CHAPTER 67

# On the Methodology of Selecting Design Wave Height

Yoshimi Goda*, M. ASCE

## ABSTRACT

A statistically-rational method of extreme wave data analysis is presented. A combination of the Fisher-Tippett type I and the four Weibull distributions is proposed as the candidates of distribution functions. The least square method is used for data fitting. The best plotting position formula for each function is determined by the Monte Carlo method with 10,000 simulations per sample size.

Confidence intervals of estimated extreme wave heights for given return periods are evaluated by simulations and expressed in the form of empirical formulas, for both the cases when the true distribution is known and unknown. An example of extreme wave data analysis is given.

## 1. INTRODUCTION

Design waves must be decided upon by the responsible engineer in charge of a maritime project. He looks for various data source on storm waves and asks statisticians to make extreme wave analysis. Based on the statistical estimate of wave height for a certain return period, he makes a final decision. The statistical procedure for this purpose is rather confused at present, however, with several different methods being employed in various occasions. Some procedure is not recommendable from the statistical point of view, and some other is quite complicated to use. There is a need for a clear and sound statistical method for extreme wave data analysis.

Another problem in extreme data analysis is the lack of information on the statistical uncertainty or confidence interval of estimated extreme value for a given return period (hereinafter called "return value" for the sake of simplicity). For the Fisher-Tippett type I distribution (abbreviated as FT-I), a formula is given in Gumbel's book (1958, Sec. 6.2.3). Lawless (1974) also gave integral form-

* Professor, Yokohama National University, Department of Civil Engineering, 156 Tokiwadai, Hodogaya-ku, Yokohama, Japan 240.

ulas for the FT-1 distribution for the case when the parameters are estimated by the maximum likelihood method. Both formulas are applicable for the annual maximum series data but inapplicable for the partial-duration series data. For the Weibull distribution, there is practically no formula available for the confidence interval of return values. The problem of confidence interval when the true distribution is unknown has only been mentioned by Petruaskas and Aagaard (1970), but they did not give any formula for its estimation.

The present paper tries to give a clear view of the practical procedure of extreme wave data analysis, which is statistically sound and easily applicable. The confidence interval of return values will also be estimated for several distribution functions used in the analysis. Detailed descriptions can be found in Goda (1988) in Japanese.

## 2. CLASSIFICATION OF EXTREME WAVE STATISTICS

### *Total Sample Method versus Peak Value Method*

Two different methods are currently employed to prepare extreme statistics of storm waves. The first one may be called the total sample method which employs the all wave data recorded at a regular interval of a few hours. The second one may be called the peak value method which picks up the peak wave heights of individual storms and thus composes a set of extreme wave data.

The total sample method was proposed by Draper (1966), when wave observation projects at various countries were at their initial stages and the accumulation of wave data was short in terms of the time span of observation. This method is applicable even when the observation is a few years long and it is easy to use. However, the method violates the condition of statistical independence between individual data, because the regularly recorded wave heights are mutually correlated; the correlation coefficient remains over 0.3 to 0.5 with the time lag of 24 hours (see a review by Goda 1979). Therefore, use of the total sample method should be refrained in the present days when longer wave data by observations and/or hindcasting are available.

The total sample method also has an ambiguity in the selection of unit time in assessing return wave heights. One may take the recording interval of a few hours, but he may use one day as the unit time instead. Depending on the unit time, the return wave heights vary. Medina and Aguilar (1986) have proposed to employ the mean period between successive storms in their attempt to compromise the total sample method with the statistical condition of independent data.

### *Annual Maximum Series versus Partial-Duration Series*

The data set for the peak value method can be either the annual maximum series or the partial-duration series. The latter refers to the series of peak storm wave heights in

the order of appearance. The annual maximum series data is clear in definition and easy to deal with. Most of wave data, however, cover rather short time spans, say less than 40 years by hindcasting and 10 years by instrumental observations. Therefore, the partial-duration series is the usual data source for extreme wave analysis.

In the analysis of partial-duration series data, the mean number of storm waves per year is an important parameter. It is called here as the **mean rate** and denoted with $\lambda$. The annual maximum series data can be treated as the case with $\lambda = 1$, although the meaning is slightly different. In practice of extreme wave analysis, the data above a certain threshold height are usually adopted to constitute the data set. Such data is called the **censored data** in contrast with the **uncensored data** which is composed of all storm wave data. The ratio of the number of adopted data N to the whole number of storm data $N_T$ is hereby called the **censoring parameter** and denoted by $\nu$ : i.e., $\nu = N/N_T$. Burcharth (1988) calls the method utilizing the censored partial duration series data as the Peak-over-threshold method.

### *Methods of Data Fitting to Distribution Functions*

There are four methods for fitting a set of extreme wave data to some distribution function. They are the graphical method, the method of moments, the least square method, and the maximum likelihood method. The graphical method is susceptible to subjective judgment and not recommended except for initial analysis. The moment method requires fewer calculation than the least square method, but cannot deal with the censored data. The maximum likelihood method is theoretically rigorous and favored by many statisticians, but the calculation is cumbersome even though the present computers can handle it with ease.

The least square method is a sophistication of the graphical method, but it has been neglected by most of statisticians except for a few such as Blom (1963) by reasons unknown. It is simple and clear in the calculation process, and therefore it is adopted in the present paper for extreme wave data analysis.

## 3. DISTRIBUTION FUNCTIONS AND PLOTTING POSITION FORMULAS

### *Candidates of Distribution Functions*

In the extreme data analysis of various environmental phenomena such as rainfall, flood discharge, strong wind, and storm waves, selection of distribution functions to be fitted to the data is always a problem. No theoretical justification is provided for the selection except that the FT-I should be applicable if the basic population to define the maximum of a group of data belongs to the exponential distribution. In some occasion, a particular distribution such as the FT-I or the log-normal is assumed as applicable to the data set even without any theoretical support. In many occasions, however, several candidate functions are tested and the best fitting distribution is adopted.

Petruaskas and Aagaard (1970) have proposed to choose among a set of the FT-I and the seven Weibull distributions with the shape parameter fixed at certain values. As will be seen later, the capability of the extreme data analysis to recognize the parent distribution among several candidates is rather low when the sample size is less than 100 or so. Therefore, it is better to restrict the number of candidate distributions. The present paper proposes the following set of five functions:

FT-I distribution:

$$F(x) = \exp\{-\exp[-(x-B)/A]\} \tag{1}$$

Weibull distribution:

$$F(x) = 1 - \exp\{-[(x-B)/A]^k\} \tag{2}$$

$$: k = 0.75,\ 1.0,\ 1.4,\ \text{and}\ 2.0$$

The distribution function F(x) represents the probability of nonexceedance of the variate x, and A, B, and k are called the location, scale, and shape parameters, respectively. It is often useful in the extreme statistics to introduce the reduced variate y which is defined by

$$y = (x-B)/A \tag{3}$$

The log-normal distribution is not adopted here, because its characteristics are quite similar with those of the Weibull distribution with k = 2.0 and it requires the calculation of the error function, which is not easy to make with a pocket calculator.

### *Selection of Plotting Position Formulas*

A drawback of the least square method for extreme data analysis is the necessity of choosing a right plotting position formula for each distribution function. Figure 1 is an example of the histograms of the ordered extreme data and its nonexceedance probability. The parent distribution is the FT-I (A = 1.0 and B = 0), and 100 samples with 10 data each are randomly drawn from the population. In each sample, the largest data $x_{(1)}$ are sorted out and its histogram is shown in the right. The nonexceedance probability $F_1$ corresponding to $x_{(1)}$ is shown in the top as a histogram.

The probability to be assigned to each ordered extreme data, which depends upon the ordered number and the total data number only, is called the plotting position. Cunnane (1978) has made a critical review of this problem and denounced the recommendation by Gumbel (1958) of the Weibull formula, which is expressed as

$$\hat{F}_m = 1 - m/(N+1) \tag{4}$$

where m is the descending ordered number from the largest.

A plotting position formula should be so selected to

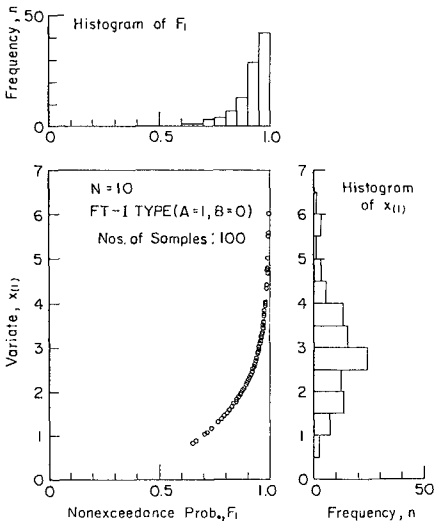


Fig. 1 Distribution of the largest data in a sample from FT-I.

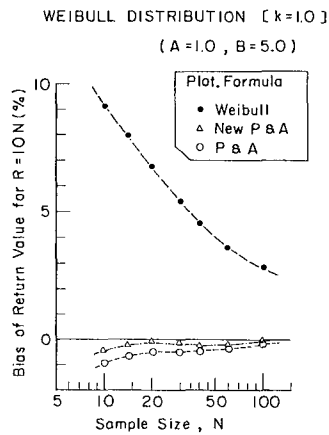


Fig. 2 Mean bias of return values due to plotting formulas in Weibull distrib. (k=1.0).

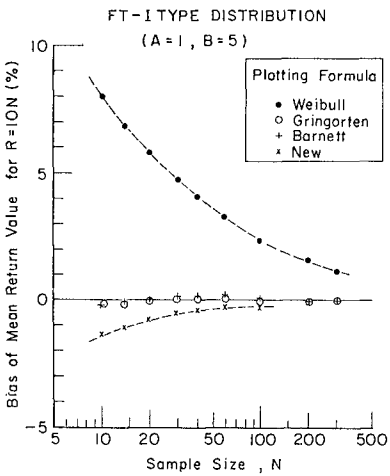


Fig. 3 Mean bias of return values due to plotting formulas in FT-I distribution.

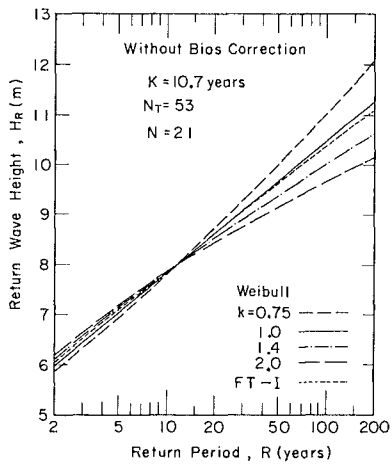


Fig. 4 Example of return wave height estimation with various distribution functions.

yield no bias on the mean and the least root-mean-square error of the return values. The first choice would be the use of the probability corresponding to the expected value of the ordered variate, or $F\{E[x_{(m)}]\}$. Gringorten (1961) has derived the following formula for the FT-I distribution:

$$\text{FT-I:} \quad \hat{F}_m = 1 - (m - 0.44)/(N_T + 0.12) \qquad (5)$$

$$: m = 1, 2, \ldots, N$$

in which $N_T$ is the number of whole extreme data including those below a threshold value. Use of $N_T$ instead of N (number of adopted data) is to enhance the possibility of recognizing the parent distribution for the case of censored data, after the suggestion by Muir and El-Shaarawi (1986). For the Weibull distribution, Petruaskas and Aagaard (1970) have presented the following formula:

$$\text{Weibull:} \quad \hat{F}_m = 1 - (m - \alpha)/(N_T + \beta) \qquad (6)$$

where

$$\alpha = 0.30 + 0.18/k, \quad \beta = 0.21 + 0.32/k \qquad (7)$$

Figure 2 shows the bias of return values caused by use of various plotting position formulas for the case of FT-I distribution. The sample size is varied from 10 to 300, and 10,000 uncensored samples are simulated by the Monte Carlo method at each sample size. The bias is defined here as $100 \times (\hat{x}_R / x_R - 1)$, where $\hat{x}_R$ and $x_R$ refer to the estimated and true return values, respectively. The return values are evaluated at the return period being 10N.

It is clear in Fig. 2 that the Weibull formula yields conspicuous positive bias, indicating a tendency of overestimation. The formula by Barnett (1975) though not presented here gives practically no bias just as same as Gringorten's formula. The formula denoted as "New" has the expression same as Eq. 6 but with the constants $\alpha = 0.51$ and $\beta = 0.18$. This formula has been derived as to yield the best fitting to $F\{E(x_m)]\}$, but it yields slightly negative bias. Figure 2 clearly indicates the Gringorten formula to be used for the FT-I distribution. Poor performance of the Weibull formula has also been reported by Carter and Challenor (1983) with a small scale Monte Carlo simulation, in in which they compared the effectiveness of various methods of extreme data analysis.

Figure 3 shows a similar comparison of plotting position formulas for the Weibull distribution with k = 1.0. The Weibull formula again exhibits large positive bias. The formula by Petruaskas and Aagaard (denoted as "P & A") shows a tendency to yield slightly negative bias. A modification is made here by changing the constants as

$$\alpha = 0.20 + 0.27/\sqrt{k}, \quad \beta = 0.20 + 0.23/\sqrt{k} \qquad (8)$$

The new formula is denoted as "New P & A" in Fig. 3. This plotting formula is used in this paper hereinafter.

For the log-normal distribution, the appropriate plotting position formula is that by Blom (1958) in the form of Eq. 6 with $\alpha = 3/8$ and $\beta = 1/4$. Use of the Weibull formula yields positive bias as demonstrated in a simulation study by Earle and Baer (1982). The amount of bias reported by them has been confirmed in the author's simulation (Goda, 1988); use of the Blom formula for the same condition has yielded no bias.

## 4. DATA FITTING AND CALCULATION OF RETURN VALUES

For a given set of extreme wave data, the first step of data processing is to rearrange the data in the descending order from the largest to the smallest. Then the nonexceedance probability $\hat{F}_m$ is assigned to each data by Eqs. 5, 6, and 8 by assuming the FT-I and the four Weibull distributions as the candidate functions. For each $\hat{F}_m$, the following reduced variate $y_{(m)}$ is then calculated:

$$\text{FT-I:} \qquad y_{(m)} = -\ln[-\ln \hat{F}_m] \tag{9}$$

$$\text{Weibull:} \qquad y_{(m)} = [-\ln(1 - \hat{F}_m)]^{1/k} \tag{10}$$

Because there should exist a linear relation between the ordered variate $x_{(m)}$ and its reduced variate $y_{(m)}$, the following equation is assumed and solved by the least square method to yield the estimates $\hat{A}$ and $\hat{B}$ of the scale and location parameters:

$$x_{(m)} = \hat{A}\, y_{(m)} + \hat{B} \tag{11}$$

Table 1 is an example of the above procedures applied for an extreme data of typhoon waves measured at the depth of 50 m at a location facing the Pacific. The measurement has been continued for more than 15 years, and the portion of successful recording covering K = 10.74 years is used in the analysis. There were 53 typhoon waves during this effective observation period of 10.74 years, and thus N = 53 and $\lambda$ = 4.93 per year. The wave heights exceeding 4.0 m have been chosen for the analysis, and there were 21 such data; i.e., N = 21 and $\nu$ = 0.396.

The nonexceedance probabilities $\hat{F}_m$ listed in Table 1 are all above 0.61, because the total number of typhoon waves $N_T$ = 53 is used in evaluating $\hat{F}_m$ by Eqs. 5 and 6. Even if the threshold height is changed to 4.5 m and the number of data is reduced to N = 18, the probability $\hat{F}_m$ remains at the same value so long as $N_T$ is the same.

The return value or the expected extreme wave height for a given return period R is then calculated by the following equation with the estimated parameters $\hat{A}$ and $\hat{B}$:

$$x_R = \hat{A}\, y_R + \hat{B} \tag{12}$$

where

$$\left.\begin{aligned} y_R &= -\ln\{-\ln[1 - 1/(\lambda R)]\} \quad &&: \text{FT-I} \\ y_R &= [\ln(\lambda R)]^{1/k} &&: \text{Weibull} \end{aligned}\right\} \tag{13}$$

Table 1. Example of Extreme Wave Data Fitting to Several Distribution Functions [units in meters] - N = 21, $N_T$ = 53, K = 10.7 years -

| m | $x_m$ | FT-I $\hat{F}_m$ | FT-I $y_m$ | Weibull(0.75) $\hat{F}_m$ | Weibull(0.75) $y_m$ | Weibull(1.0) $\hat{F}_m$ | Weibull(1.0) $y_m$ | Weibull(1.4) $\hat{F}_m$ | Weibull(1.4) $y_m$ | Weibull(2.0) $\hat{F}_m$ | Weibull(2.0) $y_m$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.36 | 0.9895 | 4.55 | 0.9909 | 7.86 | 0.9901 | 4.61 | 0.9893 | 2.95 | 0.9886 | 2.12 |
| 2 | 7.02 | 0.9706 | 3.51 | 0.9722 | 5.48 | 0.9714 | 3.55 | 0.9706 | 2.46 | 0.9699 | 1.87 |
| 3 | 6.94 | 0.9518 | 3.01 | 0.9535 | 4.46 | 0.9527 | 3.05 | 0.9518 | 2.21 | 0.9511 | 1.74 |
| 4 | 6.85 | 0.9330 | 2.67 | 0.9348 | 3.82 | 0.9339 | 2.72 | 0.9331 | 2.04 | 0.9324 | 1.64 |
| 5 | 6.74 | 0.9142 | 2.41 | 0.9161 | 3.35 | 0.9152 | 2.47 | 0.9144 | 1.90 | 0.9136 | 1.57 |
| 6 | 6.20 | 0.8953 | 2.20 | 0.8974 | 3.00 | 0.8965 | 2.27 | 0.8957 | 1.79 | 0.8945 | 1.50 |
| 7 | 5.92 | 0.8765 | 2.03 | 0.8787 | 2.71 | 0.8778 | 2.10 | 0.8769 | 1.70 | 0.8762 | 1.45 |
| 8 | 5.68 | 0.8577 | 1.87 | 0.8599 | 2.46 | 0.8591 | 1.96 | 0.8582 | 1.61 | 0.8574 | 1.40 |
| 9 | 5.57 | 0.8389 | 1.74 | 0.8412 | 2.26 | 0.8404 | 1.84 | 0.8395 | 1.54 | 0.8387 | 1.35 |
| 10 | 5.42 | 0.8200 | 1.62 | 0.8225 | 2.08 | 0.8216 | 1.72 | 0.8207 | 1.47 | 0.8199 | 1.31 |
| 11 | 5.34 | 0.8012 | 1.51 | 0.8038 | 1.92 | 0.8029 | 1.62 | 0.8020 | 1.41 | 0.8012 | 1.27 |
| 12 | 5.10 | 0.7824 | 1.41 | 0.7851 | 1.78 | 0.7842 | 1.53 | 0.7833 | 1.35 | 0.7825 | 1.24 |
| 13 | 5.09 | 0.7636 | 1.31 | 0.7664 | 1.65 | 0.7655 | 1.45 | 0.7646 | 1.30 | 0.7637 | 1.20 |
| 14 | 4.95 | 0.7447 | 1.22 | 0.7477 | 1.53 | 0.7468 | 1.37 | 0.7458 | 1.25 | 0.7450 | 1.17 |
| 15 | 4.81 | 0.7259 | 1.14 | 0.7290 | 1.43 | 0.7281 | 1.30 | 0.7271 | 1.21 | 0.7262 | 1.14 |
| 16 | 4.77 | 0.7071 | 1.06 | 0.7103 | 1.33 | 0.7093 | 1.24 | 0.7084 | 1.16 | 0.7075 | 1.11 |
| 17 | 4.63 | 0.6883 | 0.99 | 0.6916 | 1.24 | 0.6906 | 1.17 | 0.6896 | 1.12 | 0.6888 | 1.08 |
| 18 | 4.61 | 0.6694 | 0.91 | 0.6729 | 1.16 | 0.6719 | 1.11 | 0.6709 | 1.08 | 0.6700 | 1.05 |
| 19 | 4.41 | 0.6506 | 0.84 | 0.6542 | 1.08 | 0.6532 | 1.06 | 0.6522 | 1.04 | 0.6513 | 1.03 |
| 20 | 4.34 | 0.6318 | 0.78 | 0.6355 | 1.01 | 0.6345 | 1.01 | 0.6335 | 1.00 | 0.6325 | 1.00 |
| 21 | 4.11 | 0.6130 | 0.71 | 0.6168 | 0.95 | 0.6158 | 0.96 | 0.6147 | 0.97 | 0.6138 | 0.98 |
| $\bar{x}$=5.585 | | A = 1.091 | | A = 0.614 | | A = 1.147 | | A = 2.084 | | A = 3.560 | |
| $\sigma_x$=1.101 | | B = 3.617 | | B = 4.029 | | B = 3.374 | | B = 2.334 | | B = 0.786 | |
| | | r = 0.9842 | | r = 0.9621 | | r = 0.9790 | | r = 0.9878 | | r = 0.9910 | |

Figure 4 exhibits the result of estimating the return wave heights for the data of Table 1. Depending on the distribution functions fitted to the data, the return wave heights vary and the difference increases as the estimate is made to longer return periods.

There is no absolute criterion to choose a particular function among several candidates fitted to an extreme data. For practical purposes, the correlation coefficient r between the ordered variate $x_{(m)}$ and its reduced variate $y_{(m)}$ can serve as the basis of selection. In the example of Table 1, the Weibull distribution with k = 2.0 indicates the highest correlation with r = 0.9910 and is chosen as the distribution best fitted to the data.

## 5. STANDARD DEVIATION OF RETURN VALUE WHEN THE TRUE DISTRIBUTION IS KNOWN

A set of extreme wave data being analyzed is regarded as a sample drawn from an unknown population of storm waves. Another set of extreme waves to be obtained in the coming several tens of years will form a sample different from the present sample, even if no long-term climatic change exists. The two sets of extreme data will surely yield different distribution functions and the return wave heights will eventually be different.

Figure 5 shows some results on the statistical variability of return values. A population with the FT-I distribution (A = 1 and B = 5) is assumed, and Monte Carlo simula-

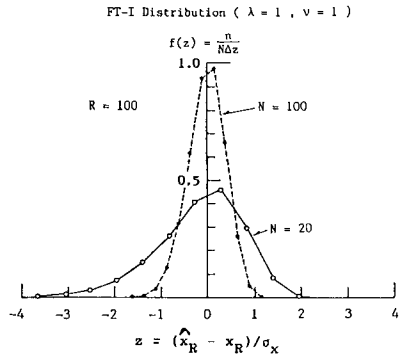


Fig. 5 Distribution of estimated return values in dimensionless form.

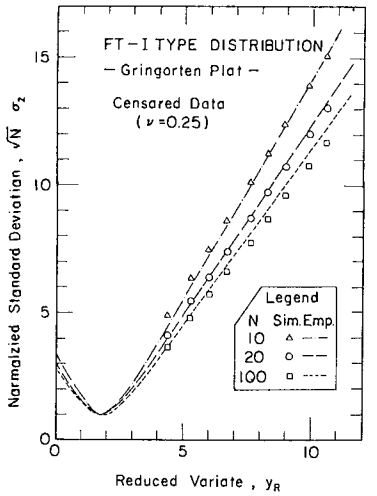


Fig. 6 Example of standard deviation of return values in censored samples from FT-I.

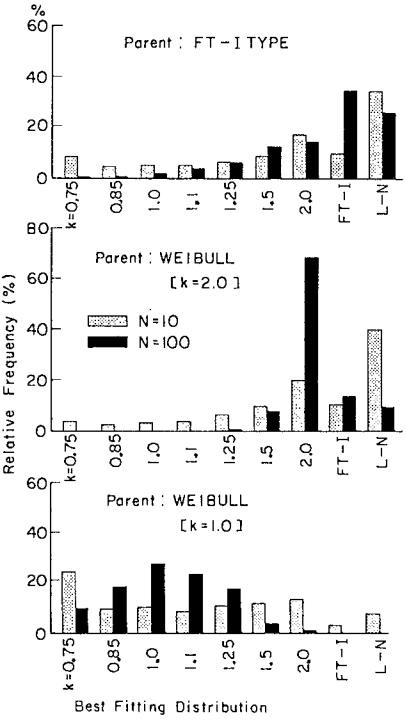


Fig. 7 Relative frequencies of best-fitted distributions for random samples.

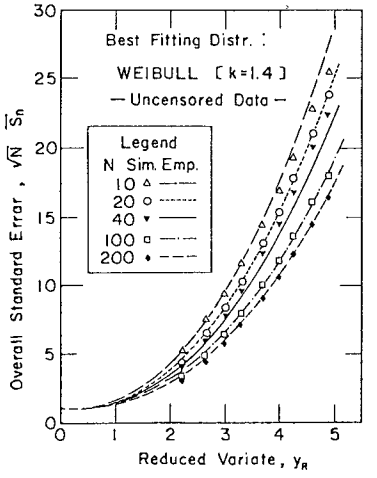


Fig. 8 Example of standard error of return values when the true distribution is unknown.

tions are carried out to yield 10,000 uncensored samples each for the sample size $N = N_T = 20$ and 100. By fitting the FT-I function to each sample, the return value at the return period $R = 100$ is estimated. The difference between the estimated and true return value is normalized with the standard deviation of each sample $\sigma_x$. Figure 5 presents the histograms of this normalized deviation in the form of probability density. As seen clearly, the deviation may exceed $2\sigma_x$ in case of the sample with $N = 20$ and $1\sigma_x$ for $N = 100$.

Monte Carlo simulations have been carried out for various combinations of the distribution functions, the sample size, and the censoring parameter, in order to obtain sufficient data on the statistical variability of the return values. For each combination, the standard deviation of the dimensionless variate $z = (\hat{x}_R - x_R)/\sigma_x$ has been evaluated. Figure 6 is an example of the standard deviation $\sigma_z$ multiplied by $\sqrt{N}$ in order to reduce the influence of the sample size on data presentation. The data are those of censored ones ($\nu = 0.25$) sampled from the FT-I distribution.

Empirical formulations of the standard deviations of the return values have been tried by referring to the formula cited by Gumbel (1958) as in the following form:

$$\sigma_z = [1.0 + a\,(y_R - c + \varepsilon \ln\nu)^2]^{1/2}/\sqrt{N} \qquad (14)$$

where

$$a = a_1 \exp[a_2\, N^{-1.3} + \kappa\,(-\ln\nu)^{1/2}] \qquad (15)$$

The coefficients in the above equations are assigned the values listed in Table 2. The dashed and dotted curves in Fig. 6 represent the empirical predictions of the standard deviations of the return values.

Table 2. Coefficients of Empirical Formulas for Standard Deviation of Return Values When the True Distribution Is Known

| Distribution | $a_1$ | $a_2$ | $\kappa$ | c | $\varepsilon$ |
|---|---|---|---|---|---|
| FT-I | 0.64 | 9.0 | 0.93 | 0.0 | 1.33 |
| Weibull (k=0.75) | 1.65 | 11.4 | -0.63 | 0.0 | 1.15 |
| Weibull (k=1.0) | 1.92 | 11.4 | 0.00 | 0.3 | 0.90 |
| Weibull (k=1.4) | 2.05 | 11.4 | 0.69 | 0.4 | 0.72 |
| Weibull (k=2.0) | 2.24 | 11.4 | 1.34 | 0.5 | 0.54 |

Once the dimensionless deviation $\sigma_z$ is evaluated, the standard deviation of the return value can be approximately estimated as $\sigma(\hat{x}_R) = \sigma_z\,\sigma_x$. The confidence interval of the return value is then constructed with $\sigma(\hat{x}_R)$.

## 6. PROBLEM OF FINDING TRUE DISTRIBUTION

The capability of finding out the true distribution from a random sample of extreme data has been tested with Monte Carlo simulations. Figure 7 demonstrates some results. A parent distribution is assumed, and 10,000 uncensored random samples with the size 10 and 100 were prepared. For

each sample, a group of nine candidate distributions were applied and the best fitting one was selected by the criterion of the largest correlation coefficient between $x_{(m)}$ and $y_{(m)}$. The ordinates in Fig. 7 shows the relative frequency of the best-fitted function, while the abscissa stands for the distribution functions: k = 0.75 to 2.0 denote the Weibull distributions and "L-N" stands for the log-normal distribution. The set of Weibull distributions is same as that used by Petruaskas and Aagaard (1970).

It is clear in Fig. 7 that a small sample has only a small chance of selecting the parent distribution as the best-fitting one. Even at the sample size N = 100, samples from the FT-I distribution can be mistaken as those belonging to the log-normal distribution. Such competitive power of the log-normal against the FT-I and the Weibull distributions is one reason for the rejection of the former from the candidates of the distributions functions in the present proposal of extreme data analysis.

The low capability of a statistical data analysis in recognizing the true distribution function is due to the statistical variability of random samples. The standard deviation of an uncensored sample with the size 40 from the Weibull distribution (k = 1.0), for example, can be less than 0.5 times or more than 1.7 times the population value. A sample with a small deviation tends to be better fitted by the distribution with narrower spreading, and vice versa. Use of the statistical method other than the least square method may slightly enhance the capability of recognizing the true distribution, but the amount of enhancement will be small.

## 7. BIAS CORRECTION TO RETURN VALUE WHEN THE TRUE DISTRIBUTION IS UNKNOWN

Fitting of an extreme data to a distribution other than the true one causes a certain bias to the estimated return values, even though the right plotting position formula is employed. For example, if a sample from the Weibull with k = 1.0 is judged to be best fitted to the Weibull with k = 0.75, the return values will be overestimated because of the latter's characteristic of having a longer tail.

An attempt for the correction of this kind of bias is made in this paper, by assuming the five candidate distributions have the equal probability of existence in the nature. Another series of Monte Carlo simulations have been made with 2,000 to 10,000 samples for each combination of the parent distribution, sample size, and censoring parameter. Each sample is analyzed with the procedure described in this paper, and the return values are estimated and compared with the true values. The difference is normalized as $z = (\hat{x}_R - x_R)/\sigma_x$ and its ensemble mean is calculated for each distribution function which is best fitted to samples taken from various parent distribution functions. The ensemble mean of the dimensionless difference $\bar{Z}_n$ is thought to represent the bias due to not-knowing true distribution and is empirically formulated as in the following:

Table 3. Example of the Estimation of Return Wave Heights with Bias Correction and Their Standard Errors

| Return Period (years) | FT-I $H_R$ (m) | FT-I $\sigma$ (H) (m) | Weibull(0.75) $H_R$ (m) | Weibull(0.75) $\sigma$ (H) (m) | Weibull(1.0) $H_R$ (m) | Weibull(1.0) $\sigma$ (H) (m) | Weibull(1.4) $H_R$ (m) | Weibull(1.4) $\sigma$ (H) (m) | Weibull(2.0) $H_R$ (m) | Weibull(2.0) $\sigma$ (H) (m) |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.0 | 6.2 | 0.4 | 5.9 | 0.3 | 6.0 | 0.4 | 6.2 | 0.5 | 6.4 | 0.5 |
| 5.0 | 7.3 | 0.7 | 6.8 | 0.5 | 7.1 | 0.6 | 7.3 | 0.7 | 7.6 | 0.8 |
| 10.0 | 8.1 | 0.9 | 7.6 | 0.7 | 7.9 | 0.8 | 8.1 | 0.9 | 8.4 | 1.0 |
| 20.0 | 8.9 | 1.2 | 8.3 | 1.0 | 8.8 | 1.1 | 8.9 | 1.2 | 9.3 | 1.3 |
| 50.0 | 10.0 | 1.6 | 9.4 | 1.4 | 9.9 | 1.5 | 10.0 | 1.5 | 10.4 | 1.7 |
| 100.0 | 10.8 | 1.9 | 10.1 | 1.6 | 10.7 | 1.8 | 10.6 | 1.9 | 11.2 | 2.1 |

$$\overline{Z_n} = \begin{cases} A_c\ (y_R + \alpha \ln\nu)^p & : y_R > -\alpha \ln\nu \\ 0 & : y_R \leqq -\alpha \ln\nu \end{cases} \quad (16)$$

The coefficients $A_c$ and $\alpha$ and the exponent p are assgined the following values depending on the distribution functions and the censoring parameter:

FT-I :

$\nu = 1.0$ ;

$$\left.\begin{aligned} A_c &= \begin{cases} 0.046 - 0.40[\log_{10}(60/N)]^3 & : N < 60 \\ 0.046 \exp\{-2.5[\log_{10}(N/60)]^2\} & : N \geqq 60 \end{cases} \\ \alpha &= 0.9, \quad p = 1.0 \end{aligned}\right\} \quad (17)$$

$\nu = 0.5$ & $0.25$ ;

$$\left.\begin{aligned} A_c &= 0.01 - 0.044[\log_{10}(N/300)]^4 \\ \alpha &= 0.9, \quad p = 1.0 \end{aligned}\right\} \quad (18)$$

Weibull (k = 0.75) :

$$\left.\begin{aligned} A_c &= \begin{cases} 0.030 \exp\{-0.6[\log_{10}(N/4)]^2\} & : \nu = 1.0 \\ 0.025 \exp\{-0.7[\log_{10}(N/15)]^2\} & : \nu = 0.5 \text{ \& } 0.25 \end{cases} \\ \alpha &= 2.7, \quad p = 1.6 \end{aligned}\right\} \quad (19)$$

Weibull (k = 1.0) :

$$\left.\begin{aligned} A_c &= \begin{cases} -0.028\ N^{-0.25} & : \nu = 1.0 \\ -0.0022 - 0.0006[\log_{10}(N/50)]^2 & : \nu = 0.5 \text{ \& } 0.25 \end{cases} \\ \alpha &= 1.0, \quad p = 2.1 \end{aligned}\right\} \quad (20)$$

Weibull (k = 1.4) :

$$\left.\begin{aligned} A_c &= \begin{cases} -0.40\ N^{-0.8} & : \nu = 1.0 \\ -0.10\ N^{-0.4} & : \nu = 0.5 \text{ \& } 0.25 \end{cases} \\ \alpha &= 0.5, \quad p = 2.7 \end{aligned}\right\} \quad (21)$$

Weibull (k = 2.0) :

$$\left.\begin{aligned} A_c &= \begin{cases} -0.50\ N^{-0.7} & : \nu = 1.0 \\ -0.64\ N^{-0.6} & : \nu = 0.5 \text{ \& } 0.25 \end{cases} \\ \alpha &= 0.35, \quad p = 3.4 \end{aligned}\right\} \quad (22)$$

The bias correction can be made by using $\overline{Z}_n$ as

$$(x_R)_{cor} = \hat{x}_R - \overline{Z}_n\ \sigma_x \quad (23)$$

Table 3 lists the return wave heights with bias correction for the data of Table 1. The bias correction by the

above formula may have been excessive, being judged from the fact that the Weibull distribution with k = 2.0 predicts the largest return heights after bias correction. This is due to the assumption of the equal probability of the five distribution employed. A responsible engineer may adjust the amount of bias correction in his selection of design waves.

## 8. STANDARD ERROR OF RETURN VALUE WHEN THE TRUE DISTRIBUTION IS UNKNOWN

When the true distribution is unknown, the standard error of the estimated return value differs from the standard deviation formulated as in the form of Eq. 14. By using the simulation data for the above bias correction, empirical formulation has been made for the standard error of return value with bias correction. The standard error is defined as $\overline{S_n} = \sigma [(\hat{x}_R - x_R)/\sigma_x]$ and is formulated as

$$\overline{S_n} = (1.0 + A_s |y_R + \alpha \ln\nu|^p)/\sqrt{N} \qquad (24)$$

The coefficient $A_s$ is expressed as below,

$$A_s = b_1 + b_2 [\log_{10}(N/N_c)]^2 \qquad (25)$$

and the coefficients $\alpha$, $b_1$, $b_2$, and $N_c$ and the exponent p are assigned the values as in Table 4: $\alpha$ is common with Eqs. 16 to 22. An example of the comparison between the simulation data and empirical formula is shown in Fig. 8.

Table 4. Coefficients of Empirical Formulas of Standard Error of Return Values when the True Distribution Is Unknown

| Distribution | $\nu$ | $b_1$ | $b_2$ | $N_c$ | p | $\varepsilon$ |
|---|---|---|---|---|---|---|
| FT-I | 1.0 | 0.24 | 0.36 | 80 | 1.6 | 0.9 |
| | 0.5, 0.25 | 0.46 | 0.14 | 50 | 1.6 | 0.9 |
| Weibull (k=0.75) | 1.0 | 0.57 | 0.18 | 20 | 1.2 | 2.7 |
| | 0.5, 0.25 | 0.41 | 0.22 | 20 | 1.2 | 2.7 |
| Weibull (k=1.0) | 1.0 | 0.55 | 0.15 | 15 | 1.7 | 1.0 |
| | 0.5, 0.25 | 0.38 | 0.17 | 20 | 1.7 | 1.0 |
| Weibull (k=1.4) | 1.0 | 0.37 | 0.08 | 1000 | 2.3 | 0.5 |
| | 0.5, 0.25 | 0.46 | 0.09 | 20 | 2.3 | 0.5 |
| Weibull (k=2.0) | 1.0 | 0.30 | 0.36 | 80 | 3.2 | 0.35 |
| | 0.5, 0.25 | 0.56 | 0.20 | 100 | 3.2 | 0.35 |

The absolute magnitude of the standard error can be estimated as $\sigma(\hat{x}_R) = \overline{S_n}\,\sigma_x$. The columns of $\sigma$ (H) in Table 3 list the estimated standard errors of return wave heights for the extreme wave data listed in Table 1. By taking the one-sigma or two-sigma criterion, one can easily construct the confidence interval of the return wave heights.

## 9. PROBLEM OF MULTI-POPULATIONS

Any statistical data analysis must satisfy the condition of homogenuity: i.e., all the data under analysis should belong to the same population. Extreme waves are generated by hurricanes, monsoons, frontal systems, and others. Strictly speaking, these meteorological disturb-

ances create different populations of storm waves and they should be analyzed separately. Resio (1978), for example, demonstrated the necessity of separate analysis of extreme waves off Cleveland in Lake Erie. Carter and Challenor (1981) also discussed the effect of seasonal variations of wind speeds and wave heights on the estimation of their return values.

Separation of extreme waves according to their generating sources is feasible and the extreme data analysis can be made for each type of storm waves. For each population data, the best distribution function is fitted and the return wave height is estimated separately. To estimate the overall return value, the method by Carter and Challenor (1981) can be modified as follows. First, the distribution function for the partial-duration series data is converted to that corresponding to annual maximum series data by assuming the Poisson distribution. Then, the overall distribution function is evaluated as the product of all the individual distribution functions. Thus,

$$F(x) = \prod_{j=1}^{n} \exp\{-\lambda_j [1 - F_j(x)]\}$$

$$= \exp\{-\sum_{j=1}^{n} \lambda_j [1 - F_j(x)]\} \qquad (26)$$

where n denotes the number of populations, and $\lambda_j$ and $F_j$ are the mean rate and the best-fitted distribution function of the j-th population, respectively. The overall distribution function thus evaluated represents the nonexceedance probability of annual maximum wave height. The return wave height for the return period R is numerically evaluated as the height corresponding to the probability of $F = 1 - 1/R$.

## 10. CONCLUDING REMARKS

The present paper is essentially an extension of the work by Petruaskas and Aagaard (1970). The plotting position formulas have been examined and modified. The number of candidate distribution functions is reduced from eight to five, in recognition of low capability of the extreme data analysis in separating the true distribution from other candidates. The magnitude of errors owing to not-knowing the true distribution is estimated in terms of both the bias and the standard deviation. Empirical formulations are presented to estimate the amounts of bias and standard error of return wave heights.

It should be emphasized that any estimation of return value is accompanied by the statistical variability due to sampling error. In other words, an extreme wave data under analysis is but a sample taken from an unknown population of storm waves. Depending on the characteristics of a particular sample relative to those of population, the result of extreme data analysis might be an overestimate or an underestimate compared with the population value,

which remains unknown. A responsible engineer should take this uncertainty into account when he has to make selection of design waves. The formula of standard error presented as Eq. 24 would serve as a guide to measure the magnitude of uncertainty.

The uncertainty of return wave height can only be reduced through the increase of the time span of wave observation or wave hindcasting. Sampling of many storms as much as possible within a given time span is also helpful in reducing the amount of standard error of return wave height. In this sense, further continuous efforts of instrumental wave observations of longer duration should be encouraged.

## REFERENCES


Barnett, V. (1975): Probability plotting methods and order statistics, Applied Statistics, **24**(1), pp. 95-108.

Blom, G. (1958): Statistical Estimates and Transformed Beta-Variables, John Wiley & Sons, New York, Chap. 12.

Blom, G. (1962): Nearly best estimates of location and scale parameters, Contributions to Order Statistics (ed. by A.E. Sharman and B. G. Greenberg, John Wiley & Sons, New York, pp. 34-46.

Burcharth, H.F. (1988): Draft Report of Sub-group B "Uncertainty related to environmental data and estimated extreme events," PIANC PTC II Working Group 12 on Analysis of Rubble Mound Breakwaters.

Carter, D.J.T. and Challenor, P.G. (1981): Estimating return values of environmental parameters, Quart. J. R. Met. Soc., **107**, pp. 259-266.

Carter, D.J.T. and Challenor, P.G. (1983): Methods of fitting the Fisher-Tippett type I extreme value distribution, Ocean Engng., **10**(3), pp. 191-199.

Cunnane, C. (1978): Unbiased plotting positions - a review, J. Hydrology, **37**, pp. 205-222.

Draper, L. (1966): The analysis and presentation of wave data - a plea for uniformity, Proc. 10th ICCE, pp. 1-11.

Earle, M.D. and Baer, L. (1982): Effects of uncertainties on extreme wave heights, J. Wat., Port, Coast, and Ocn. Div., Proc. ASCE, **108**(WW4), pp. 456-478.

Goda, Y. (1979): A review on statistical interpretation of wave data, Rept. Port & Harb. Res. Inst., **18**(1), pp. 5-32.

Goda, Y. (1988): Numerical investigations on plotting formulas and confidence intervals of return values in extreme statistics, Rept. Port & Harb. Res. Inst., **27**(1), pp. 31-92. (*in Japanese.*)

Gringorten, I.I. (1963): A plotting rule for extreme probability paper, J. Geophys. Res., **68**(3), pp. 813-814.

Gumbel, E.J. (1958): Statistics of Extremes, Columbia Univ. Press, New York.

Lawless, J.F. (1974): Approximation to confidence intervals for parameters in the extreme value and Weibull distributions, Biometrika, **61**(1), pp. 123-129.

Medina, J.R. and Aguilar, J. (1986): Discussion to "Wave statistical uncertainities and design of breakwater" by B. Le Mehaute and S. Wang (Sept. 1985), J. Wat., Port, Coast, & Ocn. Div., ASCE, **112**, pp. 609-612.

Muir, L.R. and El-Shaarawi, A.H. (1986): On the calculation of extreme wave heights: a review, Ocean Engng., **13**(1), pp. 93-118.

Petruaskas, C. and Aagaard, P.M. (1970): Extrapolation of of historical storm data for estimating design wave heights, Prepr. 2nd Offshore Tech. Conf., OTC1190.

Resio, D.T. (1978): Some aspects of extreme wave prediction related to climatic variations, Prepr. 10th Offshore Tech. Conf., OTC3278.