

## CHAPTER 71

### Intercomparison of Parameter Estimation Methods in Extremal Wave Analysis

Masataka Yamaguchi<sup>1</sup>

#### Abstract

Based on statistical analysis of extreme wave height data generated with a Monte-Carlo simulation technique for the prescribed parent probability distributions, a preferable method for the parameter estimation was determined for each of 8 distributions. It is also verified that a jackknife method is applicable to the correction of bias and the estimation of variance irrespective of parameter estimation method in most parent distributions, and that the information matrix methods inherent to the maximum likelihood method give generally satisfactory results in the estimation of variance of return wave height for samples of size greater than around 50.

#### 1. Introduction

In the statistical analysis of extreme wave height data, several kinds of theoretical probability distributions and fitting methods for the parameter estimation have been employed, because the population distribution is not known a priori. Many attempts (for instance, Goda et al., 1993) have been made to find what kind of fitting method is preferable for the parameter estimation of each probability distribution to obtain a reliable estimate of return wave height and how the sampling variability could be evaluated, but the answer is still uncertain, because the class of parent distribution and the parameter condition investigated are limited.

This study uses 8 kinds of probability distributions including the Gumbel and Weibull distributions and 4

---

1 Prof. of Civil and Environmental Eng., Ehime Univ.  
Bunkyocho 3, Matsuyama 790, Ehime Pref., Japan

kinds of parameter estimation methods. Based on the statistical analysis with use of the 4 methods for data generated by a Monte-Carlo simulation technique, in which case the parent probability distribution is taken from one of the 8 distributions, the advantage of a parameter estimation method over the other methods is investigated from the view points of bias and variance of return wave height. Also, applicability of a jackknife method to the correction of bias and the estimation of variance of return wave height, and that of information matrix methods usable in the maximum likelihood method to the estimation of variance of return wave height are discussed.

## 2. Parent Distributions and Estimation Methods of Parameter and Variance

### 2. 1 Parent distributions

The probability distributions investigated are the Gumbel, Weibull, GEV, Lognormal, Gamma, Loggamma, Hypergamma (Generalized Gamma) and Poisson-square root exponential-type maximum (SQRT) distributions. These distributions except for the Gumbel and SQRT distributions have three parameters respectively. Each probability distribution  $F(x)$  is written as follows.

(a) Gumbel distribution (Greenwood et al., 1979; Goda, 1988)

$$F(x) = \exp[-\exp\{-(x-B)/A\}] ; -\infty < x < \infty \quad (1)$$

where  $x$  is the random variable,  $A$  the scale parameter and  $B$  the location parameter.

(b) Weibull distribution (Greenwood et al., 1979; Goda, 1988)

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

where  $k$  is the shape parameter.

(c) GEV distribution (Fisher-Tippett type II (FT-II) distribution for  $k > 0$ ) (Hosking et al., 1985; Phien and Emma, 1989; Goda, 1990)

$$F(x) = \exp[-\{1 + (x-B)/kA\}^{-k}] ; \begin{array}{l} B - kA < x < \infty, k > 0 \\ -\infty < x < B - kA, k < 0 \end{array} \quad (3)$$

(d) 3-parameter Lognormal distribution (Takeuchi and Tsuchiya, 1988)

$$F(x) = (1/\sqrt{\pi}) \int_{-\infty}^y \exp(-y^2) dy$$

$$\begin{array}{l} y = k \cdot \log\{(x-B)/A\} ; B < x < \infty, C_S > 0 \\ y = k \cdot \log\{A/(B-x)\} ; -\infty < x < B, C_S < 0 \end{array} \quad (4)$$

where  $C_s$  is the skewness coefficient.

(e) 3-parameter Gamma distribution (Bobée, 1975; Takeuchi and Tsuchiya, 1988)

$$\begin{aligned} F(x) &= \gamma\{k, (x-B)/A\} / \Gamma(k) & ; B < x < \infty, A > 0 \\ F(x) &= 1 - \gamma\{k, (x-B)/A\} / \Gamma(k) & ; -\infty < x < B, A < 0 \end{aligned} \quad (5)$$

where  $\Gamma(k)$  is the gamma function and  $\gamma(k, x)$  the incomplete gamma function of the first kind defined by

$$\gamma(k, x) = \int_0^x \exp(-t) t^{k-1} dt \quad (6)$$

(f) 3-parameter Loggamma distribution (Condie, 1977)

$$\begin{aligned} F(x) &= \gamma\{k, (\log x - B)/A\} / \Gamma(k) & ; B < \log x < \infty, A > 0 \\ F(x) &= 1 - \gamma\{k, (\log x - B)/A\} / \Gamma(k) & ; -\infty < \log x < B, A < 0 \end{aligned} \quad (7)$$

(g) 3-parameter Hypergamma distribution (Suzuki, 1964)

$$\begin{aligned} F(x) &= \gamma(k, t) / \Gamma(k), \quad t = (x/A)^C & ; 0 < x < \infty, C > 0 \\ F(x) &= 1 - \gamma(k, t) / \Gamma(k) & ; 0 < x < \infty, C < 0 \end{aligned} \quad (8)$$

(h) SQRT distribution (Etoh et al., 1986)

$$F(x) = \exp\{-k(1 + \sqrt{x/A}) \exp(-\sqrt{x/A})\} ; 0 < x < \infty \quad (9)$$

This is one of the compound distributions, and  $k$  signifies yearly-averaged occurrence rate of event rather than shape property of the distribution.

## 2. 2 Parameter estimation methods

The parameter estimation methods used in this study are the moment method (MOM), the probability weighted moment (PWM) method and the maximum likelihood method (MLM) and the least square method (Goda, 1988, 1990) (LSM). Sample mean, unbiased variance and skewness are used in the moment method. In the parameter estimation with the moment method for the Loggamma distribution, two methods based on mean, unbiased variance and skewness of log-transformed sample data (MOM1) and cumulants of sample data (MOM2) are applied. PWM solutions are not derived in the cases of Lognormal distribution for negative skewness, Loggamma distribution, Hypergamma distribution and SQRT distribution. The parameter estimation for SQRT distribution is only derived from the maximum likelihood method.

The least square method is based on the model by Goda (1988, 1990). A set of candidate distributions is the Gumbel and Weibull distribution whose shape parameter is either of 0.75, 1.0, 1.4 or 2.0. The other set consists of the Gumbel and FT-11 type distribution whose

shape parameter is either of 2.5, 3.33, 5.0 or 10.0. A distribution with the largest correlation coefficient between the ordered data of sample and its reduced variate is selected as the best fitting distribution.

### 2. 3 Index of goodness of fit

The SLSC (Takasao et al., 1986) is used as an index of goodness of fit. It is defined by

$$SLSC = \left\{ \sum_{i=1}^N (x_i - s_i)^2 / N \right\}^{1/2} / |s_{0.99} - s_{0.01}| \quad (10)$$

where  $N$  is the sample size,  $x_i$  the ordered data,  $s_i$  the variate which is calculated from a theoretical probability distribution for designated probability such as 0.99 or 0.01. The Weibull plotting position formula is used as a standard formula to estimate non-exceedance probability  $F(x)$  of sample data, but in the least square method, the distribution-dependent plotting position formula is applied. The least square method eventually gives smaller SLSC than the other methods owing to its definition.

### 2. 4 Methods of bias correction and variance estimation

A jackknife method (Miller, 1974) is applied for bias correction and variance estimation of return wave height estimated using either method of MOM, PWM or MLM. The formulas are given by

$$H_J = N\hat{H} - (N-1)\hat{H}_*, \quad \hat{H}_* = \sum_{i=1}^N H_{*i} / N, \quad \sigma_J^2 = (N-1) \sum_{i=1}^N (H_{*i} - \hat{H}_*)^2 / N \quad (11)$$

where  $\hat{H}$  is the estimate of return wave height,  $H_{*i}$  the estimate of return wave height based on  $N-1$  data excluding  $x_i$ ,  $H_J$  the bias-corrected estimate of return wave height and  $\sigma_J^2$  the jackknife estimate of variance indicated by  $\sigma_{JM}^2$  for MOM,  $\sigma_{JP}^2$  for PWM and  $\sigma_{JY}^2$  for MLM.

In the application of the maximum likelihood method for the parameter estimation of the probability distributions except for SQRT distribution, the methods based on variance-covariance matrix of the maximum likelihood estimator (Suzuki, 1964; Phien and Emma, 1989) can be used for the asymptotic evaluation of variance of return wave height. It is defined as

$$A_{ij} = -E \left[ \frac{\partial^2 L(X; \theta_1, \dots, \theta_r)}{\partial \theta_i \partial \theta_j} \right] ; \quad i, j = 1, 2, \dots, r \quad (12)$$

where  $E$  means the expected value operator,  $L$  the log-transformed maximum likelihood and  $\theta_i$  the parameter of a probability distribution.  $A_{ij}$  is called the Fisher

information matrix, and if the expected value operator is dropped in eq. (12), it is called the observed information matrix. The variances estimated with both methods are indicated by  $\sigma_{FM}^2$  and  $\sigma_{OM}^2$  respectively.

In the case of the least square method, the standard deviation of return wave height  $\sigma_{LSM}$  is estimated with the empirical formula derived from numerical experiments by Goda (1988, 1990).

### 3. Monte-Carlo Simulation

As the inverse forms of the Gumbel, Weibull, GEV and lognormal distributions are known analytically, a sample of extreme wave height data is simulated sequentially by giving uniformly-distributed numbers between 0 and 1 generated by computer as input. In the cases of the other distributions such as the Gamma distribution, a sample of wave height is made with use of a numerical table given as the relation between equally-divided non-exceedance probability  $F(x)$  and random variable  $x$ . The number of samples is 5,000 and sample size in each sampling  $N$  ranges from 10 to 1000, i.e. 10, 20, 30, 40, 50, 70, 100, 200, 500 and 1000. Value of SLSC, the parameters and the resulting 5 return wave heights  $H(n)$  from 50 to 1000 years ( $n=50, 100, 200, 500$  and 1000) and their variances are estimated with the above-mentioned methods from each sample. By averaging the results of 5000 simulations for each data size, mean values for SLSC, return wave height  $H(n)$ , jackknife-corrected return wave height  $H_J(n)$ , variances ( $\sigma_{JM}^2, \sigma_{JP}^2, \sigma_{JY}^2, \sigma_{FM}^2, \sigma_{OM}^2$ ) and standard deviation  $\sigma_{LSM}$ , and variance of 5000 return wave height data  $Var(n)$  are obtained. Then two kinds of bias  $\Delta H(n)$  and  $\Delta H_J(n)$  are respectively as

$$\Delta H(n) = \overline{H(n)} - H_{tr}(n), \quad \Delta H_J(n) = \overline{H_J(n)} - H_{tr}(n) \quad (13)$$

where  $H_{tr}(n)$  is the true return wave height corresponding to  $n$  years,  $\Delta H_J(n)$  the residual bias after jackknife correction, and ' $\bar{\phantom{x}}$ ' means the average value. These quantities are called error statistics. Error statistics  $\Delta H(n)$ ,  $\Delta H_J(n)$ ,  $Var^{1/2}(n)$  are normalized with use of the true return wave height  $H_{tr}(n)$ , and square root mean variances and mean standard deviation are divided by  $Var^{1/2}(n)$ . The normalized error statistics are expressed with the notation ' $\sim$ ' and the figures are shown in the case of return period of 100 years.

### 4. Consideration of Results

A set of parameters is given to every parent dis-

tribution as input condition in the simulation study to find the advantage of one parameter estimation method over the other methods. Four shape parameters with the other parameters fixed are used in the simulation study to investigate the effect of shape of the distribution on the bias and variance estimated with the optimum parameter estimation method for each parent distribution.

Fig. 1 shows the relation between normalized error statistics and sample size  $N$  in the case of the Gumbel distribution, in which simulation is conducted under the condition of  $A=1.39$  m and  $B=4.5$  m, and the bias by the LSM given in both figures of  $\Delta H$  and  $\Delta H_j$  is the same one. It can be seen that bias of each, especially bias after jackknife correction by any methods is small and that the jackknife method and the information matrix methods give proper estimates of variance. Although the MLM with the jackknife correction is the optimum method for samples of size greater than about 30 from view-points of bias and variance, the PWM method is more proper from general view-points, when goodness of fit is taken into account. The LSM naturally produces the smallest SLSC, but gives greater bias and variance than the other methods. Also, the LSM yields poor estimate of standard deviation. This may be due to the fact that the empirical formula for the estimation of standard deviation is

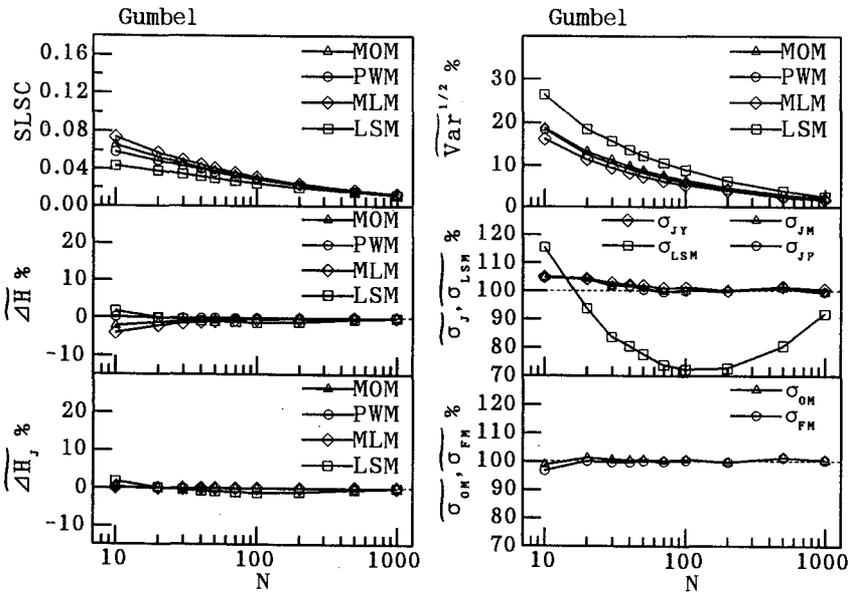


Fig. 1 Relation between error statistics and sample size (Gumbel distribution).

derived on the basis of numerical simulation for a fixed shape parameter, without taking a procedure of selecting the best fitting distribution. Similar tendencies are observed for the Weibull and FT-II type distributions.

Fig. 2 describes the relation between error statistics and sample size for the Weibull distribution. Simulation corresponding to the upper figures is conducted for the condition of  $k=1.8$ ,  $A=4.0$  m and  $B=1.0$  m to find a preferable parameter estimation method, and simulations corresponding to the lower figures are made by giving either of  $k=0.75, 1.0, 1.4$  or  $2.0$  under the fixed values of

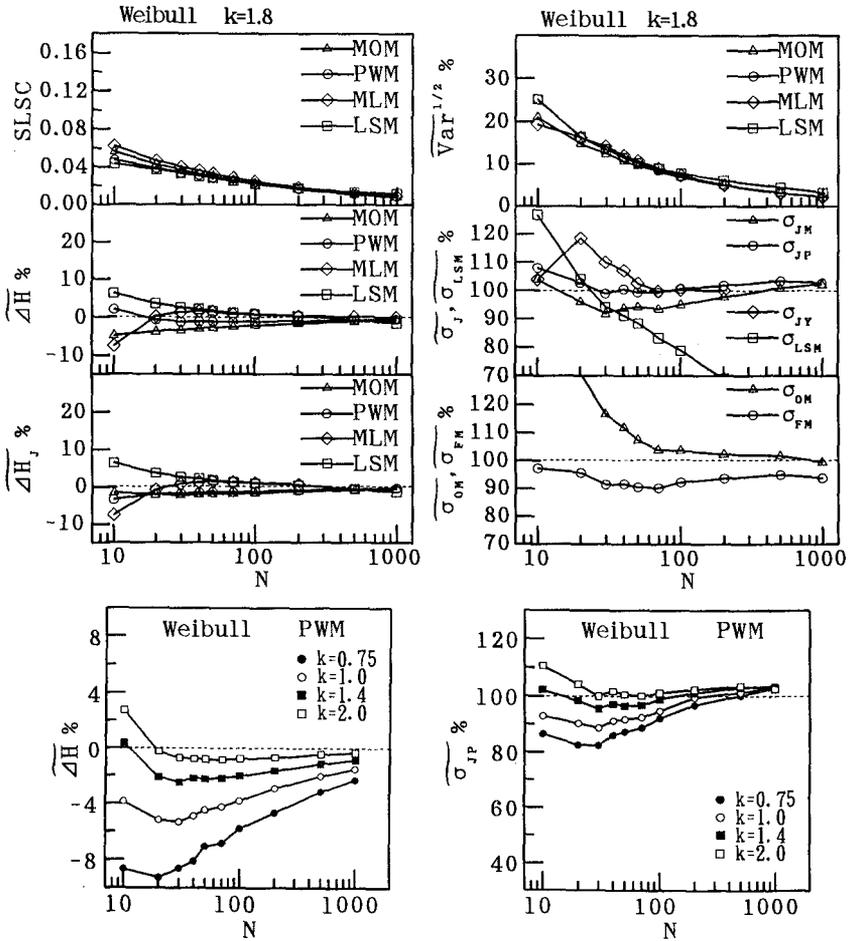


Fig. 2 Relation between error statistics and sample size (Weibull distribution).

$A=4.0$  m and  $B=1.0$  m to investigate the effect of shape parameter on the error statistics. The PWM method is seen to be the optimum method from a view point of bias, although it yields a slightly larger estimate of variance than the MOM. In the PWM method, the jackknife method does not always work effectively for the bias correction, but it gives close estimate of variance. The MLM is a recommendable method in the case of sample size greater than 50 or 70. It is seen that the use of the observed information matrix method (OIMM) to the estimation of variance is possible for sample data greater than 30, if overestimation less than 10 % is allowed and that the Fisher information matrix method (FlMM) is applicable with underestimation less than 10 %. The OIMM usually gives greater estimate of variance than the FlMM. The effects of shape parameter on bias and estimate of variance are not negligible. Negative bias and degree of underestimation of variance increase with decrease of shape parameter. These reflect the widening of the Weibull distribution with decrease of shape parameter. Therefore, the application of the PWM method is preferably restricted for the case of shape parameter less than 1.0 to properly estimate return wave height and its variance.

Results for the GEV distribution are shown in Fig. 3. Conditions of simulation for finding a preferable parameter estimation method and for investigating the effect of shape parameter are  $k=5.0$ ,  $A=1.0$  m,  $B=4.0$  m, and either value of  $k=2.5$ , 3.33, 5.0 or 10.0 for the fixed values of  $A=1.0$  m and  $B=4.0$  m respectively. The PWM method with the jackknife method produces excellent estimates of return wave height and its variance. Small bias is also brought about by the LSM which uses the adjusted plotting position formula, but the accuracy of estimation of variance is not so high for the reason mentioned above. The MLM with the jackknife correction gives small bias, but it does not yield good results on variance for small sample size. The information matrix methods are applicable for sample of size greater than about 50 or 70. According to the results of the lower figures, bias based on the PWM method is small except for  $k=2.5$ , distribution of which is widest in the investigated distributions, and the jackknife method gives fairly proper estimate of variance.

Fig. 4 illustrates the results for the Lognormal distribution. Parameter conditions in the simulation are taken as  $k=3.4$ ,  $A=8.0$  m,  $B=-2.9$  m and  $k=1.5$ , 2.0, 2.5, 3.0,  $A=3.5$  m,  $B=-2.5$  m for each purpose mentioned above. The MOM is a more preferable method than the other methods from view points of bias and variance, and the jackknife method gives proper correction to bias and good

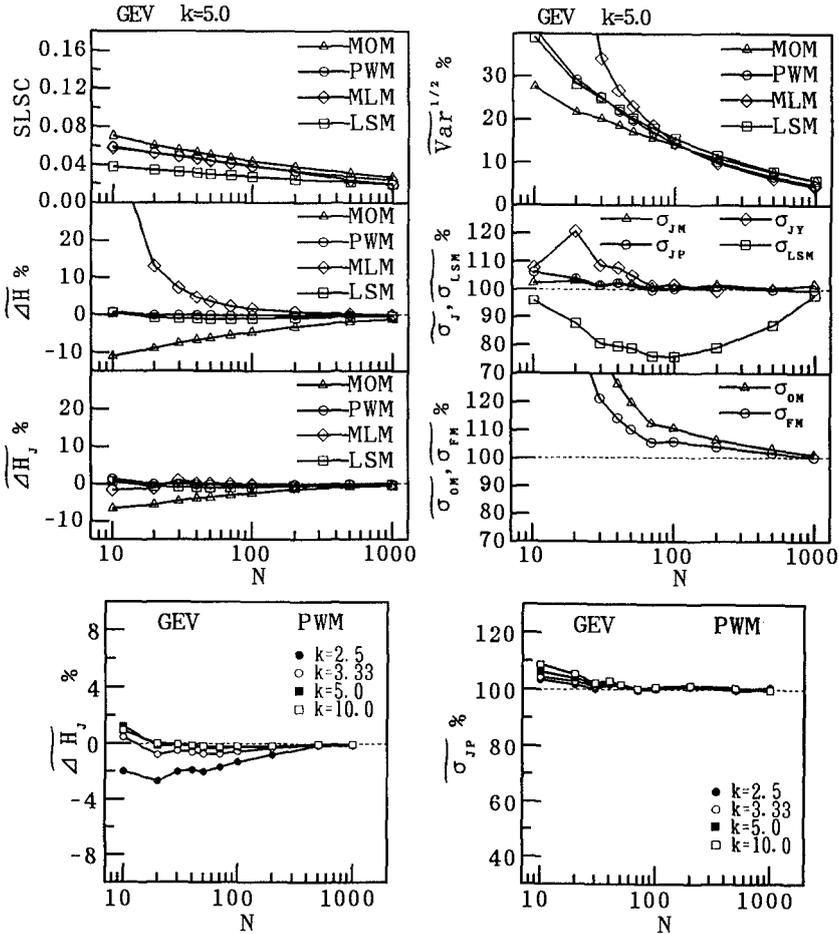


Fig. 3 Relation between error statistics and sample size (GEV distribution).

estimate of variance. Also, the MLM is a preferable method, in particular, for sample of size greater than 50 or 70. The influence of shape parameter on error statistics is seen in a diagram of bias. Even if bias-correction with the jackknife method is made, negative bias for small shape parameters is still at significant level. On the other hand, the jackknife method yields proper estimate of variance irrespective of the value of shape parameter.

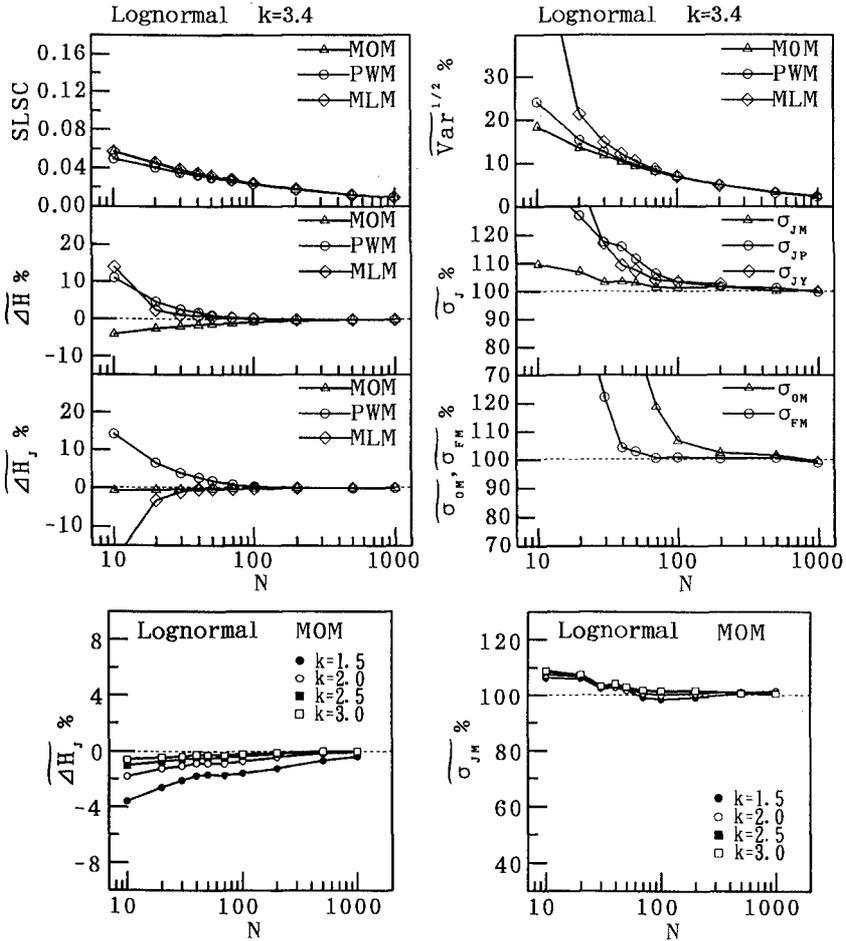


Fig. 4 Relation between error statistics and sample size (Lognormal distribution).

It can be said from similar investigation that the optimum parameter estimation method is the PWM method for the Gamma distribution, the MOM2 for the Loggamma distribution and the MOM for the Hypergamma distributions. Fig. 5 illustrates the effect of shape parameter on bias and estimate of variance in the cases of Gamma ( $A=1.0$  m,  $B=4.5$  m), Loggamma ( $A=-0.1$  m,  $B=3.5$  m) and Hypergamma ( $A=0.447$  m,  $C=2.0$ ) distributions. The shape parameters used are indicated in the corresponding figure. In the case of the Gamma distribution, even the PWM method which was judged to be the optimum method of the three methods pro-

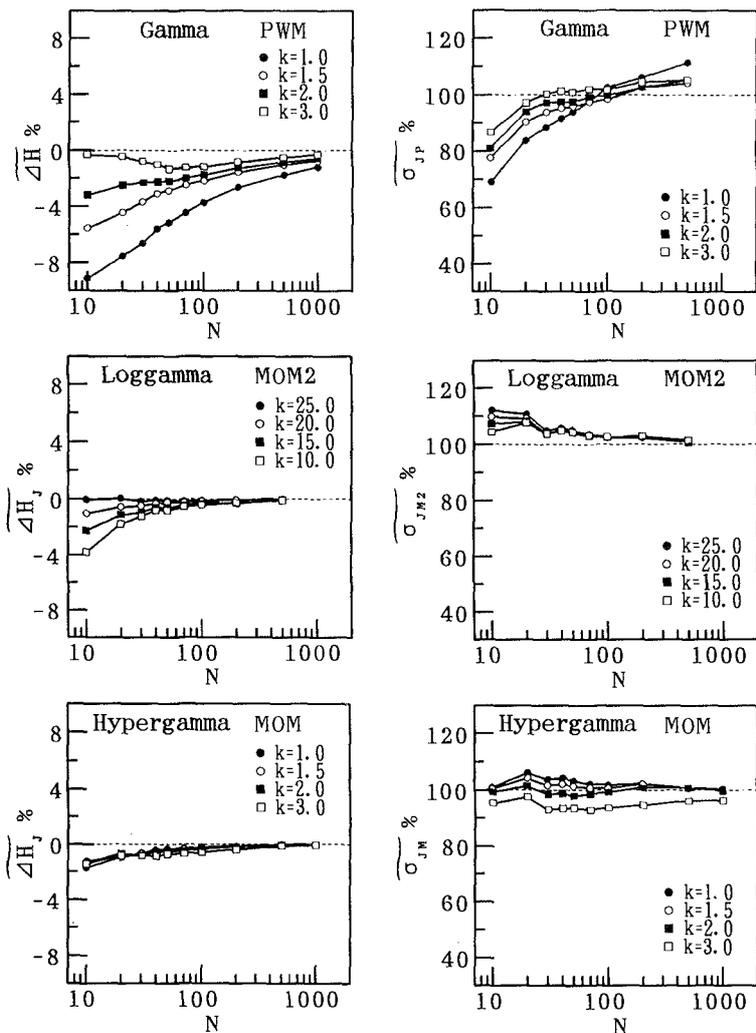


Fig. 5 Effect of shape parameter on error statistics (Gamma, Loggamma and Hypergamma distributions).

duces negative bias increasing with decrease of shape parameter, and the jackknife method does not work efficiently on the bias correction and the variance estimation. In the cases of the Loggamma and Hypergamma distribution, the jackknife method is effective for the correction to bias and the estimation of variance, although deviation from the true value slightly increases for wider distribution.

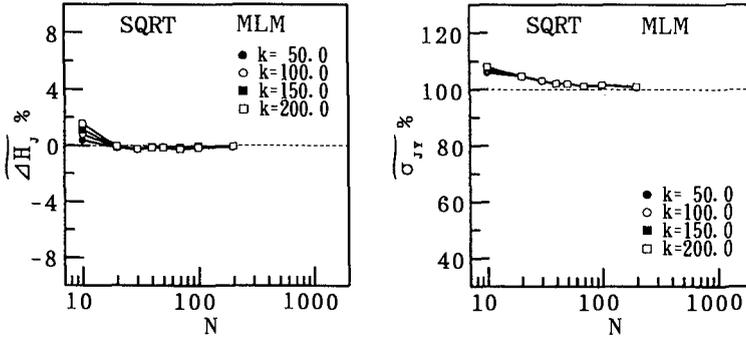


Fig. 6 Effect of shape parameter on error statistics (SQRT distribution).

Finally, Fig. 6 shows the effect of shape parameter on bias and estimate of variance in the case of the SQRT distribution. The scale parameter  $\lambda$  is fixed as  $1/13$  m and the shape parameter  $k$  is either of 50, 100, 150 or 200. At present, the only method applicable to the parameter estimation is the MLM. As mentioned above, the parameter  $k$  means yearly-averaged occurrence rate of event. Change of shape parameter  $k$  in this case brings shift of a peak position of the distribution rather than variation of shape of the distribution. In the usage of the MLM, the jackknife method yields excellent correction to bias and proper estimate of variance for the given cases.

### 5. Conclusions

Conclusions in this study are summarized as follows.

(1) A jackknife method is applicable to the correction of bias and the estimation of variance irrespective of parameter estimation methods in most parent probability distributions.

(2) A preferable method to the parameter estimation in each distribution is determined as

- Gumbel distribution : PWM with jackknife correction
- Weibull distribution : PWM without jackknife correction
- GEV distribution : PWM with jackknife correction
- Lognormal distribution : MOM with jackknife correction
- Gamma distribution : PWM without jackknife correction
- Loggamma distribution : MOM2 with jackknife correction
- Hypergamma distribution: MOM with jackknife correction
- SQRT distribution : MLM with jackknife correction

- (3) The LSM proposed by Goda is a preferable method from the view points of bias and goodness of fit.
- (4) The MLM with the jackknife method usually gives satisfactory estimates of return wave height and its variance for samples of size greater than about 50.
- (5) The information matrix methods are effective as variance estimators in the MLM for samples of size greater than about 50.

#### References

- Bobée, B.(1975): The log Pearson type 3 distribution and its application in hydrology, Water Resour. Res., Vol. 11, No. 5, pp.681-689.
- Condie R.(1977): The log Pearson type 3 distribution: the T-year event and its asymptotic standard error by maximum likelihood theory, Water Resour. Res., Vol. 13, No. 6, pp.987-991.
- Etoh, T. et al.(1986): Frequency of record-breaking large precipitation, Proc. JSCE, No. 369/11-5, pp.165-174 (in Japanese).
- Goda, Y.(1988): Numerical investigations on plotting formulas and confidence intervals of return values in extreme statistics, Rept. of the Port and Harb. Res. Inst., Vol. 27, No. 1, pp.31-91 (in Japanese).
- Goda, Y. and M. Onozawa(1990): Characteristics of the Fisher-Tippett type 11 distribution and their confidence intervals, Proc. JSCE, No. 417/11-13(note), pp.289-292 (in Japanese).
- Goda, Y. et al.(1993): Intercomparison of extremal wave analysis methods using numerically simulated data: a comparative analysis, Proc. WAVES'93 Conf., pp.963-977.
- Greenwood, J. A. et al.(1979): Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, Water Resour. Res., Vol. 15, No. 5, pp.1049-1054.
- Hosking, J. R. M. et al. (1985): Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, Technometrics, Vol.27, No. 3, pp.251-261.
- Miller, R. G.(1974): The jackknife - review, Biometrika, Vol. 61, No. 1, pp.1-15.

Phien, H. N. and F. T. Emma(1989): Maximum likelihood estimation on the parameters and quantiles of the general extreme-value distribution from censored sample, Jour. Hydrol., Vol.105, pp.139-155.

Suzuki, E.(1964): Hypergamma distribution and its fitting to rainfall data, Papers in Meteorol. and Geophy., Vol. 15, pp.31-51.

Takasao, T. et al.(1986): A basic study on frequency analysis of hydrologic data in the Lake Biwa basin, Annual Rept. of Disas. Prev. Res. Inst., Kyoto Univ., No. 29 B-2, pp.157-171 (in Japanese).

Takeuchi, K. and K. Tsuchiya(1988): PWM solutions to normal, lognormal and Pearson-III distributions, Proc. of JSCE, Vol. 393/II-9, pp.95-112 (in Japanese).