Journal of Hydroscience and Hydraulic Engineering
Vol. 18, No.1 May, 2000, 45-52

EFFECT OF INTRODUCING UNCERTAIN HISTORICAL HYDROLOGIC DATA ON QUANTILE ESTIMATION ACCURACY

by

Kenjiro Sho

Research Associate, Department of Systems Management and Engineering,
Nagoya Institute of Technology, Nagoya, Japan

Seiichiro Iwasaki

Fujita Corporation, Shibuya, Tokyo, Japan

Masashi Nagao

Professor, Faculty of Urban Science, Meijo University, Kani, Gifu, Japan

and

Akihiro Tominaga

Professor, Department of Architecture and Civil Engineering,
Nagoya Institute of Technology, Nagoya, Japan

SYNOPSIS

When the length of systematic gauged hydrological record is not enough, historical flood information is practically useful to improve the accuracy of flood-quantile estimation. Historical flood data, which can be utilized for flood frequency analysis by the adjusted-moment method or the maximum likelihood method, often have much larger errors than systematic gauged data. Thus, it is important to evaluate the effect of the error on the precision of flood-quantile estimators. Monte Carlo simulations using the Gumbel distribution show that flood-quantile estimates contain positive bias when the standard error of historical data exceeds a certain level (in this study, approximately 1/6 of the average for the systematic gauged data). Moreover, when it exceeds 1/4 to 1/3 of the average, improvement of the accuracy cannot be confirmed for calculated flood-quantile values.

INTRODUCTION

In the planning and design of flood-control projects, the design flood corresponding to a T-year return period of is determined from hydrological data. When the systematic gauged record is not long enough to estimate a T-year flood accurately, historical flood information is practically useful. Methods of utilizing historical flood data, which contain both quantitative values and categorical information (e.g., above or below a certain threshold), have been investigated for flood frequency analysis. Furthermore, the effects of increasing record lengths on improvement of the accuracy of flood-quantile estimation have been evaluated by computer simulations (e.g., Cohn and Stedinger (1); Hosking and Wallis (2); Ikebuchi and Maeda (3); Stedinger and Cohn (4)). However, historical data often contain much larger errors than systematic gauged data because those are estimated from indirect evidences, such as sediment deposits, botanical evidences or subjective written records. When the errors of additional historical data are too large, the reliability of obtained flood-quantile estimates is likely to decrease than at the case when only systematic data are used.

Thus, it is necessary to simulate the relation between both factors. This paper examines the effect of the errors of historical data on the accuracy of flood-quantile estimation by incorporating the process of error generation in the Monte Carlo procedure, and investigates the acceptable size of errors in flood frequency analysis for various estimators.

METHODS OF UTILIZING HISTORICAL FLOOD INFORMATION FOR FLOOD FREQUENCY ANALYSIS

In historical years, the larger floods tend to leave more information about their magnitudes and dates. We define "a historical period" as the period such that all floods greater than a threshold left a record that is currently available, and "a systematic period" as the period such that all quantitative gauged data are available for that period. Historical flood data can be categorized into three classes: "censored" data, where the magnitudes of historical flood peaks are known, "binomial" data, where only threshold exceedance information is available and "range" data, where the magnitudes of historical flood peaks have errors within a predetermined range. In this section we describe methods of estimating parameters of probability distributions using systematic and historical flood data.

In this study, the Gumbel distribution is used as the probability distribution of flood peaks. Its distribution function and probability density function are respectively,

$$F(x) = \exp\left[-\exp\left\{-\alpha(x-\mu)\right\}\right]$$

$$f(x) = \alpha \cdot \exp\left[-\alpha(x-\mu) - \exp\left\{-\alpha(x-\mu)\right\}\right]$$

with the mean value m and the standard deviation σ. The parameters μ and α are denoted as

$$\mu = m - 0.455005\,\sigma \quad ; \quad \alpha = 1.28255/\sigma \tag{1}$$

Parameters μ and α can be estimated by the method of moments or the maximum likelihood method.

Adjusted-Moment Estimator

Let x_i (i=1, 2, ···, s) be annual flood peaks in a s year systematic period, y_i (i=1, 2, ···, k) be flood peaks that exceeded a perception threshold U in a h year historical period and k be the number of threshold exceedance data of the historical period. For remaining (h–k) years, the magnitudes of flood peaks are unknown except for the fact that those are below the threshold U. Then we assume that the mean and variance of historical data below the threshold U are the same as those of systematic data below U, respectively. When the number of systematic data below U is n_{xy} and those values are x_{yi} (i=1, 2, ···, n_{xy}), the mean value M_h and the variance σ_h^2 of total samples for the historical period are

$$M_h = \frac{1}{h}\left[(h-k)\frac{1}{n_{xy}}\sum_{i=1}^{n_{xy}} x_{yi} + \sum_{i=1}^{k} y_i\right]$$

$$\sigma_h^2 = \frac{1}{h}\left[(h-k)\frac{1}{n_{xy}}\sum_{i=1}^{n_{xy}} (x_{yi}-M)^2 + \sum_{i=1}^{k} (y_i-M)^2\right]$$

where M is the mean value for the total period,

$$M = (sM_s + hM_h)/(s+h) \tag{2}$$

The variance for the total period σ^2 is denoted as

$$\sigma^2 = (s\sigma_s^2 + h\sigma_h^2)/(s+h) \tag{3}$$

where M_s and σ_s^2 are the mean and variance for the systematic period,

$$M_s = \frac{1}{s}\sum_{i=1}^{s} x_i \quad ; \quad \sigma_s^2 = \frac{1}{s}\sum_{i=1}^{s}(x_i - M)^2$$

Thus, the values of μ and α can be obtained by substituting Eqs. 2 and 3 into Eq. 1.

Maximum Likelihood Estimator

For the probability density function $f(x;\theta)$, the likelihood function L is defined as

$$L = \prod_{i=1}^{n} f(x_i;\theta)$$

where θ is a vector of unknown parameters and n is the number of data. Given sample data x_i (i=1, 2, ···, n), L is the function of θ, and the maximum likelihood estimate $\hat{\theta}$ can be calculated by partially differentiating in θ, equating to zero and solving for θ. Since the Gumbel distribution has two parameters μ and α, two equations are obtained;

$$\frac{\partial}{\partial\mu}L(\mu,\alpha) = 0 \quad ; \quad \frac{\partial}{\partial\alpha}L(\mu,\alpha) = 0$$

Considering the nature of historical data, there can be four types of likelihood function as follows.

a) Censored Data

If sample data z_i (i=1, 2, ···, s) are known as quantitative values, the likelihood function L_1 is given by

$$L_1 = \prod_{i=1}^{s} f(z_i;\mu,\alpha)$$

All of systematic data and the historical data estimated as quantitative values are categorized into this case.

b) Binomial Data (below threshold)

In the historical period, if there are no records of flood during a specified year, the flood magnitudes are treated as less than a censoring threshold. Let n_j (j=1, 2, ···, k) be the number of data below a threshold U_j, and the likelihood function L_2 is

$$L_2 = \prod_{j=1}^{k} F(U_j;\mu,\alpha)^{n_j}$$

c) Binomial Data (above threshold)

For m_j (j=1, 2, ···, l), the number of historical data whose magnitudes cannot be estimated but are known to have exceeded a threshold T_j, the likelihood function L_3 is given by

$$L_3 = \prod_{j=1}^{l}\left[1 - F(T_j;\mu,\alpha)\right]^{m_j}$$

d) Range Data

Occasionally, one must treat historical data as having errors within a predetermined range because of their unreliability. If there are n historical data every one of whose value is known to range between Y_i and Z_i (i=1, 2, ···, m), then the likelihood function L_4 is

$$L_4 = \prod_{i=1}^{m} [F(Z_i;\mu,\alpha) - F(Y_i;\mu,\alpha)]$$

Since every systematic or historical datum can be categorized into one of the above, the likelihood function for all data is denoted as the combination of these four cases, say,

$$L = L_1 L_2 L_3 L_4$$

The maximum likelihood method is effective in being appropriate for any kind or combination of data and easy to alter thresholds.

MONTE CARLO EXPERIMENT

In this section we examine various methods of estimating flood-quantile values using systematic and historical data by Monte Carlo experiments.

Estimation Models

a) Maximum Likelihood Method Case 1 (MLE1)

In this case, in addition to systematic data x_{0i} (i=1, 2, ···, k_0), the historical data x_{1i} (i=1, 2, ···, k_1), which happened to exceed a threshold U during the h_1 year historical period, are available as quantitative values. The remaining (h_1–k_1) historical data are assumed to be below the threshold. The likelihood function is given by

$$L = \prod_{i=1}^{k_0} f(x_{0i}) \cdot \prod_{i=1}^{k_1} f(x_{1i}) \cdot F(U)^{h_1 - k_1}$$

b) Maximum Likelihood Method Case 2 (MLE2)

In this case, in addition to systematic data x_{0i} (i=1, 2, ···, k_0), each of the historical data x_{1i} (i=1, 2, ···, k_1), which exceeded a threshold U during the h_1 year historical period, can be estimated with a $\pm R$ error margin. The likelihood function is

$$L = \prod_{i=1}^{k_0} f(x_{0i}) \cdot \prod_{i=1}^{k_1} [F(x_{1i} + R) - F(x_{1i} - R)] \cdot F(U)^{h_1 - k_1}$$

c) Maximum Likelihood Method Case 3 (MLE3)

In this case, in addition to systematic data x_{0i} (i=1, 2, ···, k_0), only the fact that a threshold U was exceeded k_1 times in historical h_1 years is known. The likelihood function is

$$L = \prod_{i=1}^{k_0} f(x_{0i}) \cdot F(U)^{h_1 - k_1} \cdot [1 - F(U)]^{k_1}$$

d) Adjusted-Moment Method (MOM)

For this case, the procedure as described in the previous section is applied.

Simulation Procedure

For the above four cases, we examine the effect of historical flood data containing errors on the precision of flood-quantile estimators by the Monte Carlo experiment. Estimators are evaluated using root mean square errors (RMSE) of the 100-year flood. The procedure is described as follows:

Step 1 : Assume a population distribution

We use the Gumbel distribution as the population distribution, with the parameters μ= 300 and α= 0.01. These values have been derived by considering the actual distribution of annual-maximum 30-day precipitation for Lake Biwa area.

Step 2 : Take a sample of size (s+h) from the population

Generate (s+h) random numbers of the distribution given in Step1, where s and h are the length of systematic and historical periods, respectively. In this study, we assume s = 80 and vary h from 0 to 200, in order to examine the performance of estimators by increasing historical information in addition to systematic gauged data.

Step 3 : Generate errors of historical samples

Generate h random numbers of a normal distribution $N(0, \varepsilon^2)$, and add them to each of historical samples generated in Step2. The errors of systematic data are assumed to be negligibly small. In this study we examine various cases of $\varepsilon = 0 \sim 120$.

Step 4 : Fit a flood frequency distribution and calculate quantile estimates

Apply each of the four methods a) ~ d) to the samples obtained in Step3 and estimate the parameters μ, α and T-year floods. Here we calculate under the condition of the return period T = 100 and the censoring threshold of the historical period U = 500 (in this case, the ratio of threshold exceedace data is approximately 1/8 of total samples). For the presumed range of errors R in the MLE2 model, we examine two cases of R = 20, 60 and express them by "MLE2-1" and "MLE2-2", respectively.

Step 5 : Repeat Steps 1~4 many times, and calculate the mean value and RMSE of estimates of parameters and T-year quantiles. In this study, we repeat Steps 1~4 1,000 times.

RESULTS AND CONCLUSIONS

Figure 1 shows the effect of increase of historical information on the mean and RMSE of estimated 100-year floods, for various standard errors (ε) of historical data. Moreover, Fig. 2 shows the same information as in Fig. 1 and compares four estimators, for ε = 0, 20, 60 and 120. The following conclusion can be drawn from Fig. 1 and Fig. 2.

1. When ε is less than 60 (approximately 1/6 of the mean value of the population), the accuracy of estimated 100-year floods is improved with the increase of historical information. Especially when ε is less than 40 (approximately 1/9 of the mean value of the population), it shows almost the same performance as the case of ε = 0, for all estimators (see Fig. 1).
2. When ε is 60 or more, the estimates tend to have positive bias with the increase of historical data without a considerable reduction in RMSE. Thus, for these cases, the historical period should not be too long (in this study, approximately less than 1/2~2/3 of the length of systematic period).
3. From the cases of ε = 20 and 60 in Fig. 2, it can be seen that the cases of MLE2 having the presumed range of errors R corresponding to ε show smaller bias and RMSE than the other estimators. Therefore, if the errors of historical data can be estimated approximately, in that case, the best result can be obtained by using MLE2, such that the value of R equal to the estimated errors.
4. When ε exceeds 100 (approximately 1/4~1/3 of the mean value of the population), RMSE increases with the number of historical data. In these cases, the method of maximum likelihood using only systematic data is most appropriate.
5. The estimation accuracy of MOM (the method of moments) is clearly lower than that of MLE (maximum likelihood method) estimators. And in the case of MLE3, the RMSEs are larger than those of other MLE cases.

In this study, the simulation was performed under the simple conditions that the threshold and accuracy of data are constant during the historical period and the errors of the historical data have the normal distribution with a mean value of 0. Further analyses on error distribution of the historical data, temporal variation of accuracy or censoring threshold, and their nonstationarity remain to be addressed.

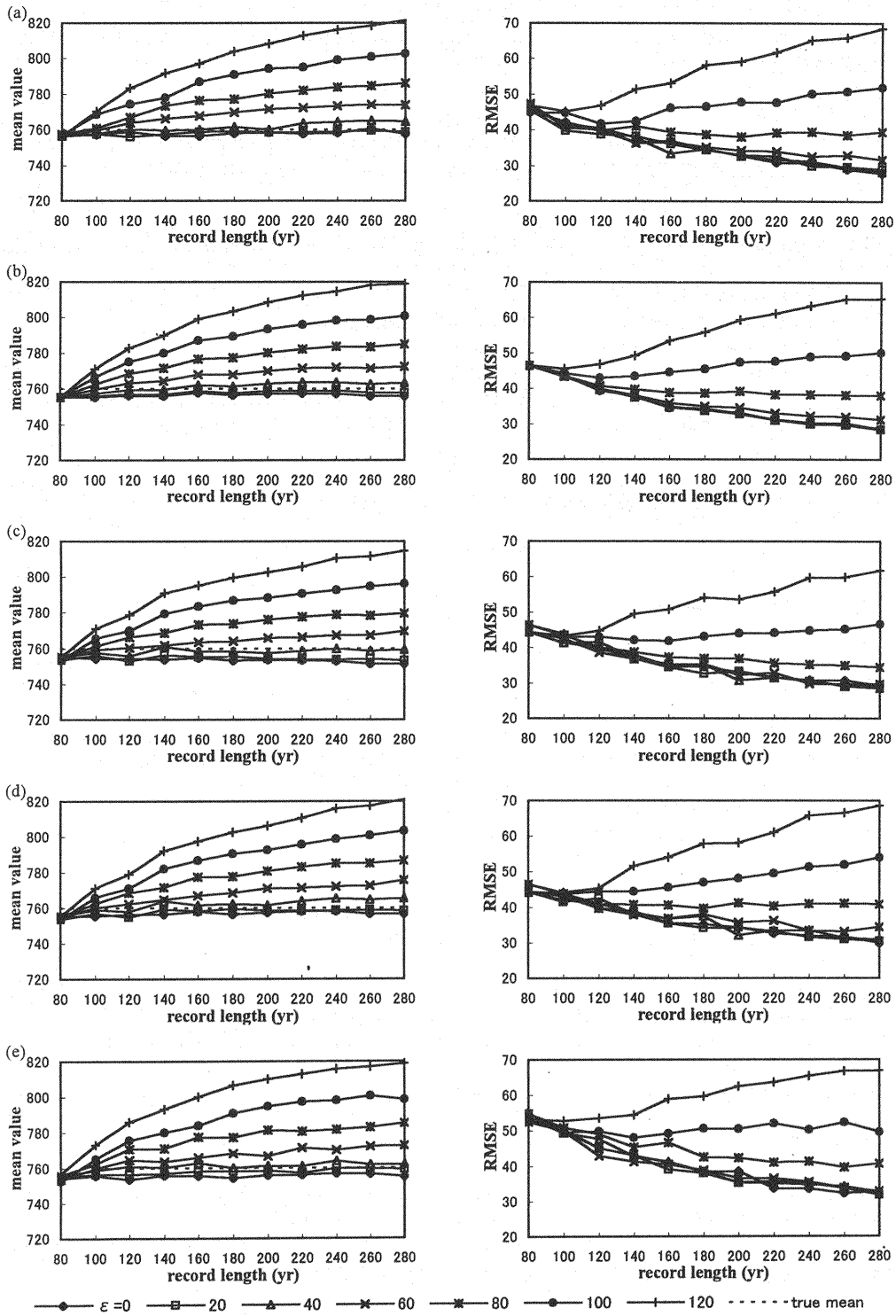

Fig. 1 Mean value (left) and RMSE (right) of 100-year flood estimates for various standard errors of historical
(a) MLE1, (b) MLE2-1, (c) MLE2-2, (d) MLE3, (e) MOM

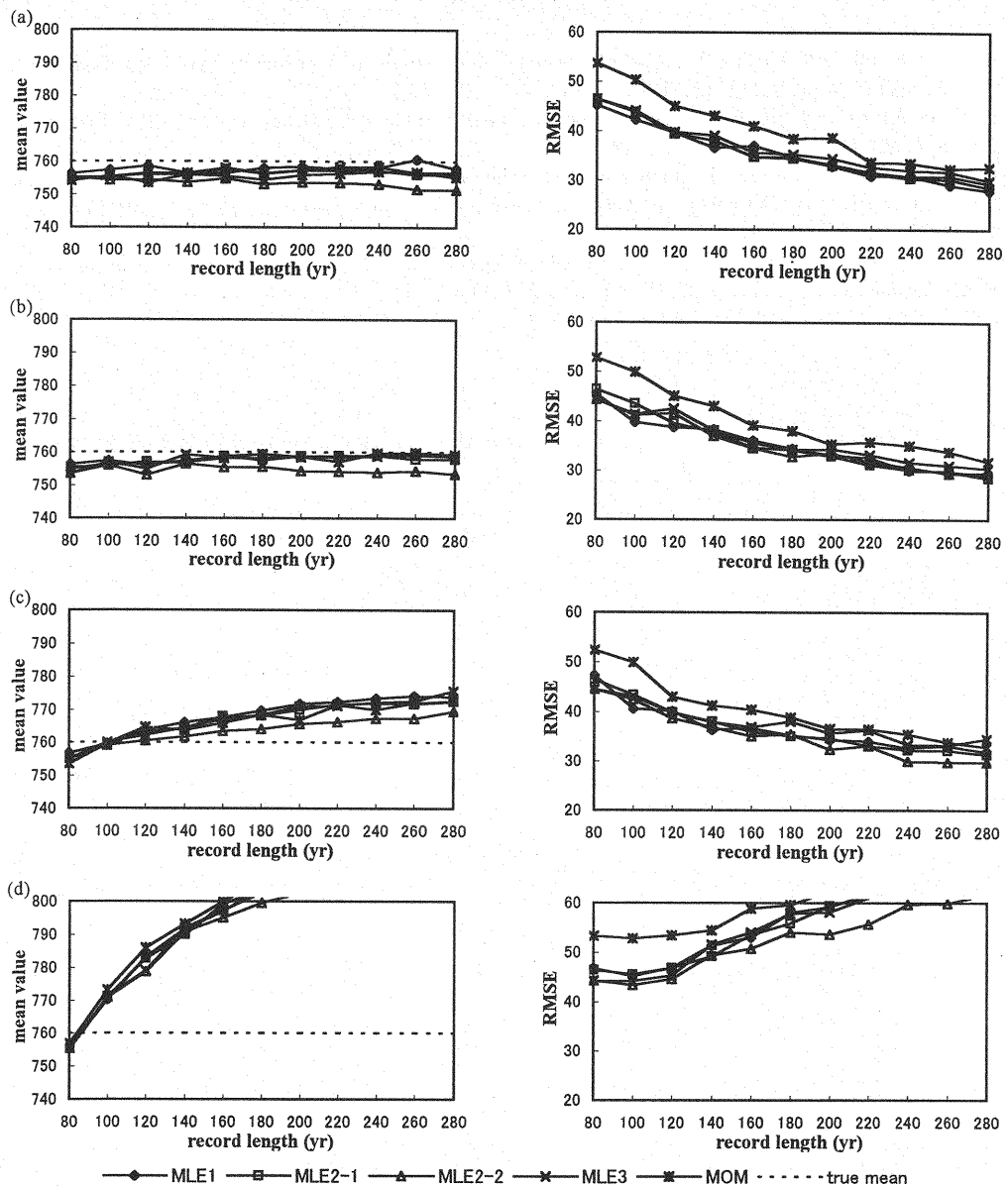

Fig. 2 Mean value (left) and RMSE (right) of 100-year flood estimates (comparing the four estimaters).
(a) $\varepsilon = 0$, (b) $\varepsilon = 20$, (c) $\varepsilon = 60$, (d) $\varepsilon = 120$

REFERENCES

1. Cohn, T.A. and J.R. Stedinger : Use of historical information in a maximum-likelihood framework, *Journal of Hydrology*, Vol.96, pp.215-223, 1987.
2. Hosking, J.R.M. and J.R. Wallis : Paleoflood hydrology and flood frequency analysis, *Water Resources Research*, Vol.22, No.4, pp.543-550, 1986.
3. Ikebuchi, S. and M. Maeda : Estimation of design rainfall and water level with systematic and historical flood information (in Japanese), *Annuals of the Disaster Prevention Research Institute Kyoto University*, No.34 B-2, pp.103-125, 1991.
4. Stedinger, J.R. and T.A. Cohn : Flood frequency analysis with historical and paleoflood information, *Water Resources Research*, Vol.22, No.5, pp.785-793, 1986.

(Received November 25, 1999 ; revised February 22, 2000)