Flood Probability Estimation

By Samuel O. Russell,[1] M. ASCE

Abstract: A procedure is developed for estimating design floods in situations where data are insufficient for a conventional frequency analysis. It involves the use of a compound distribution which is a weighted combination of individual two parameter distributions. Initial values are assigned to the parameters of the component distributions and their weights on the basis of subjective estimates of the mean and standard deviation of the flood peaks. The weights of the component distributions are updated with Bayes' theorem in the light of any additional information such as recorded floods, the largest flood in a number of years or a flow which has not been exceeded in a given number of years. Numerical methods are used. A computer program is developed to carry out the necessary computations and it is described to illustrate how the method could be used by practicing hydrologists. Examples are presented to illustrate the procedure.

Introduction

Estimates of peak flows are required for the design of many projects. For major projects, or where the consequences of failure could be catastrophic, it is customary to use the maximum probable flood as the design flood, and there are generally accepted procedures for the derivation of such floods. For other projects, floods with much smaller return periods (generally in the range from 2–200 years) are used as a basis for design. Where adequate data are available, it is a simple matter to make a frequency analysis and derive the flood with the specified average return period (1,5,6,12). Where the data are sparse, the hydrologist must weigh what evidence there is and use his or her judgement. Often there is a surprising amount of information available: Perhaps there are a few recorded floods; perhaps it is known that a bridge or culvert with known capacity has not been overtopped in a number of years; there may be evidence about the largest flood in memory; usually there are envelope curves for the area; it may be possible to estimate the mean annual flood from the river cross-section, and so forth. Although much of this information cannot be used directly in a traditional frequency analysis, an experienced hydrologist will use all available information, although in an intuitive rather than an explicit way. Bayesian statistics offer a framework for combining different types of information and making best use of what is available. Writers such as Tang

[1] Assoc. Prof., Dept. of Civ. Engrg., Univ. of British Columbia, Vancouver, B.C., Canada.

Note.—Discussion open until June 1, 1982. To extend the closing date one month, a written request must be filed with the Manager of Technical and Professional Publications, ASCE. Manuscript was submitted for review for possible publication on September 4, 1980. This paper is part of the Journal of the Hydraulics Division, Proceedings of the American Society of Civil Engineers, ©ASCE, Vol. 108, No. HY1, January, 1982. ISSN 0044-796X/82/0001-0063/$01.00.

(9) and Wood and Rodriguez-Iturbe (11) have suggested applications to flood flow estimation, and Rascon and Villarreal (8) to design wave estimation. However, these do not deal with different types of data and they tend to involve rather sophisticated mathematics which may be somewhat intimidating to the practicing engineer or hydrologist. In this paper, a procedure is presented for combining the types of information likely to be available and relevant for peak flood estimation. It involves the use of a computer program but, after the program has been set up, the user need only worry about the hydrology.

Approach

The basic concept is to use a compound probability distribution and update the weights of the component distributions with new information using Bayes' theorem. A compound distribution has been described as one where the parameters are themselves random variables (3), but it is more convenient to consider it as a weighted mixture of individual distributions. It is not possible to update

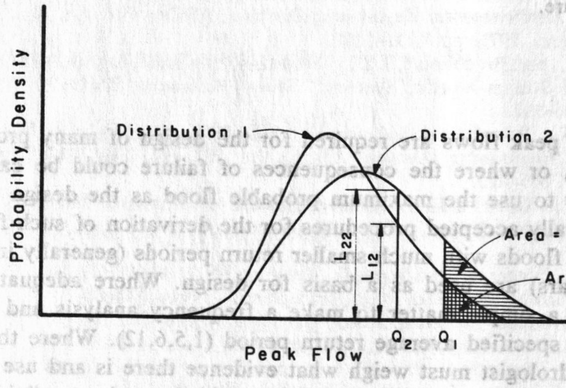

FIG. 1.—Compound Distribution

a probability distribution directly, but it is possible to update the weights of the underlying component distributions (3).

As an example, suppose that the peak flood regime on a river can be characterized by one or other of the probability distributions shown in Fig. 1, and that, on the basis of prior information, we are prepared to assign probabilities that either distribution is the correct one. (Two are used to keep the explanation simple.) If this represents the full extent of our knowledge, then we could say that the probability of a flood exceeding Q_1 is given by $w_1 P_{11} + w_2 P_{12}$ in which w_1 = probability that distribution 1 is correct; w_2 = probability that distribution 2 is correct; P_{11} = probability that Q_1 will be exceeded using distribution 1; and P_{12} = probability that Q_1 will be exceeded using distribution 2. The probabilities, w_1 and w_2, are probably best considered as "weights" and as such they add up to 1.0. The probability distribution of Q thus is a compound one.

Now, if the peak flow in one year is known to be Q_2, the weights can be updated using Bayes' theorem which states (in words) that

$$\text{Posterior probability} = \frac{\text{Prior probability} \times \text{Likelihood}}{\text{Normalizing factor}} \quad (1a)$$

or $$w_{12} = \frac{w_1 L_{12}}{N} \quad (1b)$$

and $$w_{22} = \frac{w_2 L_{22}}{N} \quad (1c)$$

in which w_{12} = posterior probability (weight) that distribution 1 is correct (given that Q_2 has occurred); L_{12} = likelihood that Q_2 would occur if distribution 1 is correct (it is given by the height of the probability density curve as shown); w_{22} and L_{22} are similar but refer to distribution 2; and $N = w_1 L_{12} + w_2 L_{22}$, and is simply a normalizing factor to make the sum of the posterior probabilities add up to 1.0. Now, in the light of this latest evidence, we can say that the probability of Q_1 being exceeded in any one year $= w_{12} P_{11} + w_{22} P_{21}$. The procedure can, of course, be repeated if more evidence becomes available. Similarly, the procedure is not limited to two distributions. Any number can be used although, of course, the calculations increase with the number.

Note that the main steps in the procedure, as shown in the example, are the following:

1. Select the form of the underlying distribution (the curves in Fig. 1. In practice, a standard distribution which can be specified by no more than two parameters is preferred).
2. Assign prior probabilities to each of the component distributions (w_1 and w_2 in the example).
3. Compute the relative likelihood of obtaining the data for each of the component distributions (L_{12} and L_{22} in the example).
4. Update the prior probabilities, w_1 and w_2, by multiplying them by the likelihoods to obtain posterior probabilities, w_{12} and w_{22}, of the component distributions.
5. Compute the peak flood corresponding to any required probability of exceedence or, alternatively, the probability of exceedence of any specified flood from the compound distribution.

In theory, it is possible to apply the above concept to continuous distributions using "conjugate pairs." These are pairs of distributions where the sample likelihood function is compatible with the prior distribution and the resulting compound distribution (which is called a Bayesian distribution) has a known mathematical form. They are given in standard texts such as Raiffa and Schlaiffer (6) and, where applicable, they can greatly simplify computations. However, with the possible exception of the normal distribution (which is used for comparison later in this paper), these are too restrictive in form to be useful to a practicing hydrologist.

It therefore becomes necessary to use numerical methods and carry out the calculations by computer. An interactive computer program is particularly well suited to the application of the concept to the estimation of design floods. It can be designed to accept both subjective estimates and hard data, and to allow the hydrologist to observe the influence of each additional piece of information on the design flood estimates as they develop. Such a program has been developed and copies can be obtained at nominal cost from the writer. The main features of this program are described in the following section, first to show how such a program can be used, and then to illustrate the type of assumptions made and the level of detail involved in applying the concept to the development of a program.

Steps in Using Computer Program

The program has been designed for interactive use from a computer terminal. The computer prompts the user with questions such as "Do you wish to use a Gumbel or lognormal distribution?" and gives instructions: "Enter 1 for Gumbel, 2 for lognormal." It also tells the user when and how to enter the data. When using the program, the user need concern himself only with the flood estimation problem and need not worry about the mathematical computations.

The following describes the program from the point of view of the user who is assumed to be a hydrologist:

1. Select the underlying distribution to be used, depending on previous experience with flood analyses in the area of the stream in question. This particular program allows either lognormal or Gumbel distributions.
2. Input information on the parameters of the selected distribution in the form of low, probable, and high values of the mean and standard deviation. The low value is the one which the user is 90% sure will be exceeded, the probable is his best guess, and the high is the value he is 90% sure will not be exceeded. Estimates of the mean can be made from an examination of the actual river channel or from generalized estimates of the mean annual flood discharge per unit area. Estimates of the standard deviation can be made from generalized estimates for the area. The program accepts information in several forms, including the coefficient of variation; the hydrologist can use whichever he is most comfortable with. Alternatively, the user can allow the program to compute means and standard deviations from data he supplies.
3. Input data on peak floods recorded on the stream being analyzed.
4. Input data on the *m*th largest flood in *n* years. Often there is information on the rank of a particular flood, such as the largest in the last 50 years, or the 2nd largest, the smallest and so on.
5. Input data on floods not exceeded in *n* years. Culverts or bridges with known discharge capacity may have successfully passed all floods for a number of years.
6. Input weighted data such as estimates obtained by correlation from an adjacent stream.

Any of the data in aforementioned steps 3–5 can be input in terms of low, probable, and high estimates, rather than single valued estimates. This allows the hydrologist to use information which may be uncertain but yet may be important and relevant. The degree of uncertainty is indicated by the width of the band between the low and high estimates.

After each new piece of information, the program outputs its latest estimate of the 2, 5, 10, 20, 50, 100, and 200 year floods, as well as the floods with any other required average return periods. Also, if desired, the program outputs its latest estimate of the probabilities and the parameters of the component distributions. In this way, the analyst can see the effect of each new piece of data.

PROGRAM

The program has been set up to use either the lognormal or Gumbel distribution as the underlying distribution. These are the two-parameter distributions which are most widely used in hydrology. Three parameter distributions could be used but they would involve much more computation for probably little gain in accuracy.

The following paragraphs describe the main features of program operation and, to avoid confusion, give the equations for the lognormal distribution only. Corresponding equations for the Gumbel distribution are readily available (5,11).

In the program, the mean and the standard deviation of the logarithms of the annual flood peak values are each specified by five discrete values: The

TABLE 1.—Initial Probability Matrix

σ^m (1)	L (2)	$L + P/2$ (3)	P (4)	$P + H/2$ (5)	H (6)	Marginal probability (7)
L	0.028	0.035	0.042	0.035	0.028	0.168
$L + P/2$	0.035	0.042	0.052	0.042	0.035	0.206
P	0.042	0.052	0.064	0.052	0.042	0.252
$P + H/2$	0.035	0.042	0.052	0.042	0.035	0.206
H	0.028	0.035	0.042	0.035	0.028	0.168
Marginal probability	0.168	0.206	0.252	0.206	0.168	

low, probable, and high values input by the user, a value half way between the low and the probable, and a value half way between the probable and the high. These are assigned initial probabilities of 0.169, 0.206, 0.250, 0.206, and 0.169 corresponding to the areas under the normal curve between 0 and 20%, 20% and 40%, 40% and 60%, 60% and 80%, and 80% and 100%.

The compound distribution is made up of 25 component distributions, each specified by a mean and standard deviation and weighted according to its probability. The five discrete values of each of the mean and the standard deviation give a total of 25 combinations. The probability of any particular combination of mean and standard deviation is obtained by multiplying the individual probabilities and normalizing to make the sum of all the probabilities add up to 1.0. The initial probability matrix is shown in Table 1. When additional data is inputted, the relative likelihood of obtaining this data is computed for each of the 25 sets of parameters and stored in a 5 × 5 likelihood matrix. The likelihood matrix can also be looked upon as showing the relative likelihood of each of the component distributions being correct, given the data.

The likelihood matrix is used to update the original probability matrix by multiplying corresponding elements and then renormalizing to make the new (posterior) probabilities add up to 1.0, i.e.

$$P'_{i,j} = \frac{P_{ij} L_{i,j}}{N} \qquad (2)$$

in which $P'_{i,j}$ = posterior probability (weight) assigned to combination i, j; i = index of value of mean (from 1–5); j = index of standard deviation value (from 1–5); P_{ij} = prior probability of combination i, j; L_{ij} = likelihood of combination of i, j given the data; and N = normalizing factor = $\Sigma_{i=1}^{5} \Sigma_{j=1}^{5} P_{ij} L_{ij}$.

The program has the capability to fit a bivariate normal distribution to the posterior probability distribution. The user can elect to have the program refit distributions after each updating, only after the probability of one cell in the matrix goes to zero, or not at all. Refitting allows the program to "follow the evidence" and change the parameter values as well as the weights of the component distributions. This prevents it from being "locked in" on the initially selected parameter values.

The likelihood depends on the type of data. In the case of a single recorded peak flood, it is given by the height of the probability density function, i.e.:

$$L_{i,j} = \frac{1}{\sigma_j} \exp\left[-\frac{1}{2}\left(\frac{\log x - m_i}{\sigma_j}\right)^2\right] \qquad (3)$$

in which $L_{i,j}$ = likelihood of mean m_i and standard deviation σ_j given data x (the constant term $1/2\pi$ is omitted).

When there are several recorded flows, the likelihood of each combination is the product of the likelihoods computed from the individual pieces of data, i.e.

$$L_{i,j}(x) = \Pi L_{ij} x_k \quad k = 1, \ldots n \qquad (4)$$

When the user does not wish to enter prior estimates of the mean and standard deviation of the peak flows, these can be estimated from the peak flow data directly using standard formulas (2,3,10). In this case, there is no need to update the probability matrix with the recorded flood data since it has already been taken into account.

For information on flows exceeded or not exceeded a given number of times, the likelihoods are based on the probability of a flow being exceeded in any one year:

$$P = 1 - \frac{1}{2\pi\sigma_j} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{\log x - m_i}{\sigma_j}\right)^2\right] dx \qquad (5)$$

in which P = probability that the flow, x, will be exceeded in any one year given the distribution with mean m_i and standard deviation σ_j.

The likelihood of the flow, x, not being exceeded in n years is $(1 - P)^n$, and of a flow, x, being exceeded m times in n years is $P^m(1 - P)^{n-m}$.

When flow data are provided in the form of low, probable, and high values the likelihood is given by

$$\sum_{k=1}^{k=5} w_k L_{i,j}(Q_k) \quad \ldots \quad (6)$$

in which w_k = the weights, as given previously; Q_k = the five discrete values L, $1/2(L + P)$, P, $1/2(P + H)$, H for k = 1–5; and $L_{ij}(Q_k)$ = likelihood of data Q_k, given mean m_i and standard deviation σ_j.

When weighted data are provided, the likelihood is given by $(L_{i,j})^w$ in which w = weight assigned to the data; and L_{ij} = likelihood of data given mean m_i and standard deviation σ_j. In the case of data obtained by regression, R^2 (in which R = coefficient of correlation) provides a good approximation to the weight which should be used.

Examples

Fig. 2 shows the difference between peak floods estimated by the preceding program and corresponding values computed from the theoretical Bayesian

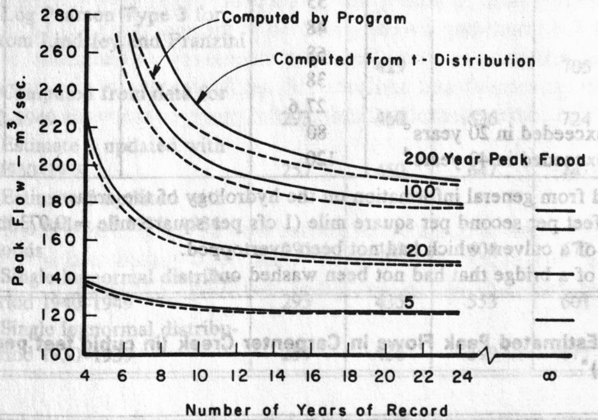

FIG. 2.—Variation in Peak Flow Estimate with Length of Record

distribution, for a hypothetical case with a mean of 100 m^3/s and a standard deviation of 25%. For the Bayesian distribution, both the mean and the standard deviation are assumed to be unknown random variables with a normal gamma distribution (7). The resulting Bayesian distribution of the peak floods is a t-distribution with $n - 2$ degrees-of-freedom (3). For both cases, it was assumed that there was no prior information (which meant that means and standard deviations had to be estimated from the data), that the underlying distribution was lognormal, and that the mean of the logarithms was always 2.0 (log 100), and the standard deviation 0.0969 (log 1.25), regardless of how many pieces of data (years of record) there were. From Fig. 2 it can be seen that the discrepancy between the two sets of graphs is quite small. This shows that relatively little error is introduced by "discretizing" the distributions of the mean and the standard deviations, except when there are few records in which case the results are not reliable anyway. Fig. 2 also shows that the compound distribution automatically introduces a degree of conservatism due to parameter uncertainty, when the parameters have been estimated from a small sample.

As another example, the program was used as an aid in estimating the 100 year flood which was required for designing a diversion on Carpenter Creek, a small creek in British Columbia, Canada. The available data are shown in Table 2.

TABLE 2.—Carpenter Creek Peak Flood Data

Flood variable[a] (1)	Peak[b] (2)	Low (3)	Probable (4)	High (5)
Estimated mean annual flood[a]	—	25	40	60
Estimated coefficient of variation	—	0.2	0.3	0.45
Year of record				
1914	14.9			
1915	38			
1916	75.6			
1954	55			
1955	48			
1956	68			
1957	38			
1958	27.6			
Flow not exceeded in 20 years[c]	80			
Flow not exceeded in 40 years[d]	120			

[a]Estimated from general information on the hydrology of the area.
[b]In cubic feet per second per square mile (1 cfs per square mile = 0.073 $m^3/s/km^2$).
[c]Capacity of a culvert which had not been overtopped.
[d]Capacity of a bridge that had not been washed out.

TABLE 3.—Estimated Peak Flows in Carpenter Creek (in cubic feet per second per square mile)[a]

Average return period, in years (1)	2 (2)	5 (3)	10 (4)	50 (5)	100 (6)	200 (7)
Estimate 1[b]	40.0	55.5	63.5	82.7	91.9	101.4
Estimate 2	38.8	54.4	64.8	87.9	97.9	107.8
Estimate 3[c]	31.3	45.3	63.4	74.9	83.6	92.2
Estimate 4	31.3	45.2	63.3	74.9	83.5	91.1

[a]From the estimates of the mean and coefficient of variation.
[b]Updated with flow data; updated with flow data including relative flood magnitudes (such as largest in 60 years). Updated with information on flows not exceeded in a number of years.

Experience with the area indicated the Gumbel distribution to be suitable, so it was selected for use and the program was run with the aforementioned data. Then, following a closer examination of other gage data from the surrounding area, it was concluded that the 1916 flood was probably the biggest in 60 years, the 1956 flood probably the biggest in 30 years, and the 1914 flood probably the second lowest in 56 years. The program was then rerun using this information (results are shown in Table 3).

As a third example, tests were run on peak flood data on the Susquehanna River, as given by Linsley and Franzini (6). The lognormal distribution was selected and the program was run using data for the period 1940–1949. Estimates of the mean and standard deviation were computed by the program, so there was no subjective input. Additional runs were made updating the estimates with data from the 1930–1939 period, first by simply inputting this additional data, and then by taking into account the fact that the 1934 flood was the lowest, the 1931 flood the second lowest, and the 1936 flood the highest in 76 years of record (results are given in Table 4).

TABLE 4.—Estimated Peak Floods at Susquehanna River

Average return period, in years (1)	2 (2)	10 (3)	50 (4)	100 (5)	200 (6)
Estimate 1—Gumbel distribution for 76 years, from Lindsley and Franzini (6)		435		640	
Estimate 2—Log Pearson Type 3 for 76 years, from Lindsley and Franzini (6)		429		705	
Estimate 3—Computed from data for period 1940–1949	293	460	636	724	820
Estimate 4—Estimate 3 updated with data from 1930–1939	257	450	647	741	843
Estimate 5—Estimate 3 updated with data, including order of 1931, 1934 and 1936 floods	295	455	604	672	745
Estimate 6—Single lognormal distribution for period 1940–1949	293	435	553	601	649
Estimate 7—Single lognormal distribution for period 1930–1939	257	436	599	671	743

Analysis

The preceding examples all demonstrate that when data are scarce, use of the compound distribution which accounts for parameter uncertainty, provides conservative estimates especially with low frequency floods, a reassuring feature for practicing hydrologists. The second and third examples show that additional information on the rank of recorded floods helps to refine the estimates and generally reduce the degree of conservatism, thus demonstrating the value of additional information. The second example also shows that accurate but irrelevant information, such as the fact that very large floods were not exceeded in a limited number of years, has little effect on the results.

Some psychologists (4) maintain that we tend to think along Bayesian lines which suggests that the results of the procedure described in this paper should agree with our intuition. The writer has found the described program to be very helpful in dealing with practical problems. When results turned out as anticipated, they provided reassurance and useful documentation. When they did not, closer examination usually disclosed discrepancies in the input data or additional information that was not being taken into account. When adjusted accordingly, the program gave intuitively satisfying results. The program is not a substitute for engineering judgement but a helpful supplement to it, and a useful learning tool.

Conclusion

A procedure has been presented for estimating peak floods in situations where data are sparse and all available information must be used. The basic concept is to use a compound distribution and update the parameters of the underlying distribution using Bayes' theorem. This allows subjective information derived from the hydrologist's experience to be combined, in an explicit way, with information on recorded floods. This information can be in a variety of forms, such as actual recorded floods or flows which have not been exceeded in a given number of years. Available analytical procedures have too many limitations to be useful for practicing hydrologists, so numerical methods have to be used. A computer program to carry out the necessary calculations is described. Examples are presented and analyzed. A useful feature of the approach is that the sparser and more uncertain the data, the more conservative the results.

Acknowledgment

The writer wishes to thank John Bonser who developed the computer program described in this paper, and Andrew Chan who tested and applied it to the estimation of design floods in the Yukon Territory, Canada.

Appendix I.—References

1. "A Uniform Technique for Determining Flood Flow Frequencies," *Bulletin No. 15*, United States Water Resources Council, Washington, D.C., 1967.
2. Ang, A. H-S., and Tang, W. H., *Probability Concepts in Engineering Planning and Design, Vol. I. Basic Principles*, John Wiley and Sons, Inc., New York, N.Y., 1975.
3. Benjamin, J. R., and Cornell, C. A., *Probability, Statistics, and Decision for Civil Engineers*, McGraw Hill Book Co., Inc., New York, N.Y., 1971.
4. Broadbent, D. S., *In Defense of Empirical Psychology*, Methuen, London, England, 1973.
5. "Guidelines for Determining Flood Flow Frequency," *Bulletin No. 17*, Hydrology Committee, United States Water Resources Council, Washington, D.C., Mar., 1976.
6. Linsley, R. K., and Franzini, J. B., *Water-Resources Engineering*, McGraw-Hill Book Co., Inc., New York, N.Y., 1979.
7. Raiffa, H., and Schlaifer, R., *Applied Statistical Decision Theory*, Massachusetts Institute of Technology Press, Cambridge, Mass., 1968.
8. Rascon, O. A., and Villarreal, A. G., "On a Stochastic Model to Estimate Tsunami Risk," *Journal of Hydraulic Research*, ASCE, Vol. 13, No. 4, 1975, pp. 383–404.
9. Tang, W. H., "Bayesian Frequency Analysis," *Journal of the Hydraulics Division*, ASCE, Vol. 106, No. HY7, July, 1980, pp. 1203–1218.
10. Walpole, J., and Myers, K., *Probability and Statistics for Engineers and Scientists*, The MacMillan Co., New York, N.Y., 1972.
11. Wood, E. F., and Rodriquez-Iturbe, I., "Bayesian Influence and Decision Making

for Extreme Hydrologic Events," *Water Resources Research*, Vol. 11, No. 4, Aug., 1975, pp. 533–542.
12. Yevjevich, V., *Probability and Statistics in Hydrology*, Water Resources Publications, Fort Collins, Colo., 1972.

Appendix II.—Notation

The following symbols are used in this paper:

i = index;
j = index;
k = index;
L = likelihood;
m = mean;
N = normalizing constant;
P = probability;
Q = peak flow;
R = correlation coefficient;
x = peak flow data; and
σ = standard deviation.