THE WORTH OF DATA IN HYDROLOGIC DESIGN

M. E. Moss, U.S. Geological Survey, Washington, D.C.; and
D. R. Dawdy, U.S. Geological Survey, Menlo Park

Reviews of current methodologies for the determination of the worth of hydrologic data indicate that each method has certain shortcomings. The simulation approach requires information concerning the statistical properties of the data that are not usually known. The Bayesian approach often leads to mathematically intractable relations. The two approaches may be combined by defining the data worth through simulation under the condition of assumed statistical properties and releasing the conditioning by making use of the prior distributions of the unknown statistics obtained from the Bayesian approach. This combined approach circumvents the problems encountered when either the Bayesian or the simulation approach is used exclusively. An example illustrates the use of the combined approach in evaluation of the worth of flood data in the design of highway crossings.

THE CONCEPT of the monetary worth of data is one basis for answering questions such as how many data are sufficient to make an economically optimal design decision. This concept is an aspect of hydrologic analysis that has only recently received direct attention (1, 2, 5). Because a methodology for the determination of hydrologic data worth has remained unstructured for so long, the designer in most cases uses available data to formulate his design without assessing the possibility that collection of additional data might lead to a better design in the long run. As more and more pressures are exerted on land and water resources by the expanding economies of the world, design decisions must be improved, and one way to do this is through the collection of optimal quantities of data on which the decisions will be based.

Data in general have value or worth only if they are used in a decision process. Hydrologic data are no exception. Anticipation of the uses of data is thus a prerequisite of any statement concerning their worth. Those data that are never used or are only used improperly in hydrologic design have a maximum worth of zero.

The randomness of streamflow data tends to cloud the definition of their worth. Because the data are governed by the laws of probability, only probabilistic inferences can be made concerning their reliability in the design process, and the true optimal design is never known. However, as more data are collected, the design based on the sample approaches the true optimum in a statistical sense. Each data point that is collected has a current worth equal to its degree of improvement of the design. Because the optimal design, which is the measure by which improvement is gauged, is unknown, only probabilistic statements can be made concerning data worth. The expected value of data worth is a parameter that can be used to define the optimal alternative between additional data collection and immediate design.

The worth of a data set may be defined as the benefits foregone if it were not available. In the context of design, this definition is interpreted as the difference in net benefits between designs including and omitting the data being evaluated. Dawdy, Kubik, and Close (2) have used the benefits-foregone definition in conjunction with synthetic hydrology to evaluate data worth in reservoir design. They synthesized a 500-year realization of streamflow from which the optimal storage capacity of the reservoir was determined. The 500-year record was then partitioned into fifty 10-year records, twenty 25-year records, ten 50-year records, and five 100-year records. Reservoir capacities were designed for each record, and net benefits based on the 500-year record were computed for each design. Average net benefits for each class of record lengths were computed. The differences in average net benefits between two successive classes of record length were computed and equated with the expected value of the increment of data between the pairs of classes. Dawdy and others found that the unit worth of data decreases with increasing record length; i.e., the marginal value of data decreases with an increase in the level of information gained. Similar findings were reported by Tschannerl (5).

The simulation method just described can be adapted to any hydrologic design problem by the use of pertinent design procedures and the related cost and benefit functions. It does, however, have one serious restriction: The statistical moments of the distribution of the hydrologic variables must be known prior to the analysis. The method is thus limited to either an after-the-fact analysis, when sufficient data have already been collected, or a conditional analysis with the understanding that the answers have only a certain probability of being correct.

Davis (1) developed a methodology, based on decision theory and subjective probabilities, that avoids the inherent restriction of the simulation approach. The subjective approach is statistically elegant, but it often requires very complex mathematics in order to arrive at an answer. It treats the unknown population parameters as random variables with probability distributions defined by prior knowledge and available data. From these probability distributions a design is formulated that takes into account the randomness of the unknown parameters. The design and prior distributions are used to estimate the expected opportunity loss that is caused by imperfect knowledge about the parameters. The opportunity loss, which is also an unknown that is treated as a random variable, is the equivalent of the worth of an added record of infinite length. The prior distributions are next used to evaluate the expected value of the expected opportunity loss after the addition of one more unit of data. The difference between this expected expected opportunity loss and the expected opportunity loss is the expected worth of the added unit of data. The subjective approach thus estimates data worth for a single step into the future. Attempts to extend the method to multiple steps are thwarted by the complexity of the mathematics, which increases in degree equal to the number of steps considered.

Each of the methodologies thus have their benefits and their deficiencies. Because none of the deficiencies mentioned above is common to both, it is possible to combine the two methodologies and arrive at a blend that is superior to either used alone. This is accomplished by defining the expected values of data worth under the condition of known statistical parameters by the simulation method. The conditioning is then released by use of the prior distribution of the parameters, which is derived from the subjective approach. This procedure is illustrated in the following hypothetical example.

Suppose that a new highway is to be constructed between two towns in western Oregon and that this highway must cross 10 water courses that have basins upstream of the proposed crossings that are described by the information given in Table 1. The engineer's design manual tells him to design the crossing structure so that it will pass the discharge of the 50-year flood without incurring any damage. There are no discharge records on any of the streams, so the engineer decides to use regression estimates (4) of the 50-year floods as his design discharges. The regression estimates, of course, contain errors, in this case a 40 percent standard error of estimate; and the resulting designs will be suboptional. The suboptional designs result in benefits foregone in the form of added construction costs for overdesign and added maintenance and replacement costs in the case of underdesign. These benefits foregone are the potential worth of the uncollected data.

Assume that the present value at some specified discount rate of benefits foregone at each crossing is related to the difference between the logarithms of the design discharge Q_D and the true 50-year flood Q_{50} by the function shown in Figure 1, which could be defined by simulation of long flood records and by computing the resulting costs. The present cost C_D of the structure if Q_D were an optimal design is related to the design discharge in the manner shown in Figure 2.

Implicit in the logarithmic regression analysis that was used to define the value of Q_D is the assumption that the residuals or in this case the errors of the logs of the dependent variables are distributed normally about the logs of the regression estimates.

Under such an assumption, Hardison (3) has provided the means for converting from standard error of estimate in log units to equivalent years of record; i.e., a certain standard error from the regression analysis can be expressed as the number of years of actual data that would have to be analyzed to give an estimate of the parameter that had an accuracy equal to that of the regression estimate. Making use of this analogy reduces the simulation phase to the definition of Figure 1.

The assumption of log normality for the residuals from the regression designates that the log-normal distribution should be the family of the prior distribution of unknown true values of Q_{50}. The prior distribution of the logarithms of Q_{50} would thus be normal with a mean of log Q_D and a standard deviation equal to the standard error of estimate of the regression analysis. The standard error of estimate rather than the standard error of prediction is used only for computational convenience and for demonstration purposes.

The expected opportunity loss can now be computed for each of the crossings by the integral equation

$$XOL(i) = \int_0^{\infty} P_i(Q_{50}, Q_D)\, f_i(Q_{50})\, d\,Q_{50} \qquad (1)$$

where

$XOL(i)$ = expected opportunity loss at the ith crossing,
$P_i(Q_{50}, Q_D)$ = penalty function or benefits foregone for a design at the ith crossing based on Q_D when the true value is Q_{50}, and
$f_i(Q_{50})$ = probability density function of the prior distribution of Q_{50}.

Transforming the values of Q_{50} and Q_D to their logarithms and taking note of the form of the penalty function

$$P_i(Q_{50}, Q_D) = C_D(i)\, p(\log Q_D - \log Q_{50}) \qquad (2)$$

expand Eq. 1 to

$$XOL(i) = C_D(i) \int_{-\infty}^{\infty} p(\log Q_D - \log Q_{50})\, f_N(\log Q_D - \log Q_{50})\, d(\log Q_{50}) \qquad (3)$$

where

$p(\log Q_D - \log Q_{50})$ = function of Figure 1 and
$f_N(\log Q_D - \log Q_{50})$ = normal prior distribution of log Q_D.

The integral of Eq. 3 is independent of the value of log Q_D because the distribution of log Q_{50} is always symmetrically distributed about it. The value of the integral is thus a function only of the standard error of estimate of log Q_{50}. Equation 1 can be further simplified

$$XOL(i) = C_D(i)\, I(s) \qquad (4)$$

where I(s) = integral of Eq. 3, which depends on s, the standard error of log Q_{50}. The relation of I(s) to s is shown in Figure 3 along with the equivalent lengths of record.

The total expected opportunity loss for all 10 sites is the sum of the individual values. Because the integral of Eq. 3 has a common value at each crossing, the total expected opportunity loss is expressed as

$$XOL^{*} = I(s) \sum_{i=1}^{10} C_D(i) \qquad (5)$$

Table 1. Basin parameters, design discharges, and theoretical costs for stream crossings.

Station	Drainage Area (square miles)	Storage (percent + 1)	Elevation (10^3 ft)	Annual Precipitation (in.)	Temperature Index (deg F)	Soils Index[a]	Q_D[b] (cfs)	C_D (thousands of dollars)
1	22.3	1.67	2.44	96	29	5.6	3,830	49.0
2	52.7	1.06	4.09	58	24	5.6	4,600	56.0
3	73.0	3.56	2.99	120	29	5.6	13,400	113.0
4	107	2.77	2.62	115	29	5.6	17,600	136.0
5	113	2.15	4.46	60	21	5.5	5,750	64.0
6	117	1.51	4.14	62	24	5.0	9,870	92.0
7	246	5.23	3.76	58	25	4.6	13,200	112.0
8	258	1.18	4.38	60	24	5.4	20,900	152.0
9	392	1.34	4.08	58	24	5.2	27,300	182.0
10	924	2.38	3.97	58	25	4.9	56,000	294.0
Total								1,250.0

[a] Compiled from U.S. Soil Conservation Service maps.

[b] $Q_D = 1.70 \times 10^{-4} A^{0.903} S_t^{-0.310} E^{0.741} P^{1.94} T_t^{2.714} S_i^{-0.660}$.

Figure 1. Penalty function for inaccuracies in the 50-year flood.

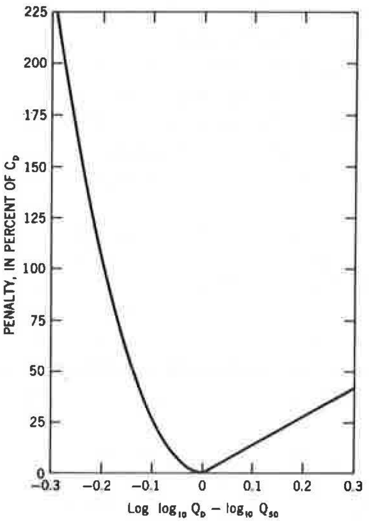


Figure 2. Cost function for optimal design.

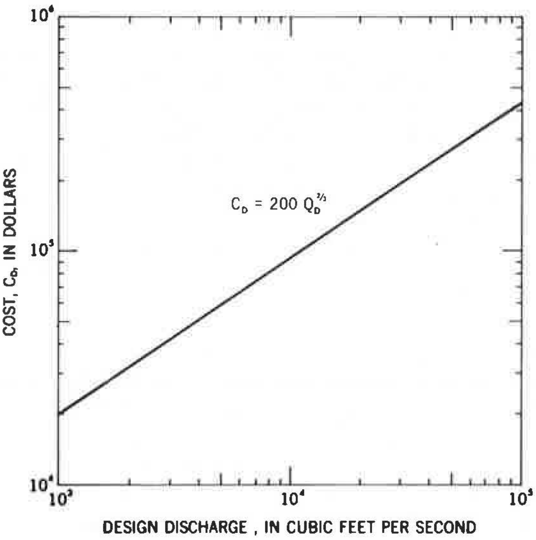
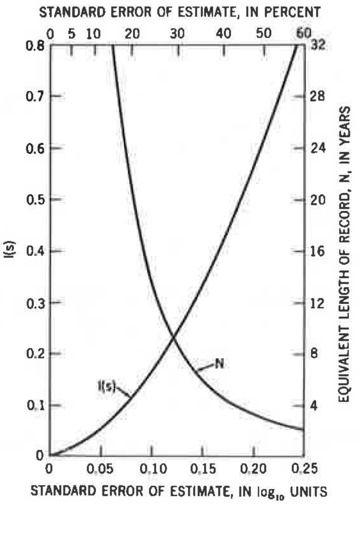


Figure 3. Equivalent length of record and integral from Eq. 3 as functions of standard error of estimate.



The regression analysis of Lystrom (4) yields a standard error of estimate for the 50-year flood of 40 percent, which is equivalent to 0.17 log units. The regression estimates of Q_{50} that were derived from the basin parameters of Table 1 are also given in that table. If these data are used in conjunction with Figures 2 and 3, Eq. 5 can be evaluated for the example. The expected opportunity loss is $524,000, which is the expected present worth of perfect knowledge concerning values of Q_{50} at each crossing.

Because perfect knowledge is an unattainable state, the expected opportunity loss is only a gauge by which the worth of additional data can be measured. The expected change in expected opportunity loss at any point in time of a data collection program is the expected worth of the additional data that might be added to that already collected up to that time. For instance, if a data program that provided the equivalent of 10 years of data were instigated, the 50-year flood would have an accuracy of 28 percent, according to Lystrom (4); a 25-year equivalent record would have an accuracy of 16 percent. Using these two figures in the analysis gives expected expected opportunity losses of $286,000 and $115,000 for the 10- and 25-year records. When each of these expected expected opportunity losses is subtracted from the expected opportunity loss of the previous paragraph, the expected worth of the 10-year records is found to be $238,000, and the 25-year records are found to be worth $409,000. The difference between these two values, $171,000, is the added worth of extending the records from 10 to 25 years.

The above analysis can be repeated for each consecutive year of data added beyond the currently available 5-year equivalent. By the use of 1-year increments the marginal worth of each added year of record can be computed. Comparison of the marginal worth with the cost of obtaining the data indicates the optimum record length or its equivalent that should be collected for the design. Data should be collected as long as their expected marginal worth exceeds their expected marginal cost. The expected marginal costs must include the costs of benefits foregone by delaying the design to obtain the added data.

It would be very rare, indeed, if the value of additional hydrologic data were sufficient to delay the construction of a highway. However, because the design floods contain certain unknown errors, it may be economically efficient to consider the collection of data in order to improve the estimates for those structures that will have to be replaced in the future. If the designer decides to make an investment in data, he has at least two alternatives at his disposal. He can invest in a gauging program at the present crossing sites, or he can help finance improvement of the regression analysis, thereby making better estimates available in the future.

The at-site gauging program could be accomplished by the use of relatively economical crest-stage gauges; but, because it is not known at which sites the information will be needed, the program might result in a gauge at each site. This in not necessarily the case, however, because there usually is some transferability contained in the information gathered at nearby sites. If the transferability is sufficient, a less-than-saturation gauging program will be optimal.

Improvement of the regression estimates is probably the most effective means of investment. It not only yields better estimates at the sites in question but also improves those throughout the region in which the regression analysis is valid. Therefore, other suboptimal crossing structures in the region may be improved as they are replaced, and new structures will also benefit from the added level of regional information.

Improvement of the current version of the regression model may be accomplished by reducing any one of the three types of errors that are contained therein: time-sampling, space-sampling, or modeling errors. Reduction in time-sampling errors is the result of extending the streamflow records at the sites that were used in the current analysis, thereby improving the estimates of the streamflow statistics, in this example Q_{50}, that are used as dependent variables.

Decreased space-sampling error is effected by the initiation of new gauging stations at sites that will provide a broader range of variables for the analysis. The new stations should not be included in the analysis, however, until their time-sampling errors are small enough so that they will not add more noise than information. In the current example, installation of gauges at one or more of the proposed crossings would serve the double purpose of reducing space-sampling error in the regression analysis and providing at-site records for the chosen crossings.

Modeling errors can be reduced only by improving the structure of the model. This may often be done by including additional independent variables, in which case data describing the new variables are required.

At the present time it is not possible to state which source of error should be attacked first. However, the economic efficiencies of data collection as a means of reduction of space- and time-sampling error are being investigated, while further developments in the understanding of the physical processes of hydrology indicate that improvements in the structure of the regression models also are forthcoming.

REFERENCES

1. Davis, D. R. Decision Making Under Uncertainty in Systems Hydrology. Hydrology and Water Resources Dept., Univ. of Arizona, Tucson, Tech. Rept. 2, 1971.
2. Dawdy, D. R., Kubik, H. E., and Close, E. R. The Value of Streamflow Data for Project Design—A Pilot Study. Water Resources Research, Vol. 6, No. 4, 1970, pp. 1045-1050.
3. Hardison, C. H. Accuracy of Streamflow Characteristics. In Geological Survey Research 1969, U.S. Geological Survey, Prof. Paper 650-D, 1969, pp. D210-D214.
4. Lystrom, D. J. Evaluation of the Streamflow-Data Program in Oregon. U.S. Geological Survey, Open-file report, 1970.
5. Tschannerl, G. Designing Reservoirs With Short Streamflow Records. Water Resources Research, Vol. 7, No. 4, 1971, pp. 827-833.

DISCUSSION

K. C. Wilson and W. E. Watt, Queen's University, Kingston, Ontario

Matters of considerable practical importance are dealt with in the paper by Moss and Dawdy. In particular, the basic concept that the worth of hydrologic data should be expressed in design-directed economic terms has not received adequate attention in the past, and this concept is well worth stressing. The point was expressed succinctly by Linsley (6) in 1965: "The collection of data is not an end in itself. Data are collected for use and the requirements of this use should govern network design."

Not long afterward, the Canadian Department of Energy Mines and Resources (now the Department of the Environment) commissioned various studies directed, in part, toward developing a use-oriented, cost-benefit approach to hydrometric network planning. Techniques developed in Canada, largely as a result of these studies, have recently been given a broader audience through papers presented in the United States. In several of these papers (7, 8, 10) it has been remarked that on a regional basis the economic loss resulting from uncertainties in hydrologic data varies approximately as the square of the relative uncertainty.

In this regard, it is interesting to note that the same parabolic relation between relative economic loss and relative uncertainty can be shown to apply to the results of Moss and Dawdy. Their function I(s) can be taken as the relative loss (Eq. 5), and the standard error of estimate, in percent (the upper abscissa of Fig. 3, which is assigned the symbol S_r for the purposes of this discussion), can be taken as the relative uncertainty. The parabolic approximation $I(s) = KS_r^2$, with a K-value of about 2.7×10^{-4}, provides a reasonable approximation to the curve labeled I(s) in Figure 3.

It may also be remarked that in the paper by Moss and Dawdy the method of calculating expected opportunity loss would appear to depend on a preexisting knowledge of the optimal return period as given by "the engineer's design manual" or the subsequent assumption by which the "true 50-year flood" is taken as a datum. However, if one wishes to obtain a scientific basis for the design of hydraulic structures, the optimal return period must itself be taken as a dependent variable. In a previous paper concerned with design strategy for small flood-control structures (9), we showed that the optimal return period would differ according to whether the "perfect knowledge" or "real-world" case was being considered. In our paper, an equation was first obtained that expresses the perfect knowledge return period in terms of factors such as interest rate, damage ratio (cost of typical damage event/cost of optimum structure), and flood-frequency exponent. This case is denoted as perfect knowledge because it assumes that the true values of the various factors are known. The analysis was then extended to the real-world case where uncertainties in the component factors combine to give an overall uncertainty in estimating the optimum. It was found that the penalty resulting from underdesign caused by this uncertainty is greater than that for an equivalent overdesign, and hence the real-world optimum should be larger than the estimated perfect knowledge value.

The example given in Moss and Dawdy's paper will also impose a larger penalty for underdesign, as indicated by the relatively greater steepness of the left limb of Figure 1. On the basis of our paper (9), it is suggested that in this case the authors have used a perfect knowledge design strategy for a real-world case. If the real-world strategy had been used, the design discharge would have been progressively increased with increasing uncertainty. This apparent overdesign ensures that the probability of occurrence of very large damage (points far to the left on Fig. 1) remains reasonably small. It follows that the use of this real-world design strategy will produce a smaller estimated opportunity loss than the strategy adopted by Moss and Dawdy.

REFERENCES

6. Linsley, R. K. Symposium on Hydrometeorological Networks: A Summary. International Association of Scientific Hydrology, Publication 68, 1965, pp. 809-814.
7. Solomon, S. I. Natural Regime Hydrological Network Planning: A Methodological Discussion. Proc. International Symposium on Uncertainties in Hydrologic and Water Resources Systems, Tucson, 1972, pp. 473-492.
8. Watt, W. E., and Wilson, K. C. An Economic Approach for Evaluating the Adequacy of Hydrologic Data. Proc. Second International Hydrology Symposium, Fort Collins, 1972.
9. Watt, W. E., and Wilson, K. C. Optimal Design Strategy for Small Flood-Control Structures. Proc. International Symposium on Uncertainties in Hydrologic and Water Resource Systems, Tucson, 1972, pp. 903-916.
10. Wilson, K. C. Cost-Benefit Approach to Hydrometric Network Planning. Water Resources Research, Vol. 7, No. 5, 1972, pp. 1347-1353.

Gene E. Willeke, Environmental Resources Center, Georgia Institute of Technology

Moss and Dawdy have presented an approach to evaluating the worth of data in hydrologic design that is generally appropriate. As is usually the case in such studies, the difficulty lies in applying the conceptual approach to a real problem. An additional difficulty is that the person appraising the worth of data is rarely in a position to influence data collection programs in any significant way.

One example illustrates the difficulty in applying the approach to a real situation. The authors use a regression analysis that yields values of Q_{50}. Then the expected opportunity loss is calculated. It is difficult to decide whether such a calculation is better than no calculation at all, because the losses that would occur from floods whose peak discharge exceeds Q_{50} depend more on the hydrograph than on the peak discharge. A stream crossing, bridge or culvert, does not automatically fail when a given peak discharge occurs. Ponding, duration of discharge above that peak discharge, and site characteristics must be considered. This does not imply that regression equations have no place in design. They may provide a quick and adequate means of sizing a waterway opening. A full analysis that included consideration of the full hydrograph would yield a better design, from an economic standpoint, but not necessarily a safer design.

Data collection programs are in need of optimization, especially in the area of coordinating the collection of precipitation and runoff data, perhaps our greatest weakness in data collection at the present time. The worth of the data should have some bearing on decisions to optimize those data collection programs. However, it still seems that our larger needs are to find logical, defensible methods of using the data we have in design of hydraulic structures. This would allow a much better appraisal of the worth of the data. Then, the approach suggested by the authors (which may be ahead of its time) can be fruitfully used.

AUTHORS' CLOSURE

We appreciate the discussions of our paper by Wilson and Watt and Willeke. In general we agree with their comments, which may be considered as extensions of our ideas. Wilson and Watt point out that the loss function relation is quadratic. This was a basic assumption, rather than a conclusion, because a quadratic average loss function was considered as relatively realistic. This can in no way, however, then be used to justify that assumption. We agree that the use of a 50-year design flood was an oversimplification. The design discharge itself should be a decision variable, but we felt that explicit consideration of that problem would have obscured the major point, which was that data networks should be judged in economic terms. The uncertainty in knowledge of the proper design flow affects the results of our study, just as Wilson and Watt outline, by causing hedging against underdesign. Introducing the design discharge as a decision variable will reduce the estimated expected opportunity loss (11).

As stated earlier, the loss function is an assumption and an oversimplification of the facts. As Willeke pointed out, it may be unrealistic. However, it is a first step. We should not refuse to walk merely because we cannot win an Olympic Gold Medal. More realistic loss functions will lead to better decisions. Until they are developed, any "optimization" in fact is suboptimal. We hope we can start to converge toward more defensible and, perhaps, better information networks.

REFERENCE

11. Close, E. R., Beard, L. R., and Dawdy, D. R. An Objective Determination of Safety Factor in Reservoir Design. Jour. Hydraulics Div., Proc. American Society of Civil Engineers, No. HY5, May 1970, pp. 1167-1177.