Proceedings of The Sixth (2004) ISOPE Pacific/Asia Offshore Mechanics Symposium
Vladivostok, Russia, September 12–16, 2004

ISBN 1-880653-63–X

A Study of the Possible Effects of Missing Values in a Wave Record

John Z. Yim[a], Chung-Ren Chou[a] & Wei-Po Huang[b]
[a]Department of Harbour & River Engineering, National Taiwan Ocean University
Keelung, Taiwan
[b] Dept. of Hydraulic Engineering, Sinotech
Taipei, TAIWAN, China

ABSTRACT

Possible effects of missing data on the statistics of wave climate are considered here in this paper. Gaps between recorded wave heights are filled in using estimates either from interpolation, or through ARMA modelling. Statistical models found in the literature are then used to study the possible fits of both the original and the full-length records. Preliminary results have shown that the amount of the missing data seems to have only limited effects on the statistics of the wave heights. However, since only one data set is used for the study, the present conclusion is inconclusive.

АННОТАЦИЯ

В данной статье рассматривается возможное влияние отсутствия статистических данных по режиму волнения. Отсутствие данных по высоте волн восполняется с помощью данных, полученных методом интерполяции или АРМА-моделирования. Далее используются обнаруженные в литературе статистические модели для изучения возможного соответствия оригинальных и полных данных. Предварительные результаты показали, что количество пропущенных данных, похоже, имеет ограниченное влияние на высоту волн. Тем не менее, поскольку в работе используется только один ряд данных, данные выводы не являются окончательными.

KEY WORDS: Missing data, extreme wave heights, long-term wave height distribution.

INTRODUCTION

When handling with measurements, one has to deal with the problem of missing data as a rule. Data missing can occur due to a lot of reasons. In social science, when a survey is conducted, the person who is filling the questionnaire may refuse to answer some of the questions. This non-response will then cause gaps. On the other hand, in environmental studies, there are several reasons that can cause the missingness of the data. The measuring instruments, for example, may be malfunctioned; or the weather condition was too severe to conduct measurements. It can even be the case that some data in a record were entered erroneously and have to be deleted. In all the possible causes mentioned above, the invaluable data are lost forever.

Depending on the circumstances of the missing data, the effects on the subsequent analyses may or may not be serious. For example, when the amount of data missing in a file is not large, the effect may not be severe. However, when a substantial amount of data is missing, the evaluated results may then be biased. On the other hand, when the recording devices strike during normal (weather) conditions, it is probable that no severe effects will be caused. This may not be the case for extraordinary happenings where data may become vital for the researchers.

The problem of missing data is sometimes called irregular sampling in statistics. Researchers have studied this problem in many fields of statistical analyses. When missing data occurred in a record, the most direct way is to ignore or delete them. This then results in a data set without missing data. This method is called listwise or casewise deletion in the social sciences. The remaining data is then treated as a complete data set and the conventional statistical methods of analysis are then applied. However, as have been pointed out by many researchers, the result may be associated with the loss of valuable information. As remedy, methods of imputation have been proposed by many researchers. For a short review, see, for example, Allison (2001), while Little & Rubin (1987) wrote a textbook on this topic.

There is a large amount of research papers in the literature which deal with the problem of missing data. Nevertheless, the methods proposed by the researchers are often inadequate for the analyses of time series. This is because that when compared with other data, time series have some unique properties. This is especially true for meteorological records, where natural phenomena are reflected. For example, when a typhoon emerges from South Pacific, the sea surface around Taiwan will be roughened with long waves at first. As the typhoon passes by, the sea surface will then be filled with locally generated wind waves. The whole process will probably be lasting for two to three days, and all these will be reflected in the wave record measured by a nearby station. In this respect, time series can be considered as "*continuous*". To deal with the problem of missing data in time series, techniques other than those proposed in the textbook are probably needed.

There seems to have few books which deal with the problem of missing data in time series. To the best knowledge of the present authors, Brillinger et al. (1984) have published a conference proceedings concerning of handling the problem of missing data, or irregular

sampling, in time series. On the other hand, it is also rather strange that, albeit its importance, there are only a few researchers in the field of coastal/ocean engineering who addressed this problem. Noticeably are a series of papers published by Guedes Soares and his coworkers, see, e.g., Guedes Soares & Ferreira (1995, Guedes Soares et al., 1996; Hidalgo et al., 1995; see also Stefanakos & Athanassoulis, 2001). In these papers, the authors discussed the possible methods that can be used to estimate the values of the missing data.

In this paper, we aim at studying experimentally the possible effect(s) of missing data on the long-term statistical properties of wave heights. Two sets of field data will be used for the purpose. Each data set contains further three records. Among them, the first one is the original data, where the missing data is simply omitted. In the second record, only the small gaps between the original data will be filled through interpolation, while leaving large gaps unattended. The third record then has its full length where the larger gaps are further replaced through estimates obtained from ARMA modelling. These records are then fitted with long-term statistical models found in the literature. The results of the fits are then compared to find possible differences. In the following, we further divide the remaining of this article into three parts. In Section II we describe shortly the data source and the methods used for the analyses. Section III contains results of our analyses, and a short conclusion in Section IV then closes this paper.

THE DATA AND METHOD OF ANALYSES

Two data sets of wave records were used for the analyses. These are the records from wave measuring stations of Long-Dong and Long-Men. Both stations are located in the northeast edge of Taiwan. A schematic graph shown in Fig. 1 reveals the relative position of these two measuring stations. While wave heights and periods are recorded by a buoy at Long-Dong in a water of 32 m, a pressure type wave gauge was used at Long-Men, where the water depth is 14 m. The lengths of the records are also different. Measuring station Long-Dong is maintained by the Central Weather Bureau (CWB) and the record started from October 13, 1998 to the present day. The measuring station Long-Men is project-oriented and started from April 23 2002 and ended on July 4 2003.

Not only are the lengths of these two data sets different, but also are the recording times. The data from Long-Dong contains only the significant wave heights, $H_{1/3}$, recorded every two hours, and those from Long-Men contains, besides the significant wave heights, also the mean, H_{mean}, the maximum wave heights, H_{max}, as well as $H_{1/10}$. Huang et al. (2004) have also analyzed these two data sets. They have found that the wave records are highly correlated. This is rather natural, considering the short distance between these two measuring stations.

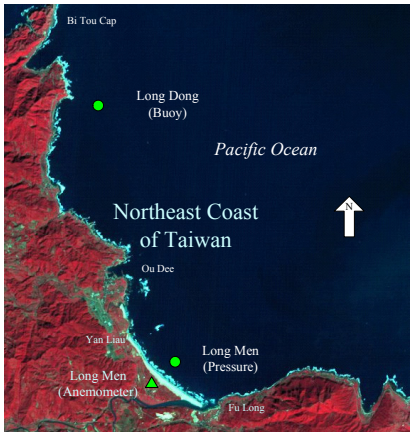

Fig. 1. Schematic sketch of the locations of the measuring stations Long-Dong & Long-Men

The data files were first checked for outliers. Possible outliers were then replaced with zeros, and treated as if they were missing data. We then count the total amount of missing data in the files. From the 26280 should-be data points for the measuring station Long-Dong, a total of 9988 data were missing. The percentage of the missing data is therefore 38%. It should be noted that this high percentage of the data missing rate is due to the fact that we have used zeros to '*elongate*' the file so that it starts from at 00:00 on January 1 1998 and ends at 23:00 on December 31 2003. On the other hand, no additional zeros were added to the data file of Long-Men, which has, on the average, a 22% out of the 11029 should-be data points.

Depending upon the amount of data missing, there can be small or large gaps between recorded wave heights/periods. When less than three consecutive data points were not recorded, a small gap exists and we then used interpolation to estimate the values of these missing data points. When the amount of data missing is larger than three, a large gap exists. We then use the record of the previous year to estimate the parameters of an ARMA model. The missing data in these large gaps are then replaced by estimates from ARMA modelling. Similar procedures were also used by Hidalgo et al. (1995). However, due to the fact that only a little of more than a year's data are available for the measuring station Long-Men, no ARMA simulations were carried out, and the large gaps were left unchanged.

Long-term statistical models found in the literature were used to estimate the possible statistical distributions of the wave heights. Since, as pointed out by many researchers, there is no theoretical justification for the choice of any specific model, we have decided to test all the possible models. Models used for the analyses are: (a) the two-parameter (2-P) lognormal; (b) the three-parameter (3-P) lognormal; (c) the beta; (d) the gamma; (e) the exponential; (f) the Fisher-Tippett Type I; (g) the Pearson type III; (h) the log-Pearson; (i) the Generalized Extreme Value (GEV); (j) the Weibull and (k) the Pareto distribution. Among the 11 models considered, the lognormal, the Weibull, and the Fisher-Tippett Type I (the Gumbel), distributions are often used by researchers in studying the long-term distribution of ocean waves. It is noted that, both the Pearson Type III and the log-Pearson distributions have been used successfully to model flood frequency distributions (See, for example, U. S. Army Civil Works Engineer Manuals, 1993, and Bulletin 17B of the US Water Resources Council, 1982). However, we have found that the curves of the Pearson family failed to fit the wave heights of Long-Men totally, while only the Pearson type III can be used to model the wave height distributions in Long-Dong. Except probably for the last two distributions, the mathematical expressions for all other models can be found in textbooks concerning statistics (See, for example, Hahn & Shapiro, 1967; Haan, 1991), and will not be repeated here for brevity. Detailed descriptions concerning the properties of the GEV distribution and the Pareto distribution can be found in Rao & Hamed (2000). According to them, the GEV distribution can be written as:

$$f(x)=\frac{1}{\alpha}\left(1-k\frac{x-u}{\alpha}\right)^{\frac{1}{k}-1}\exp\left[-\left(1-k\frac{x-u}{\alpha}\right)^{\frac{1}{k}}\right] \qquad (1)$$

where α, k, and u are the parameters of the distribution. The variable x here is the dimensionless wave height, $x = H_j / H_{j_{mean}}$, with the subscript j denotes the specific type of wave heights, i.e., H_{mean}, H_{max}, $H_{1/3}$, or $H_{1/10}$, under consideration. Generally speaking, when the parameter $k \leq 0$ the GEV can be used to model extreme value distribution.

The probability density function of the Pareto distribution for dimensionless wave height x can be expressed as:

$$f(x)=\frac{1}{\alpha}\left[1-\frac{k}{\alpha}(x-\varepsilon)\right]^{\frac{1}{k}-1} \qquad (2)$$

where α, k, and ε are parameters. Similar as for the GEV, the value of the random variable, x, depends on the value of the parameter k. When $k \leq 0$, x have values in the range $\varepsilon \leq x < \infty$, and which makes it more probable to model the distribution of wave heights obtained from the so-called POT (Peaks-Over-Threshold) method (Ferreira & Guedes Soares, 1998). This method, however, was not attempted here in this paper.

RESULTS AND DISCUSSION

Figs. 2 & 3 show the results of fitting measured wave heights with the models mentioned above. Fig. 2 considers the date from Long-Men, while the data from Long-Dong are used in Fig. 3. Among these models considered, the exponential distribution always has its peak at the origin, which makes it unsuitable for the present purpose. As can be seen from Fig. 2, which shows the results of Long-Men, the Pareto distribution overestimates the distribution of the peak. In the following, the curves for these distributions will be excluded from the figure to avoid crowdedness. While models of the Pearson family are always found fail to fit wave heights measured in Long-Men, reasonable results are found for the data from Long-Dong (Fig. 3). Notice also that the three-parameter lognormal distribution follows the peak of the empirical data quite well in Fig. 2, whereas the result shown in Fig. 3 is not so. However, in both figures, the results of the two-parameter lognormal and the GEV distribution are seen to follow the trend of the measured data.

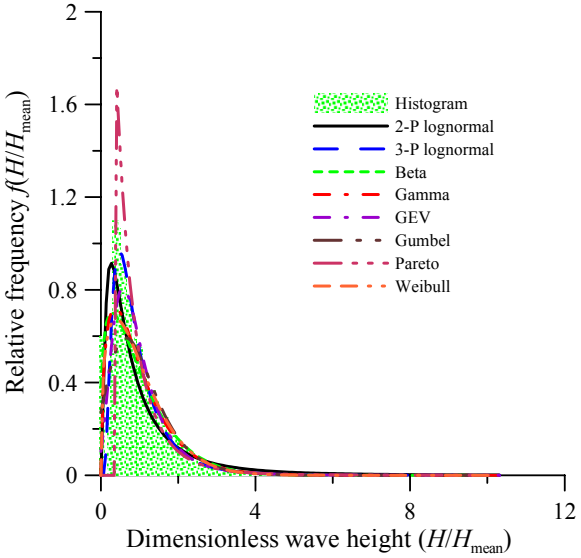

Fig. 2. Results of fitting the statistical models for measured maximum wave heights of Long-Men.

As mentioned earlier, only more than a year's data are available for the measuring station Long-Men. It is considered that the sampling variability may be too large to draw any definite conclusion. We will, therefore, only consider the results obtained for the data of Long-Dong in the following. Fig. 4 shows the results of fitting the wave heights where only the small gaps were replaced with estimates from interpolation. Notice that the two-parameter lognormal distribution is the only model that can fit the peak of this data set. Similar results are also obtained for the completed data set where all the gaps are imputed either through interpolation, or through simulations. This will not be presented here for space reasons.

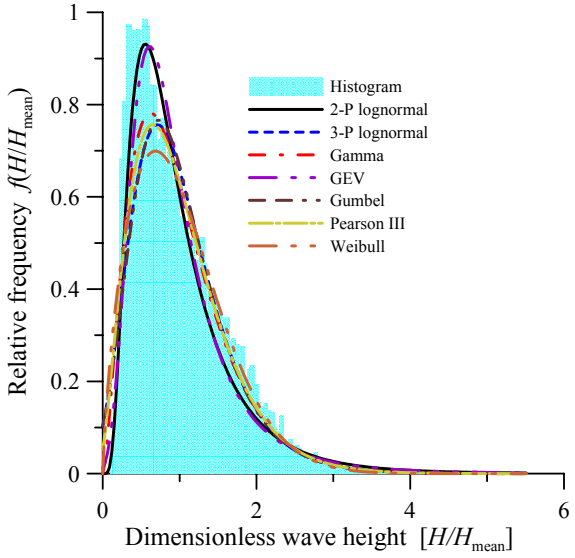

Fig. 3. Results of fitting the statistical models for measured significant wave heights of Long-Dong.

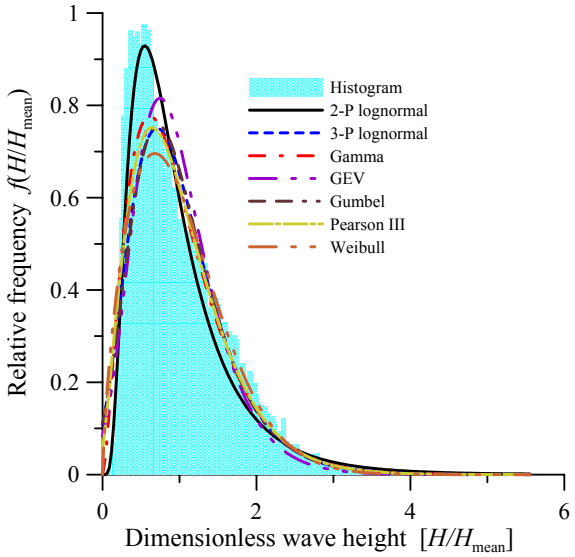

Fig. 4. Results of fitting the statistical models for measured and interpolated significant wave heights of Long-Dong

Some researchers have argued that, since the main objective of studying the statistical properties of extreme wave heights is to be able to predict the design wave height, the ability of a model to fit the tail of large wave heights is, therefore, more important than to fit the peak. This is shown as cumulative distribution in Fig. 5 where the original data is used. As can be seen from Fig. 5, the two lognormal distributions, as well as the GEV distribution failed to follow the measured results. Fig. 6 shows the results of fitting the whole data set, i.e., where the values of the missing data were estimated either through interpolation or through ARMA simulation. It is seen that, besides the three models mentioned above, the Pearson Type III also failed to match the '*measured'* data set.

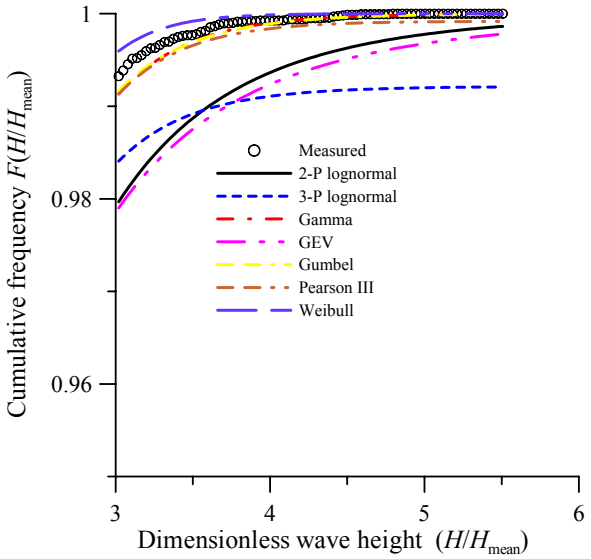

Fig. 5. Comparison of the empirical and fitted cumulative probability distributions for the original data of Long-Dong

The reason for the departure of the Pearson Type III from the tails of the extreme wave heights for the completed data set is unclear. We plot the cumulative distributions of the three data sets, i.e., the original, the original with small gaps filled in through interpolation, and the completed data set, in Fig. 7 for comparison. It can be seen from Fig. 7, there seems to have no large differences for the three "*empirical*" cumulative distributions. More studies are needed to clarify this.

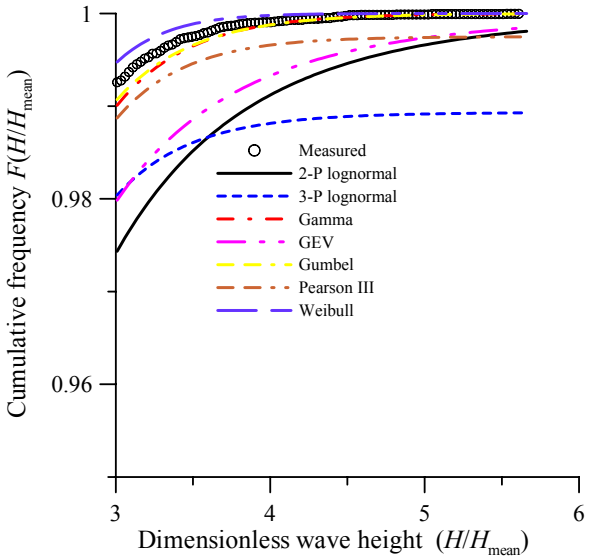

Fig. 6. Comparison of the empirical and fitted cumulative probability distributions for the complete data set of Long-Dong

SUMMARY AND CONCLUSION

Studies on the possible effects of missing data were carried out. In this paper, we originally started with the intention to use two data sets and to draw some conclusions by comparing the results. The data sets were obtained from two adjacent measuring stations, both located at the northeast coast of Taiwan. The wave height data measured at Long-Dong has a length of three and half years; while measurements at Long-Men were conducted for only a little more than a year. It is felt that the data from Long-Men was too short to draw any reasonable conclusions. We have then decided only to use the data set from Long-Dong for further study. It is found that:

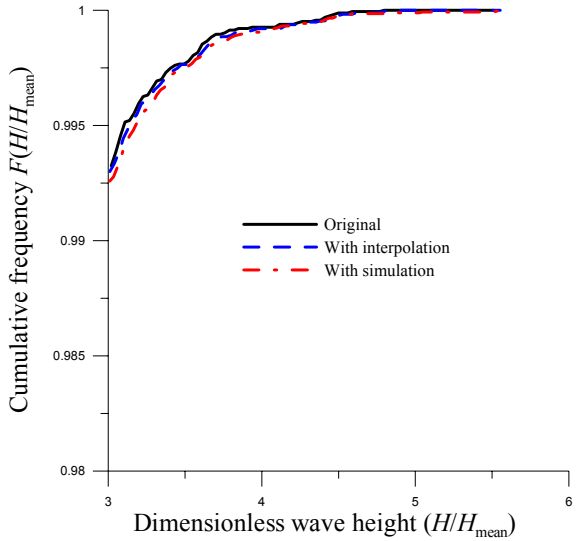

Fig. 7. Comparison of the empirical cumulative probability distributions. Line: original data; dashed line: original + interpolation; dash-dot line: original + interpolation + simulation

- As far as the probability distributions of wave heights are concerned, statistical models such as the two- and three-parameter lognormal, the gamma, the Fisher-Tippett Type I, the Pearson Type III, the GEV, and the Weibull distributions can be used. The log-Pearson distribution, which was found to be applicable for flood frequency analysis failed to yield reasonable results for both data sets.
- However, only curves of the two-parameter lognormal and the GEV distributions were found to match the peaks for all the cases studied.
- If the tails of the high wave heights were to be considered important, the gamma, the Fisher-Tippett Type I (Gumbel), and the Weibull distributions outperform all other models used in this study.
- The above findings seem also to hold when missing data were replaced with estimates either from interpolation or from ARMA simulation.

However, it is stressed that the last conclusion is drawn only from one data set. This could be misleading, considering the sampling variabilities that might be associated with records of nature phenomena.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude for the financial aids of the National Science Council, Republic of China. Project No. NSC-92-2611-E-019-001 (JZY).

REFERENCES

Allison, PD (2001). *Missing Data*, Sage Pub., Inc., Thousand Oaks, 93p.

Brillinger, D, Fienberg, S, Gani, J, Hartingan, J & Krickeberg K (eds.) (1984). "*Time series analysis of irregularly observed data*," Springer

Verlag, New York, 363p.
Ferreira, JA & Guedes Soares, C (1998). "An application of the peaks over threshold method to predict extremes of significant wave height," *J Offshore Mech. & Arc Eng.,* Vol 120, pp 165-176.
Guedes Soares, C & Ferreira, AM (1995). "Analysis of the seasonality in non-stationary time series of significant wave height," in "*Comput Stochastic Mech*", Spanos PD (ed.), Balkema, Rotterdam. pp 559-568.
Guedes Soares, C, Ferreira, AM, Cunha, C (1996). "Linear models of the time series of significant wave height in the Southwest Coast of Portugal," *Coastal Eng*, Vol 29, pp 149-167.
Haan, CT (1991). "*Statistical methods in hydrology*," Iowa State Univ Press, 5th Printing, 378 p.
Hahn, GJ & Shapiro, SS 1967) "*Statistical models in engineering*," John Wiley & Sons, New York, 355 p.
Hidalgo, OS, Nieto Borge, JC, Cunha, CC & Guedes Soares, C (1995). "Filling missing observations in time series of significant wave height," *Proc Offshore Mech&Arc Eng, Vol. II,* pp 9-17.
Huang, WP, Yim, JZ, Chou, CR and Kung, CS (2004). "Short Term Wave Climate of Northeast Taiwan," *Proc 6th.Pacific/Asia Offshore Mechanics Symp,* Vladivostok, Russia, ISOPE.
Little, RJA & Rubin, DB (1987). "*Statistical analysis with missing data*," John Wiley & Sons, New York, 278 p.
Rao, AR & Hamed, KH (2000). *Flood Frequency Analysis*, CRC Press, Boca Raton, 350 p.
Stefanakos, ChN & Athanassoulis, GA (2001). "A Unified Methodology for the Analysis, Completion and Simulation of Nonstationary Time Series with Missing Values, with Application to Wave Data," *Appl Ocean Res*, Vol 23, pp 207-220.
U.S. Army Corps of Engineers, (1993) "*Hydrologic frequency analysis*" U. S. Army Civil Works Engineer Manuals, EM 1110-2-1415, 149p.
Water Resources Council, (1982) "*Guidelines for determining flood flow frequency*," Bull 17B, Hydrology Comm, 183p.