WATER RESOURCES RESEARCH, VOL. 35, NO. 1, PAGES 203–209, JANUARY 1999

Approximate confidence intervals for design floods for a single site using a neural network

Robert Whitley
Department of Mathematics, University of California, Irvine

T. V. Hromadka II
Department of Mathematics and Environmental Studies, California State University, Fullerton
Exponent Failure Analysis Associates, Costa Mesa, California

Abstract. A basic problem in hydrology is the computation of confidence levels for the value of the T-year flood when it is obtained from a log Pearson III distribution using the estimated mean, estimated standard deviation, and estimated skew. Here we give a practical method for finding approximate one-sided or two-sided confidence intervals for the 100-year flood based on data from a single site. These confidence intervals are generally accurate to within a percent or two, as tested by simulations, and are obtained by use of a neural network.

1. Introduction

A basic problem in hydrology is the estimation, for design purposes, of the 100-year flood. One major source of uncertainty in this estimation is the choice of underlying distribution for maximum discharge [*Bobee et al.*, 1993; *Cohon et al.*, 1988; *World Meteorological Organization*, 1989]. The U.S. Water Resource Council's *Bulletin 17B Advisory Council on Water Data* [1982] recommends the use of a log Pearson III distribution, fit to yearly maximum discharge data, for the prediction of T-year events. Other distributions and methods have been proposed (see, for example, discussions by *Bobee et al.* [1993], *Cohon et al.* [1988], and *World Meteorological Organization* [1989]), but in practice, because of the authority of the U.S. Water Resource Council, the log Pearson III distribution is used extensively.

An important source of uncertainty in using the log Pearson III distribution to calculate the T-year event is that caused by the estimation of the parameters of that distribution. To give a more realistic estimate of the level of risk involved in a chosen level of flood protection, it is necessary to quantify this uncertainty by means of one-sided or two-sided confidence intervals for the T-year flood estimates. The log Pearson III distribution contains three parameters that, using the procedure of *Bulletin 17B*, are estimated by use of the estimated mean, estimated standard deviation, and estimated skew of the distribution of logarithms of the yearly maximal discharge data, as described in more detail below. The estimators used for the mean, standard deviation, and skew are, except for the scaling factor appearing in front of the bracket in (2) for the skew, the usual ones.

The simplest case to consider is when the mean and standard deviation are estimated but the skew is known to be zero. Then, since the skew is zero, the log Pearson III distribution is actually a lognormal distribution, and confidence intervals can be obtained from the noncentral t distribution [*Advisory Committee on Water Data*, 1982; *Resnikof and Lieberman*, 1957; *Stedinger*, 1983].

The case of a known nonzero skew is more complicated than the case of known zero skew [*Bobee and Robitaille*, 1975; *Hu*, 1987; *Kite*, 1975; *Phien and Hsu*, 1985]. For this case *Stedinger* [1983] showed that the method of computing confidence intervals suggested by the *Advisory Committee on Water Data* [1982] is not satisfactory (also see the general discussion by *Chowdhury and Stedinger* [1991]). *Stedinger* [1983] gave an approximate expression for confidence intervals for the quantiles of the log Pearson III distribution using an asymptotic variance formula [*Bobee*, 1973; *Kite*, 1976], the accuracy of which was discussed by *Whitley and Hromadka* [1986b, 1987]. And *Whitley and Hromadka* [1986a] showed how to obtain confidence levels for the T-year flood by means of simulations in the case of known skew.

For the realistic problem of computing confidence intervals for the T-year flood when the mean, the standard deviation, and the skew are all estimated, the only approaches available that give confidence limits that are accurate for certain parameter ranges are those given by *Ashkar and Bobee* [1988] and *Chowdhury and Stedinger* [1991]. The method of *Ashkar and Bobee* [1988] applies to Pearson III distributions with positive skew and is tested in their paper for 90% and 95% confidence intervals for $T = 100$ and $T = 500$ year floods; $m = 10$ ($T = 100$ only), $m = 25$, and $m = 50$ site data points; and skews of 0.5(0.5)3.0, with absolute errors ranging from 0.13% to 2.6%. For the (small) set of values common to the tests of our method and the tests of *Ashkar and Bobee* [1988], the accuracy is similar; it would be interesting to test their method over a wider range of confidence levels, in particular for those small return periods that would be needed to extend their results to negative skews. The other available method is that suggested in the interesting paper by *Chowdhury and Stedinger* [1991], in which an approximate formula for confidence intervals is derived on the basis of work by *Stedinger* [1983]. These approximate formulas were tested by *Whitley and Hromadka* [1997] for the case of data from a single site by means of simulations and were found to be accurate when the unknown values of skew were zero but, roughly summarizing the contents of a 19×10

Paper number 1998WR900016.
0043-1397/99/1998WR900016$09.00

Table 1. Errors $m = 10$

	q								
γ	2.5	5	10	25	50	75	90	95	97.5
−1.00	−0.4	−0.9	−0.4	−0.4	+0.2	+0.6	+0.5	−1.2	−1.3
−0.75	+0.4	+0.2	+0.7	+0.5	−0.5	−0.3	+0.3	−0.9	−0.8
−0.50	+0.4	+0.4	+1.0	+0.2	−0.9	−1.6	−0.3	−0.8	−0.6
−0.25	+0.3	+0.2	+0.8	+0.4	−0.7	−1.7	−0.6	−0.9	−0.5
+0.00	+0.2	+0.1	+0.8	+0.3	−0.9	−1.7	−0.5	−0.8	−0.6
+0.25	−0.1	−0.1	+0.7	+0.3	−0.7	−1.0	−0.4	−0.8	−0.5
+0.50	−0.4	−0.2	+0.2	+0.0	−0.8	−0.2	−0.1	−0.6	−0.3
+0.75	−0.3	−0.3	+0.6	+0.3	+0.3	−0.2	+0.0	−0.2	−0.3
+1.00	−0.6	−0.4	+0.1	−0.2	+0.9	+1.2	+1.0	+0.5	−0.1

Entries are $100[\hat{Q}(\gamma) - q]$ for $m = 10$.

table, were in absolute error by about 4% (e.g., giving a confidence interval of 46% rather than the correct 50%) for a skew of 1/2 and in error by about 6% for a skew of −1/2. *Bulletin 17B* also recommends weighting at-site skew with a regional skew, and *Chowdhury and Stedinger* [1991] report better results for this case.

The approach we take below involves finding approximate confidence intervals for the 100-year flood based on data from a single site by means of a nonstandard use of a neural network applied to a large set of simulated data. The resulting method, using a family of curves for the cases of 10, 20, or 30 data points at the site and for confidence levels of 2.5, 5, 10, 25, 50, 75, 90, 95, and 97.5% is shown by simulations to be generally accurate (see Tables 1–3).

2. Basic Formulas

In fitting yearly maximum discharge by a log Pearson III distribution, for the prediction of T-year events, the logarithm of the yearly peak discharge is assumed to have a density function given by $1/(|a|\Gamma(b))[(x - c)/a]^{b-1} \exp(-[(x - c)/a])$, where if a is positive, the density is given for $x > c$ and is zero for $x < c$, while if a is negative, the density is given for $x < c$ and is zero for $x > c$. From the above the mean μ, the standard deviation σ, and the skew γ are seen to be related to the parameters a, b, and c, by

$$\sigma^2 = a^2 b$$
$$\gamma^2 = 4/b \tag{1}$$
$$\mu = c + ab,$$

where a has the same sign as γ.

The case of zero skew is taken to be the limiting case when the positive parameter b tends to infinity; the above density function then converges to the density for the normal distribution.

The recommendation of *Bulletin 17B* is that the parameters a, b, and c be estimated using (1) and using the usual moment estimators for μ, σ, and γ, but with the moment estimator for γ scaled to make it less biased [*Bobee and Robitaille*, 1975; *Lettenmaier and Burges*, 1980]. If there are m observations at the site, then

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\hat{\sigma} = \left[\frac{m}{m-1}\right]^{1/2}\left[\frac{1}{m}\sum_{i=1}^{m} x_i^2 - \hat{\mu}^2\right]^{1/2} \tag{2}$$

$$\hat{\gamma} = \frac{1}{m-2}\left(\frac{m}{m-1}\right)^{1/2}\frac{[(1/m)\sum_{i=1}^{m} x_i^3 - 3\hat{\mu}\hat{\sigma} - \hat{\mu}^3]}{\hat{\sigma}^3}$$

The formula for the density function (above) shows that if X denotes the logarithm of the maximum yearly discharge, then $(X - c)/a$ has a gamma distribution with parameter b and density $[1/\Gamma(b)]x^{b-1}e^{-x}$ for $x > 0$ and zero for $x < 0$.

A consequence of this (see (8)) is that the parameters a and c can be scaled out of the problem by considering the random variable $(X - \hat{\mu})/\hat{\sigma}$, in the same way that the mean and standard deviation can be scaled out when constructing confidence intervals for a normal distribution. However the pres-

Table 2. Errors $m = 20$

	q								
γ	2.5	5	10	25	50	75	90	95	97.5
−1.00	+0.6	−0.3	+0.1	+1.8	+1.2	−2.2	+1.0	−0.8	−0.9
−0.75	+1.0	+0.8	+1.0	+1.6	+1.6	+0.7	+0.6	+0.2	−0.3
−0.50	+0.7	+0.4	+0.3	+0.7	+1.0	+1.4	+0.1	+0.5	−0.1
−0.25	+0.2	−0.1	−0.4	−1.0	+0.3	+0.9	−0.5	+0.4	−0.1
+0.00	−0.4	−0.4	−1.1	−1.7	−0.4	+0.1	−0.9	+0.0	+0.0
+0.25	−0.6	−0.7	−1.1	−1.9	−0.9	−0.6	−1.1	+0.1	+0.0
+0.50	−0.5	−0.4	−0.8	−1.7	−1.3	−0.2	−1.0	+0.4	+0.2
+0.75	−0.2	+0.0	+0.3	−0.7	+0.0	+0.3	−0.3	+0.8	+0.5
+1.00	+0.2	+0.6	+1.0	+1.3	+1.6	+0.8	+0.1	+1.3	+1.1

Entries are $100[\hat{Q}(\gamma) - q]$ for $m = 20$.

Table 3. Errors $m = 30$

	q								
γ	2.5	5	10	25	50	75	90	95	97.5
−1.00	+0.7	+1.4	+1.7	−0.3	−2.5	−2.2	+0.3	−1.0	−1.0
−0.75	+0.7	+1.4	+1.7	−0.6	+1.7	+1.2	+0.9	+0.3	+0.1
−0.50	+0.4	+0.6	+0.8	−1.1	+2.1	+1.5	+0.7	+0.8	+0.5
−0.25	−0.1	−0.4	−0.2	−1.6	+1.2	+0.7	+0.3	+0.3	+0.0
+0.00	−0.4	−0.8	−1.4	−1.8	−0.1	−0.3	−0.3	+0.0	−0.2
+0.25	−0.6	−1.2	−1.2	−2.1	−1.5	−0.9	−0.8	−0.1	−0.6
+0.50	−0.2	−0.5	+0.7	−1.0	−1.7	−1.3	−0.8	+0.0	−0.7
+0.75	+0.2	+0.2	+0.1	+0.2	−0.7	−0.9	−0.4	+0.4	−0.5
+1.00	+0.6	+1.1	+1.4	+1.8	+0.6	+0.2	+0.5	+1.0	−0.2

Entries are $100[\hat{Q}(\gamma) - q]$ for $m = 30$.

ence of the parameter b, or equivalently a nonzero skew, makes the problem of obtaining confidence intervals much more difficult than for the textbook case when the random variable is normally distributed [*Bowman and Shenton*, 1988].

Let X denote the logarithm of yearly maximum discharge, which has a Pearson III distribution. For any return period $T > 1$, corresponding to the T-year flood, set $p = 1 - (1/T)$. The T-year flood value for X is the number x_p having the property that

$$P(X < x_p) = p. \tag{3}$$

The interpretation of (3) is that over a long period of time the maximal yearly discharge will not exceed the value x_p in a fraction p of the total yearly values; for example, for the $T = 100$ year flood the yearly maxima will be less than $x_{0.99}$ in (approximately) 99% of the years of record.

For a site having m years $t_1, t_2, \cdots, t_m$ of logarithms of yearly maximal discharge data, a function $g(m, p, q, t_1, t_2, \cdots, t_m)$ is said to provide a $100q\%$ one-sided confidence interval if it has the property that if it were to be used repeatedly on a large number of sites, each of which has m values of yearly maxima (with a log Pearson III distribution), then the inequality

$$x_p < g(m, p, q, t_1, t_2, \cdots, t_m) \tag{4}$$

holds (approximately) $100q\%$ of the time. For example, if $q = 0.9$, then $g(m, p, q, t_1, t_2, \cdots, t_m)$ is a 90% safe estimate for x_p, in the following sense: Using (4) at a large number of independent sites satisfying the basic assumption that their maximal yearly discharges have a log Pearson III distribution will provide the desired protection from the true but unknown flood value x_p 90% of the time. It is convenient to write this limiting value of the ratio of the number of times $x_p < g(m, p, q, t_1, t_2, \cdots, t_m)$ holds divided by the total number of samples, as one repeatedly samples more and more times from independent sites as

$$\text{Prob}\,(x_p < g(m, p, q, t_1, t_2, \cdots, t_m)) = q \tag{5}$$

The probability Prob, specifying the limiting results of repeated sampling, is to be distinguished from the probability measure P for the space of events on which the random variable X is defined, as in (3).

All manner of estimators $g(m, p, q, t_1, t_2, \cdots, t_m)$ could be considered because there is no theory that indicates whether it is possible to find an estimator using only the m sample data points from the site which will provide a function with the desired property of (5) or, if such a function does exist, what form it takes. In what follows we will use an estimator which depends only on the site sample mean $\hat{\mu}$, sample standard derivation $\hat{\sigma}$, and sample skew $\hat{\gamma}$, and has the form

$$g(m, p, q, t_1, t_2, \cdots, t_m) = \hat{\mu} + \hat{\sigma} f(m, p, q, \hat{\gamma}), \tag{6}$$

where f is an unknown function to be determined.

Rewrite (5) using (6):

$$\text{Prob}\left(\frac{x_p - \hat{\mu}}{\hat{\sigma}} < f(m, p, q, \hat{\gamma})\right) = q. \tag{7}$$

The key to finding confidence intervals in the case of known skew for γ, not zero [*Whitley and Hromadka*, 1986a], is that in sampling m points from a Pearson III distribution X the random variable $(x_p - \hat{\mu})/\hat{\sigma}$ appearing in (7) has the same probability distribution as

$$h = \text{sgn}\,(\gamma)[(z_p - \hat{\mu}_z)/\hat{\sigma}_z] \tag{8}$$

where Z denotes a gamma distribution with the density as given above, z_p is the T-year value for Z, and $\hat{\mu}_z$ and $\hat{\sigma}_z$ denote the random variables which are the sample mean and the sample standard deviation using m points independently sampled from Z's gamma distribution. (In the case $\gamma = 0$ the random variable of (8) has a noncentral t distribution since X is normally distributed). Then one can consider the random variable in (8) that depends only on the one parameter γ as opposed to X, which depends on the three parameters a, b, and c.

The choice of the functional form of (6) was strongly influenced by the fact that the right-hand side of (8) depends only on the skew of the distribution X and not on its mean or its standard deviation. Therefore when we wish to estimate the function f by means of simulations, (7) and (8) show that X can be taken to have mean 0 and standard deviation 1, and γ can be varied over a suitable range. If, instead, the function f had not been chosen so as to be invariant under the choice of mean and standard deviation, then the simulation would need to be performed over a range of possible mean values, possible standard deviation values, and possible skew values, which would be much more difficult to do than simulating only over a range of possible skew values. The risk in choosing g to have the form of (6), although it is easiest to work with, is that there is no guarantee this simple choice of form for the confidence interval function g will provide reasonable accuracy; the only way to know is to do the simulations.

3. The Neural Network

In the calculations below the T-year flood value of $T = 100$ was chosen as being the value commonly used in flood control design. The range of parameter values $m = 10$, 20, and 30, and confidence levels for q of 2.5, 5, 10, 25, 50, 75, 90, 95, 97.5% were chosen as being representative, and which would allow the computation of two-sided confidence intervals as well as the indicated one-sided confidence intervals. For each of these values of m and q, a separate function f was determined, so that each f was a function only of the estimated skew $\hat{\gamma}$.

The neural network used to approximate f has the form $c_0 + \sum_{i=1}^{n} c_i \tau(a_i \hat{\gamma} + b_i)$, the transfer function τ being the common choice [*Hassoun*, 1995; *Haykin*, 1994]:

$$\tau(t) = \frac{-1}{2} + \frac{1}{1 + \exp(-t)} = \tanh\left(\frac{t}{2}\right) \tag{9}$$

Preliminary calculations showed that it was sufficient to take the number of nodes $n = 3$, for which the neural network is a function of the 10 variables c_0, c_1, c_2, c_3, a_1, a_2, a_3, b_1, b_2, and b_3.

The use of this neural network requires the minimization over the 10 variables above of an objective function chosen so that small values of this objective function correspond to the network being able to better predict some target events, that is, the network "learning" whatever one tries to "teach" by means of examples.

Let $p = 1 - (1/100)$, m, and q be fixed. The objective function that the neural network computation will attempt to minimize is constructed as follows. For a given value γ of skew, there is a 100-year value $x_p = x_p(\gamma)$, depending on γ. By sampling m independent points from a gamma distribution [*Devroye*, 1986], a sample mean $\hat{\mu}$, a sample standard deviation $\hat{\sigma}$, and sample skew $\hat{\gamma}$ can be computed. Repeating these calculations N times gives the collection $(\hat{\mu}_j, \hat{\sigma}_j, \hat{\gamma}_j)$, for $j = 1, \cdots, N$. These sample statistics can used to define

$$\hat{Q}(\gamma) = \frac{1}{N}[\text{the number of } j \text{ for which } \hat{\mu}_j + \hat{\sigma}_j f(\hat{\gamma}_j) > x_p(\gamma)] \tag{10}$$

For the given value of skew γ, there is a constant value for f which makes (10) hold with the required confidence value of q. The problem arises from our not knowing the value of γ but knowing only the value of the variable estimator $\hat{\gamma}$.

It would be desirable for (10) to have the approximate value q for all γ, but it is probably impossible to find an f for which this is true. However it is possible to find an f for which (10) is approximately equal to q for a range of values of γ which include a sufficiently wide range of skews so as to be applicable to real world problems. The range of skew considered by *Chowdhury and Stedinger* [1991] is $[-0.75, 0.75]$, while the range considered here will be $[-1, 1]$. Thus the conclusion reached here will be that if the unknown value of the skew lies between -1 and 1, a plausible assumption for most regions in the United States, then the curves constructed will supply confidence levels with an accuracy discussed below.

To have (10) be the approximate value q for a wide range of skew values, the objective function will be chosen to have this true for a selected set of skew values; specifically, for $\gamma_1 = -1$, $\gamma_2 = -1/2$, $\gamma_3 = 0$, $\gamma_4 = 1/2$, and $\gamma_5 = 1$. For these values, a good choice of objective function is $(1/5)\sum_1^5 (\hat{Q}(\gamma_j) - q)^2$. In order to smooth out the distribution of errors, the actual objective function used was chosen by trial and error to be 0.9 times the function above plus 0.1 times $\max[|\hat{Q}(\gamma_j) - q| : j = 1, \cdots, 5]$.

Ordinarily [*Hassoun*, 1995; *Haykin*, 1994], the use of a neural network, with function f, involves a training set S of vectors x and for each x, a real target value Target(x). The training, in batch mode, consists of minimizing $\sum_{x \in S} (f(x) - \text{Target}(x))^2$. The function f so determined is then tested on an independent set of input vectors to ascertain how well it computes the target values for that set. An interesting aspect of our use of a neural network is that the target response for an input value $\hat{\gamma}$ is not known; the desired network response cannot be described in terms of its response to an individual input but only in terms of its behavior over a large data set.

In minimizing the objective function the technique used is a modification of a conjugate gradient method due to M. Powell [*Press et al.*, 1992, chap. 10.5; *Bazaraa et al.*, 1993]; the modifications include periodically resetting the orthogonal search directions to randomly chosen orthonormal directions and using a simple robust linear search algorithm.

Some trial and error was required to find workable coefficient values for the case $m = 20$. At first 1000 sites were considered (each consisting of 20 sample gamma variates), but to obtain a neural network with consistent predictive ability ultimately 15,000 sites were used. Each objective function evaluation then requires 1.5 million random gamma deviates, since it involves 5 values of skew, each with 15,000 sites, and each site has a sample mean, sample standard deviation, and sample skew, based on 20 values of a gamma random variable. Because the minimization for each value of q requires thousands of objective function evaluations, it is prohibitive to repeatedly generate the random gamma deviates. Instead one file of 225,000 numbers consisting of 15,000 sets of a sample mean, sample standard deviation, and sample skew, for each of the skew values $-1(1/2)1$, was generated, read into memory, and used repeatedly when calculating the objective function.

A typical calculation of this type, for one given value of m, but for $q = 2.5$, 5, 10, 25, 50, 75, 90, 95, 97.5%, takes about 10 hours using a 200 MHz Pentium Pro. (It is interesting to note that on the original IBM PC, which was used for the simulation work of *Whitley and Hromadka* [1986a], this calculation would have taken about 2 months.)

In practice, large values of skew are usually regarded with suspicion and are combined with other smaller estimates of skew from nearby sites or are adjusted by use of a regional skew value obtained from a map of regional skews [*McCuen*, 1979], or are reduced in value some other way [*Tasker and Stedinger*, 1986]. To reflect this practice, as well as for some technical reasons, the sample skews used by *Chowdhury and Stedinger* [1991] were truncated to lie in the interval $[-2, 2]$ for random population skews. In work by *Whitley and Hromadka* [1997] the sample skews used were truncated to lie in the interval $[-2, 2]$, and that will also be done here: Whenever a sample skew value $\hat{\gamma}$ is computed, either for training the neural network or for testing it, if $\hat{\gamma} > 2$, then it is set to the value $\hat{\gamma} = 2$ and if $\hat{\gamma} < -2$, then it is set to the value $\hat{\gamma} = -2$. To get some idea of the magnitude of this effect, if $\gamma = 1$, then the percentages of the time that $\hat{\gamma} > 2.0$ are 3.4%, 2.9%, 2.7% for $m = 10$, $m = 20$, and $m = 30$, respectively, and if $\gamma = 0.75$, then the corresponding percentages are 2.1%, 1.5%, and 1.1%. By symmetry the same percentages hold for the percentages of the time that $\hat{\gamma} < -2$ for $\gamma = -1$ and $\gamma = -0.75$. In the purely mathematical problem of finding approximate confidence in-

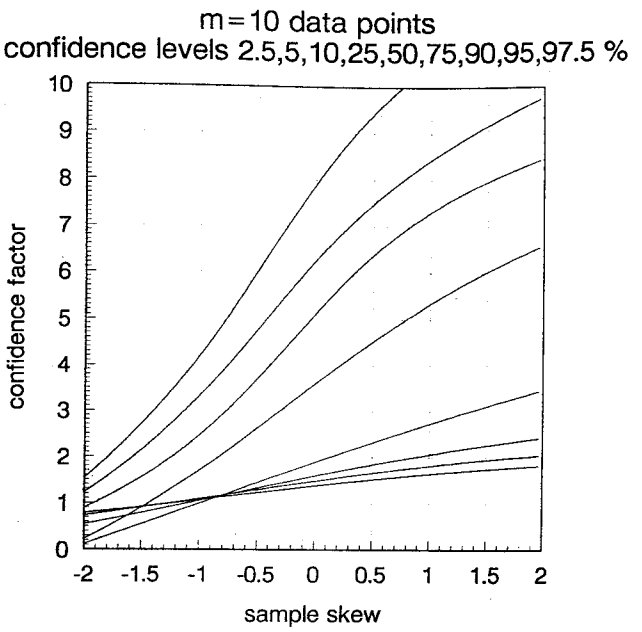

Figure 1. Confidence factor $m = 10$.

tervals for the T-year flood value, truncating the sample skew values involves some loss of information and therefore makes for a less accurate solution, but truncation better reflects how the data is actually used.

4. Results

The numbers reported in Tables 1–3 are each the result of an independent simulation for 50,000 sites. Using one of the indicated values of γ, and for a specified value of m = 10(10)30 and a specified confidence level q = 2.5, 5, 10, 25, 50, 75, 90, 95, 97.5%, the corresponding neural network confidence factor function f was tested as follows: For each site independent samples of the size m were taken from a gamma distribution with skew γ; a sample mean $\hat{\mu}$, sample standard deviation $\hat{\sigma}$, and sample skew $\hat{\gamma}$ were computed; and a count was kept of how many times $\hat{\mu} + \hat{\sigma}f(\hat{\gamma}) > x_p(\gamma)$ held, from which the value of $\hat{Q}(\gamma)$, given in (10), was computed for N = 50,000. For each of the values of m = 10(10)30 a table is given. For the indicated values of q and γ, these tables report the values of $100[\hat{Q}(\gamma) - q]$ for independent simulations of 50,000 sites each. For example, Table 1 shows that the neural network function f for $m = 10$ and $q = 90\%$ gave confidence limits of 90.3% if the true skew value was $\gamma = -0.75$, while f gave 89.9% confidence limits if $\gamma = +0.50$.

Recall that the neural network was optimized to give accurate confidence intervals for skew values of −1.0, −0.5, 0, 0.5, and 1.0 for a specific data set of 50,000 sites. The tests in the table include these values of skew and the additional values −0.75, −0.25, 0.25, and 0.75, all for entirely different data sets than the one over which f was optimized. Thus the tabulated values indicate the probable accuracy obtained by using the neutral network functions f if the unknown site skew lies between −1 and 1, a plausible assumption for many regions of the United States. Furthermore, additional testing for some of the m and q values, using skews not in [−1, 1] but in the wider range [−2, 2], gave curves that were indistinguishable from the curves given below, showing that little accuracy was lost by restricting the range of skew to [−1, 1] when constructing the neural network curves.

The neural network function f is plotted in Figures 1–3 for $m = 10$, $m = 20$, and $m = 30$, respectively. In each figure the family of lines plotted intersects the vertical line $\hat{\gamma} = 2$, with the function f for the value $q = 2.5\%$ being the lowest curve, $q = 5\%$ being the next highest curve, and so on. For example, suppose a $q = 90\%$ confidence interval is desired for a site of $m = 20$ points, with the logarithm (base 10) of maximal discharges having sample mean $\hat{\mu} = 3.4$ (log feet^3/s), a sample standard deviation $\hat{\sigma} = 0.2$ and a sample skew $\hat{\gamma} = 0.5$. The value of f for $q = 90\%$ can be read from Figure 2 for $m = 20$ as $f(0.5) = 4.7$ approximately. Thus the 90% confidence limit for the logarithm of the 100-year flood in log space would be the number 3.4 + 4.7(0.2) = 4.34 for a discharge value of $10^{4.34}$ = 21,900 feet^3/s (62,000 m^3/s); that is, the probability, in the sense of repeated sampling, is 0.90 that the true 100-year flood value is no larger that 21,900. Some of the curves in Figures 1–3 show rather odd behavior for values of skew less than −1. The principle reason for this will be discussed below, but in spite of this anomalous behavior, using these curves gives the accuracy reported in Tables 1–3. There would be perhaps a small improvement in accuracy, and certain sets of values would definitely be more consistent, if the errant curves were interpolated by hand in a smooth fashion.

An interesting comparison concerns the confidence values for known skew given by *Stedinger* [1983] and *Whitley and Hromadka* [1986a, 1986b]. The neural network function f is plotted in Figure 4 for the illustrative case $m = 20$ and $q = 90\%$. The solid circles represent the values of the 90% confidence factor if the skew is not estimated but is actually known to have the values of skew indicated. So if in the numerical example discussed above the skew was known to be equal to 0.5, and not estimated, then the 90% confidence value can be calculated by using the circle value of 3.7 for a skew of 0.5, and would be 3.4 + 3.7(0.2) = 4.14, or $10^{4.14}$ = 13,800 feet^3/s (39,100 m^3/s).

The curve in Figure 4 lies below the circles of the curve for

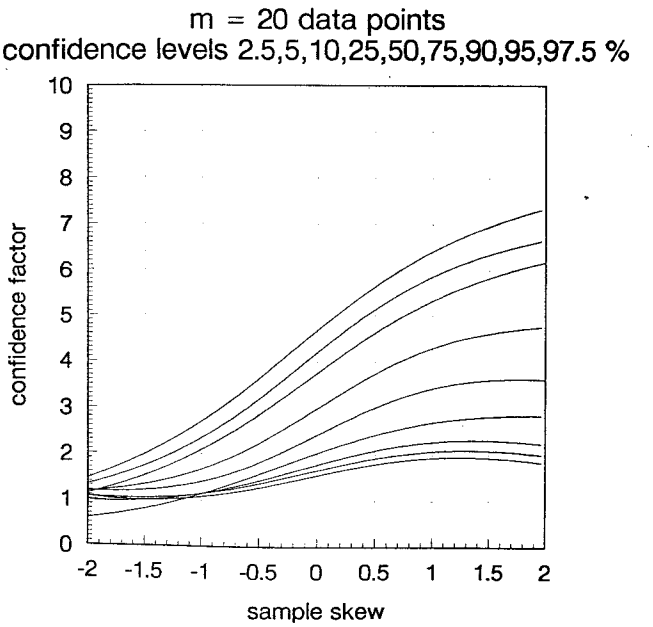

Figure 2. Confidence factor $m = 20$.

known skew if the sample skew is less than -1. This seems to imply that one is worse off (the confidence interval is larger) if one knows that the skew is, say, -2, than if one has only the estimate -2 for the skew. This is puzzling since one should be better off knowing the skew exactly. To understand why this interpretation of the curves is incorrect, we must look at the problem of prediction that the neural network is designed to solve.

Consider the following sequence of skews γ followed in parentheses by the approximate percentile of the number γ itself in the distribution of sample skews: 0 (50%); 0.5 (60%); 0.75 (66%); 1.00 (70%). From the point of view of the predicting neural network, the sample skews tend to be less than the underlying unknown positive skew; for example, 70% of the time the sample skew from a distribution with an actual skew of 1 will be less than 1. Since the distribution of sample skews for a Pearson III distribution with negative skew γ (and mean zero, which is the case for our normalized simulations) can be obtained as the negative of the distribution of the sample skews of a Pearson III distribution with skew $-\gamma$, the percentiles given above for positive skews are reversed for negative skews; for example, only 30% of the sample skew from a distribution with skew -1 are less than -1. The neural network can be thought of as a two-step process. First, use the sample skew to estimate the unknown skew. Second, use this estimated skew to calculate the confidence limits. In the first step there are two competing factors: If the sample skew came from a distribution with a positive skew, then since the sample skews are usually less than the actual positive skew, a tendency which increases with the magnitude of the skew, the network needs to estimate a true skew to the right of the sample skew to be used in the second step. A larger skew means a larger confidence value, and this raises the f curve about the curve for known skew. If, on the other hand, the sample skew came from a distribution with negative skew, the estimate for the actual skew should be to the left of the sample skew, making the confidence value smaller. The way in which the distributions of the sample skews for various values of true skew interact is complicated but is summed up by the graph of the neural network function f. With this in mind, reconsider the extreme example in which the sample skew in Figure 4 is -2. The graph shows that for this relatively large negative value there is very little contamination of sample skews by negative skews from distributions with an actual positive skew, and therefore since the unknown true skew is (probably) negative, the corresponding estimate for the unknown skew should be to the left of the sample skew, which makes the confidence value smaller than if the sample skew were used as an estimate for the unknown skew.

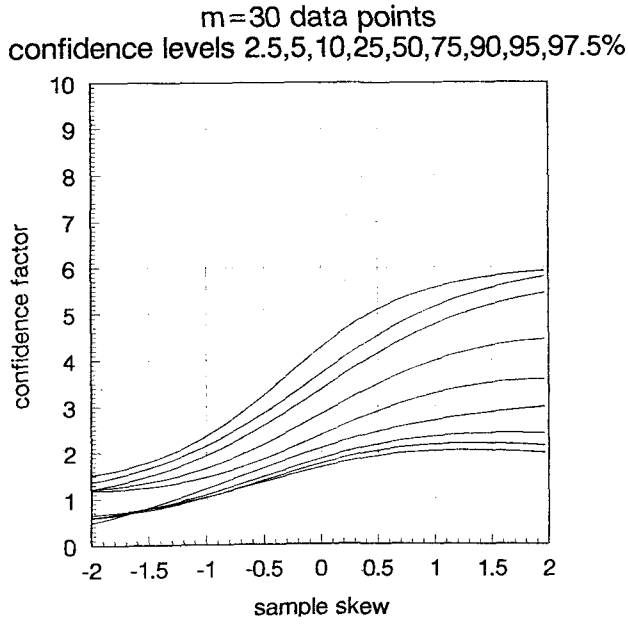

Figure 3. Confidence factor $m = 30$.

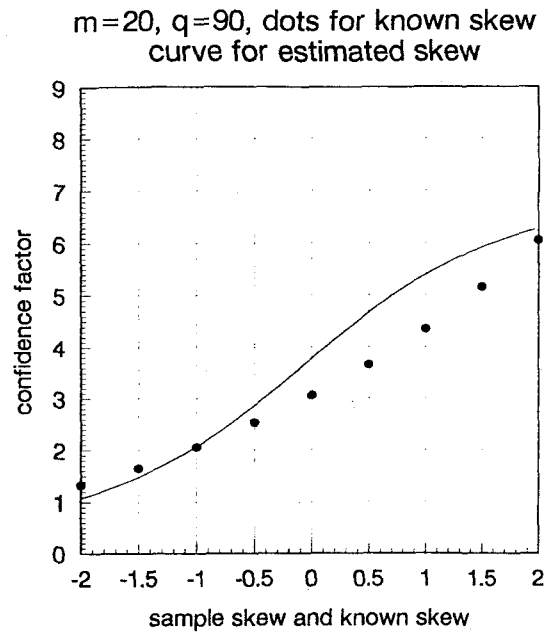

Figure 4. Sample skew versus known skew.

5. Conclusion

Confidence level curves f are calculated for upper confidence limits for the 100-year flood when the yearly maximal discharge data is from a log Pearson III probability distribution. This is done by means of an unusual application of neural networks. These curves have been calculated for confidence levels of 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 97.5% for sites with 10, 20, and 30 data values of yearly maximal discharges. The user chooses the appropriate curve f (two curves if a two-sided interval is desired); interpolates if the number of data points m lies between 10 and 20 or 20 and 30; and computes a sample mean $\hat{\mu}$, sample standard deviation $\hat{\sigma}$, and sample skew $\hat{\gamma}$ for the logarithms of yearly maximal discharges, from which the confidence level for the logarithm of maximal yearly discharges is given by $\hat{\mu} + \hat{\sigma}f(\hat{\gamma})$. The accuracy obtained in using this estimate for the desired confidence level has been tested and shown to be have the good accuracy displayed in Tables 1–3 under the assumption that the unknown value of the site skew lies in the interval $[-1, 1]$.

It would be possible to extend the range of skew used in the neural network simulations so as to cover a broader range of unknown skew values, perhaps with not much loss in accuracy. It would also be of value to extend the range of data points m considered. Since a strength of neural networks is in approximating functions of several variables, it might be possible to extend the neural network and find one function of the three

variables of confidence level q, number of points at the site m, and sample skew $\hat{\gamma}$ and allow a single f to be used for a broad range of m, q, and $\hat{\gamma}$ values, avoiding interpolations. In fact, we had this in mind when we began, but had troubles enough with the one variable $\hat{\gamma}$.

While the accuracy obtained from the neural network is, we feel, remarkable, computing the neural network curves involves a considerable amount of trial and error to fit training data sets of 5000 to 15,000 points accurately, although once that is done the resulting curve always gives satisfactory test results for 50,000 independent data points. Not only is this process somewhat tedious, but even with these large data sets we were not able to obtain the accuracy we wanted for the 99.5% and 0.5% confidence limits which are needed to compute 99% two-sided confidence intervals. However, work is in progress, using other methods that are less computationally intensive and appear to be even more accurate than the neural network, which we hope will allow the computation of more confidence levels and a wider range of skew values.

Acknowledgments. We were fortunate to have knowledgeable referees whose helpful advice was graciously given: F. Ashkar, J. Chowdhury, and J. Stedinger.

References

Advisory Committee on Water Data, Guidelines for determining flood flow frequency, *Bull. 17B*, Hydrol. Subcomm., Off. of Water Data Coord., U.S. Geol. Surv., Reston, Va., 1982.

Ashkar, F., and B. Bobee, Confidence intervals for flood events under a Pearson 3 or log-Pearson distribution, *Water Resour. Bull.*, *24*(3), 639–650, 1988.

Ashkar, F., and T. Ouarda, Assessment of flood magnitude estimator uncertainty: Tolerance limits for the gamma and generalized gamma distributions, paper presented at ASCE Water Power '95, Am. Soc. of Civ. Eng., San Francisco, Calif., July 25–28, 1995.

Bazaraa, M., H. Sherali, and C. Shetty, *Nonlinear Programming*, 2nd ed., John Wiley, New York, 1993.

Bobee, B., Sample error of T-year events computed by fitting a Pearson type 3 distribution, *Water Resour. Res.*, *5*, 1264–1270, 1973.

Bobee, B., and R. Robitaille, Correction of bias in the estimation of the coefficient of skewness, *Water Resour. Res.*, *11*, 851–854, 1975.

Bobee, B., et al., Towards a systematic approach to comparing distributions used in flood frequency analysis, *J. Hydrol.*, *142*, 121–136, 1993.

Bowman, R., and L. Shenton, *Properties of Estimators for the Gamma Distribution*, Marcel Dekker, New York, 1988.

Chowdhury, J., and J. Stedinger, Confidence intervals for design floods with estimated skew coefficient, ASCE *J. Hydraul. Eng.*, *11*, 811–831, 1991.

Cohon, J., et al., *Estimating Probabilities of Extreme Floods*, Natl. Acad. Press, Washington, D. C., 1988.

Devroye, L., *Non-Uniform Random Variable Generation*, Springer, New York, 1986.

Hassoun, M., *Fundamentals of Artificial Neural Networks*, MIT press, Cambridge, Mass., 1995.

Haykin, S., *Neural Networks*, IEEE Press, Piscataway, N. J., 1994.

Hu, S., Determination of confidence intervals for design floods, *J. Hydrol.*, *96*, 201–213, 1987.

Kite, G., Confidence intervals for design events, *Water Resour. Res.*, *11*, 48–53, 1975.

Kite, G., Reply to comments on "Confidence limits for design events," *Water Resour. Res.*, *12*, 826, 1976.

Lettenmaier, D., and S. Burges, Correction for bias in estimation of the standard deviation and coefficient of skewness of the log Pearson 3 distribution, *Water Resour. Res.*, *16*, 762–766, 1980.

McCuen, R., Map skew???, *J. Water Resour. Plann. Manage.*, *105*(2), 269–277, 1979.

Phien, H., and L. Hsu, Variance of the T-year event in the log-Pearson type 3 distribution, *J. Hydrol.*, *77*, 141–158, 1985.

Press, W., et al., *Numerical Recipes in Fortran*, 2nd ed., Cambridge Univ. Press, New York.

Resnikoff, G., and G. Lieberman, *Tables of the Non-Central t-Distribution*, Stanford Univ. Press, Stanford, Calif., 1957.

Stedinger, J., Fitting lognormal distributions to hydrologic data, *Water Resour. Res.*, *16*(3), 481–490, 1980.

Stedinger, J., Confidence intervals for design events, *J. Hydraul. Eng.*, *109*, 13–27, 1983.

Tasker, G., and J. Stedinger, Regional skew with weighted LS regression, *J. Water Resour. Plann. Manage.*, *112*, 709–722, 1986.

Whitley, R., and T. Hromadka II, Computing confidence intervals for floods, I, *Microsoftware Eng.*, *2*(3), 138–150, 1986a.

Whitley, R., and T. Hromadka II, Computing confidence intervals for floods, II, *Microsoftware Eng.*, *2*(3), 151–158, 1986b.

Whitley, R., and T. Hromadka II, Estimating 100-year flood confidence intervals, *Adv. Water. Resour.*, *10*, 225–227, 1987.

Whitley, R., and T. Hromadka II, Chowdhury and Stedinger's approximate confidence intervals for design floods for a single site, *Stochastic Hydrol. Hydraul.*, *11*, 51–63, 1997.

World Meteorological Organization, Statistical distributions for flood frequency analysis, *Oper. Hydrol. Rep. 33*, Geneva, Switzerland, 1989.

T. V. Hromadka II, Failure Analysis Associates, 3187 Redhill Blvd., Suite 100, Costa Mesa, CA 92626.

R. Whitley, Department of Mathematics, University of California, Irvine, CA 92697-3875. (rwhitley@math.uci.edu)

(Received June 30, 1998; revised September 9, 1998; accepted September 9, 1998.)