for wave tracking, retrieval and prediction. Pacific Oceanography 3 (1), 21–30.

Makarynskyy, O., Pires-Silva, A.A., Makarynska, D., Ventura-Soares, C., 2005. Artificial neural networks in wave predictions at the west coast of Portugal. Computers and Geosciences 31 (4), 415–424.

US Army, 1984. Shore Protection Manual. Coastal Engineering Research Centre, Washington, DC.

Whigham, P.A., Crapper, P.F., 2001. Modeling rainfall-runoff using genetic programming. Mathematical and Computer Modeling 2001 (33), 707–721.

WMO, 1988. Guide to Wave Analysis and Forecasting, Pub. No. 702, World Meteorological Organization, WMO Secretariat, Geneva.