Predicting wave heights in the north Indian Ocean using genetic algorithm

Sujit Basu, Abhijit Sarkar, K. Satheesan, and C. M. Kishtawal
Meteorology and Oceanography Group, Space Applications Centre, Ahmedabad, India

Received 2 June 2005; revised 2 August 2005; accepted 4 August 2005; published 9 September 2005.

[1] The present work reports the development of a nonlinear technique based on genetic algorithm (GA) for the prediction of wave heights in the north Indian Ocean. Time series observations of surface wind speed and significant wave height at three locations in the Arabian Sea (AS) and the Bay of Bengal (BOB) have been used for developing and testing the technique. The predictions have been compared with persistence forecasts and it has been found that the prediction by GA is always superior to persistence forecast and thus represents a net information gain. The predicted wave heights have been found to be consistent with the results obtained with autocorrelation analysis applied on the respective time series of wave heights. **Citation:** Basu, S., A. Sarkar, K. Satheesan, and C. M. Kishtawal (2005), Predicting wave heights in the north Indian Ocean using genetic algorithm, *Geophys. Res. Lett.*, *32*, L17608, doi:10.1029/2005GL023697.

1. Introduction

[2] Prediction of the future state of a fluid system is the central problem in fluid dynamics. Traditionally, prediction is carried out using numerical models, which use equations of motion derived from first principles. However, such an approach is not always possible in practice. One may not have access to required computing resources for carrying out numerical wave modeling or may not have access to the required initial and forcing data. Also, the requirement may be to forecast the wave height only at a single location, e.g., at a buoy location. In such a case, approximate equations governing the time evolution of the local system (e.g., buoy observed wave heights) can be obtained by model-fitting approaches based on the observed variability of the system evolution. Forecast thus can be achieved by deterministic models directly built from the observations. One such powerful modern approach is based on GA, which is programmed to approximate the equation, in symbolic form, that describes the time series [*Szpiro*, 1997; *Álvarez et al.*, 2001]. A detailed description of the algorithm has recently been provided by *Kishtawal et al.* [2003, 2005]. For the sake of self-consistency of this paper, we sketch the salient features of the algorithm. The GA considers an initial population of potential solutions, which is subjected to an evolutionary process, by selecting from the initial population those equations (individuals) that best fit the data. The strongest strings choose a mate for reproduction whereas the weaker strings become extinct. The newly generated population is subjected to mutations that change fractions of information. Mutation is applied to the individuals of the population, except to the top ranked equation strings in order to avoid inadvertent loss of information. The evolutionary steps are repeated with the new generation. Once the desired fitness strength (defined later in the paper) is reached, the iterations are stopped.

[3] The works of *Takens* [1981], *Casdagli* [1989] and many others have established the methodology for nonlinear modeling of chaotic time series. Explicitly, Takens' theorem [*Takens*, 1981] establishes that given a deterministic time series $\{x(t_k)\}$, $t_k = k\Delta t$, $k = 1 \ldots N$, there exists a smooth map P satisfying:

$$x(t) = P[x(t-\Delta t), x(t-2\Delta t), \ldots x(t-m\Delta t)] \quad (1)$$

where m is called the embedding dimension obtained from a state-space reconstruction of the time series [*Abarbanel et al.*, 1993] and Δt is the sampling time interval, 1 day, in our case. A GA basically tries to obtain the function P(.) in equation (1) that best represents the amplitude function of a chaotic time series, which can then be used to predict the future state of the system. Generally the evolution of a natural dynamical system is not restricted to a single variable, and a nonlinear interaction among several variables is quite common. Such a situation demands the use of multivariate or vectorial time series to obtain the fittest model that can explain a process. The model of connection between different variables, e.g. x, y, and z can be written as:

$$x(t) = P[y(t-\Delta t), y(t-2\Delta t), \ldots y(t-m\Delta t) \ldots \cdot z(t-\Delta t), z(t-2\Delta t) \ldots z(t-m\Delta t)] \quad (2)$$

where $m + 1 \leq t \leq T$, where T is the length of vector time series.

[4] The GA works in the following manner. First, for an amplitude function x(t), a set of candidate equations for P(.) is randomly generated. An equation is stored in the computer as a set of characters that define the independent variables, $y(t-\Delta t)$, $y(t-2\Delta t) \ldots z(t-\Delta t)$, $z(t-2\Delta t)$, etc. in equation (2), and four elementary arithmetic operators (+, −, ×, and /). A criterion that measures how well the equation strings perform on a training set of the data is its fitness to the data, defined in equation (3). The strongest individuals (equations with best fits) are then selected to exchange parts of the character strings between them (reproduction and crossover) while individuals poorly fitted to the data are discarded. Finally, a small percentage of the equation strings' most basic elements, single operators and variables, are mutated at random. The process is repeated a large number of times to improve the fitness of the evolving

0094-8276/05/2005GL023697$05.00

Table 1. Location Details and Observation Duration of the Three Buoys

Buoy	Depth (m)	Latitude	Longitude	Observation Duration
DS1	3800	15.5°N	69.2°E	10 Feb 2001–25 June 2001
SW3	22	15.4°N	73.8°E	11 Mar 2003–15 Aug 2003
DS3	3170	12.1°N	90.8°E	05 Jul 2004–27 Nov 2004

population of equations. The fitness strength of the best scoring equation is defined as:

$$R^2 = 1 - \left[\Delta^2 / \Sigma(x_o - \langle x_o \rangle)^2\right] \quad (3)$$

where $\Delta^2 = \Sigma(x_c - x_o)^2$, x_c is parameter value estimated by the best scoring equation, x_o is the corresponding "true" value, $\langle x_o \rangle$ is the mean of the "true" values of x.

[5] *Szpiro* [1997] showed the robustness of GA to forecast the behavior of one-dimensional chaotic dynamical system. Later, *Álvarez et al.* [2000] applied the GA to real physical systems and used this algorithm for the prediction of space-time variability of the sea surface temperature (SST) in the Alboran Sea. *Kishtawal et al.* [2003] used this algorithm for forecasting summer rainfall over India, whereas, *Álvarez et al.* [2004] used this algorithm to forecast Sea Surface Temperature (SST) and Sea Level Anomaly (SLA) of the Ligurian Sea. Very recently, *Kishtawal et al.* [2005] used the same algorithm to estimate tropical cyclone intensity from satellite observations. In the present study we applied the GA to predict wave heights at different buoy locations in the north Indian Ocean one, two and three days in advance.

2. Data

[6] Several deep-sea and shallow water moored buoys have been functional in the north Indian Ocean since 1997 [*Premkumar et al.*, 2000] under the National Data Buoy Program. We have selected data from three of the buoys (DS1, DS3 and SW3) in the years 2001, 2004 and 2003 representing different oceanic conditions. The DS1 and SW3 are in the AS while DS3 is in the BOB. Location details, depth and period covered are provided in Table 1. The number of observations for the DS1, DS3 and SW3 buoys are 136, 146 and 158 respectively. The reported wind magnitudes and wave heights are available at 3 hour intervals. It is clear that the number of observations in each time series is very few (of the order of 100–150). This number may not be adequate for training other nonlinear data-fitting algorithms like artificial neural network. However, the beauty of the GA lies in the fact that it is able to generate prediction equations from only a few observations as will be shown in this study.

3. Results

[7] GA was first applied to the univariate time series at the DS3 buoy location. It is well known that skill of any prediction method can be judged by comparing its performance with the performance of persistence model [*Álvarez et al.*, 2004] defined by the equation A(t + 1) = A(t). A predictor system showing better performance than persistence indicates a net information gain versus the hypothesis that the best forecast is provided by the present state. In our case, however, it was found that the GA was able to improve upon the performance of the persistence model only marginally in spite of various combinations of parameters employed by the algorithm, like time lag, maximum number of symbols allowed. etc. It was thus thought that it would probably be better to predict the fluctuation after subtracting a nonlinear trend. Accordingly, a nonlinear trend was fitted to the data using a 1-2-1 recursive filter [*Hartmann and Michelson*, 1989]. This trend was subtracted from the data time series to obtain the time series of fluctuations to which the GA was applied. It was found that GA applied to the univariate time series of fluctuations is unable to significantly improve upon the persistence forecast either. Hence, GA was applied to the multivariate time series. The predictors employed were past values of winds and past values of wave fluctuations. This was found to lead to a substantial improvement of the forecast. The trend was also forecast using the same algorithm. This time, however, it was sufficient to apply the algorithm to the univariate time series of nonlinear trend. This is probably due to the fact that the time series of trend is very smooth unlike the time series of fluctuations. The two forecasts (of the fluctuation and the trend) were added to provide forecasts of SWH with various time leads like one day, two days and three days. It was found that the forecasts were significantly better than the persistence forecast in each of the cases studied. This was found out by computing the root mean square errors of forecast by persistence model and by GA (Table 2).

[8] We now digress a bit and comment about the GA parameters employed by us. We used 200 numbers of individuals in each population in majority of the cases studied. In a few individual cases 120 numbers of individuals in the population were used. Total number of arguments and operators allowed was 24 in each of the cases studied. These parameters were largely selected by trial and error. Of course we were also guided by the works of previous researchers, mostly by *Álvarez et al.* [2001]. The

Table 2. Comparison of Forecast by GA With Persistence Forecast[a]

	Forecast (Days in Advance)								
	1 Day			2 Days			3 Days		
Buoy	RMS in m (Persistence)	RMS in m (Genetic)	R^2	RMS in m (Persistence)	RMS in m (Genetic)	R^2	RMS in m (Persistence)	RMS in m (Genetic)	R^2
DS1	0.38	0.19	0.99	0.72	0.36	0.94	1.02	0.70	0.80
DS3	0.22	0.09	0.98	0.36	0.09	0.96	0.41	0.22	0.89
SW3	0.21	0.08	0.99	0.33	0.17	0.96	0.35	0.14	0.92

[a]RMS means root mean square error of forecast in meters. R^2 is the square of the coefficient of correlation between forecasts and observations. This is shown only for the case with GA.

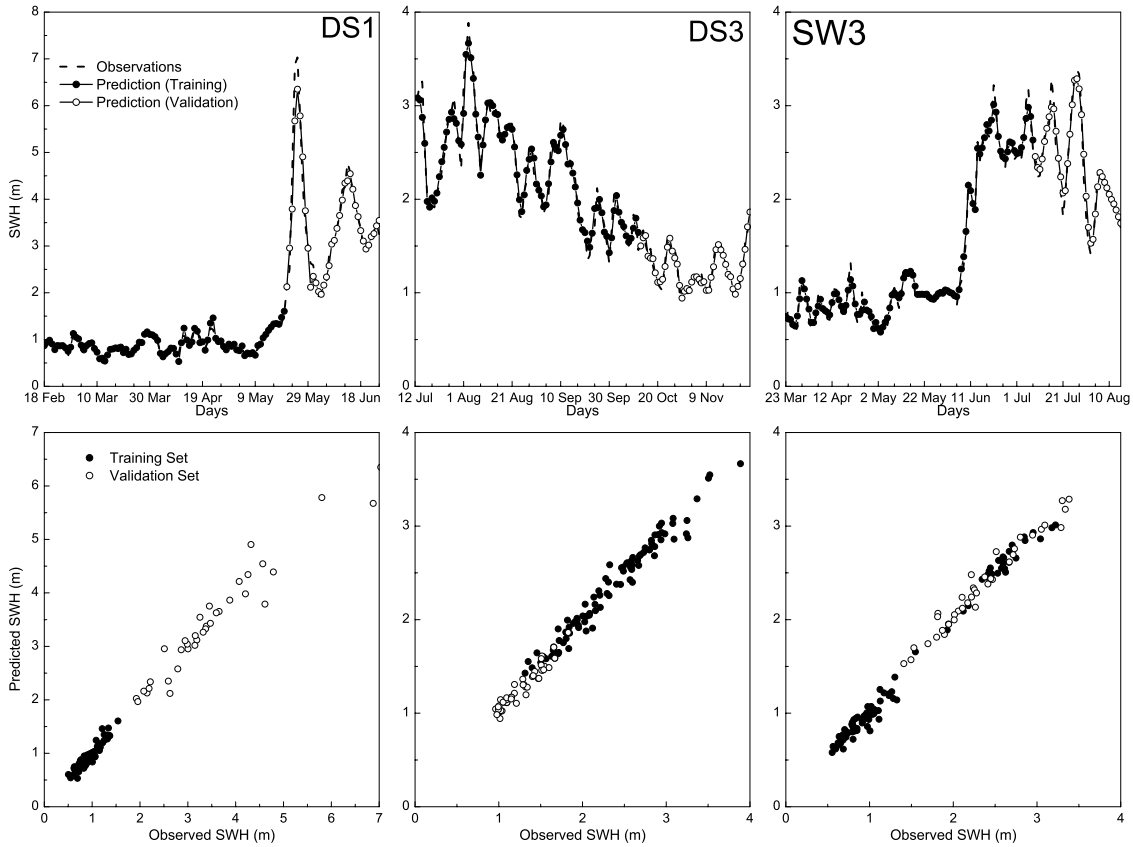

Figure 1. The upper panels show the time series of observed and predicted SWH in meters at the three buoy locations. The predictions are one day ahead forecast. Lower panels show the scatter diagrams of predicted and observed SWH at the corresponding locations.

selection criterion used was the achievement of optimum fitness strength which was fixed at 0.9 for the trend part and $0.35 \sim 0.4$ for the fluctuation part. These choices reflect the fact that trends are smooth functions of time whereas fluctuations are noisier and hence more difficult to predict. The number of generations employed was governed by a simple criterion. Originally this number was assigned the value of 5000. However, iterations were stopped as soon as the optimum fitness strength was reached. This number varied from case to case, but mostly it was of the order of 1000 to 1500. The only parameter that could be theoretically fixed was the number of lags, which is actually equal to the embedding dimension m in equation (3). Estimation of m is possible if a large amount of data is available using a specific algorithm [*Grassberger and Procaccia*, 1983]. Unfortunately in our case, the lengths of the buoy observed wave height time series are too short. Hence the value of m must be fixed ad-hoc. It is known [*Álvarez et al.*, 2004] that small values of m would avoid the system from getting enough information from the past, while big values would degrade the performance of the GA due to dimensional increasing of the searching space [*Álvarez et al.*, 2004]. In our work, m has been fixed mainly by trial and error. It has been seen that m varies from 8 to 12 for the time series of trends while it is as large as 24 or 36 in individual cases of fluctuation time series. This is probably due to the fact that time series of trend is quite smooth, requiring small m for reasonably good prediction, while large values of m are required for predicting fluctuating part, which is not as smooth as the trend.

[9] We repeated the procedure for DS1 and SW3. The result was similar to the case of DS3 buoy. The results are summarized in Table 2. It can be seen that the GA always performs better than the persistence model. In Figure 1, we demonstrate the performance of the method by showing the time series of observations, 1-day ahead predictions (top panel) and the scatter plots of observations and 1-day ahead predictions (bottom panel). In Figure 1, data used for the training process is shown with black dots while data used for the validation process with unfilled dots. For the DS1 and DS3 buoys, the training period consists of first 100 observations, and the remaining observations are used for the validation. For the SW3 buoy, the training period consists of 120 observations. For the sake of economy of space, we are not showing the 2 days and 3 days ahead predictions. The detailed statistics is given in Table 2. It is also quite interesting to see that the validation data cover the entire range of wave heights (high in DS1, medium in SW3 and low in DS3). This points to the capability of GA to predict wave heights of all range.

[10] We further computed the autocorrelation of wave height time series for all the three buoys. As seen in Figure 2, the autocorrelation values for all the cases are falling with increasing lag but are displaying temporal coherence characteristics differing in nature. The fall in autocorrelation is most pronounced in DS1. The shallow AS

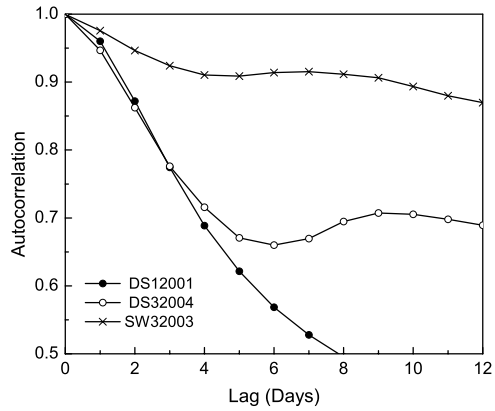

Figure 2. Variation of autocorrelation of SWH with lag at the three buoy locations.

location displays the longest lag. This is possibly due to the fact that shallow water dynamics is very different from deep water dynamics. The shallow water waves are nondispersive, whereas deep water waves are dispersive. Deep water waves are also exposed to winds, swells and currents from all around. In contrast, shallow water waves are somewhat protected from the coastal side. Perhaps, this explains the steadier trend of the autocorrelation of the shallow buoy waves. It is thus not surprising that autocorrelation in the shallow water exhibits different characteristics than deep water. We also computed the autocorrelation after removing the nonlinear trends from the data. This time the shapes of autocorrelation functions were similar at the various buoy locations. The autocorrelation pattern partly explains the results obtained by GA proposed in this work. The R^2 values at SW3 are found to maintain a relatively high magnitude (0.92) even in the prediction of 3rd day and the RMS is the lowest among the three. The buoy DS1, which has a sharper autocorrelation fall, shows a larger degradation in terms of R^2 on the third day. The number of cases discussed in this work is limited and is being reported only to demonstrate the potential of the new method in predicting ocean wave heights.

4. Conclusion

[11] An empirical technique has been developed using GA that allows the prediction of significant wave heights at different buoy locations in the north Indian Ocean, a few days in advance. The algorithm uses past values of winds and waves at various buoy locations. The major advantage of using GA over other nonlinear forecasting techniques such as artificial neural networks is that an explicit analytical expression for the dynamic evolution of the parameter concerned (wave height in our case) is obtained. Another advantage is that the algorithm is based on actual observations and does not depend on any numerical model and hence does not require auxiliary initial and forcing fields. The proposed method is expected to have direct practical applications in offshore stations engaged in oil and gas resource generation and processing. The method is also expected to be useful for coastal stations, such as ports and harbors used by the incoming and outgoing ships. For the benefit of the readers, we provide in Appendix A the analytical equations of the GA model for one-day forecast of wave heights at the DS1 and SW3 buoy locations.

[12] We have analyzed the sensitivity of GA equations at the two locations to the observations of wave and wind. We define the sensitivity in terms of the ratio of the change in predicted wave height to the change in input variables (past observations of wave height and wind). The sensitivity was analyzed only for the fluctuation part, and the magnitude of mean perturbation introduced in each input field was taken as half of the standard deviation of that particular field. Perturbations were introduced randomly at each point. We observed that the sensitivity of GA solution to wind fields is almost twice as large as that to the wave height fields in deep water. In shallow water, the GA solution is even more sensitive to winds, where the sensitivity to winds was found to be more than 5 times the sensitivity to wave height. This points to the fact that the GA is indeed able to differentiate between the deep and shallow water dynamics.

Appendix A: Equation for Wave Height Forecast

[13] Analytical Equation for 1-day forecast of wave height at the DS1 buoy location

$$\begin{aligned}\mathrm{fit(t)} = {} & (((\mathrm{x2(t-2)}/(-0.35)) + (\mathrm{x2(t-1)} + \mathrm{x2(t-1)})) \\ & /(((((8.24)/((\mathrm{x2(t-8)}/(\mathrm{x2(t-1)} + \mathrm{x2(t-1)})) \\ & + (7.98))) + \mathrm{x1(t-4)}) + (\mathrm{x2(t-8)}*\mathrm{x1(t-6)}))))\end{aligned}$$

$$\begin{aligned}\mathrm{trendfit(t)} = {} & (\mathrm{tr(t-1)} - ((\mathrm{tr(t-2)} + (\mathrm{tr(t-8)} - (\mathrm{tr(t-8)} \\ & + \mathrm{tr(t-1)})))/((\mathrm{tr(t-2)} - (\mathrm{tr(t-4)}/((-4.29) \\ & / ((\mathrm{tr(t-5)}/\mathrm{tr(t-1)})/\mathrm{tr(t-1)})))))/\mathrm{tr(t-4)})))\end{aligned}$$

wave height(t) = fit(t) + trendfit(t)
waveheight (t) = waveheight at t-th day
t = day for which prediction of wave height is made.
fit(t) = fluctuation of wave height for t-th day
x2(t − 1) = wave height fluctuation at day (t −1)
x1(t − 1) = wind at day (t − 1)
trendfit (t) = nonlinear trend for t-th day
tr(t-1) = trend at day (t − 1)
Similar notations hold good for other time steps.

[14] **Acknowledgments.** We are indebted to A. Álvarez who generously provided us the computer code of the GA used by us. We are also thankful to the National Institute of Ocean Technology (Department of Ocean Development) for the buoy data used in the study. We wish to express our sincere gratitude to two anonymous reviewers for their helpful comments and valuable suggestions.

References

Abarbanel, H. D. I., R. Brown, J. Sidorowich, and L. S. Tsimring (1993), *Rev. Mod. Phys*, *65*, 1331–1392.

Álvarez, A., C. Lopez, M. Riera, E. Hernandez-Garcia, and J. Tintoré (2000), Forecasting the SST space-time variability of the Alboran Sea with genetic algorithms, *Geophys. Res. Lett.*, *27*, 2709–2712.

Álvarez, A., A. Orfila, and J. Tintoré (2001), DARWIN: An evolutionary program for nonlinear modeling of chaotic time series, *Comput. Phys. Commun.*, *136*, 334–349.

Álvarez, A., A. Orfila, and J. Tintoré (2004), Real-time forecasting at weekly timescales of the SST and SLA of the Ligurian Sea with a satellite-based ocean forecasting (SOFT) system, *J. Geophys. Res.*, *109*, C03023, doi:10.1029/2003JC001929.

Casdagli, M. (1989), Nonlinear prediction of chaotic time series, *Physica D*, *35*, 335–356.

Grassberger, P., and I. Procaccia (1983), Measuring the strangeness of strange attractors, *Physica D*, *9*, 189–208.

Hartmann, D. L., and M. L. Michelson (1989), Intraseasonal periodicities in Indian rainfall, *J. Atmos. Sci.*, *46*, 2838–2862.

Kishtawal, C. M., S. Basu, F. Patadia, and P. K. Thapliyal (2003), Forecasting summer rainfall over India using genetic algorithm, *Geophys. Res. Lett.*, *30*(23), 2203, doi:10.1029/2003GL018504.

Kishtawal, C. M., F. Patadia, R. Singh, S. Basu, M. S. Narayanan, and P. C. Joshi (2005), Automatic estimation of tropical cyclone intensity using multi-channel TMI data: A genetic algorithm approach, *Geophys. Res. Lett.*, *32*, L11804, doi:10.1029/2004GL022045.

Premkumar, K., M. Ravichandran, S. R. Kalsi, D. Sengupta, and S. Gadgil (2000), First results from a new observational system over the Indian Sea, *Curr. Sci.*, *78*, 323–330.

Szpiro, G. (1997), Forecasting chaotic time series with genetic algorithms, *Phys. Rev. E*, *55*, 2557–2568.

Takens, F. (1981), Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence*, *Lecture Notes Math.*, vol. 898, edited by D. Rand and L. S. Young, pp. 366–381, Springer, New York.

S. Basu, C. M. Kishtawal, A. Sarkar, and K. Satheesan, Meteorology and Oceanography Group, Space Applications Centre, Ahmedabad 380015, India. (rumi_jhim@yahoo.com; cmk307@rediffmail.com; sarkar_abhi2000@yahoo.com; k_satheesan@rediffmail.com)