Bay of Bengal wave forecast based on genetic algorithm: A comparison of univariate and multivariate approaches

A.D. Rao a, Mourani Sinha a,⇑, Sujit Basu b

aCentre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
bAtmospheric and Oceanic Sciences Group, Space Applications Centre, Ahmedabad 380015, India

Article info
Article history:
Received 28 October 2011
Received in revised form 8 August 2012
Accepted 7 September 2012
Available online 18 September 2012

Keywords:
Empirical orthogonal function analysis
Genetic algorithm
Significant wave height
Real time wave forecast
Numerical wave modeling

Abstract
Prediction of significant wave height (SWH) field is carried out in the Bay of Bengal (BOB) using a combination of empirical orthogonal function (EOF) analysis and genetic algorithm (GA). EOF analysis is performed on 4 years (2005–2008) of numerical wave model generated SWH field, and analyzed fields of zonal (U) and meridional (V) winds. This is to decompose the space-time distributed data into spatial modes ranked by their temporal variances. Two different variants of GA are tested. In the first one, univariate GA is applied to the time series of the first principal component (PC) of SWH in the training dataset after a filtering with singular spectrum analysis (SSA) for effecting noise reduction. The generated equations are used to carry out forecast of SWH field with various lead times. In the second method, multivariate GA is applied to the SSA filtered time series of the first PC of SWH, and time-lagged first PCs of U and V and again forecast equations are generated. Once again the forecast of SWH is carried out with same lead times. The quality of forecast is evaluated in terms of root mean square error of forecast. The results are also compared with buoy data at a location. It is concluded that the method can serve as a cost-effective alternate prediction technique in the BOB.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Forecast of SWH fields has widespread application in ocean engineering and marine studies. This is often required for design of offshore and near shore structures, oil and fishery industries, port design and operability, sediment transport, etc. In what follows, the terms SWH and wave height will be used interchangeably. Generally, SWH prediction has been carried out by various numerical models which are being updated continuously. These models integrate the governing fluid dynamical equations. However, a major drawback of these models is that huge computing resources and data support are required for running these models. To overcome this difficulty, various data adaptive approaches have been proposed in recent times. These approaches typically require only a time series of the parameter to be predicted at a particular location. Thus, Deo and Naidu [1] and Agrawal and Deo [2] used artificial neural networks (ANN) along with buoy data for the prediction of SWH. Whereas these studies used relatively short datasets, Londhe and Panchang [3] used data from several National Data Buoy Centre (NDBC) buoys for a large number of years to demonstrate the power of ANN for wave forecast.

Another powerful data adaptive approach known as genetic programming [4] has been advocated in recent years to carry out forecast of various ocean surface parameters. Genetic programming is also known by the name of genetic algorithm (GA) in the scientific literature (e.g., [5,6]). Hereafter, the term GA will be used. In contrast to ANN, this method (GA) provides

⇑ Corresponding author. Tel.: +91 11 2659 6023.
E-mail address: mou510@gmail.com (M. Sinha).

0307-904X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.apm.2012.09.001
explicit analytical forecast equations. Thus, Alvarez et al. [6] and Alvarez [7] used GA to forecast the space-time variability of sea surface temperature (SST) in the Alboran Sea and Adriatic Sea respectively, while Neetu et al. [8] carried out a similar study in the Arabian Sea. Basu et al. [9] used GA for predicting wind speed at two shallow buoy locations in the Arabian Sea (AS) and the BOB with a lead time of 1–3 days. Later, Sharma et al. [10,11] and Basu et al. [12] extended this study for predicting spatio-temporal variability of satellite scatterometer derived ocean surface wind vector fields in the Arabian Sea, Bay of Bengal and the entire North Indian Ocean respectively.

Coming to the case of wave forecast relevant to this study, one needs to mention the pioneering study by Basu et al. [13] related to the prediction of wave heights in an area of the Arabian Sea using different time series of surface wind speeds and SWH from three different oceanographic buoys as input for a GA. Jain and Deo [14] used ANN to carry out wave forecast off western Indian coast from buoy observations. Gaur and Deo [15] used GA to carry-out real-time wave forecast in a limited area of the Gulf of Mexico. However, these three studies and the earlier mentioned studies using ANN were limited to wave forecast at isolated locations and limited areas.

Since spatially distributed wave height data over a regular grid and at regular time intervals are not available from buoys or satellites, one has to resort to numerical models for carrying out basin-scale wave forecast. However, as mentioned previously, this method has its own associated difficulties. The combination of numerical models and the data-adaptive approaches of ANN or fuzzy logic are known to improve the accuracy of the modeling approach. Thus, Makarynskyy [16] made an effort to improve wave predictions off the Irish coast using WAM outputs and ANN. However, the dataset was relatively short. More recent efforts of similar nature can be found in the studies by many investigators [17–20].

GA, in combination with numerical wave models, has been employed in relatively fewer studies. The present authors are aware of only two such studies. Thus, Canellas et al. [21] used WAM model generated SWH fields along with GA for wave height prediction in the western Mediterranean Sea at several buoy locations, while Jain et al. [22] used MIKE-21 generated wave height fields along with GA to predict wave height at four offshore sites along the east coast of India. Interestingly, in neither of these studies, there is any attempt to carry out wave forecast over an entire ocean basin, although numerical wave models generate such outputs required to train the data-adaptive approach of GA.

Application of any data-adaptive approach (which is basically an algorithm for time-series prediction, albeit nonlinear and more sophisticated than the known linear methods) to a spatio-temporally distributed dataset requires some kind of data compression in the spatial domain, if one is interested in basin scale forecast. One such powerful method is the technique of EOF analysis [23,24]. This technique compresses the spatial variability into a few eigenmodes and the temporal variability is contained in the leading principal components which could be subjected to GA in the usual fashion. In the past, this technique has been used by Alvarez et al. [6], Sharma et al. [10,11], Basu et al. [12], Neetu et al. [8] and Remya et al. [25] in their studies related to basin scale forecast of various ocean surface parameters using GA.

In the present study, the EOF technique has been used in conjunction with GA to predict SWH fields in the BOB. In a previous diagnostic study, Rao et al. [26] explored the spatio-temporal variability of SWH in the same region using the same EOF technique. Apart from EOF and GA, another novel feature of the present study is the use of SSA for noise reduction. It is known that the algorithm, in principle, can predict a strictly deterministic, albeit chaotic, time series. However, the inevitable presence of noise in any physical system introduces spurious features, not amenable to prediction. Therefore a pre-filtering of the data series using a noise reduction technique becomes absolutely necessary [12].

2. EOF, GA and SSA

The EOF technique has been elaborately described by Preisendorfer [23] and Basu et al. [24]. More recently Sharma et al. [10] and Basu et al. [12] have elegantly summarized the method in relation to forecast using GA. Nevertheless, for the sake of self-consistency of this paper, we recapitulate its salient features, borrowing heavily from [10,11]. The aim is to expand the parameter vector (model simulated wave heights, in our case), in a series of empirical orthogonal vectors, which are eigen-vectors of the covariance matrix of the data set, in the following manner:

$$H(x, y, m) = \langle H(x, y) \rangle + \sum_{n=1}^{N} C_n(m)P_n(x, y).$$

(1)

Here $H(x, y, m)$ denotes the wave height at grid location $(x, y)$ and $m$ is the time index. Further, $N$ is the total number of grid points. $\langle H(x, y) \rangle$ is the temporal mean at each grid location, $P_n(x, y)$ and $C_n(m)$ are the $n$th EOF and $n$th principal component (PC) respectively. In Eq. (1) functional notation has been used. However, in actual practice, all the wave heights are collected into a one-dimensional vector for applying matrix algebra. The eigenvalues are found numerically. Because of the symmetry and positive definiteness of the covariance matrix, the eigenvalues are all real and positive and are customarily arranged in descending order. The associated eigenvectors (denoted also by the name EOFs) are mutually orthogonal, provided the eigenvalues are distinct (found to be true in our case). After numerically solving the eigenvalue problem the eigenvectors are also routinely normalized. Once a set of such orthonormalized EOFs is available, it is trivial to find the PCs required in the expansion, given by Eq. (1). It immediately follows from Eq. (1) and the orthonormality of the EOFs that the PCs are given by the scalar product:

$$C_j(m) = \langle H_{dev}(x, y, m)P_j(x, y) \rangle,$$

(2a)
where \( H_{\text{dev}}(x, y, m) = H(x, y, m) - \langle H(x, y) \rangle \).

Thus, in effect, the EOF technique summarized in Eqs. (1) and (2) amounts to decomposing the spatiotemporal data into a set of spatial modes (EOFs) and temporal amplitude functions (PCs). The EOFs provide information about spatial patterns, while the PCs describe the dynamics and can be predicted using GA, the details of which are given below.

Alvareze et al. [27], Sharma et al. [10] and Basu et al. [12] have explained GA in great details. However, just like the case of EOF, we describe the salient features of GA for the sake of self-consistency. GA is, in effect, an approach to build a predictive model directly from observations of the parameter concerned. That such a model exists, in principle, follows from Takens’ theorem [28]. According to this theorem, given a deterministic, albeit chaotic, time series \( \{x(t_i)\}, t_i = it, i = 0,..., N \), there exists a smooth map \( P \) satisfying

\[
x(t) = P(x(t - \tau), x(t - 2\tau), \ldots, x(t - m\tau)),
\]

where \( m \) is called the embedding dimension, \( \tau \) is the sampling time interval, six hours in our case. Here \( N \) is the total number of data points. GA basically tries to obtain the function \( P(.) \) in Eq. (3) that best represents the amplitude function of the given 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 chaotic time series, which can then be used to predict the future state of the system.

The algorithm starts with an initial population of \( \text{equation strings} \) which are sequences of randomly chosen symbols that conform to a simple grammar of mathematical equations: two arguments are combined by an arithmetic operator (addition, subtraction, multiplication or division) and the resulting expression is enclosed in parentheses. The arguments are either real numbers or values from the time series (e.g., \( x(t - \tau), x(t - 2\tau), \ldots \)) of the parameter concerned, or are themselves self-contained expressions enclosed in a pair of parentheses. The only constraint is that the final individuals must be consistent mathematical expressions.

The criterion of the strength of each individual (equation string) in the population is its fitness measure. Before defining the fitness measure, we generalize Eq. (3) for addressing longer term predictability. Specifically, the prediction model

\[
x(t) = Q(x(t - \mu\tau), x(t - (\mu + 1)\tau), \ldots, x(t - (\mu + m - 1)\tau)),
\]

where, \( F(.) \) is the function to be determined.

The algorithm is programmed to approximate the mapping \( P \) or \( F \) using a technique borrowed from evolutionary biology. The algorithm starts with an initial population of \( N \) ‘equation strings’ which are sequences of randomly chosen symbols that conform to a simple grammar of mathematical equations: two arguments are combined by an arithmetic operator (addition, subtraction, multiplication or division) and the resulting expression is enclosed in parentheses. The arguments are either real numbers or values from the time series (e.g., \( x(t - \tau), x(t - 2\tau), \ldots \)) of the parameter concerned, or are themselves self-contained expressions enclosed in a pair of parentheses. The only constraint is that the final individuals must be consistent mathematical expressions.

The criterion of the strength of each individual (equation string) in the population is its fitness measure. Before defining the fitness measure, we generalize Eq. (3) for addressing longer term predictability. Specifically, the prediction model

\[
x(t) = Q(x(t - \mu\tau), x(t - (\mu + 1)\tau), \ldots, x(t - (\mu + m - 1)\tau)),
\]

where \( L = m\tau \) and \( T \) is the total length of the training set.

One can further define a strength index for each individual in the following fashion:

\[
R_j = 1 - \Delta_j^2 / \delta^2,
\]

where

\[
\delta^2 = \frac{1}{T - L + 1} \sum_{t=L+1}^{T} (x(t) - \langle x(t) \rangle)^2,
\]

and \( \langle x \rangle \) represents the mean value of the training data. It is clear that the closer the value of \( R_j \) is toward unity, the stronger is the \( j \)th individual equation string. An independent set for validation is constituted by the data not included in the training set. This set is not employed by the algorithm during the selection process.

The generalization of the multivariate GA case represented by Eq. (4) for the longer-time predictability is trivial and the relevant prediction and fitness equations analogous to Eqs. (5)–(7) will not be explicitly mentioned here. The major steps of the algorithm are elaborated below.

(a) \textit{Initialization}. The initial population consists of equation strings of the form

\[
\gamma_j = ((A * B) * (C * D)), \quad 1 \leq j \leq N,
\]
where A, B, C, and D are either real variables or variables $x_{\lambda} (1 \leq \lambda \leq L)$ and * stands for one of the four arithmetic operators, addition, subtraction, multiplication, or division. The operators and arguments are assigned at random: The numerical values are chosen from a finite set of real numbers, uniformly distributed in $[-Z, +Z]$, and the integer $\lambda$ is uniformly distributed in $1 \leq \lambda \leq L$. The value of $Z$ has been taken to be 10 from past experience.

(a) Computing the fitness. The fitness measure is computed using the appropriate equation.

(b) Ranking the agents. The equations, or agents, as they are often called, are ranked in descending order of their fitness.

(c) Choice of mates. The members of the population are organized into pairs. The fittest agent first chooses a mate randomly, with the probability of choosing a partner being proportional to its fitness. Later, the next fittest chooses a partner and so on, until N/2 agents have formed pairs. The remaining agents disappear. A total of N/4 pairs are formed.

(d) Reproduction and crossover. The “genetic information” is passed on to the next generation using this method. Each of the N/4 pairs has four offspring. The first two are identical to their parents. The equation strings of the two other offspring are formed as recombinations of their parents’ equation strings and randomly chosen self-contained parts of the parents’ strings are interchanged.

(e) Mutation. A small percentage of the equation strings’ most basic elements, which are single operators and arguments, are mutated at random. The top ranked equation strings are exempted from mutation, however, so that their information is not lost inadvertently.

These steps are run and rerun for a certain number of generations, or until some stopping criterion is satisfied, for example, when strength index no longer increases. Finally the top-ranked equation is selected and is broken down into a concise formula.

The SSA approach has been described in detail by Penland et al. [29] and Casdagli et al. [30]. This approach has been used earlier by Alvarez [7], Sharma et al. [11] and Basu et al. [12]. It is arguably the best known data-adaptive approach for noise reduction. Briefly, for the time series of the earlier by Alvarez [7], Sharma et al. [11] and Basu et al. [12].

Fractional differencing is used to deseasonalize the data. The data is then divided into segments of equal length. Each segment is then analyzed with SSA. This is performed for the first ten years of the time series which is divided into the following segments:

3. Data

The analysis region covering the BOB extends from 78°E to 98°E and 5°N to 25°N. Six hourly SWH generated in this region by NOAA WAVEWATCH III (WW3) global wave model forced with NCEP reanalyzed winds were obtained from the site (http://www.polar.ncep.noaa.gov/pub/history/waves). The six hourly zonal and meridional winds used to run the model were also obtained from the same site. The model is a multi-grid forecast model with two way nesting capability with 0.5° global grid. There are 960 ocean points in the study region.

The training dataset covers the period of approximately 4 years from 1st March 2005 to 31st December 2008 consisting of 5604 points. For validating the algorithm, an independent data set from 1st January to 31st December 2009 with 1460 points has been considered.

Validation of forecast (with various lead times) has also been effected by using data from the buoy OB10 located at (83.5°E and 14°N) for certain months of 2009.

4. Results

4.1. EOF analysis

The EOF analysis has been carried out separately on all the three variables using 4 years of model data. The first eigenmode accounts for 81.3%, 75.2% and 71% of the total variability of SWH, U and V, respectively. The other modes contribute insignificantly and have not been considered further.

Stability of EOFs is an important criterion that should be tested before using these EOFs in any analysis [24]. If the EOFs are stable, this signifies that the dominant spatiotemporal variability has been well captured by the leading modes and addition of further data would not lead to any significant improvement. By definition, an EOF is stable if $(\delta x_k / \Delta x_k)$ is significantly smaller than one. Here $\delta x_k \approx x_k (2 / M)^{1/2}$ is the sampling error of the $k$th eigenvalue $x_k$ and $\Delta x_k$ is the difference between the $k$th and $(k + 1)$th eigenvalues. Here, $M$ denotes the total number of observations. These ratios were found to be 0.02, 0.02, 0.02 for the first eigenmodes of SWH, U and V respectively. This definitely signifies that even four years of data are sufficient for establishing the forecast algorithm. It may also be worth mentioning here that a study of similar kind, but concerned with forecast of scatterometer-derived winds, was accomplished with just three years of data [12]. The reason was the same, i.e., the EOFs were found to be stable.

Fig. 1(a) depicts the first spatial mode (EOF1) of SWH. The maximum loading is observed in the northeast part of the central bay. This may be due to the northeastward propagation of the waves during the southwest monsoon with its accompanying strong southwesterly winds. Another factor favoring this loading is the northeastward propagation of swells
originating in the southern Indian Ocean. The minimum amplitude is seen along the southeast and southwest coasts of India. The amplitude is increasing towards north along the coast. The time series for PC1 corresponding to the first EOF is shown in Fig. 1. The first spatial eigenmodes and first principal components of SWH, zonal and meridional winds (a) EOF1 for SWH in m (b) PC1 for SWH (c) EOF1 for zonal wind in m/s (d) PC1 for zonal wind (e) EOF1 for meridional wind in m/s (f) PC1 for meridional wind. The PCs are unitless.
Fig. 1(b). The annual signal is clearly visible in this figure. Maxima of PC1 occur during the summer because the wave heights are maximum during southwest monsoon months as the high winds prevail during this season.

EOF analysis has been carried out separately for zonal (U) and meridional (V) wind fields. Fig. 1(c) and (e) depict the first EOF of the U and V component respectively. The maximum loading is confined to the lower latitudes for the zonal wind, whilst the maximum loading for the meridional wind is restricted to the central bay. Analogous to the case of SWH, the PC corresponding to the first EOF of both the components describes annual periodicity (Fig. 1(d) and (f)). These patterns clearly indicate that the annual maxima are associated with seasonal reversal of the wind in the BOB. However, locations of the maxima are different for zonal and meridional components.

4.2. Forecast of the principal component

The PCs obtained are inherently noisy and for the GA forecast to be effective, a preliminary noise reduction is absolutely necessary. Accordingly, SSA filter has been applied to the PCs of the dominant eigenmodes for reducing noise. The SSA filter has been applied separately to the first PCs of SWH, U and V in the training dataset. The size of the time window of the filter is equal to 365 and the number of eigenvalues used to reconstruct the filtered series is 20. The window size has been fixed somewhat heuristically by trial and error and only the eigenvalues above the noise floor (seen from the eigenvalue diagram) have been retained. The time series of the original and filtered PCs for all the three parameters are shown in Fig. 2. Subsequently GA has been applied to the filtered time series of the PCs. The forecast consists of two different approaches, univariate and multivariate. In the univariate case, only the past values of the first PC of SWH are used in the forecast, while in the multivariate case the forecast is effected by finding a relation between the first PC of SWH and time-lagged values of the first PC of U- and V-components of wind. As mentioned earlier, the forecast equations are obtained by applying univariate and multivariate GA on a training set of the data consisting of 5604 data points (from 1st March 2005 to 31st December 2008) after SSA filtering. In both the methods the following parameters have been used for training the algorithm. The number of equation strings varied from 60 to 120. The total number of arguments and operators allowed is 20 in each case. The embedding dimension varied from 2 to 15 and the optimum dimension was found by trial and error. The mutation rate has been chosen to be 0.01. The number of iterations required to achieve maximum strength index varied from case to case, but was of the order of a few thousands. Predictions with lead times of 6, 12, 18, 24, 30, 36, 42 and 48 h have been carried out. The strength index for the univariate GA varied from 0.7 and 0.9 whilst, it varied from 0.6 and 0.8 for the multivariate GA. Generally speaking, univariate GA performed better than the multivariate GA. The analytical forecast equations obtained from the two approaches are given in the Appendices A and B. Although the performance of the univariate method is somewhat better, the importance of the second method stems from the fact that prediction of SWH fields requires just a few past values of analyzed wind fields. Knowledge of analyzed SWH data is not required. In other words, one can do without a numerical wave prediction model. The accuracy of the forecast is verified by applying the prediction equations on an independent validation dataset from 1st January till 31st December 2009 consisting of 1460 fields of the concerned parameters. The PCs for the independent dataset are computed by taking scalar product of the corresponding fields with the EOFs of the training set. Then, these PCs and the analytical forecast equations are used to obtain the genetically forecasted PCs.

5. Discussion

The first judgment of the quality of GA forecast can be made from the scatter plots of actual and forecasted PCs for forecasts with lead times of 06, 12, 18, 24, 30, 36, 42 and 48 h. Fig. 3 depicts the scatter plots of the actual first PC and the genetically forecasted first PC for the independent data set using univariate GA. It can be seen from the figure that the predictive power of the algorithm decreases as the length of the forecast horizon increases, which is a hallmark of deterministic chaos with a very large domain of the study area for forecasts with lead times of 06, 12 and 18 h. As expected, with the increase in forecast lead time, this area gradually shrinks.

We come now to the results for the multivariate case. Fig. 5 depicts the scatter plots of the actual PC1 and the genetically forecasted PC1 for the independent data in the case of multivariate GA. The value of $R^2$ between the actual and predicted PC is greater than 0.9 for all the lead times from 06 to 42 h and greater than 0.85 for the 48 h forecast. Based on the value of $R^2$, it can be noted that the performance of univariate GA is excellent for forecasts with mentioned lead times. The forecast fields of SWH in the independent validation dataset are now reconstructed using EOFs of the training set and the genetically forecasted PCs, separately for forecasts with lead times of 06, 12, 18, 24, 30, 36, 42 and 48 h. Fig. 4 shows the spatial distribution of the root mean square error (RMSE) of forecast for each lead time for the univariate GA. The RMSE is less than 0.3 m over a very large domain of the study area for forecasts with lead times of 06, 12 and 18 h. As expected, with the increase in forecast lead time, this area gradually shrinks.

Further validation of the GA based forecast has been carried out at an independent buoy location. Comparison is made between observed and GA forecasted SWH for the buoy OB10 (83.5°E and 14°N) from January to September 2009.
Figs. 7a–7c depicts the results for the buoy OB10 for different lead times (06, 24, 48 h) of forecast. From the figure it can be concluded that the univariate forecasts perform marginally better (particularly in the earlier part of the year with low wave heights). In the later part of the year the performances of the two variants are comparable. Comparatively poor performance of multivariate GA (in an overall sense) was noted previously also (Fig. 6) and the possible reason for this has been pointed out earlier. Of course the match between the univariate forecast and buoy data is also not always equally good. This is because the GA forecast algorithm has been trained using numerical model generated wave data. Hence this forecast can at best be close to model forecasts, being always marginally worse (because of the error inherent in the forecast algorithm).
That the forecasts compare poorly with buoy data at certain times signify the fact that the numerical model itself is performing poorly at those times. This has been checked independently by comparing the GA forecasts with model forecasts at all times, but the model forecasts are not shown in Figs. 7a–7c to avoid clumsiness. Ideally, of course the training should have been performed using gridded buoy dataset, in which case the GA forecasts would have compared much better with independent buoy data in the validation set. Unfortunately, such a dense network of buoys simply do not exist in the region of our interest.

6. Conclusion

Wave forecast has widespread applications in studies related to ocean engineering, marine structure, port operability, etc. Traditionally this forecast has been carried out by suitable numerical wave prediction models. However, these models suffer
from the drawback that they require huge computational and data support. To overcome this difficulty, various data-adaptive approaches have been proposed in recent times for carrying out this task. Typically these approaches require only the past values of a time series of the wave height at a particular location as input to the prediction algorithm.

Historically artificial neural networks have been used to predict wave height at isolated locations of the world's oceans. They have also been used in conjunction with numerical wave models to increase the accuracy of the latter. Of late, interest has shifted to genetic algorithm for carrying out forecast of various ocean surface parameters, possibly due to the capability of the algorithm to provide explicit analytical forecast equation, which is quite handy for use in operational environment.

Fig. 4. Spatial distribution of the root mean square error of forecast (in m) in the univariate case.
Although there are many studies related to basin-scale forecast of sea surface temperature and sea surface winds using genetic algorithm, there is no such study related to basin-scale forecast of ocean surface waves.

In the present study, a combination of EOF analysis and GA is used for the first time to carry out basin-scale forecast of SWH in the Bay of Bengal. The EOF analysis is applied on four years of numerical model generated SWH field and analyzed zonal and meridional wind fields separately. Two different approaches, viz., univariate GA and multivariate GA have been tested. Univariate approach requires only past values of SWH field, while multivariate approach requires past values of analyzed wind fields.

Forecast has been evaluated in terms of root mean square error of forecast in an independent validation dataset and also at one independent buoy location. The evaluation suggests that the univariate GA performs somewhat better than the multivariate one. On the other hand, the multivariate GA is very handy in the absence of any numerical wave model, since apart from GA generated prediction equations, the prediction of SWH data requires only a few past values of good quality analyzed wind fields, nowadays readily available from several numerical weather prediction centers.
Fig. 6. Spatial distribution of the root mean square error of forecast (in m) in the multivariate case.

Fig. 7a. Comparison of buoy (OB10) observed and 06 hourly GA forecasted SWH for 2009.
In unison with Jain et al. [22], we stress that our study is not motivated by a desire to prove the superiority of genetic algorithm over a tested numerical wave model. Rather, this study should be looked upon as an attempt to provide an alternate technique of prediction for those scientists and research centers who may not possess the necessary computational and data support to run a sophisticated wave model. Apart from the pre-computed EOFs of the concerned parameters (with very little storage requirements), one only needs the analyzed wave heights for a particular time and for few past steps (generally available from websites) to carry out basin scale forecast of wave heights in the univariate case. Even if these analyzed wave heights are not available due to unforeseen circumstances, the multivariate approach could be used to forecast wave height using only present and past observations of gridded analyzed wind fields, available from numerical weather prediction centers or from a constellation of satellite-borne scatterometers. Summarizing, it can be said that the method advocated in the study appears to be a cost-effective alternate technique for effecting basin scale forecast of wave height in the Bay of Bengal.

Acknowledgements

The authors are thankful to the Environmental Modeling Center for the archived model data from the NOAA WAVEWATCH III® runs. The authors are also indebted to Dr. Alvarez for the GA code used in the study and INCOIS, NDBP-NIOT/ MOES for the buoy data. This study forms a part of a project under Meteorology and Oceanography -2 Program of the Indian Space Research Organization (ISRO), and the first author of the study wishes to thank ISRO for sponsoring the project.

Appendix A

We provide below the analytical equations for 06, 12, 18, 24, 30, 36, 42 and 48 h ahead forecast of the first PC of SWH for the univariate GA.

1. \( pc(t) = ((pc(t-3) - ((0.01)/((pc(t-1)/pc(t-3))) + ((pc(t-1)/pc(t-3))) + ((pc(t-1)/pc(t-3))) + ((pc(t-1)/pc(t-3)))) - (pc(t-1)/pc(t-3))) \)
2. \( pc(t) = (((pc(t-2) - pc(t-2)) + (pc(t-4) - (pc(t-3) - pc(t-2)))) + (pc(t-2) - (pc(t-4) + (pc(t-3) - pc(t-2)))) \)
3. \( pc(t) = (((pc(t-3) - (pc(t-3) + (pc(t-3) - pc(t-3)))) + (pc(t-3) - (pc(t-3) - pc(t-3)))) - ((pc(t-3) + (pc(t-3)) - pc(t-3))) \)
4. \( pc(t) = (((pc(t-3) - (pc(t-5) - (pc(t-4) - pc(t-5)))) + (pc(t-4) - (pc(t-5) - (pc(t-4) - pc(t-5)))) + (pc(t-4)) \)
5. \( pc(t) = (pc(t-6) - (pc(t-6) + (pc(t-6) + (pc(t-6) + (pc(t-6) - pc(t-5)) - pc(t-5)) - pc(t-5))) \)
6. \( pc(t) = (pc(t-10) - (pc(t-7) + (((pc(t-9) + (pc(t-8) - (pc(t-6) - pc(t-8)) - pc(t-6)) - pc(t-6)) - pc(t-6))) \)
7. \( pc(t) = (pc(t-12) - (pc(t-9) + (pc(t-12) - (0.01))) - (pc(t-7) - (pc(t-9) + (((pc(t-9) - pc(t-7)) - pc(t-7))) - pc(t-7)))) \)
8. \( pc(t) = (pc(t-8) - (pc(t-11) + ((-1.77)*(pc(t-8)/pc(t-12) - (pc(t-12) - (1.78)))) + ((-4.62) - (1.78)) - (-3.77)))) \)

Fig. 7b. Comparison of buoy (OB10) observed and 24 hourly GA forecasted SWH for 2009.

Fig. 7c. Comparison of buoy (OB10) observed and 48 hourly GA forecasted SWH for 2009.
Appendix B

We provide below the analytical equations for 06, 12, 18, 24, 30, 36, 42 and 48 h ahead forecast of the first PC of SWH for the multivariate GA.

(1) \( \text{pc(t)} = (\text{pcu(t-1)}) - (pcv(t-1)) + \left(\frac{pcu(t-1)}{(6.42)} + (0.92) + (9.04)(-2.48))\right) + (5.26)(-2.55) + (4.70) - (2.90)\)

(2) \( \text{pc(t)} = (\frac{-8.53}{((pcu(t-2))(6.53)/pcu(t-6)) - (pcv(t-6)) - (5.77)/((0.72)(6.72)) + (pcu(t-3))(6.72))\)

(3) \( \text{pc(t)} = (pcu(t-3))((2.52) + (pcu(t-3))(-7.4)) + (9.91)) + ((pcu(t-3))(-8.05)) - (3.61)) - (7.74))\)

(4) \( \text{pc(t)} = (\frac{-7.48}{((1.95)) + ((pcu(t-4))(6.72)) - (4.25)) + ((pcu(t-5))(3.03)) + (4.25)\)

(5) \( \text{pc(t)} = (\frac{-0.90}{((6.89)) - (5.49)) + ((pcu(t-5))(7.81)) - (2.19) + (9.42) + (1.36)) + (6.23)\}

(6) \( \text{pc(t)} = (((pcu(t-6)) - (8.58)) + (pcu(t-8)) + ((pcu(t-6)) - (8.18)) - (7.6)) - (8.18)) - (pcu(t-6)) - (7.97)\)

(7) \( \text{pc(t)} = ((pcu(t-7))(5.97) + ((pcu(t-7))(7.61)) - (7.3)) + (((pcu(t-7))(8.59)) + (6.80) - (pcu(t-7)) + (7.28)) - (2.15)\)

(8) \( \text{pc(t)} = ((pcu(t-8))(7.23)) - ((-5.01) + ((pcu(t-8)) - (4.32)) - (4.32)) - (4.32)\)

References