UNIT HYDROGRAPHS AS A MULTIVARIATE NORMAL DISTRIBUTION

Robert J. Whitley[1] and Theodore V. Hromadka II, M.ASCE[2]

ABSTRACT

The use of a given effective rainfall and a stochastic integral equation formulation of the well-known unit hydrograph method gives criterion design variables, such as volume or maximum discharge, which are random variables depending on the stochastic variation in the unit hydrographs. This variation can be probabilistically modeled by means of a multivariate normal distribution. With this approach, the total runoff volume is normally distributed and confidence intervals for this design variable can then be directly obtained. A computer simulation can be used to obtain confidence intervals for the maximum discharge estimate. Similarly, probabilistic simulation can be used to develop confidence intervals for other criterion variables.

[1] Professor, Department of Mathematics, University of California, Irvine, California

[2] Professor, Department of Applied Mathematics, California State University, Fullerton, and Director of Water Resources Engineering, Williamson and Schmid, Irvine, California

DISCUSSION OF THE MODEL (STOCHASTIC INTEGRAL EQUATION METHOD)

The unit hydrograph method is a widely used rainfall-runoff modeling technique. In this paper, we consider how to include uncertainty in the predictions of runoff obtained by this modeling approach. We consider a variant of the unit hydrograph method which relates the effective rainfall $e(\cdot)$ (i.e., rainfall less losses) and the discharge $Q(\cdot)$ via the stochastic integral equation,

$$Q(t) = \int_0^t e(t-s)\, \eta(s)\, ds \tag{1}$$

where $\eta(\cdot)$ is a realization of a stochastic process distributed as $[\eta(\cdot)]$.

Our analysis begins by dividing the study time interval $[0,T]$ into N equal subintervals $I_n = (t_{n-1}, t_n)$, with $T_n = nT/N$ for $n = 0,1,\ldots,N$, and approximate $e(\cdot)$ by a step function with the constant value e_n on the interval I_n. Letting $\chi(a,b)$ be the characteristic function of the interval (a,b) defined by

$$\chi_{(a,b)}(t) = \begin{cases} 1, \text{ if } a < t < b \\ 0, \text{ otherwise} \end{cases}$$

$e(\cdot)$ can be written:

$$e(t) = \sum_{n=1}^{N} e_n \chi_{I_n}(t). \tag{2}$$

In the same fashion, approximate the realization $\eta(\cdot)$ from $[\eta(\cdot)]$ by a function with constant value η_n on the interval I_n. Substituting these approximations for $e(\cdot)$ and $\eta(\cdot)$ into Eq. (1) gives

$$Q(t) = \sum_{n=1}^{N} e_n \left[S(t-a_{n-1}) - S(t-a_n) \right] \tag{3}$$

where S(t) is the S-graph

$$S(t) = \int \eta(s)\, ds.$$

Thus Q(t) can be seen to be continuous and piecewise linear, with the derivative Q'(t) taking on a constant value, say q'_n, on the interval I_n.

To determine the values $\{q'_n\}$, differentiate Eq. (3) and choose t to be a point in I_n for n = 1,2,...,N, giving N equations:

$$\begin{aligned}
q'_1 &= e_1 (\eta_1 - \eta_o) \\
q'_2 &= e_1 (\eta_2 - \eta_1) + e_2 (\eta_1 - \eta_o) \qquad (4) \\
&\cdots \\
q'_N &= e_1 (\eta_N - \eta_{N-1}) + e_2 (\eta_{N-1} - \eta_{N-2}) + \cdots + e_N (\eta_1 - \eta_o)
\end{aligned}$$

where $\eta_o = 0$ is used in the formulas for symmetry. These equations can also be rewritten in the form:

$$\begin{aligned}
q'_1 &= (e_1 - e_o)\, \eta_1 \\
q'_2 &= (e_2 - e_1)\, \eta_1 + (e_1 - e_o)\, \eta_2 \qquad (5) \\
&\cdots \\
q'_N &= (e_N - e_{N-1})\, \eta_1 + (e_{N-1} - e_{N-2})\, \eta_2 + \cdots + (e_1 - e_o)\, \eta_N
\end{aligned}$$

with $e_o = 0$.

The problem of modeling the statistical variation in each of the parameter sets $\{q'_1,...,q'_N\}$, $\{\eta_1,...,\eta_N\}$, and $\{e_1,...,e_N\}$, can be considered for various cases; the one which we will consider here is where e(·) is a future storm event; e.g., it is a given design storm effective rainfall. Even for an idealized set of effective rainfall events with identical patterns and magnitudes, there would still be variations in the effective rainfall over the catchment, which would yield observed variations in the associated $Q(\cdot)$, and thereby variations in $\eta(\cdot)$. Consequently, there would be an unique realization, $\eta(\cdot)$, for each data pair of $\{e(\cdot), Q(\cdot)\}$. Because of the random variations in the effective rainfall over the catchment, and the many random processes occurring in any hydrologic rainfall-runoff model, $\eta(\cdot)$ is a stochastic process.

Each value η_n of $\eta(\cdot)$ on the interval I_n is itself a random variable and so the vector $\mathbf{E} = (\eta_1,\ldots,\eta_N)$ is a multivariate random variable. Moreover for small time intervals, say unit periods of five minutes, there will be some dependence between the unit values of $\eta(\cdot)$. This important mutually dependency in the set of components of **E** makes the problem of probabilistic modeling much more difficult. With no strong evidence to the contrary, an appeal to the central limit theorem for multivariate random variables (Brieman, 1968, chap. 11; Billingsley, 1986, page 398) suggests that **E** can be modeled with a multivariate normal distribution. And, in fact, this distribution is one of the few multivariate distributions which is simple enough to allow basic calculations to be made and yet which allows dependence between components.

As Eq. (5) indicates, q'_n is a linear combination of components of **E**. By a known property of multivariate normal distributions, this implies that $\mathbf{Q}' = (q'_1,\ldots,q'_N)$ is also a multivariate normal (Kendall and Stuart, 1977, page 375). Conversely by solving Eq. (5), if **Q'** is a multivariate normal then so is **E**.

A useful fact is that a multivariate normal distribution $X = (x_1,\ldots,x_N)$ is completely determined by its means and its covariance matrix $\gamma_{ij} = E[(x_i - \mu_i) \cdot(x_j - \mu_j)]$, where $\mu_j = E(x_j)$. In fact, in the (usual) simplest case where the covariance matrix $\Gamma = [\gamma_{ij}]$ has an inverse $A = [a_{ij}]$, the density function of **X** is

$$[(2\pi)^N \det(\Gamma)]^{-\frac{1}{2}} \exp\left(-0.5 \sum_{i,j=1}^{N} a_{ij}(x_i - \mu_i)(x_j - \mu_j)\right) \qquad (6)$$

Consequently this density can be estimated by estimating the covariances.

Thus under the model assumptions that either **E** or **Q** is multinormal, the other is also multinormal, and therefore one distribution can be estimated, by using Eq. (5) and estimating the covariance matrix for the other distribution.

We will use this technique to study the statistical properties of predicted Q(·), which are a consequence of the statistical properties of **E**, and the choice of design storm effective rainfall e(·), and so will be able to study some of the statistics of the stochastic integral equation representation of the unit hydrograph model under assumptions which allow realistic dependencies between random components of the processes.

CRITERION VARIABLES

As an example of a runoff criterion variable, consider the total volume of runoff, V. The trapezoidal rule with partition points $t_o, t_1, \ldots, t_N$, is exact for the piecewise linear Q and gives

$$V = \sum_{k=1}^{N} (Q(t_{k-1} + Q(t_k))/2. \qquad (7)$$

Since

$$Q(t_k) = (\delta^2/2) \sum_{j=1}^{k} q'_j, \qquad (8)$$

where δ is the width T/N of the intervals, I_i, V is seen to be a linear combination of q'_1, $q'_2,\ldots,q'_N$, and so therefore is normally distributed. Hence to find confidence intervals for design values of V (which is a prediction of the random variable V given a future effective rainfall), the ordinary statistical methods for a normally distributed random variable apply.

Note that the variation in V, which is characterized as normal is that produced by a fixed given design storm and a variable set of $\eta(\cdot)$ used in the stochastic integral equation formulation of the unit hydrograph method. This is distinct from the variation in V which would be found in runoff volume data from a specific catchment, (should such data be used directly); rather, this observed variation is, to a large extent, due to the variation in the effective rainfall over the catchment with respect to the assumed effective rainfall, among other factors. (As in many applications of the normal distribution, this model is not perfect in that it predicts discharge with negative volumes, but this is only with insignificant probability.)

A criterion design variable of great interest is the peak flow rate. Unlike the case of the total volume of runoff, for the peak flow rate there is no simple derivation of its distribution. To analyze a specific case requires a statistical simulation.

Consider a set of $\eta(\cdot)$, each approximated by constants on the time intervals I_n as was done following Eq. (2). On each interval I_n the values η_n are normally distributed, but that information alone is not enough to determine the joint distribution of the $\eta(\cdot)$ because of the dependence between values on different intervals. The values $\{Q(t_i)\}$, among which the peak flow rate is to be found, are each a linear combination of η_1, $\eta_2,\ldots,\eta_N$ since, as was noted in

Eq. (4), they are linear combinations of $q'_1,\ldots,q'_N$ which, from Eq. (5), are in turn linear combinations of $\eta_1,\ldots,\eta_N$:

$$Q(t_i) = \sum_{j=1}^{N} b_{ij}\,\eta_j. \tag{9}$$

The $\eta_1,\ldots,\eta_N$ will now be regarded as random variables, and we note that $E(Q(t_i)) = \sum_{j=1}^{N} b_{ij}E(\eta_j)$, so that if we subtract the expectation from each η_j this will subtract the expectation from each $Q(t_i)$ and

$$Q(t_i) - E(Q(t_i)) = \sum_{j=1}^{N} b_{ij}(\eta_j - E(\eta_j)) \tag{10}$$

Consider the random variables

$$X_i = Q(t_i) - E(Q(t_i)) \tag{11}$$

These have a multivariate normal distribution and each has zero expectation. The covariance matrix C for these $X_1,\ldots,X_N$,

$$C = [\mathrm{cov}(X_i, X_j)] \tag{12}$$

is symmetric and semidefinite. If positive definite, it has a Cholesky factorization into

$$C = LL^T \tag{13}$$

where L is lower triangular and L^T is the transpose of L; and it also has this factorization after the appropriate interchanges, which we will suppose to have been done, if only semidefinite (Wilkinson, 1978, pgs. 229-231).

If we take $Z_1, Z_2, \ldots, Z_N$ to be independent normal N(0,1) random variables, and Z to be the column vector $(Z_1, \cdots, Z_N)$, then it is easy to compute the covariance matrix of LZ and show that it is the matrix C. (This well-known fact is the basis for the characterization of zero mean multivariate normal distributions as being those whose components are linear combinations of independent N(0,1) normals (Breiman, 1968, pg. 238).) Since the multivariate distribution of $X_1, \ldots, X_N$ is determined by its covariance matrix, the X's can be simulated, if we know their covariance matrix, by simulating the Z's.

For the set of $\eta(\cdot)$ discussed below it was found that the peak flow rate occurs in only a few unit intervals, and from hydrological and statistical considerations it is unlikely that the maximum falls too far outside these few time intervals in general. So only a small number X_m, $X_{m+1}, \ldots, X_{m+r}$ of X's need be considered, which considerably reduces the complexity of the model.

EXAMPLE: COMPUTER SIMULATION FOR PEAK FLOWRATE

In the example case study considered, 12 samples $\eta(\cdot)$ were obtained from catchment rainfall-runoff data (see Table 1), each consisting of 25 unit values of flow rate, (based on the 5 minute time interval). These values of flow rate are assumed to be samples from a multivariate normal distribution. Additionally, all the $\eta(\cdot)$ were obtained from storms which are considered of similar severity (i.e., in the same storm class; see Hromadka and Whitley, 1988). The unit flow rates were visually compared with simulated values from a multivariate normal distribution as a rough check; because there are so few sample points, a more discriminating test is not feasible. The design (i.e., future) storm effective rainfall was taken to be linear increasing from 0 to 5 inches/hour at 1.5 hours, and then linear back down to zero at 3 hours, and this storm was approximated as piece-wise constant in consecutive 5 minute time intervals. The criterion variable of interest is the peak flow rate anticipated from the assumed design storm effective rainfall.

TABLE 1. $\eta(\cdot)$ SAMPLES USING MODEL OF EQ. (1)
(5-minute unit periods of flow rate, cfs)

i	$\eta^{i}(\cdot)$
1	10,30,50,185,320,450,750,900,950,1000,1200,1100,950,900,650,600,550, 500,450,400,320,240,160,80,0
2	10,85,160,230,300,400,550,850,1200,1500,1100,1080,650,400,250,200,170, 140,120,100,80,60,40,20,0
3	10,40,70,100,310,320,490,1600,1400,1550,1600,1480,750,600,350,300,270, 245,210,175,140,105,70,35,0
4	10,40,70,100,145,290,850,1200,1700,1400,1500,1260,1100,950,650,390,320, 285,245,205,165,125,85,45,5
5	10,105,205,300,550,800,870,935,1000,1200,1400,1290,1100,1000,950,300, 275,250,215,180,145,110,75,40,5
6	10,35,60,80,100,250,400,1000,1850,1700,1550,1400,1300,1200,1070,935,800, 670,535,400,325,245,165,85,5
7	10,75,140,200,390,695,1000,1175,1350,1525,1700,1500,1370,1235,1100, 850,600,350,315,280,240,200,135,70,5
8	10,105,200,400,600,750,900,1050,1200,1400,1600,1850,1900,1450,1200, 1050,900,750,600,450,300,225,150,75,0
9	10,90,170,250,350,450,625,800,1000,1200,1400,1700,1550,1400.1350,1130, 915,700,570,435,300,230,160,85,10
10	10,105,200,400,600,725,850,1025,1200,1500,1400,1250,1000,650,590,530, 470,410,350,300,245,185,125,65,5
11	10,55,100,200,400,600,690,780,870,950,1150,1300,1100,950,650,500,450, 400,350,300,245,185,125,65,5
12	10,150,300,500,700,735,770,800,950,1100,980,1350,1200,850,800,800,400, 335,270,200,165,125,85,45,5

From a calculation of the unit flow rate values for each $\eta(\cdot)$, it can be seen that the peak flow rate falls into one of the three unit time intervals [135,140], [140,145], and [145,150], time given in minutes.

The computer simulation procedure continues, as discussed above, in that the covariance matrix of the $Q(t_i)$'s is computed. Then a subset of the X_j is chosen; for example, the subset $\{X_{28}, X_{29}, X_{30}\}$ corresponds to the three intervals in which the peak flow rate occurs. Then the covariance matrix for this subset of X's is factored into a product of a lower triangular matrix L and its transpose. It is now only necessary to generate some independent N(0,1) random variables, use L, and add on the estimated means of each X, in order to develop one vector of flow rate values for the time intervals chosen. From the vector of flow rate (Q) values, the maximum value of Q is obtained, resulting in one sample point in the simulation of maximum Q's. The program does this repeatedly, and keeps track of the empirical distribution of the maximum Q (i.e., the criterion variable). As a final result, one obtains an estimated distribution of percentiles 5%(5%)95% for the maximum Q based on the subset of unit time intervals chosen.

For the given data set, this calculation was performed for the single unit value X_{28}, for $\{X_{27}, X_{28}, X_{29}\}$, and on up to $\{X_{24},\ldots,X_{34}\}$. The outcome was that all the percentiles were the same for these different subsets of X's to within a few cfs (see Table 2).

TABLE 2. PEAK FLOW RATE PERCENTILE ESTIMATES FOR VARIOUS UNIT PERIOD SETS

Unit Period #29		Unit Period 28-30		Unit Period 27-31		Unit Period 24-34	
percentile	max. Q	percentile	max. Q	percentile	max.Q	percentile	Max.Q
5	299.66	5	302.51	5	301.37	5	300.70
10	323.41	10	325.00	10	324.66	10	323.78
15	339.77	15	339.83	15	339.92	15	341.29
20	352.68	20	353.80	20	353.25	20	354.36
25	364.69	25	365.16	25	364.03	25	365.34
30	374.81	30	375.17	30	374.32	30	375.89
35	384.21	35	384.45	35	383.86	35	385.36
40	393.20	40	393.79	40	392.84	40	394.44
45	401.69	45	402.98	45	401.73	45	403.70
50	410.50	50	411.40	50	410.28	50	412.25
55	419.52	55	419.89	55	419.05	55	420.45
60	428.26	60	428.58	60	427.92	60	429.59
65	437.20	65	438.17	65	436.55	65	438.05
70	446.64	70	447.74	70	445.97	70	448.12
75	457.00	75	458.44	75	456.41	75	458.21
80	467.99	80	470.73	80	467.01	80	469.53
85	481.56	85	484.81	85	480.73	85	483.07
90	498.22	90	501.35	90	496.99	90	498.99
95	522.11	95	526.59	95	522.59	95	525.46

There are two reasons for this simple outcome. The first reason is that, as a cursory inspection of the data shown, the maxima tend to fall in a narrow range of time intervals and for those intervals the Q values have approximately the same means and standard deviations. The second reason depends on a less obvious property of the multinormal Q distribution, namely that when the covariance matrix is factored into LL^T, L puts most of its weight into one Z variable. For example, Table 3 provides the factorization for the subset $\{X_{27},\ldots,X_{31}\}$.

TABLE 3. LOWER TRIANGULAR MATRIX L
(COVARIANCE MATRIX $C = LL^T$)
FOR TIME INTERVALS 27,28,29,30,31

Row 1	61.6				
Row 2	65.1	2.6			
Row 3	68.1	5.5	0.6		
Row 4	70.5	8.4	1.4	0.3	
Row 5	72.0	11.2	2.1	0.8	0.2

The significance of this result is that, for this data, satisfactory confidence intervals for peak flow rate can be obtained merely by choosing the most common interval in which the twelve data peak flow rates occur, and then supposing that data to come from a (single) normal distribution.

DISCUSSION

The previous example problem focused upon the runoff criterion variable of peak flow rate. The above methodology can be applied to any criterion variable, A, to develop the probability distribution of $[A]$ by $[A] = \mathcal{A}[Q^D(\cdot)]$ where $[Q^D(\cdot)]$ is the stochastic process of realizations of possible runoff hydrographs, $Q^D(\cdot)$, for the assumed design storm effective rainfall, and $\mathcal{A}$ is a functional which operates on each sampled runoff hydrograph realization to develop a sample point of A.

The multivariate normal distribution, as applied to the sampled $\eta(\cdot)$ obtained from rainfall-runoff data using the model of Eq. (1), provides an estimate of the underlying probabilistic distribution of that stochastic process, which is distributed as $[\eta(\cdot)]$. Consequently, even though only 12 samples (realizations) of the η are obtained by data analysis using Eq. (1), the distribution of the stochastic process, $[\eta(\cdot)]$, can be estimated using the multivariate normal distribution analogous to fitting a probability distribution function to 12 sample points of a random variable. As a result, a continuous probability distribution of the runoff criterion variable, $[A]$, can be obtained rather than developing only a frequency-distribution of m sample points of A, where m is the number of sample realizations developed from $[\eta(\cdot)]$.

CONCLUSIONS AND FURTHER RESEARCH NEEDS

A stochastic integral equation (S.I.E.M.) formulation of the well-known unit hydrograph method is used to develop confidence intervals for runoff criterion variables (e.g., peak flow rate, volume, pipe size, etc.). The multivariate normal distribution is used with the S.I.E.M. to provide a continuous probability distribution of the selected criterion variable. Example applications to estimating a peak flow rate associated to a future effective rainfall event is considered using measured rainfall-runoff data to develop the underlying probabilistic distributions of the associated stochastic processes. Any runoff criterion variable can be evaluated for confidence interval estimates using the procedures discussed. Extension of the above probabilistic techniques to other rainfall-runoff modeling approaches can be readily achieved by analyzing the rainfall-runoff modeling error as a stochastic process (see Hromadka and Whitley, 1988).

Further research is needed in the important topic of developing regionalized multivariate normal distributions of rainfall-runoff modeling error. Regionalization would provide an estimate of the means and variances in the multivariate normal distribution estimation of modeling error, which could then be transferred to ungauged catchments where the rainfall-runoff model is to be applied. In this fashion, confidence intervals could be estimated for runoff criterion variables of interest, in order to make better design and planning decisions which include uncertainty issues and risk.

REFERENCES

1. Breiman, L., Probability, Addison-Wesley, New York, 1968.
2. Millingsley, P., Probability and Measure, Wiley, New York, 1986.
3. Kendall, M. and Stuart, A., The Advanced Theory of Statistics, Vol. I, Griffin, London, 1977.
4. Maindonald, J., Statistical Computation, Wiley, New York, 1984.
5. Wilkinson, J., The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1978.
6. Hromadka II, T.V. and Whitley, R.J., The Design Storm Concept in Flood Control Design and Planning, Stochastic Hydrology & Hydraulics, 1988, in-press.