Proceedings of The Twelfth (2002) International Offshore and Polar Engineering Conference
Kitakyushu, Japan, May 26−31, 2002

ISBN 1-880653-58-3 (Set); ISSN 1098-6189 (Set)

A Study on Simulating the Time Series of Significant Wave Heights near the Keelung Harbor

John Z. Yim, Chung-Ren Chou
Dept. of Harbor and River Engineering, National Taiwan Ocean University,
Keelung, TAIWAN, China
Ping-Chun Ho, Keelung Harbor Breau,
Keelung, TAIWAN, China

ABSTRACT

Simulation of the time series of significant wave heights near the Keelung Harbor, Taiwan was carried out using the ARMA model. The measured time series was first transformed to eliminate nonstationarity. Three methods of data transformation were used to test their adequacies for wave height data measured at specific site. As the main objective of the present investigation is to study the possibility of simulating the statistical properties of the original time series, we have further checked the statistical characteristics of simulated wave heights. The result shows that after the original data have been Box-Cox transformed, an ARMA(4,4) will be the most satisfactory results in simulating wave heights measured near Keelung Harbour.

KEY WORDS: ARMA, Box-Cox transform, significant wave height

INTRODUCTION

In their now classic book on modelling of time series, Box and Jenkins (1976) discussed the procedures of model building in detail. Three models were considered as possible candidate for a time series. The three models are, the autoregressive (AR), the moving-average (MA), and the autoregressive moving-average (ARMA) model. Box & Jenkins pointed out that, in deciding the most suitable model for the time series under consideration, a total of three steps are needed. The three steps are: (1) identification, (2) estimation, and followed by (3) diagnostic checking. Hipel et al. (1977) used the Box-Jenkins model to simulate the flow of Saint Lawrence River. Salas and Obeysekera (1982; see also Wei. 1990; Brockwell and Davis, 1991; Mills, 1996) discussed the merits and demerits of the AR, MA and ARMA models in detail.

Most often, AR, MA, and/or ARMA models are used to simulate time series of either hydrological events (Salas and Obeysekera, 1982; Thompstone et al., 1986; Beauchamp et al., 1989; Jayawardena and Lai, 1994), or economical events (Pierce, 1977; Granger and Newbold, 1986; Mills, 1996; Clenments and Hendry, 1998). However, researchers have also found that these models are also applicable for records of wave heights. Jardine and Latham (1981), for example, used the ARMA model for records of the significant wave heights near NE Atlantic. They found that an AR(1) model can be used satisfactorily to simulate the time series of significant wave heights.

Since the series of significant height is nonstationary it must be transformed into a stationary time series in order to use the ARMA models. Guedes Soares and Ferreira (1995) studied the possibility of modelling the records of significant wave heights in the North Sea. For measurements during long periods the transformation adopted in this paper follows the method proposed by Bruce (1982). They found that an AR(5) model can be used for the purpose. Later, Guedes Soares et al. (1996) used the same method to simulate the significant wave heights of the Sines and Faro near southwest coast of Portugal. The AR model was found to have orders more than 19, which means that wave height measured at present will be related to wave heights measured more than 19 time span earlier. It would be interesting to find out whether this extremely long-time memory of wave heights is location dependent.

Guedes Soares and Ferreira (1996) compared the results of data transformation methods using time series of the significant wave heights measured in the North Sea. Three data transformation methods were considered. These are: (i) the Bruce (1982) method, (ii) the Box-Cox method, and (iii) a modified Box-Cox method proposed by them – hereafter abbreviated as the GF method for brevity. AR models with orders of more than 20 were used to simulate the time series. Comparing the auto correlation functions (ACFs) of the three transformed time series with that of the original data, they concluded that results due to the GF method are the most satisfactory for simulating records of significant wave heights.

It is seen that, simulation of the time series of the significant wave heights can be carried out using one of the ARMA models. The data have to, nevertheless, be first transferred into suitable forms using one of the three different transformation methods. It is also seen that wave heights from different areas can have different orders of the AR models.

The main objective of this study is to determine the most suitable method for modeling the records of wave heights near the Keelung Harbor, Taiwan. Records consisting of maximum significant wave heights on a daily basis were used for this purpose. This is accomplished in three steps. First of all, we try to determine which one of the data transformation methods is most suitable for the wave records under consideration. To examine the aptness of the transformation, not only are the ACFs of the simulated time series and that due to the original ones were compared, but also the basic statistical properties. These include the mean, the standard deviation, the skewness, and the kurtosis of these series. Secondly, we have tried to determine the optimal order of the ARMA model through the so-called AIC test. Since it is believed that the ultimate goal of studying the wave height statistics is to be in position to make predictions concerning the extreme wave height in, say 50 or 100 years, we have further compared the statistical properties of the original and the simulated wave heights. Some preliminary results of our study are presented in the present paper.

THE WAVE DATA AND THE ARMA MODEL

The Wave Data: Wave records measured by an ultrasonic wave recorder were kindly provided by the Keelung Harbor Bureau. The instrument is anchored on the eastern breakwater, which is approximately 1000 m away from the Keelung Harbor. Significant wave heights on a daily basis were measured from May 1983 to November 1990.

Due to occasional malfunctions of the wave-recording device, as well as severe weathers such as typhoon invasion, the data are incomplete most of the time. No data imputation techniques were undertaken at this moment. In this study we use the wave height records of the year 1985 as target. This is because in this year the rate of data availability is 99.45%, which is highest among all the years of wave records.

The ARMA Model: The ARMA(p,q) due to Box and Jenkins (1976) for a stationary with approximately normally distributed random fluctuations can be written as:

$$Z_t = \sum_{j=1}^{p} \phi_j Z_{t-j} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{1}$$

where z_t is time series of wave height at time t, parameters ϕ_j, j = 1,…, p and parameters θ_j, j = 1,…,q, are the values of AR and MA, with p and q respectively the orders of the AR model and MA models. The parameters ε_t of the model are independent, normally distributed random variables, $N\left[0, \sigma_\varepsilon^2\right]$.

The Data Transformation: One of the prerequisites for the Box-Jenkins models is that the target time series is stationary. This means that the statistical properties of the data remain unchanged at different times. However, quite often, records of wave heights will not satisfy this condition. It is therefore necessary first to carry out transformation that leads to a more or less stationary time series. Three transformation techniques are used in this paper. They are the method of Bruce (1982, hereafter denoted as the B method), the method due to Box-Cox (1964, which will be called the BC method), and the modified version of the Box-Cox method due to Guedes Soares and Ferreira (1995, abbreviated as the GF method). In the following a short description of these transformation methods will be given.

The first step of the Bruce (1982) transformation method is

$$Y_{t,p}^{*} = \frac{\left(Y_{t,p} - \overline{Y_p}\right)}{S_p} \qquad t = 1,2,3.....;\ p = 1,2,3,.....12 \tag{2}$$

where t is the day in the month p, and $Y_{t,p}$ is the value of the significant wave height. $\overline{Y}_p$ and S_p are, respectively, the mean and standard deviation in month p. $Y_{t,p}^{*}$ are the resulting significant wave heights with a zero-mean and unit variance.

A second step was necessary for the Bruce transformation that can be expressed as:

$$Z_{t,p} = \Phi^{-1}\left[F_{Y_{t,p}^{*}}\left(Y_{t,p}\right)\right] \tag{3}$$

where $F_{Y_{t,p}^{*}}\left(Y_{t,p}\right)$ is the cumulative distribution of the original data. Φ^{-1} is the inverse cumulative distribution of the normal distribution. $Z_{t,p}$ is the resultant discrete wave height of day t in month p. After the transformation, the ARMA model can be used to simulate wave heights with the target values of Z_t, where the symbol for the discrete month p has been omitted for simplicity.

The formula of the Box–Cox (1964) transformation can be written as:

$$Z_t(\lambda) = \frac{Y_t^{\lambda} - 1}{\lambda} \quad \text{if} \quad \lambda \neq 0 \tag{4a}$$

$$Z_t(0) = \log(Y_t) \tag{4b}$$

where Y_t is the wave height, λ is the parameter of the transformation, Z_t is the Box-Cox transformed data.

The transformation proposed by Guedes Soares and Ferreira is a modification of the Box-Cox formula. It can be written as:

$$Z_{t,p} = \ln\left(Y_{t,p}\right) - \overline{Y}_p \tag{5}$$

where the variable have been defined earlier. After transformations were complete, building of the simulated time series then follows the usual way (see e.g., Salas et al., 1980; Wei, 1991).

In order to determine the relative merits and demerits of the respective model, statistical properties of the modelled time series were compared with those of the original ones. These include, the moments, the maximum values of the series, and the goodness-of-fit of the probability distribution functions (PDF). A total of seven probability functions were used to find the possible distributions of significant wave heights. They are, the normal, the Weibull, the Rayleigh, the exponential, the Gamma, and the two- and three-parameter lognormal distribution. As indications of the goodness-of-fit, both the χ^2- and the Kolmogorov-Smirnov tests were used.

RESULTS AND DISCUSSION

Comparisons of the plots of ACFs: Following Box & Jenkins, the orders p and q of the AR(p) and MA(q) models are first estimated from plots of the ACF and the partial autocorrelation function (PACF). The ACF of significant wave heights measured in 1985 is plotted in Figure 1. Also shown in the figure are the ACFs of several simulated series of wave heights.

As can be seen from the figure, the line of the ACF of the original series decays gradually. Furthermore, it can also be seen that, as the time lag increases the ACF does not seem to fluctuate in any cyclic way. It is therefore concluded that the original time series has no hidden periodicity.

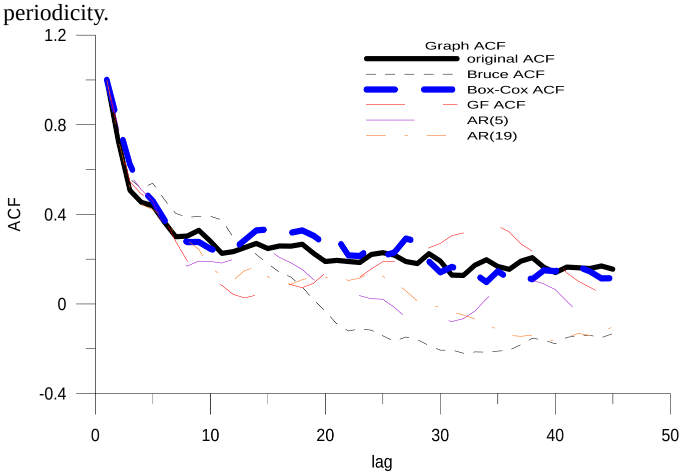

Figure 1. The ACFs of ARMA(4,4) for the original series (_________); simulated results after Bruce transformation (- - - - - -); simulated results after the BC transformation (__ _ __); and after the GF transformation (_____ _____). Also shown in the figure are the ACFs of the simulated results of AR(5) and AR(19). Both are GF transformed.

The curves of ACFs for ARMA(4,4) models after the original data were Bruce and GF transformed are seen to deviate from the ACF of the original data. This is not the case for the ACF of the BC transformed data. It can be seen that it has more resemblance to the original. It can be seen from Fig. 1 that, the ACF of the time series due to BC transformation bears more resemblance to that of the original series (than those due to other transformation methods). It is therefore conjectured that, an ARMA(4,4) model with BC transformed data is the most appropriate model to simulate significant wave heights measured near the Keelung Harbor. The reason for choosing ARMA(4,4) model will be explained in the next sub-section.

We have also used AR(5) and AR(19) models to simulate the significant wave heights. As can be discerned from Figure 1, with longer time lags, the curve of AR(5) has a slower, whereas that of AR(19) has a faster, rate of decreasing than the ACF of the original data. Both models are thus considered not suitable for wave heights measured on the specific site considered. Other tests, such as the AIC test value, the moments of the series, the goodness-of-fit test results for the probability density functions also confirm this.

The AIC and SBC test results: To determine the most appropriate orders of the AR and MA models, we used the so-called Akaike Information Criterion (1973, AIC), and the Schwartz Bayesian Criterion (1978, SBC). According to the theory, among all the possible candidates, the one that has the minimum AIC or SBC values can be considered as the model with the most appropriate order for the parameters.
We found that for p and q > 10, the AIC and SBC becomes larger than when p and q < 10. This means that when the orders of the p and q are greater than 10, the situation of over-fitting will occur. To avoid over-fitting, we have in this study used values of q and p less than 10 as the orders of the model. The values of AIC tests for the simulated wave heights due to three transformations are:

a) Bruce transformation

$$-271 \leq (AIC)_B \leq -112$$

b) BC transformation

$$-996 \leq (AIC)_{BC} \leq -811$$

c) GF transformation

$$-962 \leq (AIC)_{GF} \leq -821$$

The values of SBC tests for the simulated wave heights due to three transformations are:

d) Bruce transformation

$$-247 \leq (SBC)_B \leq -142$$

e) BC transformation

$$-986 \leq (SBC)_{BC} \leq -841$$

f) GF transformation

$$-932 \leq (SBC)_{GF} \leq -781$$

It is seen that both the AIC and SBC test results of the BC transformation are the smallest among the three transformations. The values of the p, q, and those of AIC test results are plotted in Figure 2. From the Figure 2, it can be seen that the AIC test values has a minimum when $p = q = 4$. It should be mentioned that both the Bruce and the GF transformations also has their minima of the AIC test results at $p = q = 4$. The parameters of the ARMA(4,4) model are listed in Table 1. From the series of the residuals, the ACF, the PACF and moments, there is no evidence for not considering the residual process as a Gaussian white noise process. Moreover, The Portmanteau Lack-of-fit Test, the Turning Point, the Difference Sign and the Rank Test did not reject the hypothesis for the usual significance levels. It is therefore ascertained that after the original data have been BC transformed, an ARMA(4,4) will yield the most satisfactory results in simulating wave heights measured in Keelung.

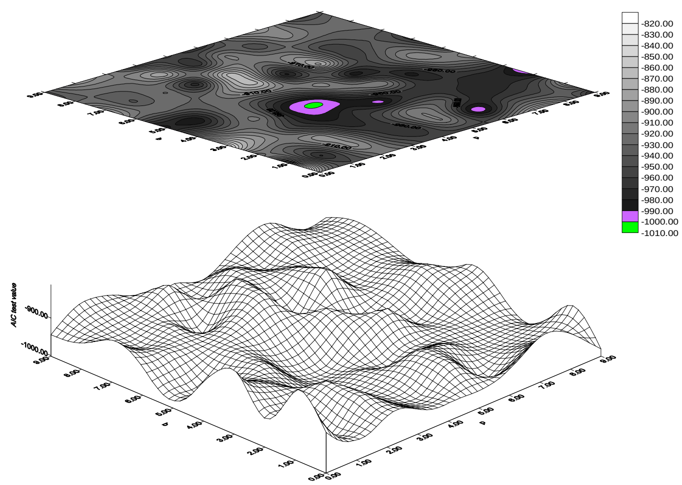

Figure 2. The AIC values of the different ARMA(p, q) after BC transformation (the abscissa: orders of the p, the abscissa: orders of the q, the ordinate: AIC value)

Table 1. The parameters of the ARMA (4, 4) model.

Orders	(1)	(2)	(3)	(4)
AR (p)	0.579	-0.277	0.665	-0.088
MA (q)	-0.184	-0.230	0.471	0.182

The Probability Distributions of Wave Heights: It was mentioned earlier that we have used 7 statistical models to find the possible statistical distribution of the maximum daily significant wave heights. These seven models are, the normal, the Weibull, the Rayleigh, the exponential, the gamma, as well as the two- and the three-parameter lognormal distribution. The results are shown in Figure 3.

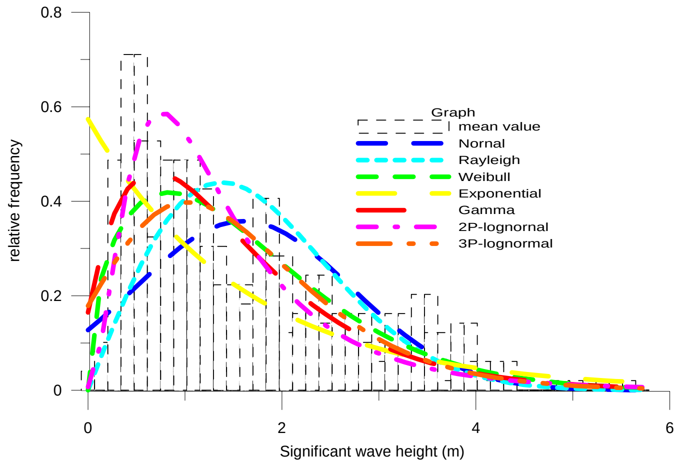

Figure 3. The histogram of the original significant wave heights fitted with different distributions.

Table 2 lists the χ^2- and the K-S goodness-of-fit test results. It can be perceived from Table 2 that, the assumptions that the significant wave heights of 1985 are Weibull-, Rayleigh-, gamma-, and/or the two-parameter lognormal distributed can not be ruled out.

The three simulated wave data using an ARMA(4,4) model for the Bruce, the BC, and the GF transformations were fitted by the seven distributions. It can be shown that except for the data based on the Box-Cox transformation, the other two data sets based on the Bruce- and the GF transformations do not have the statistical properties as the original data. We have chosen to show here the results obtained from BC-transformed data. It can be seen from Figure 4 that the pattern of the histogram has a similar appearance with that shown in Figure 3 for the original data set. The goodness-of-fit tests results are also listed in Table 2 for comparison.

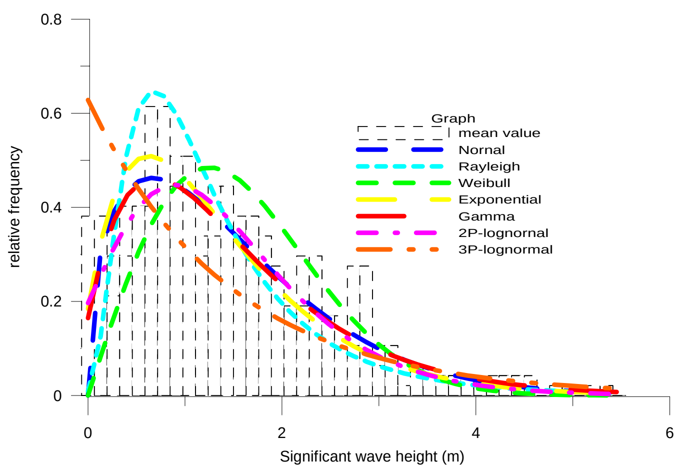

Figure 4. The histogram of the simulated significant wave heights for the ARMA(4,4) model with BC transformed data fitted with different distributions.

CONCLUSION

Judging from the results above, it can be ascertained that the best model to simulate the 1985 daily maximum significant wave heights near the Keelung Harbor is the ARMA(4,4) model with a Box-Cox transformation.

Table 2. The goodness-of-fit test results (for 95% confidence level) of the significant wave heights for both data measured in 1985 and the ARMA(4,4) simulated data with a Box-Cox transformation.

Test	χ^2-test			K-S test	
	Degrees of freedom	Original series	Simulated series	Original series	Simulated series
Distribution		Result	Result	Result	Result
Normal	40	Accepted	Accepted	Rejected	Rejected
Rayleigh	40	Accepted	Accepted	Accepted	Accepted
Weibull	40	Accepted	Accepted	Accepted	Accepted
Exponential	41	Accepted	Accepted	Rejected	Rejected
Gamma	40	Accepted	Accepted	Accepted	Accepted
2P log-normal	40	Accepted	Accepted	Accepted	Accepted
3P log-normal	39	Accepted	Accepted	Rejected	Rejected

The objective of this paper is to find the best model to simulate the daily maximum significant wave heights near the Keelung harbor. To achieve this goal, we have first transformed the nonstationary data using three methods – the Bruce-, the Box-Cox- and the GF transformation. Then we have estimated the orders of the AR and MA model using the AIC and SBC tests. Afterwards, we diagnosed the modelled data through the test for whiteness of the residuals. Since it is believed that one of the major objectives of simulating the time series is to be in a position to make predictions for future events, we proceed with fitting data sets with the statistical models found in the literature. As criteria for the goodness-of-fit both the χ^2- and the K-S tests were used.

From results presented in this paper, it is concluded that the best method to simulate the significant wave heights is to transform the data using the Box-Cox technique and thereby using an ARMA(4,4) model. The Bruce- and the GF transformation were found to be not suited for wave height data of Keelung measured in 1985. It must be stressed that, at present our finding is restricted to data for the year 1985. It is not clear whether the same conclusion is also applicable for simulation of long-term wave heights. At present, we are trying to find the most suitable way of data imputation. As soon as this has been done simulation of long-term wave heights will be carried out, and the results will be reported.

Reference

Akaike, H (1973). “Maximum likelihood identification of Gaussian auto-regressive moving average models,” *Biometrika*, Vol 60, pp 255-266.

Beauchamp, JJ, Downing DJ, and Railsback SF (1989). “Comparison of Regression and Time-series methods for synthesizing missing streamflow records,” *Water Resources Bulletin*, Vol 25, No 5, pp 961-975.

Box, GEP and Pierce DA (1970). “Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,” *Journal of the American Statistical Association*, Vol 65, pp 1509-1526.

Box, GEP and Cox DR (1964). “A analysis of transforms distribution of residual autocorrelations in autoregressive-integrated moving average time series models,” *Journal of Roy Stat Soc, B*, Vol 26, pp

211-243.
Box, GEP, and Jenkins GM (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, Calif. pp 575.
Brockwell, PJ and Davis RA (1991). *Time Series: Theory and Methods*, 2nd Ed, Springer, New York, pp 577.
Bruce, RL (1982). *On the Simulation of Long-Term Ocean Wave Data*, MSc thesis, University of Liverpool.
Clements, MP and Hendry DF (1998). *Forecasting Economic Time Series*, Printed in United Kingdom at the University Press, Cambridge, pp 368.
Granger, CWJ and Newbold P (1986). *Forecasting Economic Time Series*, Academic Press, Inc, pp 338.
Guedes Soares, C and Ferreira AM (1995). "Analysis of the seasonality in non-stationary time series of significant wave height," *Computational Stochastic Mechanics*, Spanos(ed), Balkema, Rotterdam, pp 559-568.
Guedes Soares, C and Ferreira AM (1996). "Representation of nonstationary time series of significant wave height with autoregressive models," *Probability Eng Mech*, Vol 11, pp 139-148.
Guedes Soares, C, Ferreira AM and Cunha C (1996). "Linear models of the time series of significant wave height in the southwest coast of Portugal," *Coastal Engineering*, Vol 29, pp 149-167.
Hipel, KW, McLeod AI and Lennox WC (1977). "Advances in Box-Jenkins modeling 1 Model construction," *Water Resources Research*, Vol 13, No 3, pp 567-575.
Jardine, TP and Latham FR (1981). "An analysis of wave height records for the N E Atlantic," *Quart J R Met Soc*, Vol 107, pp 415-426.
Jayawardena, AW and Lai F (1994). "Analysis and Prediction of Chaos in Rainfall and Stream Flow Time Series," *Journal of Hydrology*, Vol 153, pp 23-52.
Mills, TC (1996). *Time Series Techniques for Economists*, Printed in Great Britain at the University Press, Cambridge, pp 377.
Pierce, DA (1977). "Relationships-and the lack thereof- between economic time series, with special reference to money and interest rates," *Journal of the American Statistical Association*, Vol 72, No 357, pp 11-26.
Salas, JD and Obeysekera JTB (1982). "ARMA model identification of hydrologic time series," *Water Resources Research*, Vol 18, No 4, pp 1011-1021.
Salas, JD, Delleur JW, Yevjevich V and Lane WL (1980). *Applied Modeling of Hydrologic Time Series*, Water Resource Publications, Littleton, Colorado, pp 484.
Schwartz, G (1978). "Estimating the dimension of a model", *Ann Statist*, Vol 6, pp 461-464.
Thompstone, RM, Hipel KW, and McLeod AI (1986). "Forecasting quarter-monthly riverflow," *Water Resources Bulletin*, Vol 21, No 5, pp 731-741.
Wei, WWS (1990). *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley Publishing Company, Inc, Redwood city, California, pp 478.