Nordic Hydrology (1971), II, 186–216

Published by Munksgaard, Copenhagen, Denmark

THE STOCHASTIC APPROACH TO WATERSHED MODELING

Donn G. DeCoursey

Southern Plains Branch, Soil and Water Conservation Research Division,
Agricultural Research Service
U.S. Department of Agriculture, Chickasha, Oklahama

The stochastic approach to watershed modeling refers to the techniques used to generate synthetic hydrologic data. These data may be used either for input to a parametric watershed model or to provide directly an estimate of the output of a hydrologic process. In both cases the basic techniques of the generation processes are the same. The type of process depends primarily on the purpose for which the data are being generated and on the quality and quantity of sample data.

Techniques are presented which can be used to generate data for one or any number of variates. The data generated can be normal, skewed, or log normal, and include serial correlation. If two or more variates are involved, cross correlation may also be considered.

Brief discussions concerning missing data, statistical tests, random number generation, and interpretation of results are presented along with a review of the generation schemes that have been used in the stochastic generation of hydrologic data.

Contribution from the Southern Plains Branch, Soil and Water Conservation Research Division, Agricultural Research Service, USDA, in cooperation with the Oklahoma Agricultural Experiment Station, Chickasha, Oklahoma.

The Stochastic Approach to Watershed Modeling

The word stochastic comes from the Greek word, stochastikos, meaning skillful in aiming. Assuming that this refers to a person's ability to shoot a bow and arrow, then the indicator of his ability would be the target he was shooting at. Upon examination of the target one would find the density of points greatest near the center and least around the edge. In addition, one would find that the location of each point as it occurred would be random. Thus the word stochastic has come to refer to the random nature of a variate with respect to time. In watershed modeling, it generally refers to the random nature of hydrologic phenomena such as river flow, precipitation, wind velocity, etc.

The stochastic approach to modeling basically means that the distribution characteristics of a variate are determined from sample data and used with a random number generator to produce synthetic sequences of the variate. The distribution characteristics considered in such processes are the mean, standard deviation skew, and serial correlation. If there is more than one variate, the cross correlations among variates are considered.

The term "watershed modeling" has a very broad connotation. It is used here to refer to analytical simulation of the processes that take place in natural watersheds. The models are developed for many different reasons and therefore have many different forms. However, they are in general designed to meet one of two primary objectives. The role of stochastic processes is different in each.

One objective of watershed models is to gain a better understanding of the hydrologic phenomena operating in a watershed and how changes in the watershed may affect these phenomena. Models created for this purpose are generally physically based, deterministic models. The hydrologic phenomena they simulate are generally defined by the laws of continuity, energy, and momentum. Such models are used primarily in the analysis of individual events. As such they are seldom used to generate synthetic data. In these models stochastic processes can be used to add a random element equal to the unexplained variance.

The other objective of watershed models is to generate synthetic sequences of hydrologic data. Models created for this purpose vary from a very deterministic form using much information about the physical processes involved in converting input to output to a "black box" form in which nothing is known of the physical process involved. Most of these models are of a parametric type in which elements of the hydrologic system are combined and less detail about the internal structure of the model is known. Stochastic inputs to such models depend upon the internal structure of the model.

Models that are relatively simple, for example those that calculate annual runoff from annual rainfall, have simple stochastic requirements. In this example it would be a scheme for generating sequences of annual rainfall. As the models become more complex, the stochastic generators also become more complex. Suppose for example a watershed model designed to output the entire hydrograph of flow for a period of many years. A model such as this might use hourly rainfall, windspeed, relative humidity, variable resistance coefficients, etc., all of which could be interrelated. The stochastic generator for this model would very probably need to be a complex multivariate model.

Some watershed models designed to provide synthetic sequences of hydrologic data may be entirely stochastic. In such models absolutely nothing is assumed as to the internal structure of the model. The synthetic sequences are developed entirely by a stochastic process. There is no parametric model to distribute a set of stochastic input data. Such models are based on the statistical parameters of the historic data to which they are fitted. For example, monthly runoff at a streamflow station synthesized by a pure stochastic model could be based on the mean, standard deviation, and serial correlation of the historical data from the station.

The techniques used to develop either the output directly or the input to a parametric model are essentially the same. The degree of complexity of the stochastic generator is a function of the statistical characteristics of the data to be generated and not the type or size of the element being generated. In the following discussion several different types of stochastic generators are described, starting with the simplest and increasing in complexity. Most of the stochastic models use Monte Carlo methods to obtain the data sequences. The Monte Carlo methods are processes by which data are produced synthetically by a sampling technique or some form of a random number generator.

A SIMPLE MONTE CARLO PROCEDURE

The simplest example of a stochastic process is the random walk shown in Figure 1 where X_n is a random variable denoting the position at time n of a moving particle. The movement of the particle is either upward with a probability of .5, $Z_n = 1$, or downward with a probability of .5, $Z_n = -1$; i.e.,

$$\text{prob}(Z_n = 1) = \text{prob}(Z_n = -1) = .5 \tag{1}$$

The location of the particle is thus:

$$X_n = X_{n-1} + Z_n \tag{2}$$

where Z_n is the movement in each time interval.

A special class of stochastic processes termed Markov processes have been used in the generation of rainfall sequences. The Markov process is one in which the value of an element depends only on the probability density function, p.d.f., of the variate at the previous time point. Thus, p.d.f.'s conditional on values at two or more time points away need never be considered. The succession of p.d.f.'s is termed a Markov chain. The random walk of Figure 1 is a particular kind of single-state Markov chain. Many of the urn models developed for random sampling can be described by multistate Markov chains. An example of two-state Markov chain used in hydrologic models is described in the next paragraph.

In studies of rainfall it has been found that the two-state Markov chain gives good results in certain areas (Gabriel & Neumann 1962, DeCoursey & Seely 1969). Consider a dry state, 0, and a wet state 1, with α being the probability of a wet day following a dry day and $1-\alpha$ being the probability of a dry day following a dry day. The probability of a dry day following a wet day is β, and $1-\beta$ is the probability of a wet day following a wet day. These transition probabilities are shown in the following matrix.

$$P \equiv \begin{array}{cc} & \begin{array}{cc} \multicolumn{2}{c}{\text{Future state}} \\ \text{dry} & \text{wet} \\ 0 & 1 \end{array} \\ \begin{array}{c} \text{Present} \\ \text{state} \end{array} \begin{array}{c} \text{dry } 0 \\ \\ \text{wet } 1 \end{array} & \left[\begin{array}{cc} 1-\alpha & \alpha \\ & \\ \beta & 1-\beta \end{array} \right] \end{array} \qquad (3)$$

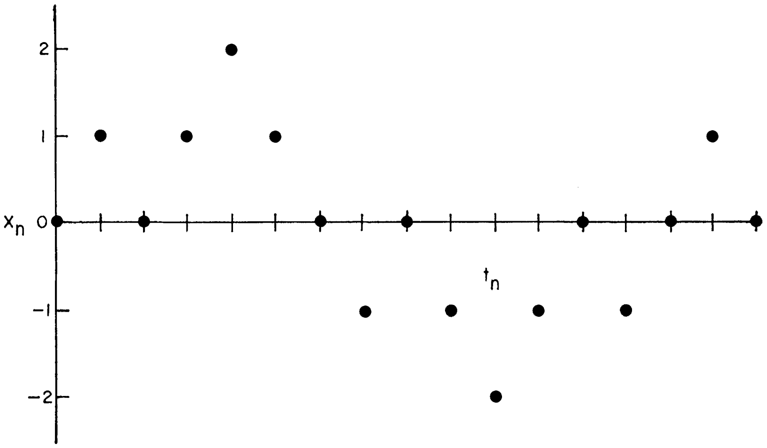

Fig. 1.
Example of a simple random walk.

By using relative frequencies over a period of about 27 years, Gabriel & Neumann found in a study of Tel Aviv rainfall data the following transition probabilities

$$P \equiv \begin{bmatrix} .750 & .250 \\ .338 & .662 \end{bmatrix}$$

Using these transition probabilities and a random sampling method, the occurrence or nonoccurrence of an event may be calculated.

Multistate Markov chains of higher order have been proposed and used in some models. Pattison (1964), for example, used a sixth-order chain for distributing the total volume of rainfall over storm periods. The length of record needed to develop high-order Markov chains is generally not available for hydrologic data. A good reference for such chains and for stochastic processes in general is the book *The theory of stochastic processes* by Cox and Miller (1965). Parzen (1962) also has a good book on the subject of stochastic processes.

Most Markov processes used in hydrologic analyses are of Gauss–Markov form in that the normal probability density function is used. Mandelbrot & Wallis (1968, 1969 a–e) in a series of articles have shown that these systems have a short memory and thus cannot account for the extremes both in size of event and in length of wet and dry cycles found in hydrologic history. They have found a new technique of data analysis called "R/S analysis" to be very effective in handling the extremes. The technique can reproduce the Hurst (1951) phenomena observed in long periods of hydrologic data. Gauss–Markov and multiple-lag linear autoregressive models cannot do this. The technique appears to be particularly useful in the analysis and reproduction of long periods of hydrologic record. Since the primary objective of this work is to present the stochastic approach used in watershed modeling, periods of record of from 10 to 100 years in length are of major interest. Under these circumstances, the more traditional techniques presented below are satisfactory. However, anyone interested in a vigorous analysis of his data should give consideration to "R/S analysis" even though only a short period of record, less than 100 years, is to be considered.

The pattern of a stochastic process, time series, can take several forms depending upon the hydrologic process being described and can vary from the very erratic form representative of daily rainfall to a fairly smooth form representative of monthly evaporation. Several generating methods are available for modeling these processes. The choice of the method is based upon how well the mathematical structure of the method conforms to the characteristics of the hydrologic process being modeled. The most common methods are:

(1) the moving average, (2) the sum of harmonics, and (3) the autoregression. Emphasis in this discussion will be placed on the last method, although the other two methods will be discussed briefly first. A tool quite often helpful in these analyses is the correlogram. It is described following the discussion on the sum of harmonics method of modeling.

THE MOVING AVERAGE

The moving average can be represented by the equation

$$y_i = b_o + b_1x_i + b_2x_{i-1} + \cdots + b_mx_{i-(m-1)} \tag{4}$$

where y_i is the value of the process being modeled at the i–th time period, x is a variable, and m is the extent of the moving average. In the above equation, y could be annual runoff and x the effective precipitation. In this case, the coefficients, b, must be positive and sum to unity. A time series generated by such a scheme is not random; however, events separated by m or more time units are independent. Julian (1961) used the moving average method in studying streamflow in the Colorado River Basin.

THE SUM OF HARMONICS

The time series of many hydrologic processes are distinguished by obvious periodicities. This suggests the possibility of using Fourier series in the mathematical representation of the series. The basic form of the most general equation is given by

$$x_{(t)} = \bar{x} + \sum_{k=1}^{S} (A_k \cos \lambda_k t + B_k \sin \lambda_k t) + \varepsilon_t \tag{5}$$

in which

$$A_k = 2/n \sum_{t=1}^{n} x_t \cos \lambda_k t, \tag{6}$$

and

$$B_k = 2/n \sum_{t=1}^{n} x_t \sin \lambda_k t. \tag{7}$$

In the above equations, x is the value of the process being modeled, k is the number of harmonics used in the representation, n is the sample size, λ_k are the frequency numbers used in the series and are functions of the cycle period, and ε_t is a ctochastic component. The λ_k are restricted to the range between 0 and π and do not have to be true fractions of the period length.

Eq. 5 has most generally been used in the analysis of hydrologic processes rather than in the generation of synthetic series. As such, it is referred to as spectral analysis or variance spectrum analysis. If

$$I_k^2 = A_k^2 + B_k^2 \tag{8}$$

is used to calculate

$$S_k = \frac{\sum\limits_{\lambda_k \leq \lambda} I_k^2(\lambda)}{\sum\limits_{\lambda_k = 1}^{S} I_k^2(\lambda)} \tag{9}$$

and S_k is in turn plotted against λ_k, the integrated periodogram is developed. In this plot, the ordinate S_k is the contribution to the variance of the system given by the harmonic of given frequency. A typical plot is shown on Figure 2 for the monthly time series of the Elk River at Clark, Colorado (Roesner & Yevdjevich 1966). The Figure also shows the effect on the variance of the system of removing the various harmonic periods. A complete description of variance spectrum analysis is given by Blackman & Tukey (1958), Matalas (1966), and Wold (1954). Its use is presented in Horn & Bryson (1960), Quimpo (1967), and Roesner & Yevdjevich (1966). Quimpo (1967) also has a good discussion and bibliography of spectrum analyses.

CORRELOGRAMS

A correlogram is a graphical representation of the autocorrelation (the correlation of an element of a series with other elements in the same series lagged by integer amounts) coefficients as functions of the lag. The correlogram can at times help in deciding what type of generating process to use in the synthesis of data. The correlogram of a process governed by a moving average will show oscillation for lagged time periods less than the length of the moving average, m, but will be zero for lags greater than m. The correlogram for a harmonic process will oscillate with a fixed period and amplitude for an indefinite period. The correlogram of autoregression methods will descrease monotonically from a value of 1 at lag 0 to zero at lag infinity if the correlation coefficient is positive at lag 1. If the correlation coefficient is negative at lag 1, the correlogram will oscillate with a period of unity above and below the abscissa with a decreasing but nonvanishing amplitude. An example of a correlogram and the effects on the correlogram of removing periods of high correlation is shown in Figure 3.

The correlogram is used primarily as a tool for the analysis of hydrologic processes rather than as a generating device. See Bhuiya & Yevdjevich (1968),

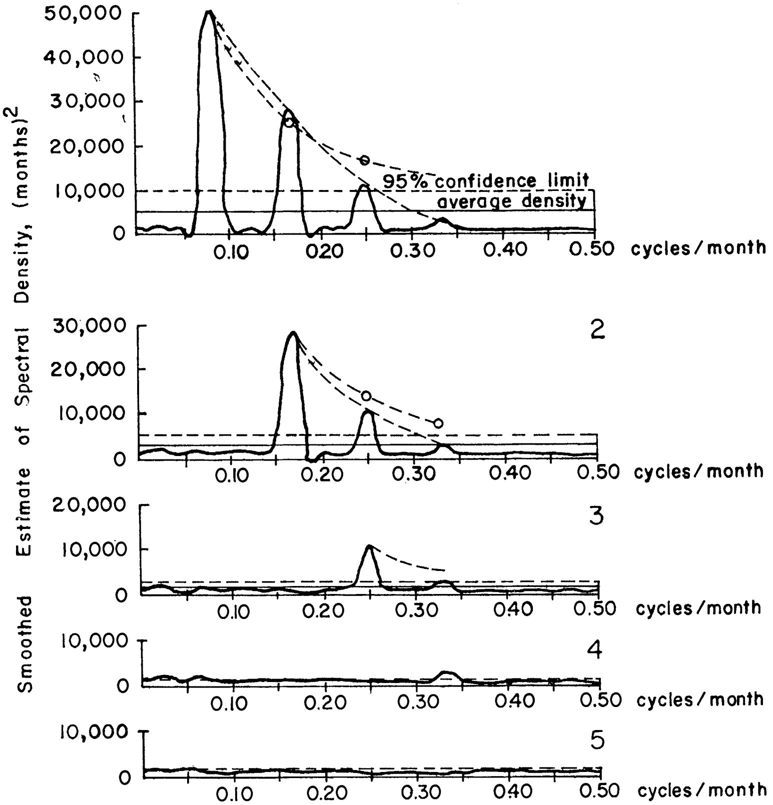

Fig. 2.

Effect of removing periods from the time series on the variance spectrum for station 9.378, Elk River at Clark, Colorado: (1) variance spectrum with 12-, 6-, 4-, and 3-month periods present, (2) variance spectrum with 6-, 4-, and 3-month periods present, (3) variance spectrum with 4- and 3-month periods present, (4) variance spectrum with 3-month period present, and (5) variance spectrum with all periods removed (Roesner & Yevdjevich 1966).

Caffey (1965), Dumas & Morel-Seytoux (1969), Roesner & Yevdjevich (1966), Todorovic & Yevdjevich (1969), Yevdjevich (1961, 1964), and Yevdjevich & Jeng (1969) for examples of its use.

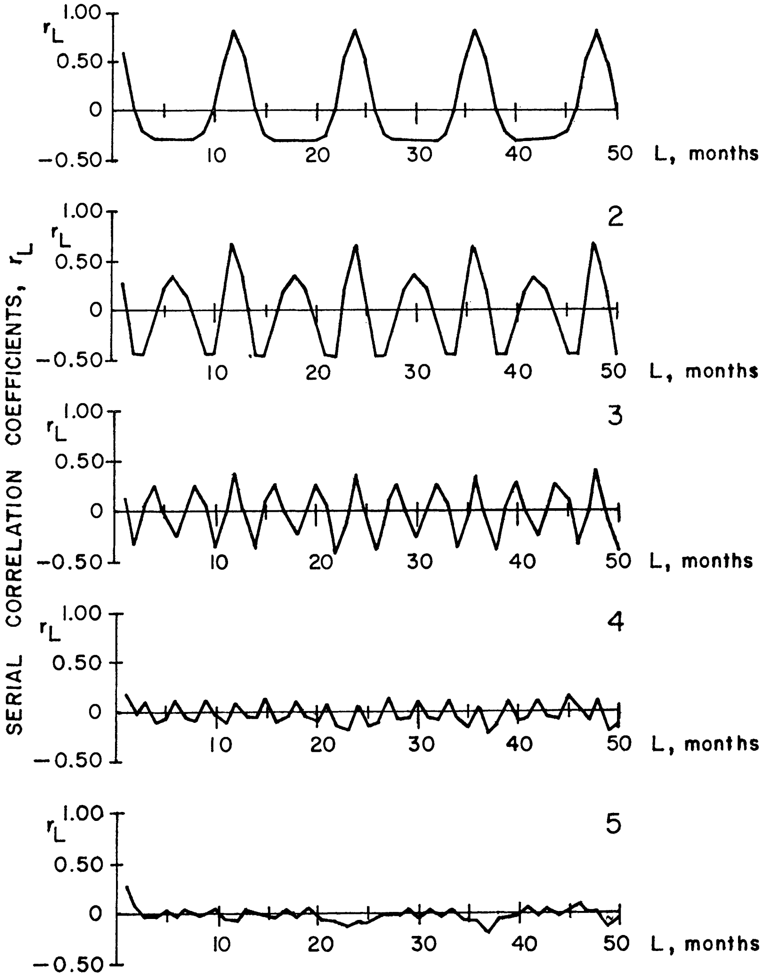

Fig. 3.

Effects of removing periods from the time series on the correlogram for station 9.378, Elk River at Clark, Colorado: (1) correlogram with 12-, 6-, 4-, and 3-month periods present, (2) correlogram with 6-, 4-, and 3-month periods present, (3) correlogram with 4- and 3-month periods present, (4) correlogram with 3-month period present, and (5) correlogram after all periods have been removed (Roesner & Yevdjevich 1966).

AUTOREGRESSIVE METHODS

Basic recursion relation

A time series can be represented by any one of several analytic functions as has been described. Most of these functions are of a form

$$x_t = f(t) + \varepsilon_t \tag{10}$$

in which x_t is the value of the series at time t, $f(t)$ is a deterministic component, and ε_t is a random component. If the series shows a long-term trend, polynomials or Fourier series may be used to represent it. If, however, the trend component is constant, then the series is derived from a stationary process. In such a process, the lagged covariances between elements in the series are functions of the absolute differences between indices of elements in the series. As a result of this argument, the linear autoregressive relation may be developed.

$$x_t = \beta_o + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots\cdot \beta_m x_{t-m} + w_t \tag{11}$$

where the β_i are autoregression coefficients and w_t is an independent random term.

Normal distributions

The series described by Eq. 11 is said to be stationary in a wide sense or weakly stationary because its mean is equal to a constant and its autocovariance is a function of only the absolute differences between elements of the series. Thus the series is assumed to be normally distributed. By taking the conditional expectation of Eq. 11 for a lag-one Markov process, $m = 1$,

$$E(x_t|x_{t-1}) = \mu_x + \varrho_x(1)(x_{t-1} - \mu_x) \tag{12}$$

and

$$\mathrm{Var}(q_t|q_{t-1}) = \sigma_x^2[1 - \varrho_x^2(1)] \tag{13}$$

the following equation may be developed

$$x_t = \mu_x + \varrho_x(1)(x_{t-1} - \mu_x) + v_t\sigma_x\sqrt{1 - \varrho_x^2(1)} \tag{14}$$

in which $\mu_x = \mu_{x,t} = \mu_{x,t-1}$ is the population mean, $\sigma_x \equiv \sigma_{x,t} \equiv \sigma_{x,t-1}$ is the population standard deviation, $\varrho_x(1)$ is the lag 1 correlation coefficient, and v_t is a normal random deviate. Eq. 14 therefore preserves both the mean, first moment, of the distribution and the variance, second moment, by combining Eqs. 12 and 13. If the mean, $\hat{\mu}_x$, standard deviation, $\hat{\sigma}_x$, and correlation coefficient, $\hat{\varrho}_x(1)$, of the historic series along with a random variate with zero mean and unit variance are substituted into Eq. 14, the following equation for the *i*–th element of the series is obtained.

$$x_i = \hat{\mu}_x + B(x_{i-1} + \hat{\mu}_x) + t_i\hat{\sigma}_x\sqrt{1-\hat{\varrho}_x{}^2(1)} \qquad (15)$$

where B is the least-squares regression coefficient equal to the correlation coefficient because $\hat{\sigma}_x = \hat{\sigma}_{x,t} = \hat{\sigma}_{x,t-1}$.

If the time series shows seasonal trends such as those of streamflow, Eq. 15 may be used to represent the *j*–th season as follows:

$$x_{i,j} \equiv \hat{\mu}_{x,j} + B_j(x_{i-1,j-1} - \hat{\mu}_{x,j-1}) + t_i\hat{\sigma}_{x,j}\sqrt{1-\hat{\varrho}_j{}^2(1)} \qquad (16)$$

in which

$$B_j \equiv \hat{\varrho}_j(1)\frac{\hat{\sigma}_{x,j}}{\hat{\sigma}_{x,j-1}} \qquad (17)$$

In Eq. 16 the index i runs sequentially and index j runs cyclically across all seasons or periods in the cycle. The equation represents the time series where there is one element per season or period.

In a situation where there may be many elements per season; for example, estimating daily streamflow where the month is the season length, the equation becomes

$$x_{i,j} \equiv \hat{\mu}_{x,j} + B_j(x_{i-1,j} - \hat{\mu}_{x,j}) + t_i\hat{\sigma}_{x,j}\sqrt{1-\hat{\varrho}_j{}^2(1)} \qquad (18)$$

where $B_j \equiv \hat{\varrho}_j(1)$.

The index i runs sequentially within season j. Eq. 18 is equivalent to Eq. 15 if Eq. 15 were developed from only data in season j.

Synthetic data may be generated by either Eqs. 15, 16, or 18 that will resemble the historic sequence in terms of $\hat{\mu}_x$, $\hat{\sigma}_x$, and $\hat{\varrho}_x(1)$.

More detail on the derivation of these equations may be found in Fiering (1964, 1967), Matalas (1967), Thomas & Fiering (1962), and Yevdjevich (1964).

Nonnormal distributions

In the above discussion, the distribution of events was assumed to be weakly stationary or normally distributed. However, in many hydrologic processes obvious departures from this assumption are common. Many phenomena, including streamflow and size of precipitation events, appear to be distributed like gamma or third-order stationary. Fiering (1967) shows that the basic lag-one Markov process, Eq. 14, may be rewritten to incorporate the effect skewness, γ, by replacing x_t by its standardized form $(x_t - \mu_x)/\sigma_x$ and then cubing the equation to obtain γ_x. By taking expectations he shows that the skewness of the random component γ_ε required to maintain the skewness of the system $\hat{\gamma}_x$ is given by

$$\gamma_\varepsilon = \hat{\gamma}_x \frac{1 - \hat{\varrho}_x{}^3(1)}{[1 - \hat{\varrho}_x{}^2(1)]^{3/2}} \tag{19}$$

The random normal deviates, t_i, must now be distributed like gamma such that $\mu_\varepsilon = 0$, $\sigma_\varepsilon{}^2 = 1$, and γ_ε is equal to Eq. 19. This is accomplished by the transform

$$\varepsilon_i = \frac{2}{\gamma_\varepsilon}\left[1 + \frac{\gamma_\varepsilon t_i}{6} - \frac{\gamma_\varepsilon{}^3}{36}\right]^3 - \frac{2}{\gamma_\varepsilon} \tag{20}$$

The equation, a third-order stationary lag-one Markov process, for generating data that will resemble the historic sequence in terms of $\hat{\mu}_x$, $\hat{\sigma}_x$, $\hat{\gamma}_x$, and $\hat{\varrho}_x(1)$ is given by

$$x_{\mathrm{i}} = \hat{\mu}_x + B(x_{\mathrm{i}-1} - \hat{\mu}_x) + \varepsilon_i \hat{\sigma}_x \sqrt{1 - \hat{\varrho}_x{}^2(1)} \tag{21}$$

Both Eqs. 16 and 18 which represent systems in which seasonal trends are present may be changed to incorporate the effect of skewness by calculating $\gamma_{\varepsilon,j}$ for each of the seasons or periods by Eq. 19. The appropriate value should then be used in Eq. 20 for calculating the random element ε_i which replaces t_i in each of the equations.

Matalas (1967) and Thomas & Fiering (1962) discuss in greater detail this method for making the generating process third-order stationary. Matalas (1967) also shows a method in which the coefficient of skewness of $\hat{\gamma}_x$ of the generated series can be approximated but must be less than $2\sqrt{2}$. The procedure involves generating a series of random variates using a lag-one Markov process which is a function of $\hat{\varrho}_x(1)$. The variates are then squared and m of them accumulated. The synthetic sequence is then a function of $\hat{\mu}_x$, $\hat{\sigma}_x$, m and the accumulated variates.

Not only are many hydrologic time series skewed but they are often found to be log-normally distributed. This can be either a 2- or 3-parameter distribution. Such a system is represented by

$$y = \ln (x - a) \tag{22}$$

where x is the random variate, a is a lower bound of the variate, and y is assumed to be a skewed normal distribution. Matalas (1967) shows how the mean, $\hat{\mu}_x$; variance, $\hat{\sigma}_x{}^2$; the coefficient of skewness, $\hat{\gamma}_x$; and the lag-one correlation coefficient, $\hat{\varrho}_x(1)$ of the random variate x are related by a set of four equations to the lower bound, a; the mean, μ_y; the variance, $\sigma_y{}^2$; and the lag-one correlation coefficient, $\varrho_y(1)$ of the random variate y. By solving the four simultaneous equations, a, μ_y, σ_y, and $\varrho_y(1)$ can be found and substituted in the equation

$$y_i = \mu_y + \varrho_y(1)\,(y_{i-1} - \mu_y) + \sigma_y t_i \sqrt{1 - \varrho_y{}^2(1)} \qquad (23)$$

The value of x_i is obtained by adding a to the antilog af y_i. The sequence of values thus generated will resemble the historic sequence in terms of $\hat{\mu}_x$, $\hat{\sigma}_x$, $\hat{\gamma}_x$, and $\hat{\varrho}_x(1)$.

The easiest method of obtaining a synthetic sequence of log-normally distributed events is to calculate the population parameters $\hat{\mu}_x$, $\hat{\sigma}_x$, $\hat{\gamma}_x$, and $\hat{\varrho}_x(1)$ from the logarithms of the variate x and then use these parameters in Eqs. 15, 16, 18, or 21. However, Matalas states that "If the values of a, μ_y, σ_y, and $\varrho_y(1)$ had been obtained from logarithms of the historic events rather than in the manner outlined above, then equation 10 [equivalent to Eq. 23 above] would lead to a synthetic sequence that would not resemble the historic sequence in terms of $\hat{\mu}_x$, $\hat{\sigma}_x$, $\hat{\gamma}_x$, and $\hat{\varrho}_x(1)$." This statement is theoretically true, but the magnitude of the error is not known. Fiering (1967) states that he found the error in maintaining serial correlation was small.

Multiple-lag Markov process

Multiple-lag Markov processes have not been used extensively in hydrologic synthesis. This has generally been because the work involved in evaluating parameters of the model does not warrant the increase in accuracy over that of the lag-one process. Fiering (1967, Chap. 3) discusses in detail the multiple-lag problem with emphasis on storage-yield functions of reservoirs. The following equation illustrates the method used to generate data with this system.

$$x_i = B_0 + B_1 x_{i-1} + \cdots B_m x_{i-m} + \sigma_x t_i \sqrt{1 - R^2} \qquad (24)$$

where B_i is the least-squares partial regression coefficient of lag i, x_i is the value of the variate at time period i, σ_x is the standard deviation of the x_i, t_i is a normal random sampling deviate, and R is the multiple correlation coefficient.

Two-variate problems

Hydrologic analyses often need two or more synthetic inputs which may be correlated with one another. In this section a technique for generating data for two interrelated variates is presented. The technique requires that the time interval be the same for both variates. Monthly streamflow on two forks of a river on which a reservoir is anticipated or monthly rainfall and evaporation for input to a water yield model are examples of synthetic data that could be generated by this technique.

Suppose that the two variates are x and y and that the means, standard deviations, skewnesses, and lag-one serial correlation coefficients for the variates are $\hat{\mu}_x$, $\hat{\mu}_y$; $\hat{\sigma}_x$, $\hat{\sigma}_y$; $\hat{\gamma}_x$, $\hat{\gamma}_y$; and $\hat{\varrho}_x(1)$, $\hat{\varrho}_y(1)$. The product-moment correlation coefficient between the data for variates x and y is $\hat{\varrho}_{x,y}$. The technique requires that one of the two variates be selected as a key variate. The selection could be based on the length of record, quality of records, etc. Assume that variate x is selected. Variate y is therefore assumed to be subordinate to x. A synthetic sequence of events is generated for x by using either Eq. 15 or 21 depending upon whether or not the skew factor is assumed to be significant. A similar cross-correlation model is used to calculate the variate values for y.

$$y_i \equiv \hat{\mu}_y + B(x_i - \hat{\mu}_x) + u_{i-1}\hat{\sigma}_y\sqrt{1-\hat{\varrho}_{x,y}{}^2} \tag{25}$$

in which B is given by $\hat{\varrho}_{x,y}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$ and u_{ji} is a standardized random variate adjusted as follows to incorporate the serial correlation coefficient for y.

$$u_{i-1} \equiv \frac{\pi}{\hat{\sigma}_y}(y_{i-1} - \hat{\mu}_y) + t_i(1-\pi)^{1/2} \tag{26}$$

in which

$$\pi = \frac{\hat{\varrho}_y(1) - \hat{\varrho}_x(1)\,\hat{\varrho}^{\,2}_{xy}}{\sqrt{1-\hat{\varrho}^{\,2}_{xy}}} \tag{27}$$

In Eq. 26, t_i is a standardized random sampling number adjusted for skew if deemed necessary. The derivation of π and Eq. 26 are described by Fiering (1964). Prior to his presentation, the common approach was to consider u_{i-1} as a standardized variate. He shows that such a practice does not keep the serial correlation of y and introduces spurious correlation into the subordinate variate.

If the variates show seasonal trends, the statistical characteristics of the populations may be determined for each of the seasons and either Eq. 16 or 18 used to generate the synthetic sequence for the key variate. The choice of equations depends upon the number of observations per season. Eq. 25 would then be used to generate the synthetic data for the subordinate variate, being sure to use the appropriate set of statistical characteristics depending upon the season being considered.

Multivariate problems

Quite often analyses aimed at appraising the water resources of a river basin need estimates of river flow at different locations within the basin. Several procedures have been proposed for generating synthetic data for such systems. Multiple regression techniques are the most commonly used method. They may or may not be the most satisfactory depending upon the objectives of the study. In general, they have two drawbacks: (1) the variates must be ranked or an ordering sequence established among them, and (2) not all cross-correlation coefficients are retained in the synthetic sequence. If these drawbacks are not considered serious, the method can be used to good advantage because it is computationally much simpler than other methods.

Assume there are m variates for which simultaneous synthetic sequences are desired. The variates are ordered using some preselected criteria. Regression equations are then calculated from the data. The equation for each variate is different from all others and is a function of all variates preceding it in the sequence of variables. The basic equation for the j-th variate is of the form

$$x_{i,j} = B_0 + B_1 x_{i-1,j} + B_2 x_{i,j-1} + \cdots\cdot B_j x_{i,1} + \hat{\sigma}_{x,j} t_i \sqrt{1 - R_j^2} \qquad (28)$$

where x is the variate value of the series, the B_i are regression coefficients, $\hat{\sigma}_{x,j}$ is the sample standard deviation of the j-th variate, t_i is a random, independent, and normally distributed variate, R_j is the sample multiple correlation coefficient of the equation for the j-th variate, the subscript i is the sequence of events, and j is the rank of the variate starting with 1 for the first variate selected. The first term following the constant is the effect of the preceding time period, and all other terms except the last are the effects from other stations preceding it in the sequence of variates. The last term is the random component required to retain the variance of the j-th variate.

Several variations of Eq. 28 are used. Quite often the variates are reduced to standard normal deviates to reduce seasonal trends. In this case B_0 becomes zero and $\hat{\sigma}_{x,j}$ is 1. If the variates are skewed, they can be made approximately normal by solving Eq. 20 for t_i and letting it equal the transformed deviate, and applying the resulting equation to the deviates. The transform is

$$x_{i,j} = t_{i,j} = \frac{6}{\hat{\gamma}_{x,j}} \left[\left[\frac{\hat{\gamma}_{x,j}}{2} \left[\frac{X_{i,j} - \hat{\mu}_{x,j}}{\hat{\sigma}_{x,j}} \right] + 1 \right]^{1/3} - 1 \right] + \frac{\hat{\gamma}_{x,j}}{6} \qquad (29)$$

where $\hat{\mu}_{x,j}$, $\hat{\sigma}_{x,j}$, and $\hat{\gamma}_{x,j}$ have the usual meaning of mean, standard deviation, and skew for the j-th variate. $X_{i,j}$ is the raw value of the variate, and $x_{i,j}$ is the standard normal deviate. If Eq. 28 is used to generate synthetic data using $x_{i,j}$, defined as in Eq. 29, then the inverse transform of Eq. 29, Eq. 20, must be used to calculate the synthetic series in raw form.

Beard (1965) has gone one step farther in using the equation to generate synthetic sequences of stream flow. Most streamflow data can be reduced to standard normal deviates by using a three-parameter log-normal distribution. Beard added a small increment to all monthly flows, 0.001 times the mean annual, then took the logarithm of the result. He then reduced the logarithms to standard normal deviates using Eq. 29. In addition to using a log normal distribution, he added one term to Eq. 28: the average of all flows from the *m* stations for the previous six months.

Matalas (1967) presents two methods of dealing with the problem of multiple variate data generation. The first method is designed to preserve the means, standard deviations, lag-one serial correlation coefficients, and lag zero cross-correlations of the data matrix. For a set of *m* variables, his model is of the form

$$\mathbf{x}_i = \mathbf{A}\,\mathbf{x}_{i=1} + \mathbf{B}\,\varepsilon_i \tag{30}$$

where $\mathbf{x}_i$ is an *m* by one matrix of variate deviates from the mean at time *i*, for the *j*-th variate, the element is $x_i^{(j)} = X_i^{(j)} - \hat{\mu}_x^{(j)}$, $\mathbf{x}_{i-1}$ is the *m* by one matrix of deviates at time *i*–1, $x_{i-1}^{(p)} = X_{i-1}^{(p)} - \hat{\mu}_x^{(p)}$, $\mathbf{A}$ and $\mathbf{B}$ are *m* by *m* matrices of coefficients, and the ε_i are random normal variates. The matrix $\mathbf{A}$ is defined by the equation

$$\mathbf{A} = \mathbf{M}_1\,\mathbf{M}_o^{-1} \tag{31}$$

in which $\mathbf{M}_1$ is the lag-one covariance matrix $\mathbf{x}_i\,\mathbf{x}_{i-1}^T$ where T indicates the transpose, $\mathbf{M}_o^{-1}$ is the inverse of $\mathbf{M}_o$, and $\mathbf{M}_o$ is the lag-zero covariance matrix $\mathbf{x}_{i-1}\,\mathbf{x}_{i-1}^T$. The matrix $\mathbf{B}$ is obtained by solution of the equation

$$\mathbf{B}\,\mathbf{B}^T = \mathbf{M}_o - \mathbf{M}_1\,\mathbf{M}_o^{-1}\,\mathbf{M}_1^T = \mathbf{C} \tag{32}$$

where $\mathbf{M}_o$ and $\mathbf{M}_1$ are as defined above. G. K. Young (1968) shows that after evaluating the elements on the right-hand side of Eq. 32, thus yielding the $\mathbf{C}$ matrix, the elements of $\mathbf{B}$ may be found by a set of recursive equations. Other solutions to $\mathbf{B}$ are available, as there are an inumerable number of matrices which will, when post multiplied by their transform, yield the $\mathbf{C}$ matrix. Matalas (1967) shows how the elements of $\mathbf{B}$ may be obtained by the techniques of principal component analysis.

The second method presented by Matalas (1967) is the multivariate extension of Eq. 23. The technique preserves the means, standard deviations, skews, lag-one serial correlation coefficients, and the lag-zero correlation coefficients.

In addition, the data $x^{(p)}$ are presumed to follow a three-parameter log-normal distribution with a lower bound $a^{(p)}$. Thus for the *p*-th variate

$$y^{(p)} = \ln\,[x^{(p)} - a^{(p)}]. \tag{33}$$

The equations referenced in the two-variate case for relating the mean, standard deviation, skew, and lag-one serial correlation coefficients of the variates $x^{(p)}$ to the log variates $y^{(p)}$ are used to find the parameters of $y^{(p)}$ for each of the p variates. The generating process is

$$\mathbf{y}_i = \mathbf{A}'\mathbf{y}_{i-1} + \mathbf{B}'\varepsilon_i \tag{34}$$

where $\mathbf{y}_i$ is an m by one matrix of variate deviates from the mean of y at time i, $\mathbf{y}_{i-1}$ is the m by one matrix of deviates at time i–1, $\mathbf{A}'$ and $\mathbf{B}'$ are m by m matrices of coefficients, and the ε_i are random normal variates. The matrices $\mathbf{A}'$ and $\mathbf{B}'$ are analogous to and evaluated by the methods used for matrices $\mathbf{A}$ and $\mathbf{B}$ in Eq. 30.

Fiering (1964) introduced the use of multivariate analysis techniques into the synthesis of streamflow data. His technique preserves the means, standard deviations, and lag-zero cross-correlation coefficients of the data matrix, but it does not preserve the lag-one serial correlation coefficients. The method is predicated on finding all the truly orthogonal or independent variates in the data. This is accomplished by calculating the lag-zero cross-correlation matrix, the elements of which are $\hat{\varrho}_x{}^{(p)(q)}(0)$, for all variates $p = 1, \cdots m$, $q = 1, \cdots m$. The diagonal elements of the matrix, $p = q$, are equal to unity and the off-diagonal elements are the cross-correlations between variates. The significant roots of this matrix, the eigenvalues, and their corresponding eigenvectors are used as independent variates in a linear regression equation on the dependent variable to generate synthetic data.

THE PROBABILISTIC ELEMENT AS A COMPONENT OF A DETERMINISTIC SYSTEM

Thus far the discussion has centered on presenting the mechanics of stochastic models where the objectives have been the generation of synthetic data sequences. Stochastic elements also play an important role in deterministic systems. For example, consider a deterministic system that takes the form of a simple linear regression of a dependent variable, y, on a set of independent variates, x_i

$$\hat{y} = f(x_1, x_2, \cdots x_n). \tag{35}$$

Let us further suppose that the multiple correlation coefficient between y and the set of x's is R. If the data set is multivariate normal or approximately so, the percentage of the variance in y explained by the linear equation is

$R^2 \leq 1.0$. If the equation is to be used subsequently for predicting new values of y from a different set of x's, and if the distribution of the new set of x's is the same as the sample set, the variance of the predicted y values will be

$$\hat{\sigma}_y{}^2 = R^2\,\sigma_y{}^2 \tag{36}$$

where $\hat{\sigma}_y{}^2$ and $\sigma_y{}^2$ are the variances of the predicted and observed dependent variates respectively. The variance of y can be retained by adding a stochastic element to Eq. 35

$$\hat{y} = f(x_1, x_2, \ldots \cdot x_n) + t\,\sigma_y\sqrt{1-R^2} \tag{37}$$

in which t is a standard normal variate.

If the data are not normally distributed or the function of the dependent variables is nonlinear such that the distribution of the error terms is not normal, then it may be necessary to study the distribution and make its parameters functions of $\hat{y}$. If the data are normally distributed, the mean of the error $\hat{\mu}_\varepsilon$ will be zero, the variance $\hat{\sigma}_\varepsilon = \sigma_y\sqrt{1-R^2}$, and the skew of the error, $\hat{\gamma}_\varepsilon$, zero. For nonnormal distributions, both $\hat{\mu}_\varepsilon$ and $\hat{\gamma}_\varepsilon$ may not be zero. If $\hat{y}_d$ represents the predicted value of y including the stochastic element, then Eq. 37 may be rewritten

$$\hat{y}_d = f(x_1, x_2, \ldots\ldots x_n) + \hat{\mu}_\varepsilon + \hat{\sigma}_\varepsilon\,\varepsilon \tag{38}$$

where ε is the random variate calculated by the transform of Eq. 20.

If, in studying the distribution of the error, it is found that its parameters are a function of $\hat{y}$, it may be necessary to stratify $\hat{y}$ and calculate the values of $\hat{\mu}_\varepsilon$, $\hat{\sigma}_\varepsilon$, and $\hat{\gamma}_\varepsilon$ for each stratum. It might then be possible to make all three parameters a function of $\hat{y}$.

MISSING DATA

In the analysis of hydrologic data one seldom finds a complete data set. The usual tendency is to fill in the missing period with synthesized data. If the observed period of record, daily runoff for example, is being used as input to a model, the calculation of flow duration curves in this example, then the use of a regression equation or some other method for filling in the missing period is advisable. If, however, the data are to be used in the development of a stochastic process, then filling in the missing period will bias the data and is not advisable.

Incomplete hydrologic records are generally a problem in the development of stochastic processes when the correlation between variables is considered.

If, for example, the periods of record for two variables do not overlap, then it would not be possible to calculate the correlation between them. The correlation matrix of incomplete data arrays may be inconsistent. Fiering (1968) describes a method for checking the consistency of a correlation matrix. If after calculating the eigenvalues of the matrix, all of them are positive and they sum to the order of the matrix, it is consistent. Fiering in the same article describes two methods of handling inconsistent matrices. One method consists of manipulating the eigenvalues according to some predetermined scheme or an algorithm which he describes. The second method of adjusting the inconsistent matrix is by a random sampling technique. Fiering considers the first method to be the most useful, but both methods lead to the development of consistent correlation matrices.

STATISTICAL TESTS USEFUL IN THE ANALYSIS OF STOCHASTIC MODELS

The reliability of stochastic models is to a great extent a function of the statistical tests and assumptions made in the development of the models. For the most part assumptions should be backed by appropriate tests, many of which are well known to the hydrologist. Some of the better known tests which are well documented in the literature are: (1) the χ^2-test for normalcy, (2) the t-test used for comparing the means of two sets of numbers, and (3) the variance ratio or F-test for comparing the variance of two distributions. The significance of multiple correlation coefficients can also be tested by comparison with tabled values. See Crow et al. (1960, pp. 159, 178, 179, 241). The test for the significance of the multiple correlation coefficient is identical to the F-test for the significance of regression as a whole.

Since many of the distributions dealt with in stochastic modeling are not normal and are often of undetermined form, the usual tests based on the assumption of normalcy are not satisfactory. There are, however, a large number of nonparametric or distribution-free tests that can be applied to nonnormal distributions. One of the oldest and probably best known nonparametric tests is the χ^2-test of fit developed by Karl Pearson. Nonparametric tests are, as their title would indicate, distribution free. That is, no assumption is made as to the form of the distribution from which the data come. Fraser's book (1966) is a good reference on the subject. Keeping (1966) reviews in an illustrated form many of the nonparametric tests useful in hydrology.

Perhaps the most useful nonparametric tests are the Kolmogorov-Smirnov one- and two-sample tests. The Kolmogorov-Smirnov one-sample test is a nonparametric goodness of fit test relating to the cumulative distribution function. It is in general more powerful than the χ^2-square test used for the same purpose because it is based on the maximum difference between the distributions rather than on the cumulative difference. The test statistic is

$$D_n = \underset{-\infty < x < \infty}{\text{Max}} |F_n(X) - F_o(X)| \tag{39}$$

where $F_o(X)$ is the cumulative population distribution being tested, and $F_n(X)$ is the cumulative empirical distribution being tested against. Eq. 39 is the test statistic for a two-tailed test because D_n is the maximum value of the absolute difference between the distributions.

For very large samples, i.e., n > 50, the critical value of D_n, $D\alpha$, is inversely proportional to the square root of the sample size. For α equal to 0.05, $D\alpha = 1.36/\sqrt{n}$ where n is the sample size.

The test is useful in hydrologic testing because the distribution of D_n is independent of the form of $F_n(X)$. The distributions of D_n have been derived and are available in tables (Hoel 1962, Lindgren & McElrath 1959).

The test is often used to test discrete population distributions, although it is based on the continuous distribution. It is on the "safe" side because the actual significance level of the resulting test is no bigger than the one assumed in using the tables.

The Kolmogorov–Smirnov one-sample test may be extended to test whether two samples of the same or different sizes are from populations with the same distribution. The test statistic is

$$D_n = \underset{-\infty < x < \infty}{\text{Max}} |G_o(X) - F_o(X)| \tag{40}$$

where $G_o(X)$ and $F_o(X)$ are independent distributions assumed to be from the same population. For large sample sizes, the critical value of D_n, $D\alpha$, is inversely proportional to the square root of $n/2$ if the samples are both of size n. If the samples are not equal in size, but both are large, $D\alpha$ is inversely proportional to the square root of $\frac{n_1 n_2}{n_1+n_2}$ where n_1 and n_2 are the two sample sizes. For α equal to 0.05, the statistic is

$$D_{0.05} = \frac{1.36}{\sqrt{\frac{n_1 \; n_2}{n_1+n_2}}} \tag{41}$$

Another useful test is one for serial correlation. The lag-L serial correlation coefficient is given by

$$\gamma_L = \frac{\sum\limits_{i=1}^{n} x_i \, x_{i-L} - n \, \hat{\mu}_{x_i} \hat{\mu}_{x_{i-L}}}{n \, \hat{\sigma}_{x_i} \hat{\sigma}_{x_{i-L}}} \tag{42}$$

where n is the sample size and the other terms have their usual meaning. Since the $\hat{\mu}_{x_i} \equiv \hat{\mu}_{x_{i-L}}$ and $\hat{\sigma}_{x_i} \equiv \hat{\sigma}_{x_{i-L}}$, the only term that changes with L is the sum

$$\gamma_L^* = \sum_{i=1}^{n} x_i x_{i-L} \tag{43}$$

Wold & Wolfowitz (1943) show that the statistic, γ_L^*, may be tested for significance by the following method if n is greater than about 20. At the 5 % level of significance, the null hypothesis which states that there is no correlation between successive observations of the variate is rejected if

$$\frac{\gamma_L^* - \hat{\mu}_{\gamma_L^*}}{\hat{\sigma}_{\gamma_L^*}} > 1.64 \tag{44}$$

in which

$$\hat{\mu}_{\gamma_L^*} = \frac{S_1^2 - S_2}{n-1} \tag{45}$$

and

$$\hat{\sigma}^2_{\gamma_L^*} = \frac{S_2^2 - S_4}{n-1} + \frac{S_1^4 - 4S_1^2 S_2 + 4S_1S_3 + S_2^2 - 2S_4}{(n-1)\,(n-2)} - \hat{\mu}^2_{\gamma_L^*} \tag{46}$$

in which

$$S_k = \sum_{t=1}^{n} x_t^K \tag{47}$$

Other nonparametric tests which are not as often used such as run and sign tests used in checking for randomness, etc., are adequately presented in Keeping (1966). Grace & Eagleson (1966) also present a brief discussion of some of the more common statistical tests in Appendix Chapter A of their report.

THE RANDOM NUMBER GENERATOR

The random number generator used in stochastic models is one of the most important elements of the process. If the system is to perform as designed, then the "random numbers" used in the model must be random numbers. Many of the standard random number generators that are available in computer program packages are not adequate. For example, the random number generator available in the IBM* scientific subroutine package was tested by the author for length of sequence and found to repeat after about 7,000 numbers. χ^2tests of the distribution showed that it was not uniform. These results are in accordance with results of tests by MacLoren & Marsaglia (1965) and Von Gelder (1967). They found that of all the random number generators using standard methods that were tested, none gave satisfactory results.

Von Gelder found several power residue (sometimes called congruential or multiplicative) generators which performed well in his tests. His tests differed somewhat from those of MacLoren & Marsaglia but were compatible with theirs. Random numbers are generated by the power residue method using the equation

$$U_{n+1} \equiv X\, U_n \pmod{10^d} \qquad (48)$$

where d is the number of significant digits in the integer, X is a constant, and U_n and U_{n+1} are consecutive integers. One of the equations, found to be satisfactory according to their tests, using d equal to 10 and X equal to 100,003 had a recycle period of over one-half billion numbers.

Random normal numbers can be calculated from the uniform numbers by using the direct method, MacLoren & Marsaglia (1965)

$$X_1 = (-2\, \ln U_1)^{1/2} \cos 2\,\pi\, U_2 \qquad (49)$$

where U_1 and U_2 are random numbers and X_1 is a random normal number.

A second method, the "sum of uniform deviates", is predicated on the fact that if a set of random integers were divided into subsets of uniform size, the means of the subsets would be normally distributed. Using this method, a minimum of 12 random integers are usually required to calculate 1 normal number. This method is therefore less efficient than the direct method.

EXAMPLES OF STOCHASTIC MODELS

In the past stochastic models have been used in the field of hydrology primarily to generate rainfall and streamflow data, although they have been used recently to generate data in the study of pollution and economics.

* Brand names are used for simplicity and easy reference and do not constitute endorsements.

Rainfall

The stochastic models used in the generation of rainfall data have varied considerably depending primarily on the type of data generated. Annual and in some cases monthly rainfall amounts have been found to be independent, i.e., persistence or serial correlations were not statistically significant (Beer 1946, Brittan et al. 1961, Hoel 1960, Kotz & Neumann 1959, Pattison 1964, Yule 1945). Under these conditions, the synthesis of data is very simple and consists only of generating a series of random numbers having the distribution of the observed record. See Beals (1954), Markovic (1965), Merriam (1941), Slade (1936), Stidd (1953), Thom (1940), and Whitcomb (1940).

It has been found that, for the most part, daily and in some cases monthly rainfall amounts exhibit a statistically significant degree of persistence (Besson 1920, 1924, Cooke 1953, Hannan 1955, Longley 1953, Namias 1952, Sellers 1955, Uttinger 1945, Williams 1952). DeCoursey (1970), however, found no significant persistence in daily rainfall in Texas. In generating daily sequences, a problem, not normally encountered in the generation of monthly sequences, is that of generating wet and dry periods. The methods proposed for handling this problem have been either a Markov process based on transition probabilities or a probability distribution fitted to lengths of wet and dry periods. DeCoursey & Seely (1969) and Gabriel & Neumann (1962) used a two-state Markov chain as exemplified by Eq. 3. Other authors such as Caskey (1963) and Weiss (1964) found the same system to be satisfactory, whereas Cooke (1953), Jorgensen (1949), and Newnham (1916) were not satisfied with the method. Feyerherm & Bark (1965) used a second-order Markov chain and suggested that the two-state Markov chain model be made a function of the time of year. Green (1964) used exponential distributions for the lengths of wet and dry periods and sampled alternately from them. Grace & Eagleson (1966) represented the distributions of the wet and dry periods by a Weibull distribution from which they sampled alternately.

Depths of daily rainfall have been calculated for the most part by serial regression-type techniques (Enger 1957, Hammerle 1951, Pattison 1964, Sellers 1955). Grace & Eagleson (1966) introduced storm duration into the regression model. Pattison (1964) also tried using transition probabilities to select rainfall amounts from probability distributions.

The distribution of rainfall amounts for short time intervals was studied by Grace & Eagleson (1966), Pattison (1964), and Ramaseshan (1964). Ramaseshan's work is also presented briefly in Chow & Ramaseshan (1965). Pattison (1964) tried two different models for daily rainfall. One method used multistate transition probabilities for both the distribution of wet and dry periods and for the selection of rainfall depths. The second method was a linear regression model which distributed total daily rainfall. Different regression equations were used for different size classes of total rainfall. He concluded that the first model was the best for his purposes.

Ramaseshan (1964) studied only annual maximum storms and tried five different regression models. He found that the simple first-order Markov model performed as well as the other more complex ones.

Grace & Eagleson (1966) used the Weibull distribution to describe the lengths of wet and dry periods. They found the storm depths to be a function of the storm duration. Rainfall throughout the storm was distributed by using a type of urn model, the parameters of which were selected by a trial and error procedure.

Runoff

The stochastic models used in the generation of runoff do not have the wide variation of type developed for rainfall. This is because periods of no flow are not so pronounced as are dry periods in the synthesis of rainfall sequences. In almost all cases, serial correlation is statistically significant and must be considered in the models selected for the generating process. All of the models previously described in the section on autoregression methods have been used at one time or another. The trend in most recent articles is toward the use of synthetic data sequences in the study of water resources systems. As such, most of the data generation described has been of a multivariate nature. Cavadias (1966), Horms & Campbell (1967), Megerian & Pentland (1968), Payne et al. (1968), Venetis (1965), Young & Pisano (1968), and Young et al. (1969) primarily discuss the use of stochastic processes for the generation of synthetic streamflow data. In general, there is very little difference between a process used for annual flows and one used for daily or monthly flows. The major difference is the use of techniques, such as those described previously, to account for seasonal trends.

Synthetic streamflow sequences have also been used in studying the operation of reservoirs both from the standpoint of flood control and water supply. The application of stochastic processes to water supply has been described by Fiering (1967) and Yevdjevich (1965). Other references to this subject may be found in Jeng & Yevdjevich (1966) and Young et al. (1969).

A group of general references which might be of interest to someone studying the stochastic generation of runoff in detail are Beard (1965), Bhuiya & Yevdjevich (1968), Fiering (1964, 1967, 1968), Hufschmidt & Fiering (1966), Kneese & Smith (1966), Loucks & Lynn (1966), Matalas (1967), Roesner & Yevdjevich (1966), Yevdjevich (1964), and Yevdjevich & Jeng (1969). Two other excellent reference works which have several individual papers on application are *Proceedings of the International Hydrology Symposium, Fort Collins, Colorado* (1967), and *Proceedings of the Thirteenth Congress of the International Association for Hydraulic Research, Japan* (1969).

In the last few years stochastic processes have come into use in other fields. Seginer (1969), Smart et al. (1967), and Surkan (1969) have been using them in studying drainage networks; Bhavagri & Bugliorello (1965) in studying flood proofing; Benson & Matalas (1967) and Matalas & Gilroy (1968) in regional analysis; and Berthouex (1969) in the simulation of industrial wastes.

OTHER CONSIDERATIONS

The parameters upon which all synthetic sequences are based, $\hat{\mu}_x^{(p)}$, $\hat{\sigma}_x^{(p)}$, $\hat{\gamma}_x^{(p)}$, $\hat{\varrho}_x^{(p)},(1)$, $\hat{\varrho}_x^{(p)(q)}(0)$, and $\hat{\varrho}_x^{(p)(q)}(1)$, p, $q = 1, \cdots m$, are unbiased estimates of the corresponding population characteristics. However, they are biased estimates with respect to the synthetic sequences as shown by Matalas (1967). The mean $\hat{\mu}_x^{(p)}$, for example, is an unbiased estimate of $\mu_x^{(p)}$ because its expected value is $\mu_x^{(p)}$. However, it is a consistent estimator because its standard error, $\sigma_x^{(p)}/\sqrt{\mathrm{n}}$, approaches zero as n approaches ∞. Very seldom if ever is the estimate $\hat{\mu}_x^{(p)}$ actually equal to the true population value $\mu_x^{(p)}$: Yet it acts as the true population value in the synthetic sequences. If several synthetic sequences of length n are generated, the means of the synthetic sequences will be distributed about $\hat{\mu}_x^{(p)}$ and will approach $\hat{\mu}_x^{(p)}$ as n tends to infinity rather than to $\mu_x^{(p)}$.

The same argument holds for all of the distribution parameters. However, parameters of the higher order moments experience larger standard errors. They are therefore more highly biased with respect to the synthetic sequences.

As an extreme example, suppose the parameters to be used in generating a synthetic streamflow sequence are taken from a 25-year period of record with a mean of 30 units and a standard deviation of 6 units. Then the estimate of the standard deviation of the sample mean about the true mean would be

$$\sigma_x^{(p)}/\sqrt{n} \cong \hat{\sigma}_x^{(p)}/\sqrt{n} = 6/\sqrt{25} \equiv 1.2 \text{ units} \tag{50}$$

If this 25-year period were in an extremely dry cycle, then it would be possible for the mean to be as much as 2 or 3 units from the true mean, i.e.. $\mu_x^{(p)} = 32$. Thus if the sample estimate were used to generate synthetic 25-year sequences, only about 1 out of 20 would be as high as the true mean. This is because the true mean is about two standard deviations greater than the sample mean.

Another thing that must be kept in mind when working with synthetic sequences is that the parameters upon which the synthetic sequences are based were developed from a truncated distribution, i.e., the sample was of finite length. Therefore, the synthetic series cannot be used to estimate extreme events. It should be used only to generate synthetic sequences of approximately the same length as the sample period of record. Many such sequences can, however, be generated for the purpose of observing the different patterns and sequences that are likely to occur. If longer sequences are desired and the data for developing the generation scheme are available, then the "R/S analysis" described by Mandelbrot & Wallis should be considered.

SUMMARY

Synthetic data may be generated either for input to a parametric model or to provide directly an estimate of the output of a hydrologic process. In both cases the basic techniques of the generation process are the same.

The type of process used depends primarily on the purpose for which the data are being generated and on the quality and quantity of the sample data. Techniques were presented which would generate data for one or any number of variates. The data could be normal, skewed, or log normal and include serial correlation. If two or more variates are involved, cross-correlation may also be considered.

A method of generating good random numbers has been described along with a series of statistical tests which are invaluable in judging the suitability of the generation technique. Several limitations of the generated data have also been described.

REFERENCES

Beals, G. A. (1954) Specification of daily precipitation through synoptic climatology. M. S. Thesis, M.I.T.

Beard, L. R. (1965) Use of interrelated records to simulate streamflow. *J. Hydraulic Div. Amer. Soc. Civil Eng. 91* (HY 5), Part I, 13–22.

Beer, A., et al. (1946) Sequences of wet and dry months and the theory of probability. *Quart. J. Roy. Meteor. Soc. 72,* 74–86.

Benson, M. A. & Matalas, N. C. (1967) Synthetic hydrology based on regional statistical parameters. *Water Resources Res. 3* (4), 931–936.

Berthouex, P. M. (1969) Monte Carlo simulation of industrial waste discharges. *J. Sanitary Eng. Div. Amer. Soc. Civil Eng. 95* (SA 5).

Besson, L. (1920) On the comparison of meteorological data with chance results. *Monthly Weather Rev. 48,* 89–94 (translated and abridged by E. W. Woolard).

Besson, L. (1924) Sur la probabilité de la pluie. *Compt. Rend. 178,* 1743–1745.

Bhavagri, V. S. & Bugliorello, G. (1965) Flood proofing in a flood plain: A stochastic model. *J. Hydraulics Div. Amer. Soc. Civil Eng. 92* (HY 4, Part I), 205–224.

Bhuiya, R. K. & Yevdjevich, V. M. (1968) *Effects of truncation on dependence in hydrological time series.* Colorado State University Paper No. 31. Fort Collins, Colorado, November 1968.

Blackman, R. B. & Tukey, J. W. (1958) *The measurement of power spectra.* Dover Publications, Inc., New York.

Brittan, M. R. et al. (1961) *Past and probable future variations in stream flow in the upper Colorado River.* 5 Parts, University of Colorado, Bureau of Economic Research, Oct. 1961.

Caffey, J. E. (1965) *Inter-station correlations in annual precipitation and in annual effective precipitation.* Colorado State University Paper No. 6. Fort Collins, Colorado, June 1965.

Caskey, J. E., Jr. (1963) A Markov chain model for the probability of precipitation occurrence in intervals of various length. *Monthly Weather Rev. 91,* 298–301.

Cavadias, C. G. (1966) River flow as a stochastic process. *Statistical methods in hydrology.* (Proceedings of Hydrology Symposium No. 5. McGill University, February 1966), pp. 315–360.

Chow, V. T. & Ramaseshan, S. (1965) Sequential generation of rainfall and runoff data. *J. Hydraulics Div. Amer. Soc. Civil Eng. 91* (HY 4, Part I), 205–224.

Cooke, D. S. (1953) The duration of wet and dry spells at Moncton, New Brunswick. *Quart. Roy. Meteorol. Soc. 79* (342), 536–538.

Cox, D. R. & Miller, H. D. (1965) *The theory of stochastic processes.* John Wiley & Sons, Inc., New York.

Crow, E. L., Davis, F. A. & Maxfield, M. W. (1960) *Statistics manual.* Dover Publications, Inc., New York.

DeCoursey, D. G. & Seely, E. H. (1969) Indirect determination of synthetic runoff. *Proc. Thirteenth Congr. Int. Ass. Hydraulic Res.,* Vol. 1, 31 August – 5 Sept. 1969. Kyoto, Japan.

DeCoursey, D. G. (1970) Use of multiple discriminant analysis to evaluate the effects of land use change on the simulated yield of a watershed. Ph. D. dissertation, Georgia Institute of Technology.

Dumas, Andre J. & Morel-Seytoux, H. J. (1969) *Statistical discrimination of change in daily runoff.* Colorado State University Paper No. 34. Fort Collins, Colorado, August 1969.

Enger, I. (1957) Some attempts at predicting a meteorological time series from its past history. M. S. thesis, M.I.T.

Feyerherm, A. M. & Bark, L. D. (1965) Statistical methods for persistent precipitation patterns. *J. Appl. Meteorol. 4,* 320–328.

Fiering, M. B. (1964) Multivariate technique for synthetic hydrology. *J. Hydraulics Div. Amer. Soc. Civil Eng. 90* (HY 5), 43–60.

Fiering, M. B. (1967) *Streamflow synthesis.* Harvard University Press, Cambridge, Mass.

Fiering, M. B. (1968) Schemes for handling inconsistent matrices. *Water Resources Res. 4* (2), 291–298.

Fraser, D. A. (1966) *Nonparametric methods in statistics.* John Wiley & Sons, Inc., New York.

Gabriel, K. R. & Neumann, J. (1962) A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Roy. Meteorol. Soc. Lond. 88,* 90–95.

Grace, R. A. & Eagleson, P. S. (1966) *The synthesis of short-time-increment rainfall sequences.* Hydrodynamics Lab. Report 91, Dept. of Civil Engineering, Mass. Inst. Tech., May 1966.

Green, J. R. (1964) A model for rainfall occurrence. *J. Roy. Stat. Soc. B 26,* 345–353.

Hammerle, J. F. (1951) Linear prediction of discrete stationary time series. M. S. Thesis, M.I.T.

Hannan, E. J. (1955) A test for singularities in Sydney rainfall. *Aust. J. Physics 8,* 289–297.

Hoel, P. G. (1960) *Elementary statistics.* John Wiley & Sons, New York.

Hoel, P. G. (1962) *Introduction to mathematical statistics.* 3rd ed., John Wiley & Sons, New York.

Horms, A. A. & Campbell, T. H. (1967) An extension to the Thomas-Fiering model for the sequential generation of streamflow. *Water Resources Res. 3* (3), 653–662.

Horn, L. H. & Bryson, R. A. (1960) Harmonic analysis of the annual march of precipitation over the United States. *Ass. Amer. Geograph. Ann. 50,* 157–171.

Hufschmidt, M. M. & Fiering, M. B. (1966) *Simulation techniques for design of water resource systems.* Harvard University Press, Cambridge, Mass.

Hurst, H. E. (1951) Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng. 116,* 770.

Jeng, R. I. & Yevdjevich, V. M. (1966) *Stochastic properties of lake outflows.* Colorado State University Hydrology Paper No. 14. Fort Collins, Colorado, August 1966.

Jorgensen, D. L. (1949) Persistency of rain and no-rain periods during the winter of San Francisco. *Monthly Weather Rev. 77,* 303–307.

Julian, P. R. (1961) *A study of the statistical predictability of stream-runoff in the upper Colorado river basin.* (*Past and probable future variations in stream flow in the upper Colorado River,* Part II). University of Colorado, Bureau of Economic Research, Boulder, Colo., October, 1961.

Keeping, E. S. (1966) Distribution-free methods in statistics. *Statistical methods in hydrology.* (Proceedings of Hydrology Symposium No. 5, McGill University, February 1966). Pp. 211–247.

Kneese, A. V. & Smith, C. S. (eds.) (1966) *Water research.* Johns Hopkins Press, Baltimore, Md.

Kotz, S. & Neumann, J. (1959) Autocorrelation in precipitation amounts. *J. Meteorol. 16,* 683–685.

Lindgren, B. W. & McElrath, G. W. (1959) *Introduction to probability and statistics.* Macmillan Co., New York.

Longley, R. W. (1953) The length of wet and dry periods. *Quart. J. Roy Meteorol. Soc. 79,* 520.

Loucks, D. P. & Lynn, W. R. (1966) Probabilistic models for predicting stream quality. *Water Resources Res. 2* (3), 593–606.

MacLoren, M. D. & Marsaglia, G. (1965) Uniform random number generators. *J. A.C.M 12,* 83–89.

Mandelbrot, B. & Wallis, J. R. (1968) Noah, Joseph, and operational hydrology. *Water Resources Res. 4,* 909–918.

Mandelbrot, B. & Wallis, J. R. (1969 a) Computer experiments with fractional gaussian noises. Part I. Averages and variances. *Water Resources Res. 5,* 228–241.

Mandelbrot, B. & Wallis, J. R. (1969 b) Computer experiments with fractional gaussian noises. Part II. Rescaled ranges and spectra. *Water Resources Res. 5,* 242–259.

Mandelbrot, B. & Wallis, J. R. (1969 c) Computer experiments with fractional gaussian noises. Part III. Mathematical appendix. *Water Resources Res. 5,* 260–267.

Mandelbrot, B. & Wallis, J. R. (1969 d) Some long run properties of geophysical records. *Water Resources Res. 5,* 321–340.

Mandelbrot, B. & Wallis, J. R. (1969 e) Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. *Water Resources Res. 5,* 967–988.

Markovic, R. D. (1965) *Probability functions of best fit to distributions of annual precipitation and runoff.* Colorado State University, Hydrology Paper No. 8, August 1965.

Matalas, N. C. (1966) Some aspects of time series analysis in hydrologic studies. *Statistical methods in hydrology,* pp. 271–314. Proceedings of Hydrology Symposium No. 5, McGill University, February 1966.

Matalas, N. C. (1967) Mathematical assessment of synthetic hydrology. *Water Resources Res. 3* (4), 937–945.

Matalas, N. C. & Gilroy, E. J. (1968) Some comments on regionalization in hydrologic studies. *Water Resources Res. 4* (6), 1361–1370.

Megerian, E. & Pentland, R. L. (1968) Simulation of Great Lakes basin water supplies. *Water Resources Res. 4* (1), 11–18.

Merriam, C. F. (1941) Long-term constancy of rainfall. Paper presented at Annual Meeting of the American Geophysical Union, Section of Hydrology.

Namias, J. (1952) The annual course of month-to-month persistence in climatic anomalies. *Bull. Amer. Meteorol. Soc. 33* (7), 279–285.

Newnham, E. V. (1916) The persistence of wet and dry weather. *Quart. J. Roy. Meteorol. Soc. 42* (179), 153–162.

Parzen, Emanuel (1962) *Stochastic processes.* Holden-Day, Inc., San Francisco, Calif.

Pattison, A. (1964) *Synthesis of rainfall data.* Stanford University, Department of Civil Engineering, Technical Report 40.

Payne, K., Newman, W. R. & Kerri, K. D. (1968) Daily streamflow simulation. *J. Hydrolics Div. Amer. Soc. Civil Eng. 94* (HY 4), 904–924.

Proceedings of the International Hydrology Symposium. 6–8 Sept. 1967, Fort Collins, Colorado.

Proceedings of the Thirteenth Congress of the International Association for Hydraulic Research, Vol. 1. 31 Aug – 5 Sept. 1969, Kyoto, Japan.

Quimpo, R. G. (1967) *Stochastic model of daily river flow sequences.* Colorado State University Paper No. 18. Fort Collins, Colorado, February 1967.

Roesner, L. A. & Yevdjevich, V. M. (1966) *Mathematical models for time series of monthly precipitation and monthly runoff.* Colorado State University Paper No. 15. Fort Collins, Colorado, October 1966.

Ramaseshan, S. (1964) A stochastic analysis of rainfall and runoff characteristics by sequential generation and simulation. Ph. D. thesis, University of Illinois, 1964.

Seginer, I. (1969) Random walk and random roughness models of drainage networks. *Water Resources Res. 5* (3), 591–607.

Sellers, W. D. (1955) Prediction of daily precipitation by using statistical methods. M. S. thesis. M.I.T.

Slade, J. J., Jr. (1936) An asymmetric probability function. *Trans: Amer. Soc. Civil Eng. 101,* 35.

Smart, J. S., Surkan, A. J. & Considine, J. P. (1967) Digital simulation of channel networks. *Proc. I.A.S.H.,* Berne, pp. 87–98.

Stidd, C. K. (1953) Cube-root-normal precipitation distributions. *Amer. Geophys. Union Trans. 34* (1), 31–35.

Surkan, A. J. (1969) Synthetic hydrographs: Effects of network geometry. *Water Resources Res. 5* (1), 112–128.

Thom, H. C. S. (1940) On the statistical analysis of rainfall data. *Amer. Geophys. Union Trans. 21* (Part II), 490.

Thomas, H. A., Jr. & Fiering, Myron B. (1962) The mathematical synthesis of streamflow sequences. *The design of water resource systems,* ed. Arthur Maass. Harvard University Press, Cambridge, Mass.

Todorovic, P. & Yevdjevich, V. M. (1969) *Stochastic process of precipitation.* Colorado State University Paper No. 35. Fort Collins, Colorado, September 1969.

Uttinger, H. (1945) Die Niederschlagsverhältnisse der Südschweiz 1901–1940. *Ann. Schweiz. Meteorol. 82,* 23.

Venetis, C. (1969) A stochastic model of monthly reservoir storage. *Water Resources Res. 5* (3), 729–734.

Von Gelder, A. (1967) Some new results in pseudo-random number generation. *J. A.C.M. 14* (1), 785–793.

Weiss, L. L. (1964) Sequences of wet and dry described by a Markov chain probability model. *Monthly Weather Rev. 92,* 169–176.

Whitcomb, Margaret (1940) A statistical study of rainfall data. M. S. Thesis, M.I.T.

Williams, C. B. (1952) Sequences of wet and dry days considered in relation to the logarithmic series. *Quart. J. Roy. Meteorol. Soc. 78* (335), 91–96.

Wold, A. & Wolfowitz, J. (1943) An exact test for randomness in the nonparametric case based on serial correlation. *Ann. Math. Stat. 14,* 378–388.

Wold, H. (1964) *A study in the analysis of stationary time series.* 2nd ed., Almquist and Widsell, Stockholm.

Yevdjevich, V. M. (1961) *Some general aspects of fluctuations of annual runoff in the upper Colorado river basin. Part III of "Past and probable future variations in stream flow in the upper Colorado River."* University of Colorado, Bureau of Economic Research, October 1961.

Yevdjevich, V. M. (1964) *Fluctuations of wet and dry years. Part II. Analysis by serial correlation.* Colorado State University Paper No. 4. Fort Collins, Colorado, June 1964.

Yevdjevich, V. M. (1965) *The application of surplus, deficit and range in hydrology.* Colorado State University Hydrology Paper No. 10. Fort Collins, Colorado, September 1965.

Yevdjevich, V. M. & Jeng, R. I. (1969) *Properties of non-homogeneous hydrologic series.* Colorado State University Paper No. 32. Fort Collins, Colorado, April 1969.

Young, G. K. (1968) Discussion of „Mathematical assessment of synthetic hydrology" by N. C. Matalas. *Water Resources Res. 4* (3), 681–682.

Young, G. K. & Pisano, W. C. (1968) Operational hydrology using residuals. *J. Hydraulics Div. Amer. Soc. Civil Eng. 94* (HY 4), 909–924.

Young, G. K., Somers, W. P., Pisano, W. C. & Fitch, W. N. (1969) Assessing upland reservoirs using a daily flow model. *Water Resources Res. 5* (2), 362–379.

Yule, G. U. (1945) On a method of studying time series based on their internal correlations. *J. Roy. Statistics Soc. 108,* 208–225.

Address:
Southern Plains Watershed Research Center, USDA-ARS-SWD,
P.O. Box 400,
Chickasha, Oklahoma 73018,
U.S.A.

Received 1 May 1971.