Technical notes

Confidence intervals for flood control design

R. J. Whitley, Dept. of Mathematics, University of California, Irvine, CA 92717, USA

T. V. Hromadka II, Director of Water Resources Engineering, Williamson and Schmid, 17782 Sky Park Blvd., Irvine, CA 92714, USA

1 Introduction

Flood control agencies at the city and county level typically develop standards and guidelines to be used in the design of local flood control facilities. Generally, a flood protection criterion is selected, such as 100-year flood protection, and this criterion becomes the design standard for tpe region. Often) this flood protection goal results in the need to estimate a T-year flood peak flowrate, such as Q_{100}, which is then used for the design of flood control channels. While many statistical methods exist for developing an estimate of Q_{100} based on an annual flood flow series, seldom do flood control agencies address the uncertainty inherent in the use of (1) the particular statistical technique or distribution, (2) the chance selection of sampling data, (3) the use of regionalized parameters (such as skew), and (4) the uncertainty in the data quality due to measurement errors, flooding problems, or due to the effects of changing catchment conditions (e.g., urbanization).

Although the Water Resources Council Bulletins 17A and 17B (USWRC 1967) utilize the log Pearson III distribution in their analysis, several questions concerning the use of this distribution have been raised (e.g., Kite 1975, Stedinger 1983). Other statistical fits are often used with the data due to preferences or ease of use. For example, oftentimes the log Pearson III analysis is replaced by a straight-line plot on log-log-paper defined by an eye-fit to the annual peak flow data, and the resulting line is then used to develop design values of peak flow.

Random variation in the data points used causes a corresponding variation in the resulting flood frequency curve. The use of regionalized parameters such as the skew parameter in Bulletin 17B is also subject to uncertainty due to variations in the skew with changing conditions, and due to variations within the region. Finally, the uncertainty in the quality of the data and its measurement introduces an uncertainty which presently cannot be quantified.

In spite of these sources of uncertainty, it is currently a common practice for flood control agencies to adopt a particular flood control goal (e.g., Q_{100} design flows) and simply utilize a flood frequency curve for developing the estimate of Q_{100}. This approach does not take into account the problem of statistical uncertainty.

In this paper we consider aspects of two of these issues. In connection with issue (1), the specific probability distribution used, it is shown that the practice of using a linear plot on log-log paper of discharge versus return period T to predict peak discharges constitutes a hidden assumption that the underlying probability distribution of the logs of the discharges is an exponential distribution. This corresponds to a log Pearson III distribution with a skew value of 2, which is a relatively large value for the skew. The value of the peak T-year discharge Q_T as predicted by the method is then compared with that predicted by assuming a log Pearson III distribution with more usual values of skew. In connection with issue (2), the chance collection of data points, tables are given which allow the computa-

tions of confidence intervals for peak Q_T values derived from the linear plot discussed above. The other issues as to the use of another choice of probability distribution, the quality of the data, the goodness of measurement, and the use of regionalized parameters are not considered; these issues must be decided on a case study basis.

2 Estimate for the T-year flood

The data for a gauged site usually consists of a collection of m values for Q: $Q_1, \ldots, Q_m$. The values of T which are associated with the values of Q via the assumed empirical relationship

$$Q = \alpha T^{\beta}, \tag{1}$$

(Q the yearly maximum discharge in cfs, T the return period in years, α and β positive numbers) are not known. Indeed, for design purposes these values of T are exactly what is needed; we want to know the magnitude of the T year floods for a representative set of values of T, and these magnitude could be obtained by a linear interpolation via Eq. (2), if the T values $\{T_i\}$ for the Q values $\{Q_i\}$ were known.

Taking logarithms in Eq. (1)

$$X = a + bZ \tag{2}$$

where $X = \log Q$, $a = \log\alpha$, $b = \beta$, and $Z = \log T$. The value of T which goes with a given Q, being unknown, can be regarded as a random variable distributed in the way that follows from the definition of T as a return period: $P(T \geq t) = 1/t$. In terms of $Z = \log T$,

$$P(Z \geq t) = e^{-t}, \tag{3}$$

so Z has an exponential distribution with mean unity and Eq. (2) represents X as a simple linear function of Z.

From Eq. (2)

$$E(X) = a + bE(Z) = a + b; \tag{4}$$

$$E(X^2) = a^2 + 2abE(Z) + b^2E(Z^2) = a^2 + 2ab + 2b^2 \tag{5}$$

It follows from Eq. (4) and Eq. (5) that

$$b^2 = \mathrm{var}(X); \quad a = E(X) - b \tag{6}$$

Estimating the mean $E(X)$ and the variance $\mathrm{var}(X)$, from the data $\log Q_1, \ldots, \log Q_m$, gives estimates for a and b. Then the magnitude of Q, for a given value of T, can be obtained from Eq. (1).

3 Confidence intervals for Q

Let μ be the mean of X and $\hat{\mu}$ be the estimator for μ:

$$\hat{\mu} = (1/m)\sum X_i = (1/m)\sum(a + bZ_i) = a + b\hat{\mu}_Z \tag{7}$$

where $\hat{\mu}_Z$ is the mean of the m values $Z_1, \ldots, Z_m$ which are, by assumption, m exponentially distributed random variables. Similarly, using the usual estimator $\hat{\sigma}$ for the standard deviation of X,

$$\hat{\sigma}^2 = (1/(m-1))\sum(X_i - \hat{\mu})^2$$

and replacing X_i by $a + bZ_i$, it follows that

$$\hat{\sigma}^2 = b^2\hat{\sigma}_Z^2 \tag{8}$$

where $\hat{\sigma}_Z^2$ is the estimator for the standard deviation of the m values $Z_1, \ldots, Z_m$. The true value of X and the estimate of X, given T, are respectively

$$X = \mu + \sigma(\log T - 1), \quad \hat{X} = \hat{\mu} + \hat{\sigma}(\log T - 1) \tag{9}$$

Thus

$$(X - \hat{X})/\hat{\sigma} = \{(\mu - \hat{\mu})/\hat{\sigma}\} + \{(\sigma - \hat{\sigma})/\hat{\sigma}\}(\log T - 1) \tag{10}$$

Equation (10) allows the construction of confidence intervals for the true value X because, using Eqs. (7) and (8), its right hand side $R_{T,m}$ can be written

$$R_{T,m} = \{(\mu_Z - \hat{\mu}_Z)/\hat{\sigma}_Z\} + \{(\sigma_Z - \hat{\sigma}_Z)/\hat{\sigma}_Z\}(\log T - 1) \tag{11}$$

For a specific value of T, corresponding to the T-year flood, and a given number m of data points $Q_1, \ldots, Q_m$, the random variable $R_{T,m}$ given by Eq. (11) has a distribution which can be simulated. Once an empirical distribution for $R_{T,m}$ is known, confidence intervals for X can be found by using Eq. (10).

A computer program was written to simulate this distribution of $R_{T,m}$ and the results are given in Table 1 as an example for sample size 50. Similar tables are available from the authors for other sample sizes.

4 Comparison with log Pearson III

The methodology of Bulletin 17B (USWRC 1967) of the Water Resources Council for determining the T-year flood Q_T involves fitting the logarithms of the peak Qs with a Pearson type III distribution. On the other hand, as we have seen, the *empirical relation given in Eq. (1) becomes, in use, an assumption about the distri-*

Table 1. Simulation of the $R_{T,50}$ distribution

Percentiles Sample size = 50

These percentiles are for the distribution of $(\log QT - e\log QT)/\text{esd}$, where QT is the peak discharge for return period T, $e\log QT$ is the statistical estimate for $\log QT$, and esd is the statistical estimate for the standard deviation of $\log QT$. The model from which these percentiles are derived is $\log QT = a + b\log QT$, T the return period. The simulation was 10 blocks of size 1000.

percentile	$T=2$	$T=5$	$T=10$	$T=25$	$T=50$
5	-0.17	-0.32	-0.47	-0.68	-0.84
10	-0.14	-0.25	-0.37	-0.54	-0.67
15	-0.11	-0.20	-0.30	-0.43	-0.53
20	-0.09	-0.16	-0.24	-0.34	-0.41
25	-0.07	-0.13	-0.19	-0.26	-0.32
30	-0.06	-0.10	-0.14	-0.19	-0.23
35	-0.04	-0.07	-0.09	-0.13	-0.15
40	-0.03	-0.04	-0.04	-0.06	-0.07
45	-0.01	-0.01	0.00	0.01	0.02
50	0.00	0.02	0.05	0.09	0.11
55	0.01	0.06	0.10	0.16	0.20
60	0.03	0.09	0.15	0.23	0.30
65	0.04	0.13	0.20	0.31	0.39
70	0.06	0.17	0.26	0.40	0.50
75	0.08	0.21	0.33	0.49	0.62
80	0.10	0.26	0.40	0.60	0.75
85	0.12	0.32	0.49	0.73	0.91
90	0.15	0.40	0.61	0.90	1.13
95	0.20	0.55	0.83	1.22	1.50

bution of log Q and this distribution is a Pearson type III but with the fixed value of skew = 2, as opposed to the region values of skew suggested in the Bulletins. This raises the question of how these two different assumptions about the skew affect the feature of major interest, namely the predicted size of the T-year flood.

As a test of the predictive difference between these distributions suppose that the distribution of peak Qs is actually given by Eq. (1) but, instead, the log Pearson III model with zero skew is used to predict the magnitude of the T-year flood. The following table gives the relative error made in using the log Pearson III distribution for a set of floods, normalized by setting the 2-year flood $Q_2 = 100$, in terms of the ratio Q_{100}/Q_2.

To understand this close agreement, note that the prediction of the T-year flood based on log Pearson III has the form $\mu + \sigma K(T)$ (Beard 1962). So using $\mu = a + b$, $\sigma = b$, and the notation of Table 2,

$$\begin{aligned}(A - B)/A &= \{a + b\log T - (a + b + bK(T))\}/(a + b\log T) \\ &= \{\log T - 1 - K(T)\}/\{(a/b) + \log T\} \\ &= \{\log T - 1 - K(T)\}/\{(\mu/\sigma) + \log T - 1\}.\end{aligned} \tag{12}$$

For $Q_{100}/Q_2 = 2$, $a/b = 25.3$; for $Q_{100}/Q_2 = 10$, $a/b = 7.1$; and for $Q_{100}/Q_2 = 100$, $a/b = 3.2$. Therefore, although in the numerator of Eq. (12) $\log T - 1$ is only a rough approximation to $K(T)$, the relatively large size of a/b (or of μ/σ) reduces the percent relative error to within acceptable limits even for such different values of skew as 0 and 2.

Recall that generally the skew parameter for the log Pearson III distribution is determined from either a map or regional skews or from a large pool of data from the area, while the other two parameters of the log Pearson III distribution are determined from the mean and standard deviation of the data; see the Bulletins (USWRC 1967). The values for $K(T)$ used in Table 3 were taken from exhibit 39 in (Beard 1962).

Table 2. Table of Relative Percent Error $100 \times (A - B)/A$

A = the true value Q_T of the T-year flood given by $\log Q = a + b\log T$

B = the predicted value from log Pearson III with zero skew

Q_{100}/Q_2	$T = 2$	$T = 5$	$T = 10$	$T = 25$	$T = 50$	$T = 100$
2	-1	-1	0	2	3	4
4	-2	-2	0	3	5	8
6	-3	-2	0	4	6	9
8	-4	-2	0	4	7	10
10	-4	-3	0	5	8	11
100	-8	-5	0	7	12	16

Table 3. Table of Relative Percent Error $100 \times (A - B)/A$

A = the true value Q_{100} of the 100-year flood given by $\log Q = a + b\log 100$

B = the predicted value from log Pearson III with skew γ as shown below

Q_{100}/Q_2	$\gamma = -0.8$	$\gamma = -0.4$	$\gamma = 0$	$\gamma = 0.4$	$\gamma = 0.8$
2	6	5	4	3	2
10	16	13	11	8	6
100	24	20	16	13	9

Note that the error diminishes as the skew approaches the value of 2.

On the other hand, if the data are actually from a log Pearson III distribution, but instead the T-year flood is predicted using Eq. (1), then, in the notation of Tables 2 and 3,

$$(B - A)/A = \{\mu + \sigma K(T) - (a + b(\log T - 1))\}/\{\mu + \sigma K(T)\}$$
$$= \{K(T) - (\log T - 1)\}/\{(\mu/\sigma) + K(T)\}.$$

But because $(B - A)/B$ is approximately $-(A - B)/A$, as Tables 1 and 2 indicate, the relative percent error $100\ (B - A)/A$ is approximately the negative of the error given in Tables 2 and 3.

We have seen that the predicted values of Q_{100} do not differ greatly for the usual range of values of skew when compared with the values for a skew of 2. However, confidence intervals do vary significantly with skew, as Table 4 indicates. In this table, it is also seen that the 85 percent confidence estimates of Q_{100} vary with skew much more than the 50 percent confidence estimates.

Table 4. Table of values $A(B)$ for
85 percent confidence intervals (A,∞) for Q_{100}
50 percent confidence intervals (B,∞) for Q_{100}
$M = 20$ data points, $Q_2 = 100$, skew γ as indicated

Q_{100}/Q_2	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$
2	180(160)	220(180)	280(210)
10	660(480)	1340(760)	3230(1200)
100	4360(2320)	17,950(5820)	103,900(14,380)

References

Beard, L. 1962: Statistical methods in hydrology. U.S. Army Engineer District, Corps of Engineers, Sacramento, CA

Kite, G.W. 1975: Confidence limits for design events. Water Resour. Res. 11, 48-53

Stedinger, J.R. 1983: Confidence intervals for design events. J. of Hydraulic Engineering, ASCE, 109, 13-27

USWRC 1967: Guidelines for determining flood flow frequency. U.S. Water Resources Council, Hydrology Committee, Bulletin 15, Washington, DC (also revised versions, Bulletin 15, 1975; Bulletin 17A, 1977; Bulletin 17B, 1981)

Accepted February 19, 1988.

On the method of maximum likelihood estimation for the log-Pearson type 3 distribution

K. Arora and V. P. Singh, Dept. of Civil Engineering, Louisiana State University, Baton Rouge, LA 70803-6405, USA

1 Introduction

Much interest has been generated in the log-Pearson type 3 (LP3) distribution since it was first recommended by the U.S. Water Resources Council (USWRC 1967), and subsequently updated in 1975, 1977 and 1981 as the base method of flood fre-