7. Concepts in Probability, Statistics and Stochastic Modelling

7 Concepts in Probability, Statistics and Stochastic Modelling

Events that cannot be predicted precisely are often called random. Many if not most of the inputs to, and processes that occur in, water resources systems are to some extent random. Hence, so too are the outputs or predicted impacts, and even people's reactions to those outputs or impacts. To ignore this randomness or uncertainty is to ignore reality. This chapter introduces some of the commonly used tools for dealing with uncertainty in water resources planning and management. Subsequent chapters illustrate how these tools are used in various types of optimization, simulation and statistical models for impact prediction and evaluation.

1. Introduction

Uncertainty is always present when planning, developing, managing and operating water resources systems. It arises because many factors that affect the performance of water resources systems are not and cannot be known with certainty when a system is planned, designed, built, managed and operated. The success and performance of each component of a system often depends on future meteorological, demographic, economic, social, technical, and political conditions, all of which may influence future benefits, costs, environmental impacts, and social acceptability. Uncertainty also arises due to the stochastic nature of meteorological processes such as evaporation, rainfall and temperature. Similarly, future populations of towns and cities, per capita water-usage rates, irrigation patterns and priorities for water uses, all of which affect water demand, are never known with certainty.

There are many ways to deal with uncertainty. One, and perhaps the simplest, approach is to replace each uncertain quantity either by its average (i.e., its mean or expected value), its median, or by some critical (e.g., 'worst-case') value, and then proceed with a deterministic approach. Use of *expected* or *median values* of uncertain quantities may be adequate if the uncertainty or variation in a quantity is reasonably small and does not critically affect the performance of the system. If expected or median values of uncertain parameters or variables are used in a deterministic model, the planner can then assess the importance of uncertainty by means of sensitivity analysis, as is discussed later in this and the two subsequent chapters.

Replacement of uncertain quantities by either expected, median or worst-case values can grossly affect the evaluation of project performance when important parameters are highly variable. To illustrate these issues, consider the evaluation of the recreation potential of a reservoir. Table 7.1 shows that the elevation of the water surface varies over time depending on the inflow and demand for water. The table indicates the pool levels and their associated probabilities as well as the expected use of the recreation facility with different pool levels.

The average pool level $\bar{L}$ is simply the sum of each possible pool level times its probability, or

$$\bar{L} = 10(0.10) + 20(0.25) + 30(0.30) + 40(0.25) + 50(0.10) = 30 \qquad (7.1)$$

This pool level corresponds to 100 visitor-days per day:

$$VD(\bar{L}) = 100 \text{ visitor-days per day} \qquad (7.2)$$

A worst-case analysis might select a pool level of ten as a critical value, yielding an estimate of system performance equal to 100 visitor-days per day:

$$VD(L_{\text{low}}) = VD(10) = 25 \text{ visitor-days per day} \qquad (7.3)$$

E021101a

possible pool levels	probability of each level	recreation potential in visitor-days per day for reservoir with different pool levels
10	0.10	25
20	0.25	75
30	0.30	100
40	0.25	80
50	0.10	70

Table 7.1. Data for determining reservoir recreation potential.

Neither of these values is a good approximation of the average visitation rate, that is

$$\begin{aligned}\overline{VD} &= 0.10\ VD(10) + 0.25\ VD(20) + 0.30\ VD(30) \\ &\quad + 0.25\ VD(40) + 0.10\ VD(50) \\ &= 0.10(25) + 0.25(75) + 0.30(100) + 0.25(80) \\ &\quad + 0.10(70) \qquad (7.4) \\ &= 78.25 \text{ visitor-days per day}\end{aligned}$$

Clearly, the average visitation rate, $\overline{VD} = 78.25$, the visitation rate corresponding to the average pool level $VD(\overline{L}) = 100$, and the worst-case assessment $VD(L_{\text{low}}) = 25$, are very different.

Using only average values in a complex model can produce a poor representation of both the average performance and the possible performance range. When important quantities are uncertain, a comprehensive analysis requires an evaluation of both the expected performance of a project and the risk and possible magnitude of project failures in a physical, economic, ecological and/or social sense.

This chapter reviews many of the methods of probability and statistics that are useful in water resources planning and management. Section 2 is a condensed summary of the important concepts and methods of probability and statistics. These concepts are applied in this and subsequent chapters of this book. Section 3 presents several probability distributions that are often used to model or describe the distribution of uncertain quantities. The section also discusses methods for fitting these distributions using historical information, and methods of assessing whether the distributions are adequate representations of the data. Sections 4, 5 and 6 expand upon the use of these mathematical models, and discuss alternative parameter estimation methods.

Section 7 presents the basic ideas and concepts of the stochastic processes or time series. These are used to model streamflows, rainfall, temperature or other phenomena whose values change with time. The section contains a description of Markov chains, a special type of stochastic process used in many stochastic optimization and simulation models. Section 8 illustrates how synthetic flows and other time-series inputs can be generated for stochastic simulations. Stochastic simulation is introduced with an example in Section 9.

Many topics receive only brief treatment in this introductory chapter. Additional information can be found in applied statistical texts or book chapters such as Benjamin and Cornell (1970), Haan (1977), Kite (1988), Stedinger et al. (1993), Kottegoda and Rosso (1997), and Ayyub and McCuen (2002).

2. Probability Concepts and Methods

This section introduces the basic concepts and definitions used in analyses involving probability and statistics. These concepts are used throughout this chapter and later chapters in the book.

2.1. Random Variables and Distributions

The basic concept in probability theory is that of the *random variable*. By definition, the value of a random variable cannot be predicted with certainty. It depends, at least in part, on the outcome of a chance event. Examples are: (1) the number of years until the flood stage of a river washes away a small bridge; (2) the number of times during a reservoir's life that the level of the pool will drop below a specified level; (3) the rainfall depth next month; and (4) next year's maximum flow at a gauge site on an unregulated stream. The values of all of these random events or variables are not knowable before the event has occurred. Probability can be used to describe the likelihood that these random variables will equal specific values or be within a given range of specific values.

The first two examples illustrate *discrete random variables*, random variables that take on values that are discrete (such as positive integers). The second two examples illustrate *continuous random variables*. Continuous random variables take on any values within a specified range of values. A property of all continuous random variables is that the probability that the value of any of those random variables will equal some specific number – any specific number – is always zero. For example, the probability that the total rainfall depth in a month will be exactly 5.0 cm is zero, while the probability that the total rainfall will lie between 4 and 6 cm could be nonzero. Some random variables are combinations of continuous and discrete random variables.

Let X denote a random variable and x a possible value of that random variable X. Random variables are generally denoted by capital letters, and particular values they take on by lowercase letters. For any real-valued random variable X, its *cumulative distribution function* $F_X(x)$, often denoted as just the cdf, equals the probability that the value of X is less than or equal to a specific value or threshold x:

$$F_X(x) = \Pr[X \leq x] \tag{7.5}$$

This cumulative distribution function $F_X(x)$ is a non-decreasing function of x because

$$\Pr[X \leq x] \leq \Pr[X \leq x + \delta] \quad \text{for} \quad \delta > 0 \tag{7.6}$$

In addition,

$$\lim_{x \to +\infty} F_X(x) = 1 \tag{7.7}$$

and

$$\lim_{x \to -\infty} F_X(x) = 0 \tag{7.8}$$

The first limit equals 1 because the probability that X takes on some value less than infinity must be unity; the second limit is zero because the probability that X takes on no value must be zero.

The left half of Figure 7.1 illustrates the cumulative distribution function (upper) and its derivative, the probability density function, $f_X(x)$, (lower) of a continuous random variable X.

If X is a real-valued discrete random variable that takes on specific values $x_1, x_2, \ldots$, then the *probability mass function* $p_X(x_i)$ is the probability X takes on the value x_i.

$$p_X(x_i) = \Pr[X = x_i] \tag{7.9}$$

The value of the cumulative distribution function $F_X(x)$ for a discrete random variable is the sum of the probabilities of all x_i that are less than or equal to x.

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i) \tag{7.10}$$

The right half of Figure 7.1 illustrates the cumulative distribution function (upper) and the probability mass function $p_X(x_i)$ (lower) of a discrete random variable.

The *probability density function* $f_X(x)$ (lower left plot in Figure 7.1) for a continuous random variable X is the analogue of the probability mass function (lower right plot in Figure 7.1) of a discrete random variable X. The probability density function, often called the pdf, is the derivative of the cumulative distribution function so that:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} \geq 0 \tag{7.11}$$

It is necessary to have

$$\int_{-\infty}^{+\infty} f_X(x) = 1 \tag{7.12}$$

Equation 7.12 indicates that the area under the probability density function is 1. If a and b are any two constants, the cumulative distribution function or the density function may be used to determine the probability that X is greater than a and less than or equal to b where

$$\Pr[a < X \leq b] = F_X(b) - F_X(a) = \int_a^b f_X(x)\mathrm{d}x \tag{7.13}$$

The area under a probability density function specifies the relative frequency with which the value of a continuous random variable falls within any specified range of values, that is, any interval along the horizontal axis.

Life is seldomly so simple that only a single quantity is uncertain. Thus, the joint probability distribution of two or more random variables can also be defined. If X and Y are two continuous real-valued random variables, their joint cumulative distribution function is:

$$\begin{aligned} F_{XY}(x, y) &= \Pr[X \leq x \text{ and } Y \leq y] \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v)\mathrm{d}u\,\mathrm{d}v \end{aligned} \tag{7.14}$$

If two random variables are discrete, then

$$F_{XY}(x, y) = \sum_{x_i \leq x} \sum_{y_i \leq y} p_{XY}(x_i, y_i) \tag{7.15}$$

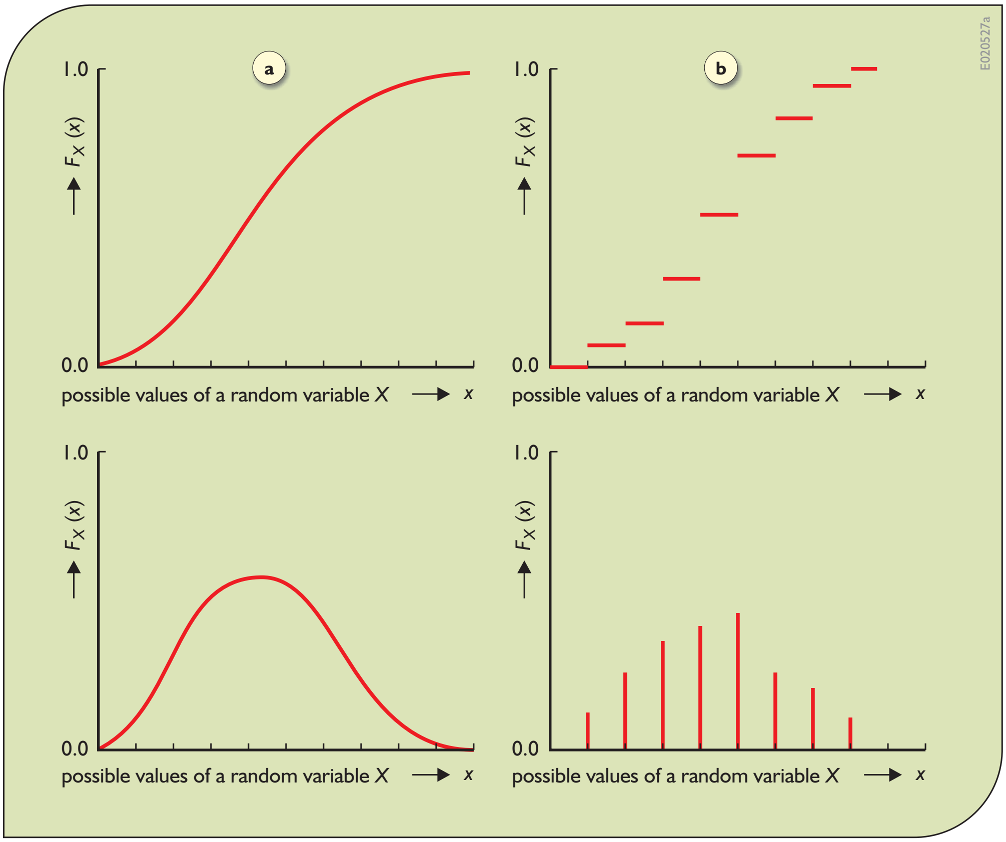

Figure 7.1. Cumulative distribution and probability density or mass functions of random variables: (a) continuous distributions; (b) discrete distributions.

where the joint probability mass function is:

$$p_{XY}(x_i, y_i) = \Pr[X = x_i \text{ and } Y = y_i] \tag{7.16}$$

If X and Y are two random variables, and the distribution of X is not influenced by the value taken by Y, and vice versa, then the two random variables are said to be *independent*. For two independent random variables X and Y, the joint probability that the random variable X will be between values a and b and that the random variable Y will be between values c and d is simply the product of those separate probabilities.

$$\begin{aligned}&\Pr[a \le X \le b \quad \text{and} \quad c \le Y \le d]\\ &\quad = \Pr[a \le X \le b] \times \Pr[c \le Y \le d]\end{aligned} \tag{7.17}$$

This applies for any values a, b, c, and d. As a result,

$$F_{XY}(x, y) = F_X(x)F_Y(y) \tag{7.18}$$

which implies for continuous random variables that

$$f_{XY}(x, y) = f_X(x)f_Y(y) \tag{7.19}$$

and for discrete random variables that

$$p_{XY}(x, y) = p_X(x)p_Y(y) \tag{7.20}$$

Other useful concepts are those of the *marginal* and *conditional distributions*. If X and Y are two random variables whose joint cumulative distribution function $F_{XY}(x, y)$ has been specified, then $F_X(x)$, the marginal cumulative distribution of X, is just the cumulative distribution of X ignoring Y. The marginal cumulative distribution function of X equals

$$F_X(x) = \Pr[X \le x] = \lim_{y\to\infty} F_{XY}(x, y) \tag{7.21}$$

where the limit is equivalent to letting Y take on any value. If X and Y are continuous random variables, the marginal density of X can be computed from

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)dy \tag{7.22}$$

The conditional cumulative distribution function is the cumulative distribution function for X given that Y has taken a particular value y. Thus the value of Y may have been observed and one is interested in the resulting conditional distribution for the so far unobserved value of X. The conditional cumulative distribution function for continuous random variables is given by

$$F_{X|Y}(x|y) = \Pr[X \le x | Y = y] = \frac{\int_{-\infty}^{x} f_{XY}(s, y)\mathrm{d}s}{f_Y(y)} \quad (7.23)$$

where the conditional density function is

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)} \quad (7.24)$$

For discrete random variables, the probability of observing $X = x$, given that $Y = y$ equals

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)} \quad (7.25)$$

These results can be extended to more than two random variables. Kottegoda and Rosso (1997) provide more detail.

2.2. Expectation

Knowledge of the probability density function of a continuous random variable, or of the probability mass function of a discrete random variable, allows one to calculate the expected value of any function of the random variable. Such an expectation may represent the average rainfall depth, average temperature, average demand shortfall or expected economic benefits from system operation. If g is a real-valued function of a continuous random variable X, the expected value of $g(X)$ is:

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\mathrm{d}x \quad (7.26)$$

whereas for a discrete random variable

$$\mathrm{E}[g(X)] = \sum_i g(x_i) p_X(x_i) \quad (7.27)$$

The *expectation operator*, E[·], has several important properties. In particular, the expectation of a linear function of X is a linear function of the expectation of X. Thus, if a and b are two non-random constants,

$$\mathrm{E}[a + bX] = a + b\mathrm{E}[X] \quad (7.28)$$

The expectation of a function of two random variables is given by

$$\mathrm{E}[g(X, Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x, y) f_{XY}(x, y)\mathrm{d}x\,\mathrm{d}y$$

or

$$\mathrm{E}[g(X, Y)] = \sum_i \sum_j g(x_i, y_i) p_{XY}(x_i, y_i) \quad (7.29)$$

If X and Y are independent, the expectation of the product of a function $g(\cdot)$ of X and a function $h(\cdot)$ of Y is the product of the expectations:

$$\mathrm{E}[g(X)\, h(Y)] = \mathrm{E}[g(X)]\, \mathrm{E}[h(Y)] \quad (7.30)$$

This follows from substitution of Equations 7.19 and 7.20 into Equation 7.29.

2.3. Quantiles, Moments and Their Estimators

While the cumulative distribution function provides a complete specification of the properties of a random variable, it is useful to use simpler and more easily understood measures of the central tendency and range of values that a random variable may assume. Perhaps the simplest approach to describing the distribution of a random variable is to report the value of several quantiles. The pth quantile of a random variable X is the smallest value x_p such that X has a probability p of assuming a value equal to or less than x_p:

$$\Pr[X < x_p] \le p \le \Pr[X \le x_p] \quad (7.31)$$

Equation 7.31 is written to insist if at some value x_p, the cumulative probability function jumps from less than p to more than p, then that value x_p will be defined as the pth quantile even though $F_X(x_p) \ne p$. If X is a continuous random variable, then in the region where $f_X(x) > 0$, the quantiles are uniquely defined and are obtained by solution of

$$F_X(x_p) = p \quad (7.32)$$

Frequently reported quantiles are the *median* $x_{0.50}$ and the *lower* and *upper quartiles* $x_{0.25}$ and $x_{0.75}$. The median describes the location or central tendency of the distribution of X because the random variable is, in the continuous case, equally likely to be above as below that value. The interquartile range $[x_{0.25}, x_{0.75}]$ provides an easily understood description of the range of values that the random variable might assume. The pth quantile is also the 100 p percentile.

In a given application – particularly when safety is of concern – it may be appropriate to use other quantiles. In floodplain management and the design of flood control structures, the 100-year flood $x_{0.99}$ is a commonly selected design value. In water quality management, a river's minimum seven-day-average low flow expected once in ten years is commonly used in the United States as the critical planning value: Here the one-in-ten year value is the 10th percentile of the distribution of the annual minima of the seven-day average flows.

The natural sample estimate of the median $x_{0.50}$ is the median of the sample. In a sample of size n where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the observations ordered by magnitude, and for a non-negative integer k such that $n = 2k$ (even) or $n = 2k + 1$ (odd), the sample estimate of the median is

$$\hat{x}_{0.50} = \begin{cases} x_{(k+1)} & \text{for } n = 2k+1 \\ \frac{1}{2}\left[x_{(k)} + x_{(k+1)}\right] & \text{for } n = 2k \end{cases} \qquad (7.33)$$

Sample estimates of other quantiles may be obtained by using $x_{(i)}$ as an estimate of x_q for $q = i/(n+1)$ and then interpolating between observations to obtain $\hat{x}_p$ for the desired p. This only works for $1/(n+1) \leq p \leq n/(n+1)$ and can yield rather poor estimates of x_p when $(n+1)p$ is near either 1 or n. An alternative approach is to fit a reasonable distribution function to the observations, as discussed in Section 3, and then estimate x_p using Equation 7.32, where $F_X(x)$ is the fitted distribution.

Another simple and common approach to describing a distribution's centre, spread and shape is by reporting the moments of a distribution. The first moment about the origin is the *mean* of X and is given by

$$\mu_X = \mathrm{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\mathrm{d}x \qquad (7.34)$$

Moments other than the first are normally measured about the mean. The second moment measured about the mean is the *variance*, denoted $\mathrm{Var}(X)$ or σ_X^2, where:

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)^2] \qquad (7.35)$$

The *standard deviation* σ_X is the square root of the variance. While the mean μ_X is a measure of the central value of X, the standard deviation σ_X is a measure of the spread of the distribution of X about μ_X.

Another measure of the variability in X is the *coefficient of variation*,

$$CV_X = \frac{\sigma_X}{\mu_X} \qquad (7.36)$$

The coefficient of variation expresses the standard deviation as a proportion of the mean. It is useful for comparing the relative variability of the flow in rivers of different sizes, or of rainfall variability in different regions when the random variable is strictly positive.

The third moment about the mean, denoted λ_X, measures the asymmetry, or *skewness,* of the distribution:

$$\lambda_X = \mathrm{E}[(X - \mu_X)^3] \qquad (7.37)$$

Typically, the dimensionless coefficient of skewness γ_X is reported rather than the third moment λ_X. The coefficient of skewness is the third moment rescaled by the cube of the standard deviation so as to be dimensionless and hence unaffected by the scale of the random variable:

$$\gamma_X = \frac{\lambda_X}{\sigma_X^3} \qquad (7.38)$$

Streamflows and other natural phenomena that are necessarily non-negative often have distributions with positive skew coefficients, reflecting the asymmetric shape of their distributions.

When the distribution of a random variable is not known, but a set of observations $\{x_1, \ldots, x_n\}$ is available, the moments of the unknown distribution of X can be estimated based on the sample values using the following equations. The sample estimate of the mean:

$$\bar{X} = \sum_{i=1}^{n} X_i/n \qquad (7.39a)$$

The sample estimate of the variance:

$$\hat{\sigma}_X^2 = S_X^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad (7.39b)$$

The sample estimate of skewness:

$$\hat{\lambda}_X = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (X_i - \bar{X})^3 \qquad (7.39c)$$

The sample estimate of the coefficient of variation:

$$\hat{C}V_X = S_X/\bar{X} \qquad (7.39d)$$

The sample estimate of the coefficient of skewness:

$$\hat{\gamma}_X = \hat{\lambda}_X/S_X^3 \qquad (7.39e)$$

The sample estimate of the mean and variance are often denoted as $\bar{x}$ and s_X^2 where the lower case letters are used when referring to a specific sample. All of these sample estimators provide only estimates of actual or true values. Unless the sample size n is very large, the difference between the estimators and the true values of $\mu_X, \sigma_X^2, \lambda_X, CV_X$, and γ_X may be large. In many ways, the field of statistics is about the precision of estimators of different quantities. One wants to know how well the mean of twenty annual rainfall depths describes the true expected annual rainfall depth, or how large the difference between the estimated 100-year flood and the true 100-year flood is likely to be.

As an example of the calculation of moments, consider the flood data in Table 7.2. These data have the following sample moments:

$$\bar{x} = 1549.2$$

$$s_X = 813.5$$

$$CV_X = 0.525$$

$$\hat{\gamma}_X = 0.712$$

As one can see, the data are positively skewed and have a relatively large coefficient of variance.

E021101b

date	discharge m ³/s	date	discharge m ³/s
1930	410	1951	3070
1931	1150	1952	2360
1932	899	1953	1050
1933	420	1954	1900
1934	3100	1955	1130
1935	2530	1956	674
1936	758	1957	683
1937	1220	1958	1500
1938	1330	1959	2600
1939	1410	1960	3480
1940	3100	1961	1430
1941	2470	1962	809
1942	929	1963	1010
1943	586	1964	1510
1944	450	1965	1650
1946	1040	1966	1880
1947	1470	1967	1470
1948	1070	1968	1920
1949	2050	1969	2530
1950	1430	1970	1490

* Value for 1945 is missing.

Table 7.2. Annual maximum discharges on Magra River, Italy, at Calamazza, 1930 – 70*.

When discussing the accuracy of sample estimates, two quantities are often considered, *bias* and *variance*. An estimator $\hat{\theta}$ of a known or unknown quantity θ is a function of the observed values of the random variable X, say in n different time periods, $X_1, \ldots, X_n$, that will be available to estimate the value of θ; $\hat{\theta}$ may be written $\hat{\theta}\,[X_1, X_2, \ldots, X_n]$ to emphasize that $\hat{\theta}$ itself is a random variable. Its value depends on the sample values of the random variable that will be observed. An estimator $\hat{\theta}$ of a quantity θ is biased if $E[\hat{\theta}] \neq \theta$ and unbiased if $E[\hat{\theta}] = \theta$. The quantity $\{E[\hat{\theta}] - \theta\}$ is generally called the *bias of the estimator*.

An unbiased estimator has the property that its expected value equals the value of the quantity to be estimated. The sample mean is an unbiased estimate of the population mean μ_X because

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu_X \qquad (7.40)$$

The estimator S_X^2 of the variance of X is an unbiased estimator of the true variance σ_X^2 for independent observations (Benjamin and Cornell, 1970):

$$E\left[S_X^2\right] = \sigma_X^2 \qquad (7.41)$$

However, the corresponding estimator of the standard deviation, S_X, is in general a biased estimator of σ_X because

$$E[S_X] \neq \sigma_X \qquad (7.42)$$

The second important statistic often used to assess the accuracy of an estimator $\hat{\theta}$ is the variance of the estimator Var $\hat{\theta}$, which equals $E\{(\hat{\theta} - E[\hat{\theta}])^2\}$. For the mean of a set of independent observations, the variance of the sample mean is:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma_X^2}{n} \qquad (7.43)$$

It is common to call $\sigma_x/\sqrt{n}$ the *standard error* of $\hat{x}$ rather than its standard deviation. The standard error of an average is the most commonly reported measure of its precision.

The bias measures the difference between the average value of an estimator and the quantity to be estimated.

The variance measures the spread or width of the estimator's distribution. Both contribute to the amount by which an estimator deviates from the quantity to be estimated. These two errors are often combined into the *mean square error*. Understanding that θ is fixed and the estimator $\hat{\theta}$ is a random variable, the mean squared error is the expected value of the squared distance (error) between θ and its estimator $\hat{\theta}$:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \theta)^2] = \text{E}\{[\hat{\theta} - \text{E}(\hat{\theta})] + [\text{E}(\hat{\theta}) - \theta]\}^2 \\ &= [\text{Bias}]^2 + \text{Var}(\hat{\theta}) \end{aligned} \qquad (7.44)$$

where [Bias] is $\text{E}(\hat{\theta}) - \theta$.

Equation 7.44 shows that the MSE, equal to the expected average squared deviation of the estimator $\hat{\theta}$ from the true value of the parameter θ, can be computed as the bias squared plus the variance of the estimator. MSE is a convenient measure of how closely $\hat{\theta}$ approximates θ because it combines both bias and variance in a logical way.

Estimation of the coefficient of skewness γ_X provides a good example of the use of the MSE for evaluating the total deviation of an estimate from the true population value. The sample estimate $\hat{\gamma}_X$ of γ_X is often biased, has a large variance, and its absolute value was shown by Kirby (1974) to be bounded by the square root of the sample size *n*:

$$|\hat{\gamma}_X| \leq \sqrt{n} \qquad (7.45)$$

The bounds do not depend on the true skew, γ_X. However, the bias and variance of $\hat{\gamma}_X$ do depend on the sample size and the actual distribution of *X*. Table 7.3 contains the expected value and standard deviation of the estimated coefficient of skewness $\hat{\gamma}_X$ when *X* has either a normal distribution, for which $\gamma_X = 0$, or a gamma distribution with $\gamma_X = 0.25, 0.50, 1.00, 2.00$, or 3.00. These values are adapted from Wallis et al. (1974 a,b) who employed moment estimators slightly different than those in Equation 7.39.

For the normal distribution, $\text{E}[\hat{\gamma}] = 0$ and $\text{Var}[\hat{\gamma}_X] \cong 5/n$. In this case, the skewness estimator is unbiased but highly variable. In all the other cases in Table 7.3, the skewness estimator is biased.

To illustrate the magnitude of these errors, consider the mean square error of the skew estimator $\hat{\gamma}_X$ calculated from a sample of size 50 when *X* has a gamma distribution with $\gamma_X = 0.50$, a reasonable value for annual streamflows. The expected value of $\hat{\gamma}_X$ is 0.45; its variance equals $(0.37)^2$, its standard deviation is squared. Using Equation 7.44, the mean square error of $\hat{\gamma}_X$ is:

$$\begin{aligned} \text{MSE}(\hat{\gamma}_X) &= (0.45 - 0.50)^2 + (0.37)^2 \\ &= 0.0025 + 0.1369 = 0.139 \cong 0.14 \end{aligned} \qquad (7.46)$$

An unbiased estimate of γ_X is simply $(0.50/0.45)\,\hat{\gamma}_X$. Here the estimator provided by Equation 7.39e has been scaled to eliminate bias. This unbiased estimator has a mean squared error of:

$$\begin{aligned} \text{MSE}\left(\frac{0.50\hat{\gamma}_X}{0.48}\right) &= (0.50 - 0.50)^2 + \left[\left(\frac{0.50}{0.45}\right)(0.37)\right]^2 \\ &= 0.169 \cong 0.17 \end{aligned} \qquad (7.47)$$

The mean square error of the unbiased estimator of $\hat{\gamma}_X$ is larger than the mean square error of the biased estimate. Unbiasing $\hat{\gamma}_X$ results in a larger mean square error for all the cases listed in Table 7.3 except for the normal distribution for which $\gamma_X = 0$, and the gamma distribution with $\gamma_X = 3.00$.

As shown here for the skew coefficient, biased estimators often have smaller mean square errors than unbiased estimators. Because the mean square error measures the total average deviation of an estimator from the quantity being estimated, this result demonstrates that the strict or unquestioning use of unbiased estimators is not advisable. Additional information on the sampling distribution of quantiles and moments is contained in Stedinger et al. (1993).

2.4. L-Moments and Their Estimators

L-moments are another way to summarize the statistical properties of hydrological data based on linear combinations of the original observations (Hosking, 1990). Recently, hydrologists have found that regionalization methods (to be discussed in Section 5) using L-moments are superior to methods using traditional moments (Hosking and Wallis, 1997; Stedinger and Lu, 1995). L-moments have also proved useful for construction of goodness-of-fit tests (Hosking et al., 1985; Chowdhury et al., 1991; Fill and Stedinger, 1995), measures of regional homogeneity and distribution selection methods (Vogel and Fennessey, 1993; Hosking and Wallis, 1997).

E021101c

expected value of $\hat{\gamma}_X$

distribution of X		sample size		
	10	20	50	80
normal $\gamma_X = 0$	0.00	0.00	0.00	0.00
gamma $\gamma_X = 0.25$	0.15	0.19	0.23	0.23
$\gamma_X = 0.50$	0.31	0.39	0.45	0.47
$\gamma_X = 1.00$	0.60	0.76	0.88	0.93
$\gamma_X = 2.00$	1.15	1.43	1.68	1.77
$\gamma_X = 3.00$	1.59	1.97	2.32	2.54
upper bound on skew	3.16	4.47	7.07	8.94

standard deviation of $\hat{\gamma}_X$

distribution of X		sample size		
	10	20	50	80
normal $\gamma_X = 0$	0.69	0.51	0.34	0.26
gamma $\gamma_X = 0.25$	0.69	0.52	0.35	0.28
$\gamma_X = 0.50$	0.69	0.53	0.37	0.31
$\gamma_X = 1.00$	0.70	0.57	0.44	0.38
$\gamma_X = 2.00$	0.72	0.68	0.62	0.57
$\gamma_X = 3.00$	0.74	0.76	0.77	0.77

Table 7.3. Sampling properties of coefficient of skewness estimator.

Source: Wallis et al. (1974b) who only divided by *n* in the estimators of the moments, whereas in Equations 7.39b and 7.39c, we use the generally-adopted coefficients of 1/(*n* – 1) and *n*/(*n* – 1)(*n* – 2) for the variance and skew.

The first L-moment designated as λ_1 is simply the arithmetic mean:

$$\lambda_1 = \mathrm{E}[X] \tag{7.48}$$

Now let $X_{(i|n)}$ be the i^{th} largest observation in a sample of size n ($i = n$ corresponds to the largest). Then, for any distribution, the second L-moment, λ_2, is a description of scale based upon the expected difference between two randomly selected observations:

$$\lambda_2 = (1/2)\ \mathrm{E}[X_{(2|1)} - X_{(1|2)}] \tag{7.49}$$

Similarly, L-moment measures of skewness and kurtosis use three and four randomly selected observations, respectively.

$$\lambda_3 = (1/3)\ \mathrm{E}[X_{(3|3)} - 2X_{(2|3)} + X_{(1|3)}] \tag{7.50}$$

$$\lambda_4 = (1/4)\ \mathrm{E}[X_{(4|4)} - 3X_{(3|4)} + 3X_{(2|4)} - X_{(1|4)}] \tag{7.51}$$

Sample L-moment estimates are often computed using intermediate statistics called *probability weighted moments* (PWMs). The r^{th} probability weighted moment is defined as:

$$\beta_r = \mathrm{E}\{X[F(X)]^r\} \tag{7.52}$$

where $F(X)$ is the cumulative distribution function of X. Recommended (Landwehr et al., 1979; Hosking and Wallis, 1995) unbiased PWM estimators, b_r, of β_r are computed as:

$$\begin{aligned} b_0 &= \bar{X} \\ b_1 &= \frac{1}{n(n-1)}\sum_{j=2}^{n}(j-1)X_{(j)} \\ b_2 &= \frac{1}{n(n-1)(n-2)}\sum_{j=3}^{n}(j-1)(j-2)X_{(j)} \end{aligned} \tag{7.53}$$

These are examples of the general formula for computing estimators b_r of $\tilde{\beta}_r$.

$$b_r = \frac{1}{n}\sum_{j=r+1}^{n}\binom{j-1}{r}X_{(j)}\bigg/\binom{n-1}{r}$$

$$= \frac{1}{r+1}\sum_{j=r+1}^{n}\binom{j-1}{r}X_{(j)}\bigg/\binom{n}{r+1} \qquad (7.54)$$

for $r = 1, \ldots, n-1$.

L-moments are easily calculated in terms of PWMs using:

$$\lambda_1 = \beta_0$$

$$\lambda_2 = 2\beta_1 - \beta_0$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \qquad (7.55)$$

Wang (1997) provides formulas for directly calculating L-moment estimators of λ_r. Measures of the coefficient of variation, skewness and kurtosis of a distribution can be computed with L-moments, as they can with traditional product moments. Where skew primarily measures the asymmetry of a distribution, the kurtosis is an additional measure of the thickness of the extreme tails. Kurtosis is particularly useful for comparing symmetric distributions that have a skewness coefficient of zero. Table 7.4 provides definitions of the traditional coefficient of variation, coefficient of skewness and coefficient of kurtosis, as well as the L-moment, L-coefficient of variation, L-coefficient of skewness and L-coefficient of kurtosis.

The flood data in Table 7.2 can be used to provide an example of L-moments. Equation 7.53 yields estimates of the first three Probability Weighted Moments:

$$b_0 = 1{,}549.20$$

$$b_1 = 1003.89$$

$$b_2 = 759.02 \qquad (7.56)$$

Recall that b_0 is just the sample average $\bar{x}$. The sample L-moments are easily calculated using the probability weighted moments. One obtains:

$$\hat{\lambda}_1 = b_0 = 1{,}549$$

$$\hat{\lambda}_2 = 2b_1 - b_0 = 458$$

$$\hat{\lambda}_3 = 6b_2 - 6b_1 + b_0 = 80 \qquad (7.55)$$

Thus, the sample estimates of the L-coefficient of variation, t_2, and L-coefficient of skewness, t_3, are:

$$t_2 = 0.295$$

$$t_3 = 0.174 \qquad (7.58)$$

Table 7.4. Definitions of dimensionless product-moment and L-moment ratios.

name	common symbol	definition
product-moment ratios		
coefficient of variation	CV_X	σ_X / μ_X
skewness	γ_X	$E[(X - \mu_X)^3] / \sigma_X{}^3$
kurtosis	κ_X	$E[(X - \mu_X)^4] / \sigma_X{}^4$
L-moment ratios		
L-coefficient of variation *	L-CV, τ_2	λ_2 / λ_1
skewness	L-skewness, τ_3	λ_3 / λ_2
kurtosis	L-kurtosis, τ_4	λ_4 / λ_2

* Hosking and Wallis (1997) use τ instead of τ_2 to represent the L-CV ratio

E021101d

3. Distributions of Random Events

A frequent task in water resources planning is the development of a model of some probabilistic or stochastic phenomena such as streamflows, flood flows, rainfall, temperatures, evaporation, sediment or nutrient loads, nitrate or organic compound concentrations, or water demands. This often requires one to fit a probability distribution function to a set of observed values of the random variable. Sometimes, one's immediate objective is to estimate a particular quantile of the distribution, such as the 100-year flood, 50-year six-hour-rainfall depth, or the minimum seven-day-average expected once-in-ten-year flow. Then the fitted distribution can supply an estimate of that quantity. In a stochastic simulation, fitted distributions are used to generate possible values of the random variable in question.

Rather than fitting a reasonable and smooth mathematical distribution, one could use the empirical distribution represented by the data to describe the possible values that a random variable may assume in the future and their frequency. In practice, the true mathematical form for the distribution that describes the events is not known. Moreover, even if it was, its functional form may have too many parameters to be of much practical use. Thus, using the empirical distribution represented by the data itself has substantial appeal.

Generally, the free parameters of the theoretical distribution are selected (estimated) so as to make the fitted distribution consistent with the available data. The goal is to select a physically reasonable and simple distribution to describe the frequency of the events of interest, to estimate that distribution's parameters, and ultimately to obtain quantiles, performance indices and risk estimates of satisfactory accuracy for the problem at hand. Use of a theoretical distribution has several advantages over use of the empirical distribution:

- It presents a smooth interpretation of the empirical distribution. As a result quantiles, performance indices and other statistics computed using the fitted distribution should be more accurate than those computed with the empirical distribution.
- It provides a compact and easy-to-use representation of the data.
- It is likely to provide a more realistic description of the range of values that the random variable may assume and their likelihood. For example, by using the empirical distribution, one implicitly assumes that no values larger or smaller than the sample maximum or minimum can occur. For many situations, this is unreasonable.
- Often one needs to estimate the likelihood of extreme events that lie outside the range of the sample (either in terms of x values or in terms of frequency). Such extrapolation makes little sense with the empirical distribution.
- In many cases, one is not interested in the values of a random variable X, but instead in derived values of variables Y that are functions of X. This could be a performance function for some system. If Y is the performance function, interest might be primarily in its mean value E[Y], or the probability some standard is exceeded, Pr{$Y >$ standard}. For some theoretical X-distributions, the resulting Y-distribution may be available in closed form, thus making the analysis rather simple. (The normal distribution works with linear models, the lognormal distribution with product models, and the gamma distribution with queuing systems.)

This section provides a brief introduction to some useful techniques for estimating the parameters of probability distribution functions and for determining if a fitted distribution provides a reasonable or acceptable model of the data. Sub-sections are also included on families of distributions based on the normal, gamma and generalized-extreme-value distributions. These three families have found frequent use in water resources planning (Kottegoda and Rosso, 1997).

3.1. Parameter Estimation

Given a set of observations to which a distribution is to be fit, one first selects a distribution function to serve as a model of the distribution of the data. The choice of a distribution may be based on experience with data of that type, some understanding of the mechanisms giving rise to the data, and/or examination of the observations themselves. One can then estimate the parameters of the chosen distribution and determine if the fitted distribution provides an acceptable model of the data. A model is generally judged to be unacceptable if it is unlikely that one could have observed the available data were they actually drawn from the fitted distribution.

In many cases, good estimates of a distribution's parameters are obtained by the *maximum-likelihood-estimation* procedure. Give a set of n *independent* observations $\{x_1, \dots , x_n\}$ of a continuous random variable X, the joint probability density function for the observations is:

$$f_{X_1,X_2,X_3,\dots,X_n}(x_1,\dots,x_n \mid \boldsymbol{\theta}) = f_X(x_1 \mid \boldsymbol{\theta}) \cdot f_X(x_2 \mid \boldsymbol{\theta}) \cdots f_X(x_n \mid \boldsymbol{\theta}) \tag{7.59}$$

where $\boldsymbol{\theta}$ is the vector of the distribution's parameters.

The maximum likelihood estimator of $\boldsymbol{\theta}$ is that vector $\boldsymbol{\theta}$ which maximizes Equation 7.59 and thereby makes it as likely as possible to have observed the values $\{x_1, \dots , x_n\}$.

Considerable work has gone into studying the properties of maximum likelihood parameter estimates. Under rather general conditions, asymptotically the estimated parameters are normally distributed, unbiased and have the smallest possible variance of any asymptotically unbiased estimator (Bickel and Doksum, 1977). These, of course, are asymptotic properties, valid for large sample sizes n. Better estimation procedures, perhaps yielding biased parameter estimates, may exist for small sample sizes. Stedinger (1980) provides such an example. Still, maximum likelihood procedures are recommended with moderate and large samples, even though the iterative solution of nonlinear equations is often required.

An example of the maximum likelihood procedure for which closed-form expressions for the parameter estimates are obtained is provided by the lognormal distribution. The probability density function of a lognormally distributed random variable X is:

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2\right\} \tag{7.60}$$

Here, the parameters μ and σ^2 are the mean and variance of the logarithm of X, and not of X itself.

Maximizing the logarithm of the joint density for $\{x_1, \dots , x_n\}$ is more convenient than maximizing the joint probability density itself. Hence, the problem can be expressed as the maximization of the *log-likelihood function*

$$\begin{aligned} L &= \ln \prod_{i=1}^{n} f[(x_i \mid \mu, \sigma)] \\ &= \sum_{i=1}^{n} \ln f(x_i \mid \mu, \sigma) \\ &= -\sum_{i=1}^{n} \ln(x_i\sqrt{2\pi}) - n\ln(\sigma) \\ &\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[\ln(x_i)-\mu]^2 \end{aligned} \tag{7.61}$$

The maximum can be obtained by equating to zero the partial derivatives $\partial L/\partial\mu$ and $\partial L/\partial\sigma$ whereby one obtains:

$$\begin{aligned} 0 &= \frac{\partial L}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}[\ln(x_i)-\mu] \\ 0 &= \frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}[\ln(x_i)-\mu]^2 \end{aligned} \tag{7.62}$$

These equations yield the estimators

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}[\ln(x_i)-\hat{\mu}]^2 \end{aligned} \tag{7.63}$$

The second-order conditions for a maximum are met and these values maximize Equation 7.59. It is useful to note that if one defines a new random variable $Y = \ln(X)$, then the maximum likelihood estimators of the parameters μ and σ^2, which are the mean and variance of the Y distribution, are the sample estimators of the mean and variance of Y:

$$\begin{aligned} \hat{\mu} &= \bar{y} \\ \hat{\sigma}^2 &= [(n-1)/n]S_Y^2 \end{aligned} \tag{7.64}$$

The correction $[(n-1)/n]$ in this last equation is often neglected.

The second commonly used parameter estimation procedure is the *method of moments*. The method of moments is often a quick and simple method for obtaining parameter estimates for many distributions. For a distribution with $m = 1$, 2 or 3 parameters, the first m moments of the postulated distribution in Equations 7.34, 7.35 and 7.37 are equated to the estimates of those moments calculated using Equations 7.39. The resulting nonlinear equations are solved for the unknown parameters.

For the lognormal distribution, the mean and variance of X as a function of the parameters μ and σ are given by

$$\mu_X = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$$
$$\sigma_X^2 = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1] \qquad (7.65)$$

Substituting $\bar{x}$ for μ_X and s_X^2 for σ_X^2 and solving for μ and σ^2 one obtains

$$\hat{\sigma}^2 = \ln(1 + s_X^2/\bar{x}^2)$$
$$\hat{\mu} = \ln\left(\frac{\bar{x}}{\sqrt{1 + s_X^2/\bar{x}^2}}\right) = \ln \bar{x} - \frac{1}{2}\hat{\sigma}^2 \qquad (7.66)$$

The data in Table 7.2 provide an illustration of both fitting methods. One can easily compute the sample mean and variance of the logarithms of the flows to obtain

$$\hat{\mu} = 7.202$$
$$\hat{\sigma}^2 = 0.3164 = (0.5625)^2 \qquad (7.67)$$

Alternatively, the sample mean and variance of the flows themselves are

$$\bar{x} = 1549.2$$
$$s_X^2 = 661{,}800 = (813.5)^2 \qquad (7.68)$$

Substituting those two values in Equation 7.66 yields

$$\hat{\mu} = 7.224$$
$$\sigma_X^2 = 0.2435 = (0.4935)^2 \qquad (7.69)$$

Method of moments and maximum likelihood are just two of many possible estimation methods. Just as method of moments equates sample estimators of moments to population values and solves for a distribution's parameters, one can simply equate L-moment estimators to population values and solve for the parameters of a distribution. The resulting method of L-moments has received considerable attention in the hydrological literature (Landwehr et al., 1978; Hosking et al., 1985; Hosking and Wallis, 1987; Hosking, 1990; Wang, 1997). It has been shown to have significant advantages when used as a basis for regionalization procedures that will be discussed in Section 5 (Lettenmaier et al., 1987; Stedinger and Lu, 1995; Hosking and Wallis, 1997).

Bayesian procedures provide another approach that is related to maximum likelihood estimation. Bayesian inference employs the likelihood function to represent the information in the data. That information is augmented with a prior distribution that describes what is known about constraints on the parameters and their likely values beyond the information provided by the recorded data available at a site. The likelihood function and the prior probability density function are combined to obtain the probability density function that describes the *posterior distribution* of the parameters:

$$f_\theta(\boldsymbol{\theta} \mid x_1, x_2, \ldots, x_n) \propto f_X(x_1, x_2, \ldots, x_n \mid \boldsymbol{\theta})\xi(\boldsymbol{\theta}) \qquad (7.70)$$

The symbol $\propto$ means 'proportional to' and $\xi(\boldsymbol{\theta})$ is the probability density function for the prior distribution for $\boldsymbol{\theta}$ (Kottegoda and Rosso, 1997). Thus, except for a constant of proportionality, the probability density function describing the posterior distribution of the parameter vector $\boldsymbol{\theta}$ is equal to the product of the likelihood function $f_X(x_1, x_2, \ldots, x_n \mid \boldsymbol{\theta})$ and the probability density function for the prior distribution $\xi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$.

Advantages of the Bayesian approach are that it allows the explicit modelling of uncertainty in parameters (Stedinger, 1997; Kuczera, 1999) and provides a theoretically consistent framework for integrating systematic flow records with regional and other hydrological information (Vicens et al., 1975; Stedinger, 1983; Kuczera, 1983). Martins and Stedinger (2000) illustrate how a prior distribution can be used to enforce realistic constraints upon a parameter as well as providing a description of its likely values. In their case, use of a prior of the shape parameter κ of a generalized extreme value (GEV) distribution (discussed in Section 3.6) allowed definition of generalized maximum likelihood estimators that, over the κ-range of interest, performed substantially better than maximum likelihood, moment, and L-moment estimators.

While Bayesian methods have been available for decades, the computational challenge posed by the solution of Equation 7.70 has been an obstacle to their use. Solutions to Equation 7.70 have been available for special cases such as normal data, and binomial and Poisson samples (Raiffa and Schlaifer, 1961; Benjamin and Cornell, 1970; Zellner, 1971). However, a new and very general set of Markov Chain Monte Carlo (MCMC) procedures (discussed in Section 7.2) allows numerical computation of the posterior distributions of parameters for a very broad class of models (Gilks et al., 1996). As a result, Bayesian methods are now becoming much more popular and are the standard approach for many difficult problems that are not easily addressed by traditional methods (Gelman et al., 1995; Carlin and Louis, 2000). The use of Monte Carlo Bayesian methods in flood frequency analysis, rainfall–runoff modelling, and evaluation of environmental pathogen concentrations are illustrated by Wang (2001), Bates and Campbell (2001) and Crainiceanu et al. (2002), respectively.

Finally, a simple method of fitting flood frequency curves is to plot the ordered flood values on special probability paper and then to draw a line through the data (Gumbel, 1958). Even today, that simple method is still attractive when some of the smallest values are zero or unusually small, or have been censored as will be discussed in Section 4 (Kroll and Stedinger, 1996). Plotting the ranked annual maximum series against a probability scale is always an excellent and recommended way to see what the data look like and for determining whether or not a fitted curve is consistent with the data (Stedinger et al., 1993).

Statisticians and hydrologists have investigated which of these methods most accurately estimates the parameters themselves or the quantiles of the distribution. One also needs to determine how accuracy should be measured. Some studies have used average squared deviations, some have used average absolute weighted deviations with different weights on under and over-estimation, and some have used the squared deviations of the log-quantile estimator (Slack et al., 1975; Kroll and Stedinger, 1996). In almost all cases, one is also interested in the bias of an estimator, which is the average value of the estimator minus the true value of the parameter or quantile being estimated. Special estimators have been developed to compute design events that on average are exceeded with the specified probability and have the anticipated risk of being exceeded (Beard, 1960, 1997; Rasmussen and Rosbjerg, 1989, 1991a,b; Stedinger, 1997; Rosbjerg and Madsen, 1998).

3.2. Model Adequacy

After estimating the parameters of a distribution, some check of model adequacy should be made. Such checks vary from simple comparisons of the observations with the fitted model (using graphs or tables) to rigorous statistical tests. Some of the early and simplest methods of parameter estimation were graphical techniques. Although quantitative techniques are generally more accurate and precise for parameter estimation, graphical presentations are invaluable for comparing the fitted distribution with the observations for the detection of systematic or unexplained deviations between the two. The observed data will plot as a straight line on probability graph paper if the postulated distribution is the true distribution of the observation. If probability graph paper does not exist for the particular distribution of interest, more general techniques can be used.

Let $x_{(i)}$ be the ith largest value in a set of observed values $\{x_i\}$ so that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The random variable $X_{(i)}$ provides a reasonable estimate of the pth quantile x_p of the true distribution of X for $p = i/(n+1)$. In fact, when one considers the cumulative probability U_i associated with the random variable $X_{(i)}$, $U_i = F_X(X_{(i)})$, and if the observations $X_{(i)}$ are independent, then the U_i have a beta distribution (Gumbel, 1958) with probability density function:

$$f_{U_i}(u) = \frac{n!}{(i-1)!\,(n-1)!}\, u^{i-1}(1-u)^{n-i} \qquad 0 \leq u \leq 1 \tag{7.71}$$

This beta distribution has mean

$$\mathrm{E}[U_i] = \frac{i}{n+1} \tag{7.72a}$$

and variance

$$\mathrm{Var}(U_i) = \frac{i(n-i+1)}{(n+1)^2\,(n+2)} \tag{7.72b}$$

A good graphical check of the adequacy of a fitted distribution $G(x)$ is obtained by plotting the observations $x_{(i)}$ versus $G^{-1}[i/(n+1)]$ (Wilk and Gnanadesikan, 1968). Even if $G(x)$ equalled to an exact degree the true X-distribution $F_X[x]$, the plotted points would not fall exactly on a 45° line through the origin of the graph. This would only occur if $F_X[x_{(i)}]$ exactly equalled $i/(n+1)$, and therefore each $x_{(i)}$ exactly equalled $F_X^{-1}[i/(n+1)]$.

An appreciation for how far an individual observation $x_{(i)}$ can be expected to deviate from $G^{-1}[i/(n+1)]$ can be obtained by plotting $G^{-1}[u_i^{(0.75)}]$ and $G^{-1}[u_i^{(0.25)}]$, where $u_i^{(0.75)}$ and $u_i^{(0.25)}$ are the upper and lower quartiles of the distribution of U_i obtained from integrating the probability density function in Equation 7.71. The required incomplete beta function is also available in many software packages, including Microsoft Excel. Stedinger et al. (1993) show that $u_{(1)}$ and $(1 - u_{(n)})$ fall between $0.052/n$ and $3(n + 1)$ with a probability of 90%, thus illustrating the great uncertainty associated with the cumulative probability of the smallest value and the exceedance probability of the largest value in a sample.

Figures 7.2a and 7.2b illustrate the use of this *quantile–quantile plotting technique* by displaying the results of fitting a normal and a lognormal distribution to the annual maximum flows in Table 7.2 for the Magra River, Italy, at Calamazza for the years 1930 – 70. The observations of $X_{(i)}$, given in Table 7.2, are plotted on the vertical axis against the quantiles $G^{-1}[i/(n + 1)]$ on the horizontal axis.

A probability plot is essentially a scatter plot of the sorted observations $X_{(i)}$ versus some approximation of their expected or anticipated value, represented by $G^{-1}(p_i)$, where, as suggested, $p_i = i/(n + 1)$. The p_i values are called *plotting positions*. A common alternative to $i/(n + 1)$ is $(i - 0.5)/n$, which results from a probabilistic interpretation of the empirical distribution of the data. Many reasonable plotting position formulas have been proposed based upon the sense in which $G^{-1}(p_i)$ should approximate $X_{(i)}$. The *Weibull formula* $i/(n + 1)$ and the *Hazen formula* $(i - 0.5)/n$ bracket most of the reasonable choices. Popular formulas are summarized by Stedinger et al. (1993), who also discuss the generation of probability plots for many distributions commonly employed in hydrology.

Rigorous statistical tests are available for trying to determine whether or not it is reasonable to assume that a given set of observations could have been drawn from a particular family of distributions. Although not the most powerful of such tests, the *Kolmogorov–Smirnov test* provides bounds within which every observation should lie if the sample is actually drawn from the assumed distribution. In particular, for $G = F_X$, the test specifies that

$$\Pr\left[G^{-1}\left(\frac{i}{n} - C_\alpha\right) \le X_{(i)} \le G^{-1}\left(\frac{i-1}{n} + C_\alpha\right)\forall i\right] = 1 - \alpha \tag{7.73}$$

where C_α is the critical value of the test at significance level α. Formulas for C_α as a function of n are contained in Table 7.5 for three cases: (1) when G is completely

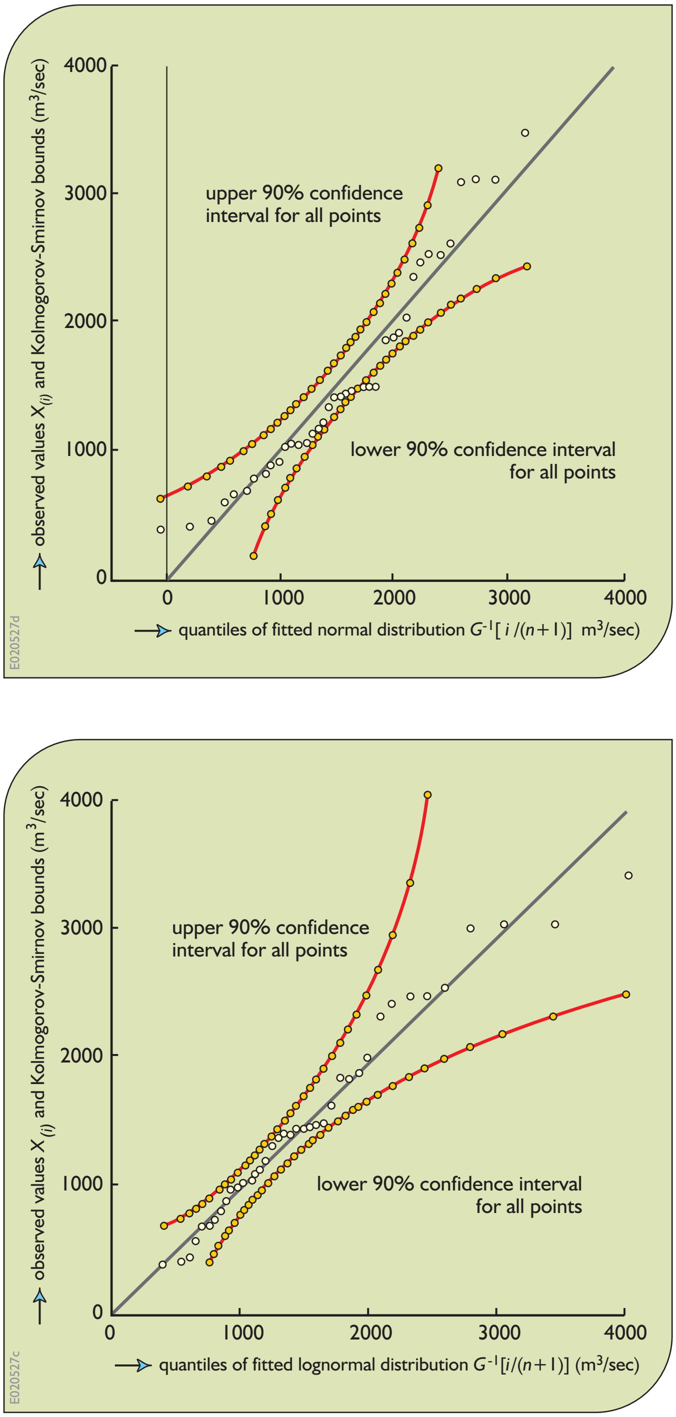

Figure 7.2. Plots of annual maximum discharges of Magra River, Italy, versus quantiles of fitted (a) normal and (b) lognormal distributions.

specified independent of the sample's values; (2) when G is the normal distribution and the mean and variance are estimated from the sample with $\bar{x}$ and s_X^2; and (3) when G is the exponential distribution and the scale parameter is estimated as $1/(\bar{x})$. Chowdhury et al. (1991) provide critical values for the Gumbel and generalized extreme value (GEV) distributions (Section 3.6) with known shape parameter κ. For other distributions, the values obtained from Table 7.5 may be used to construct approximate simultaneous confidence intervals for every $X_{(i)}$.

Figures 7.2a and b contain 90% confidence intervals for the plotted points constructed in this manner. For the normal distribution, the critical value of C_α equals $0.819/(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$, where 0.819 corresponds to $\alpha = 0.10$. For $n = 40$, one computes $C_\alpha = 0.127$. As can be seen in Figure 7.2a, the annual maximum flows are not consistent with the hypothesis that they were drawn from a normal distribution; three of the observations lie outside the simultaneous 90% confidence intervals for all the points. This demonstrates a statistically significant lack of fit. The fitted normal distribution underestimates the quantiles corresponding to small and large probabilities while overestimating the quantiles in an intermediate range. In Figure 7.2b, deviations between the fitted lognormal distribution and the observations can be attributed to the differences between $F_X(x_{(i)})$ and $i/(n + 1)$. Generally, the points are all near the 45° line through the origin, and no major systematic deviations are apparent.

The *Kolmogorov–Smirnov test* conveniently provides bounds within which every observation on a probability plot should lie if the sample is actually drawn from the assumed distribution, and thus is useful for visually evaluating the adequacy of a fitted distribution. However, it is not the most powerful test available for estimating which distribution a set of observations is likely to have been drawn from. For that purpose, several other more analytical tests are available (Filliben, 1975; Hosking, 1990; Chowdhury et al., 1991; Kottegoda and Rosso, 1997).

The *Probability Plot Correlation test* is a popular and powerful test of whether a sample has been drawn from a postulated distribution, though it is often weaker than alternative tests at rejecting thin-tailed alternatives (Filliben, 1975; Fill and Stedinger, 1995). A test with greater power has a greater probability of correctly determining that a sample is not from the postulated distribution. The Probability Plot Correlation Coefficient test employs the correlation r between the ordered observations $x_{(i)}$ and the corresponding fitted quantiles $w_i = G^{-1}(p_i)$, determined by plotting positions p_i for each $x_{(i)}$. Values of r near 1.0 suggest that the observations could have been drawn from the fitted distribution: r measures the linearity of the probability plot providing a quantitative assessment of fit. If $\bar{x}$ denotes the average value of the observations and $\bar{w}$ denotes the average value of the fitted quantiles, then

$$r = \frac{\sum (x_{(i)} - \bar{x})(w_i - \bar{w})}{\left[\left(\sum (x_{(i)} - \bar{x})^2 \sum (w_i - \bar{w})^2\right)\right]^{0.5}} \tag{7.74}$$

Table 7.5. Critical values of Kolmogorov–Smirnov statistic as a function of sample size *n* (after Stephens, 1974).

E021101e

	significance level α				
	0.150	0.100	0.050	0.025	0.010
F_X completely specified:					
$C_\alpha(\sqrt{n} + 0.12 + 0.11/\sqrt{n})$	1.138	1.224	1.358	1.480	1.628
F_X normal with mean and variance estimated as $\bar{x}$ and s_X^2					
$C_\alpha(\sqrt{n} + 0.01 + 0.85/\sqrt{n})$	0.775	0.819	0.895	0.995	1.035
F_X exponential with scale parameter b estimated as $1/(\bar{x})$					
$(C_\alpha + 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n})$	0.926	0.990	1.094	1.190	1.308

values of C_α are calculated as follows:
for case 2 with $\alpha = 0.10$, $C_\alpha = 0.819/(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$

Table 7.6 provides critical values for r for the normal distribution, or the logarithms of lognormal variates, based upon the Blom plotting position that has $p_i = (i - 3/8)/(n + 1/4)$. Values for the Gumbel distribution are reproduced in Table 7.7 for use with the Gringorten plotting position $p_i = (i - 0.44)/(n + 0.12)$. The table also applies to logarithms of Weibull variates (Stedinger et al., 1993). Other tables are available for the GEV (Chowdhury et al., 1991), the Pearson type 3 (Vogel and McMartin, 1991), and exponential and other distributions (D'Agostino and Stephens, 1986).

Recently developed L-moment ratios appear to provide goodness-of-fit tests that are superior to both the Kolmogorov–Smirnov and the Probability Plot Correlation test (Hosking, 1990; Chowdhury et al., 1991; Fill and Stedinger, 1995). For normal data, the L-skewness estimator $\hat{\tau}_3$ (or t_3) would have mean zero and Var $\hat{\tau}_3 = (0.1866 + 0.8/n)/n$, allowing construction of a powerful test of normality against skewed alternatives using the normally distributed statistic

$$Z = t_3/\sqrt{(0.1866 + 0.8/n)/n} \qquad (7.75)$$

with a reject region $|Z| > z_{\alpha/2}$.

Chowdhury et al. (1991) derive the sampling variance of the L-CV and L-skewness estimators $\hat{\tau}_2$ and $\hat{\tau}_3$ as a function of κ for the GEV distribution. These allow construction of a test of whether a particular data set is consistent with a GEV distribution with a regionally estimated value of κ, or a regional κ and a regional coefficient of variation, CV. Fill and Stedinger (1995) show that the $\hat{\tau}_3$ L-skewness estimator provides a test for the Gumbel versus a general GEV distribution using the normally distributed statistic

$$Z = (\hat{\tau}_3 - 0.17)/\sqrt{(0.2326 + 0.70/n)/n} \qquad (7.76)$$

with a reject region $|Z| > z_{\alpha/2}$.

The literature is full of goodness-of-fit tests. Experience indicates that among the better tests there is often not a great deal of difference (D'Agostino and Stephens, 1986). Generation of a probability plot is most often a good idea because it allows the modeller to see what the data look like and where problems occur. The Kolmogorov–Smirnov test helps the eye

E021101f

n	significance level		
	0.10	0.05	0.01
10	0.9347	0.9180	0.8804
15	0.9506	0.9383	0.9110
20	0.9600	0.9503	0.9290
30	0.9707	0.9639	0.9490
40	0.9767	0.9715	0.9597
50	0.9807	0.9764	0.9664
60	0.9835	0.9799	0.9710
75	0.9865	0.9835	0.9757
100	0.9893	0.9870	0.9812
300	0.99602	0.99525	0.99354
1,000	0.99854	0.99824	0.99755

Table 7.6. Lower critical values of the probability plot correlation test statistic for the normal distribution using $p_i = (i - 3/8)/(n + 1/4)$ (Vogel, 1987).

E021101g

n	significance level		
	0.10	0.05	0.01
10	0.9260	0.9084	0.8630
20	0.9517	0.9390	0.9060
30	0.9622	0.9526	0.9191
40	0.9689	0.9594	0.9286
50	0.9729	0.9646	0.9389
60	0.9760	0.9685	0.9467
70	0.9787	0.9720	0.9506
80	0.9804	0.9747	0.9525
100	0.9831	0.9779	0.9596
300	0.9925	0.9902	0.9819
1,000	0.99708	0.99622	0.99334

Table 7.7. Lower critical values of the probability plot correlation test statistic for the Gumbel distribution using $p_i = (i - 0.44)/(n + 0.12)$ (Vogel, 1987).

interpret a probability plot by adding bounds to a graph, illustrating the magnitude of deviations from a straight line that are consistent with expected variability. One can also use quantiles of a beta distribution to illustrate the possible error in individual plotting positions, particularly at the extremes where that uncertainty is largest. The probability plot correlation test is a popular and powerful goodness-of-fit statistic. Goodness-of-fit tests based upon sample estimators of the L-skewness $\hat{\tau}_3$ for the normal and Gumbel distribution provide simple and useful tests that are not based on a probability plot.

3.3. Normal and Lognormal Distributions

The normal distribution and its logarithmic transformation, the lognormal distribution, are arguably the most widely used distributions in science and engineering. The probability density function of a normal random variable is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \quad \text{for } -\infty < X < +\infty \tag{7.77}$$

where μ and σ^2 are equivalent to μ_X and σ_X^2, the mean and variance of X. Interestingly, the maximum likelihood estimators of μ and σ^2 are almost identical to the moment estimates $\bar{x}$ and s_X^2.

The normal distribution is symmetric about its mean μ_X and admits values from $-\infty$ to $+\infty$. Thus, it is not always satisfactory for modelling physical phenomena such as streamflows or pollutant concentrations, which are necessarily non-negative and have skewed distributions. A frequently used model for skewed distributions is the lognormal distribution. A random variable X has a lognormal distribution if the natural logarithm of X, ln(X), has a normal distribution. If X is lognormally distributed, then by definition ln(X) is normally distributed, so that the density function of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2\right\} \frac{d(\ln x)}{dx} = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[\ln(x/\eta)]^2\right\} \tag{7.78}$$

for $x > 0$ and $\mu = \ln(\eta)$. Here η is the median of the X-distribution.

A lognormal random variable takes on values in the range $[0, +\infty]$. The parameter μ determines the scale of the X-distribution whereas σ^2 determines the shape of the distribution. The mean and variance of the lognormal distribution are given in Equation 7.65. Figure 7.3 illustrates the various shapes that the lognormal probability density function can assume. It is highly skewed with a thick right hand tail for $\sigma > 1$, and approaches a symmetric normal distribution as $\sigma \rightarrow 0$. The density function always has a value of zero at $x = 0$. The coefficient of variation and skew are:

$$CV_X = [\exp(\sigma^2) - 1]^{1/2} \tag{7.79}$$

$$\gamma_X = 3CV_X + CV_X^3 \tag{7.80}$$

The maximum likelihood estimates of μ and σ^2 are given in Equation 7.63 and the moment estimates in Equation 7.66. For reasonable-sized samples, the maximum likelihood estimates generally perform as well or better than the moment estimates (Stedinger, 1980).

The data in Table 7.2 were used to calculate the parameters of the lognormal distribution that would describe these flood flows. The results are reported in Equation 7.67. The two-parameter maximum likelihood and method of moments estimators identify parameter estimates for which the distribution skewness coefficients

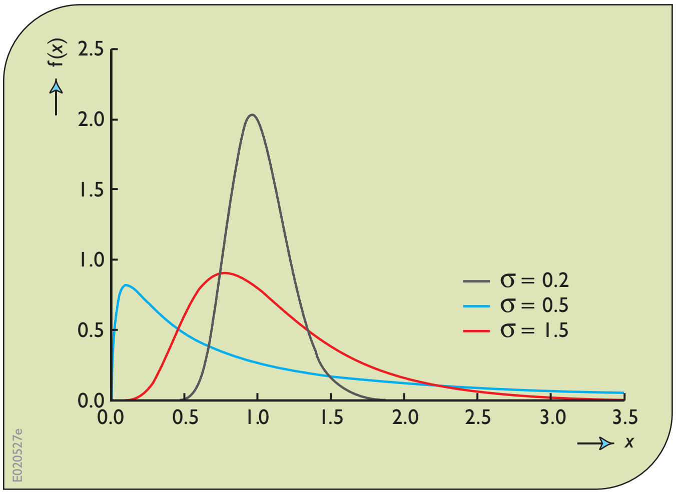

Figure 7.3. Lognormal probability density functions with various standard deviations σ.

are 2.06 and 1.72, which is substantially greater than the sample skew of 0.712.

A useful generalization of the two-parameter lognormal distribution is the shifted lognormal or three-parameter lognormal distribution obtained when $\ln(X - \tau)$ is described by a normal distribution, and $X \geq \tau$. Theoretically, τ should be positive if, for physical reasons, X must be positive; practically, negative values of τ can be allowed when the resulting probability of negative values of X is sufficiently small.

Unfortunately, maximum likelihood estimates of the parameters μ, σ^2, and τ are poorly behaved because of irregularities in the likelihood function (Giesbrecht and Kempthorne, 1976). The method of moments does fairly well when the skew of the fitted distribution is reasonably small. A method that does almost as well as the moment method for low-skew distributions, and much better for highly skewed distributions, estimates τ by:

$$\hat{\tau} = \frac{x_{(1)}x_{(n)} - \hat{x}^2_{0.50}}{x_{(1)} + x_{(n)} - 2\hat{x}_{0.50}} \tag{7.81}$$

provided that $x_{(1)} + x_{(n)} - 2\hat{x}_{0.50} > 0$, where $x_{(1)}$ and $x_{(n)}$ are the smallest and largest observations and $\hat{x}_{0.50}$ is the sample median (Stedinger, 1980; Hoshi et al., 1984). If $x_{(1)} + x_{(n)} - 2\hat{x}_{0.50} < 0$, then the sample tends to be negatively skewed and a three-parameter lognormal distribution with a lower bound cannot be fit with this method. Good estimates of μ and σ^2 to go with $\hat{\tau}$ in Equation 7.81 are (Stedinger, 1980):

$$\hat{\mu} = \ln\left[\frac{\bar{x} - \hat{\tau}}{\sqrt{1 + s_X^2 /(\bar{x} - \hat{\tau})^2}}\right]$$

$$\hat{\sigma}^2 = \ln\left[1 + \frac{s_X^2}{(\bar{x} - \hat{\tau})^2}\right] \tag{7.82}$$

For the data in Table 7.2, Equations 7.81 and 7.82 yield the hybrid moment-of-moments estimates of $\hat{\mu} = 7.606$, $\hat{\sigma}^2 = 0.1339 = (0.3659)^2$ and $\hat{\tau} = -600.1$ for the three-parameter lognormal distribution.

This distribution has a coefficient of skewness of 1.19, which is more consistent with the sample skewness estimator than were the values obtained when a two-parameter lognormal distribution was fit to the data. Alternatively, one can estimate μ and σ^2 by the sample mean and variance of $\ln(X - \hat{\tau})$ which yields the hybrid maximum likelihood estimates $\hat{\mu} = 7.605$, $\hat{\sigma}^2 = 0.1407 = (0.3751)^2$ and again $\hat{\tau} = -600.1$.

The two sets of estimates are surprisingly close in this instance. In this second case, the fitted distribution has a coefficient of skewness of 1.22.

Natural logarithms have been used here. One could have just as well use base 10 common logarithms to estimate the parameters; however, in that case the relationships between the log-space parameters and the real-space moments change slightly (Stedinger et al., 1993, Equation. 18.2.8).

3.4. Gamma Distributions

The gamma distribution has long been used to model many natural phenomena, including daily, monthly and annual streamflows as well as flood flows (Bobée and Ashkar, 1991). For a gamma random variable X,

$$f_X(x) = \frac{|\beta|}{\Gamma(\alpha)}(\beta x)^{\alpha-1}e^{-\beta x} \qquad \beta x \geq 0$$

$$\mu_X = \frac{\alpha}{\beta}$$

$$\sigma_X^2 = \frac{\alpha}{\beta^2}$$

$$\gamma_X = \frac{2}{\sqrt{\alpha}} = 2CV_X \qquad \text{for } \beta > 0 \tag{7.83}$$

The gamma function, $\Gamma(\alpha)$, for integer α is $(\alpha - 1)!$. The parameter $\alpha > 0$ determines the shape of the distribution; β is the scale parameter. Figure 7.4 illustrates the different shapes that the probability density function for a gamma variable can assume. As $\alpha \to \infty$, the gamma distribution approaches the symmetric normal distribution, whereas for $0 < \alpha < 1$, the distribution has a highly asymmetric J-shaped probability density function whose value goes to infinity as x approaches zero.

The gamma distribution arises naturally in many problems in statistics and hydrology. It also has a very reasonable shape for such non-negative random variables as rainfall and streamflow. Unfortunately, its cumulative distribution function is not available in closed form, except for integer α, though it is available in many software packages including Microsoft Excel. The gamma

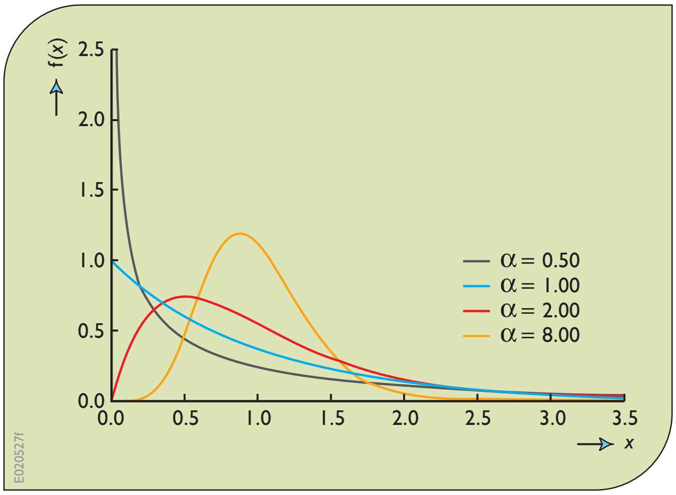

Figure 7.4. The gamma distribution function for various values of the shape parameter α.

family includes a very special case: the exponential distribution is obtained when $\alpha = 1$.

The gamma distribution has several generalizations (Bobée and Ashkar, 1991). If a constant τ is subtracted from X so that $(X - \tau)$ has a gamma distribution, the distribution of X is a three-parameter gamma. This is also called a *Pearson type 3 distribution*, because the resulting distribution belongs to the third type of distributions suggested by the statistician Karl Pearson. Another variation is the log Pearson type 3 distribution obtained by fitting the logarithms of X with a Pearson type 3 distribution. The log Pearson distribution is discussed further in the next section.

The method of moments may be used to estimate the parameters of the gamma distribution. For the three-parameter gamma distribution,

$$\hat{\tau} = \overline{x} - 2\left(\frac{s_X}{\hat{\gamma}_X}\right)$$

$$\hat{\sigma} = \frac{4}{(\hat{\gamma}_X)^2}$$

$$\hat{\beta} = \frac{2}{s_X \hat{\gamma}_X} \qquad (7.84)$$

where $\overline{x}$, s_X^2, and γ_X are estimates of the mean, variance, and coefficient of skewness of the distribution of X (Bobée and Robitaille, 1977).

For the two-parameter gamma distribution,

$$\hat{\alpha} = \frac{(\overline{x})^2}{s_X^2}$$

$$\hat{\beta} = \frac{\overline{x}}{s_X^2} \qquad (7.85)$$

Again, the flood record in Table 7.2 can be used to illustrate the different estimation procedures. Using the first three sample moments, one would obtain for the three-parameter gamma distribution the parameter estimates

$$\hat{\tau} = -735.6$$

$$\hat{\alpha} = 7.888$$

$$\hat{\beta} = 0.003452 = 1/427.2$$

Using only the sample mean and variance yields the method of moment estimators of the parameters of the two-parameter gamma distribution ($\tau = 0$),

$$\hat{\alpha} = 3.627$$

$$\hat{\beta} = 0.002341 = 1/427.2$$

The fitted two-parameter gamma distribution has a coefficient of skewness γ of 1.05, whereas the fitted three-parameter gamma reproduces the sample skew of 0.712. As occurred with the three-parameter lognormal distribution, the estimated lower bound for the three-parameter gamma distribution is negative ($\hat{\tau} = -735.6$), resulting in a three-parameter model that has a smaller skew coefficient than was obtained with the corresponding two-parameter model. The reciprocal of $\hat{\beta}$ is also reported. While $\hat{\beta}$ has inverse x-units, $1/\hat{\beta}$ is a natural scale parameter that has the same units as x and thus can be easier to interpret.

Studies by Thom (1958) and Matalas and Wallis (1973) have shown that maximum likelihood parameter estimates are superior to the moment estimates. For the two-parameter gamma distribution, Greenwood and Durand (1960) give approximate formulas for the maximum likelihood estimates (also Haan, 1977). However, the maximum likelihood estimators are often not used in practice because they are very sensitive to the smallest observations that sometimes suffer from measurement error and other distortions.

When plotting the observed and fitted quantiles of a gamma distribution, an approximation to the inverse of the distribution function is often useful. For $|\gamma| \leq 3$, the Wilson–Hilferty transformation

$$x_G = \mu + \sigma\left[\frac{2}{\gamma}\left(1 + \frac{\gamma x_N}{6} - \frac{\gamma^2}{36}\right)^3 - \frac{2}{\gamma}\right] \tag{7.86}$$

gives the quantiles x_G of the gamma distribution in terms of x_N, the quantiles of the standard-normal distribution. Here μ, σ, and γ are the mean, standard deviation, and coefficient of skewness of x_G. Kirby (1972) and Chowdhury and Stedinger (1991) discuss this and other more complicated but more accurate approximations. Fortunately the availability of excellent approximations of the gamma cumulative distribution function and its inverse in Microsoft Excel and other packages has reduced the need for such simple approximations.

3.5. Log-Pearson Type 3 Distribution

The log-Pearson type 3 distribution (LP3) describes a random variable whose logarithms have a Pearson type 3 distribution. This distribution has found wide use in modelling flood frequencies and has been recommended for that purpose (IACWD, 1982). Bobée (1975) and Bobée and Ashkar (1991) discuss the unusual shapes that this hybrid distribution may take allowing negative values of β. The LP3 distribution has a probability density function given by

$$f_X(x) = \frac{|\beta|}{x\Gamma(\alpha)}[\beta(\ln(x) - \xi)]^{\alpha-1}\exp\{-\beta(\ln(x) - \xi)\} \tag{7.87}$$

with $\alpha > 0$, and β either positive or negative. For $\beta < 0$, values are restricted to the range $0 < x < \exp(\xi)$. For $\beta > 0$, values have a lower bound so that $\exp(\xi) < X$. Figure 7.5 illustrates the probability density function for the LP3 distribution as a function of the skew γ of the P3 distribution describing $\ln(X)$, with $\sigma_{\ln X} = 0.3$. The LP3 density function for $|\gamma| \leq 2$ can assume a wide range of shapes with both positive and negative skews. For $|\gamma| = 2$, the log-space P3 distribution is equivalent to an exponential distribution function, which decays exponentially as x moves away from the lower bound ($\beta > 0$) or upper bound ($\beta < 0$): as a result the LP3 distribution has a similar shape. The space with $-1 < \gamma$ may be more realistic for describing variables whose probability density function becomes thinner as x takes on large values. For $\gamma = 0$, the two-parameter lognormal distribution is obtained as a special case.

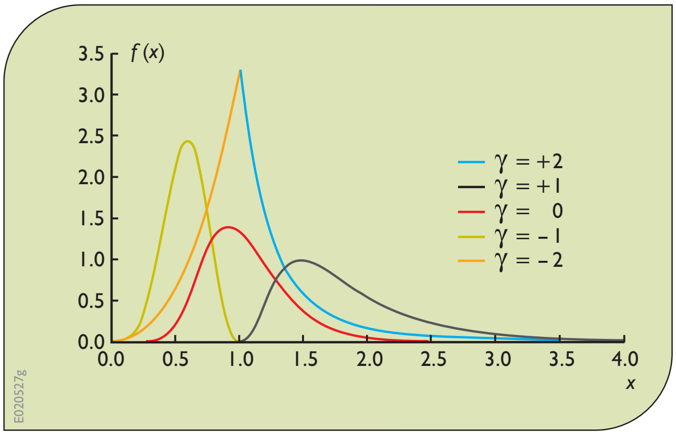

Figure 7.5. Log-Pearson type 3 probability density functions for different values of coefficient of skewness γ.

The LP3 distribution has mean and variance

$$\mu_X = e^{\xi}\left(\frac{\beta}{\beta-1}\right)^{\alpha}$$

$$\sigma_X^2 = e^{2\xi}\left\{\left(\frac{\beta}{\beta-2}\right)^{\alpha} - \left(\frac{\beta}{\beta-1}\right)^{2\alpha}\right\}$$

$$\text{for}\quad \beta > 2, \text{ or } \beta < 0. \tag{7.88}$$

For $0 < \beta < 2$, the variance is infinite.

These expressions are seldom used, but they do reveal the character of the distribution. Figures 7.6 and 7.7 provide plots of the real-space coefficient of skewness and coefficient of variation of a log-Pearson type 3 variate X as a function of the standard deviation σ_Y and coefficient of skew γ_Y of the log-transformation $Y = \ln(X)$. Thus the standard deviation σ_Y and skew γ_Y of Y are in log space. For $\gamma_Y = 0$, the log-Pearson type 3 distribution reduces to the two-parameter lognormal distribution discussed above, because in this case Y has a normal distribution. For the lognormal distribution, the standard deviation σ_Y serves as the sole shape parameter, and the coefficient of variation of X for small σ_Y is just σ_Y. Figure 7.7 shows that the situation is more

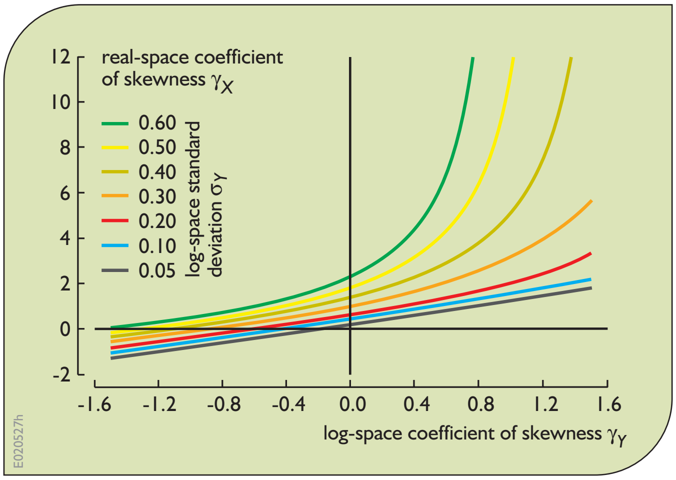

Figure 7.6. Real-space coefficient of skewness γ_X for LP3 distributed X as a function of log-space standard deviation σ_Y and coefficient of skewness γ_Y where $Y = \ln(X)$.

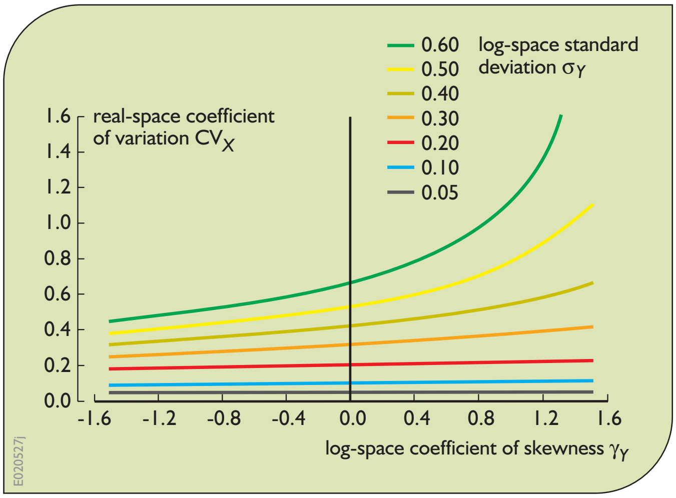

Figure 7.7. Real-space coefficient of variation CV_X for LP3 distributed X as a function of log-space standard deviation σ_Y and coefficient of skewness γ_Y where $Y = \ln(X)$.

complicated for the LP3 distribution. However, for small σ_Y, the coefficient of variation of X is approximately σ_Y.

Again, the flood flow data in Table 7.2 can be used to illustrate parameter estimation. Using natural logarithms, one can estimate the log-space moments with the standard estimators in Equations 7.39 that yield:

$$\hat{\mu} = 7.202$$

$$\hat{\sigma} = 0.5625$$

$$\hat{\gamma} = -0.337$$

For the LP3 distribution, analysis generally focuses on the distribution of the logarithms $Y = \ln(X)$ of the flows, which would have a Pearson type 3 distribution with moments μ_Y, σ_Y and γ_Y (IACWD, 1982; Bobée and Ashkar, 1991). As a result, flood quantiles are calculated as

$$x_p = \exp\{\mu_Y + \sigma_Y K_p[\gamma_Y]\} \qquad (7.89)$$

where $K_p[\gamma_Y]$ is a frequency factor corresponding to cumulative probability p for skewness coefficient γ_Y. ($K_p[\gamma_Y]$ corresponds to the quantiles of a three-parameter gamma distribution with zero mean, unit variance, and skewness coefficient γ_Y.)

Since 1967 the recommended procedure for flood frequency analysis by federal agencies in the United States has used this distribution. Current guidelines in Bulletin 17B (IACWD, 1982) suggest that the skew γ_Y be estimated by a weighted average of the at-site sample skewness coefficient and a regional estimate of the skewness coefficient. Bulletin 17B also includes tables of frequency factors, a map of regional skewness estimators, checks for low outliers, confidence interval formula, a discussion of expected probability and a weighted-moments estimator for historical data.

3.6. Gumbel and GEV Distributions

The annual maximum flood is the largest flood flow during a year. One might expect that the distribution of annual maximum flood flows would belong to the set of extreme value distributions (Gumbel, 1958; Kottegoda and Rosso, 1997). These are the distributions obtained in the limit, as the sample size n becomes large, by taking the largest of n independent random variables. The Extreme Value (EV) type I distribution, or Gumbel distribution, has often been used to describe flood flows. It has the cumulative distribution function:

$$F_X(x) = \exp\{-\exp[-(x - \xi)/\alpha]\} \qquad (7.90)$$

where ξ is the location parameter. It has a mean and variance of

$$\mu_X = \xi + 0.5772\alpha$$

$$\sigma_X^2 = \pi^2\alpha^2/6 \cong 1.645\alpha^2 \qquad (7.91)$$

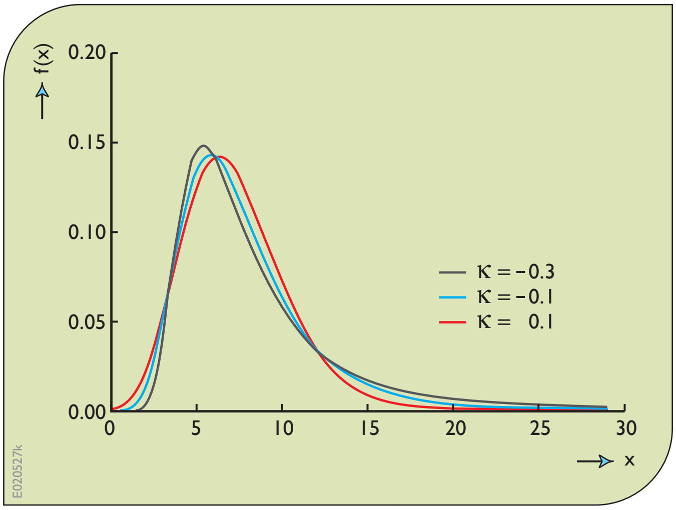

Figure 7.8. GEV density distributions for selected shape parameter κ values.

Its skewness coefficient has a fixed value equal to $\gamma_X = 1.1396$.

The generalized extreme value (GEV) distribution is a general mathematical expression that incorporates the type I, II, and III extreme value (EV) distributions for maxima (Gumbel, 1958; Hosking et al., 1985). In recent years it has been used as a general model of extreme events including flood flows, particularly in the context of regionalization procedures (NERC, 1975; Stedinger and Lu, 1995; Hosking and Wallis, 1997). The GEV distribution has the cumulative distribution function:

$$F_X(x) = \exp\{-[1 - \kappa(x-\xi)/\alpha]^{1/\kappa}\} \quad \text{for} \quad \kappa \neq 0 \quad (7.92)$$

From Equation 7.92, it is clear that for $\kappa < 0$ (the typical case for floods), x must exceed $\xi + \alpha/\kappa$, whereas for $\kappa > 0$, x must be no greater than $\xi + \alpha/\kappa$ (Hosking and Wallis, 1987). The mean, variance, and skewness coefficient are (for $\kappa > -1/3$):

$$\mu_X = \xi + (\alpha/\kappa)\ [1 - \Gamma(1+\kappa)],$$

$$\sigma_X^2 = (\alpha/\kappa)^2\ \{\Gamma(1 + 2\kappa) - [\Gamma(1 + \kappa)]^2\} \qquad (7.93)$$

$$\gamma_X = (\text{Sign } \kappa)\{-\Gamma(1 + 3\kappa) + 3\Gamma(1 + \kappa)\ \Gamma(1 + 2\kappa) \\ -2[\Gamma(1 + \kappa)]^3\}/\{\Gamma(1 + 2\kappa) - [\Gamma(1 + \kappa)]^2\}^{3/2}$$

where $\Gamma(1+\kappa)$ is the classical gamma function. The Gumbel distribution is obtained when $\kappa = 0$. For $|\kappa| < 0.3$, the general shape of the GEV distribution is similar to the Gumbel distribution, though the right-hand tail is thicker for $\kappa < 0$, and thinner for $\kappa > 0$, as shown in Figures 7.8 and 7.9.

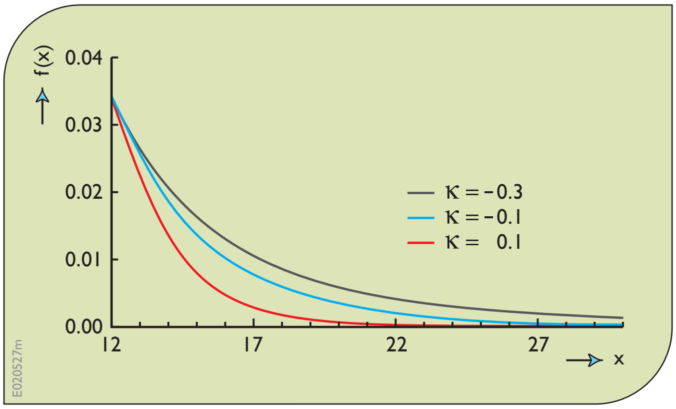

Figure 7.9. Right-hand tails of GEV distributions shown in Figure 7.8.

The parameters of the GEV distribution are easily computed using L-moments and the relationships (Hosking et al. (1985):

$$\kappa = 7.8590c + 2.9554c^2$$

$$\alpha = \kappa\lambda_2/[\Gamma(1 + \kappa)(1 - 2^{-\kappa})] \qquad (7.94)$$

$$\xi = \lambda_1 + (\alpha/\kappa)[\Gamma(1+\kappa) - 1]$$

where

$$c = 2\lambda_2/(\lambda_3 + 3\lambda_2) - \ln(2)/\ln(3)$$
$$= [2/(\tau_3 + 3)] - \ln(2)/\ln(3)$$

As one can see, the estimator of the shape parameter κ will depend only upon the L-skewness estimator $\hat{\tau}_3$. The estimator of the scale parameter α will then depend on the estimate of κ and of λ_2. Finally, one must also use the sample mean λ_1 (Equation 7.48) to determine the estimate of the location parameter ξ.

Using the flood data in Table 7.2 and the sample L-moments computed in Section 2, one obtains first $c = -0.000896$ which yields $\hat{\kappa} = -0.007036$, $\hat{\xi} = 1{,}165.20$ and $\hat{\alpha} = 657.29$.

The small value of the fitted κ parameter means that the fitted distribution is essentially a Gumbel distribution. Again, ξ is a location parameter, not a lower bound, so its value resembles a reasonable x value.

Madsen et al. (1997a) show that moment estimators can provide more precise quantile estimators. Martins and

Stedinger (2001b) found that with occasional uninformative samples, the MLE estimator of κ could be entirely unrealistic resulting in absurd quantile estimators. However the use of a realistic prior distribution on κ yielded generalized maximum likelihood estimators (GLME) that performed better than moment and L-moment estimators over the range of κ of interest.

The generalized maximum likelihood estimators (GMLE) are obtained by maximizing the log-likelihood function, augmented by a prior density function on κ. A prior distribution that reflects general world-wide geophysical experience and physical realism is in the form of a beta distribution:

$$\pi(\kappa) = \Gamma(p)\,\Gamma(q)\,(0.5 + \kappa)^{p-1}\,(0.5 - \kappa)^{q-1}/\Gamma(p+q) \tag{7.95}$$

for $-0.5 < \kappa < +0.5$ with $p = 6$ and $q = 9$. Moreover, this prior assigns reasonable probabilities to the values of κ within that range. For κ outside the range -0.4 to $+0.2$ the resulting GEV distributions do not have density functions consistent with flood flows and rainfall (Martins and Stedinger, 2000). Other estimators implicitly have similar constraints. For example, L-moments restricts κ to the range $\kappa > -1$, and the method of moments estimator employs the sample standard deviation so that $\kappa > -0.5$. Use of the sample skew introduces the constraint that $\kappa > -0.3$. Then given a set of independent observations $\{x_1, \ldots, x_n\}$ drawn for a GEV distribution, the generalized likelihood function is:

$$\ln\{L(\xi,\alpha,\kappa \mid x_1,\ldots,x_n)\} = -n\ln(\alpha) + \sum_{i=1}^{n}\left[\left(\frac{1}{\kappa}-1\right)\ln(y_i)-(y_i)^{1/\kappa}\right] + \ln[\pi(\kappa)]$$

with

$$y_i = [1-(\kappa/\alpha)(x_i-\xi)] \tag{7.96}$$

For feasible values of the parameters, y_i is greater than 0 (Hosking et al., 1985). Numerical optimization of the generalized likelihood function is often aided by the additional constraint that $\min\{y_1, \ldots, y_n\} \geq \varepsilon$ for some small $\varepsilon > 0$ so as to prohibit the search generating infeasible values of the parameters for which the likelihood function is undefined. The constraint should not be binding at the final solution.

The data in Table 7.2 again provide a convenient data set for illustrating parameter estimators. The L-moment estimators were used to generate an initial solution. Numerical optimization of the likelihood function in Equation 7.96 yielded the maximum likelihood estimators of the GEV parameters:

$$\hat{\kappa} = -0.0359, \quad \hat{\xi} = 1165.4 \text{ and } \hat{\alpha} = 620.2.$$

Similarly, use of the geophysical prior (Equation 7.95) yielded the generalized maximum likelihood estimators $\hat{\kappa} = -0.0823$, $\hat{\xi} = 1150.8$ and $\hat{\alpha} = 611.4$. Here the record length of forty years is too short to define reliably the shape parameter κ so that result of using the prior is to pull κ slightly toward the mean of the prior. The other two parameters adjust accordingly.

3.7. L-Moment Diagrams

Section 3 presented several families of distributions. The L-moment diagram in Figure 7.10 illustrates the relationships between the L-kurtosis (τ_4) and L-skewness (τ_3) for a number of distributions often used in hydrology. It shows that distributions with the same coefficient of skewness still differ in the thickness of their tails. This thickness is described by their kurtosis. Tail shapes are important if an analysis is sensitive to the likelihood of extreme events.

The normal and Gumbel distributions have a fixed shape and thus are presented by single points that fall on the Pearson type 3 (P3) curve for $\gamma = 0$, and the generalized extreme value (GEV) curve for $\kappa = 0$, respectively. The L-kurtosis/L-skewness relationships for the two-parameter and three-parameter gamma or P3 distributions are identical, as they are for the two-parameter and three-parameter lognormal distributions. This is because the addition of a location parameter does not change the range of fundamental shapes that can be generated. However, for the same skewness coefficient, the lognormal distribution has a larger kurtosis than the gamma or P3 distribution, and thus assigns larger probabilities to the largest events.

As the skewness of the lognormal and gamma distributions approaches zero, both distributions become normal and their kurtosis/skewness relationships merge. For the same L-skewness, the L-kurtosis of the GEV distribution is generally larger than that of the lognormal distribution. For positive κ yielding almost symmetric or even negatively skewed GEV distributions, the GEV has a smaller kurtosis than the three-parameter lognormal distribution.

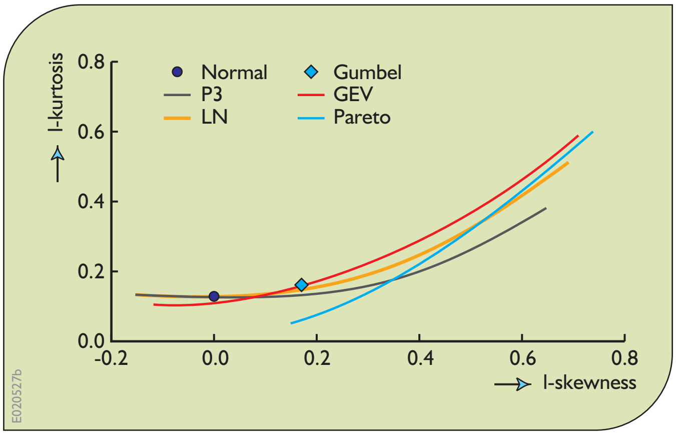

Figure 7.10. Relationships between L-skewness and L-kurtosis for various distributions.

The latter can be negatively skewed when the log normal location parameter τ is used as an upper bound.

Figure 7.10 also includes the three-parameter generalized Pareto distribution, whose cdf is:

$$F_X(x) = 1 - [1 - \kappa(x - \xi)/\alpha]^{1/\kappa} \qquad (7.97)$$

(Hosking and Wallis, 1997). For $\kappa = 0$ it corresponds to the exponential distribution (gamma with $\alpha = 1$). This point is where the Pareto and P3 distribution L-kurtosis/L-skewness lines cross. The Pareto distribution becomes increasing more skewed for $\kappa < 0$, which is the range of interest in hydrology. The generalized Pareto distribution with $\kappa < 0$ is often used to describe peaks-over-a-threshold and other variables whose probability density function has its maximum at their lower bound. In that range for a given L-skewness, the Pareto distribution always has a larger kurtosis than the gamma distribution. In these cases the α parameter for the gamma distribution would need to be in the range $0 < \alpha < 1$, so that both distributions would be J-shaped.

As shown in Figure 7.10, the GEV distribution has a thicker right-hand tail than either the gamma/Pearson type 3 distribution or the lognormal distribution.

4. Analysis of Censored Data

There are many instances in water resources planning where one encounters censored data. A data set is censored if the values of observations that are outside a specified range of values are not specifically reported (David, 1981). For example, in water quality investigations many constituents have concentrations that are reported as $<T$, where T is a reliable detection threshold (MacBerthouex and Brown, 2002). Thus the concentration of the water quality variable of interest was too small to be reliably measured. Likewise, low-flow observations and rainfall depths can be rounded to or reported as zero. Several approaches are available for analysis of censored data sets, including probability plots and probability-plot regression, conditional probability models and maximum likelihood estimators (Haas and Scheff, 1990; Helsel, 1990; Kroll and Stedinger, 1996; MacBerthouex and Brown, 2002).

Historical and physical paleoflood data provide another example of censored data. Before the beginning of a continuous measurement program on a stream or river, the stages of unusually large floods can be estimated on the basis of the memories of people who have experienced these events and/or physical markings in the watershed (Stedinger and Baker, 1987). Annual maximum floods that were not unusual were not recorded nor were they remembered. These missing data are censored data. They cover periods between occasional large floods that have been recorded or that have left some evidence of their occurrence (Stedinger and Cohn, 1986).

The discussion below addresses probability-plot methods for use with censored data. Probability-plot methods have a long history of use for this purpose because they are relatively simple to use and to understand. Moreover, recent research has shown that they are relatively efficient when the majority of values are observed, and unobserved values are known only to be below (or above) some detection limit or perception threshold that serves as a lower (or upper) bound. In such cases, probability-plot regression estimators of moments and quantiles are as accurate as maximum likelihood estimators. They are almost as good as estimators computed with complete samples (Helsel and Cohn, 1988; Kroll and Stedinger, 1996).

Perhaps the simplest method for dealing with censored data is adoption of a conditional probability model. Such models implicitly assume that the data are drawn from one of two classes of observations: those below a single threshold, and those above the threshold. This model is appropriate for simple cases where censoring occurs because small observations are recorded as 'zero,' as often happens with low-flow, low pollutant concentration, and some flood records. The conditional probability model introduces an extra parameter P_0 to describe the probability that an observation is 'zero'. If r of a total of n observations were observed because they exceeded the threshold, then P_0 is estimated as $(n - r)/n$. A continuous distribution $G_X(x)$ is derived for the strictly positive 'non-zero' values of X. Then the parameters of the G distribution can be estimated using any procedure appropriate for complete uncensored samples. The unconditional cdf for any value $x > 0$, is then

$$F_X(x) = P_0 + (1 - P_0)\, G(x) \tag{7.98}$$

This model completely decouples the value of P_0 from the parameters that describe the G distribution.

Section 7.2 discusses probability plots and plotting positions useful for graphical displays of data to allow a visual examination of the empirical frequency curve. Suppose that among n samples a detection limit is exceeded by the observations r times. The natural estimator of the exceedance probability P_0 of the perception threshold is again $(n - r)/n$. If the r values that exceeded the threshold are indexed by $i = 1, \ldots, r$, wherein $x_{(r)}$ is the largest observation, reasonable plotting positions within the interval $[P_0, 1]$ are:

$$p_i = P_0 + (1 - P_0)\, [(i - a)/(r + 1 - 2a)] \tag{7.99}$$

where a defines the plotting position that is used; $a = 0$ is reasonable (Hirsch and Stedinger, 1987). Helsel and Cohn (1988) show that reasonable choices for a generally make little difference. Both papers discuss development of plotting positions when there are different thresholds, as occurs when the analytical precision of instrumentation changes over time. If there are many exceedances of the threshold so that $r >> (1 - 2a)$, p_i is indistinguishable from

$$p_i' = [i + (n - r) - a]/(n + 1 - 2a) \tag{7.100}$$

where again, $i = 1, \ldots, r$. These values correspond to the plotting positions that would be assigned to the largest r observations in a complete sample of n values.

The idea behind the probability-plot regression estimators is to use the probability plot for the observed data to define the parameters of the whole distribution. And if a sample mean, sample variance or quantiles are needed, then the distribution defined by the probability plot is used to fill in the missing (censored) observations so that standard estimators of the mean, standard deviation and of quantiles can be employed. Such fill-in procedures are efficient and relatively robust for fitting a distribution and estimating various statistics with censored water quality data when a modest number of the smallest observations are censored (Helsel, 1990; Kroll and Stedinger, 1996).

Unlike the conditional probability approach, here the below threshold probability P_0 is linked with the selected probability distribution for the above-threshold observations. The observations below the threshold are censored but are in all other respects envisioned as coming from the same distribution that is used to describe the observed above-threshold values.

When water quality data are well described by a lognormal distribution, available values $\ln[X_{(1)}] \leq \cdots \leq \ln[X_{(r)}]$ can be regressed upon $F^{-1}[p_i] = \mu + \sigma F^{-1}[p_i]$ for $i = 1, \ldots, r$, where the r largest observations in a sample of size n are available. If regression yields constant m and slope s corresponding to the first and second population moments μ and σ, a good estimator of the pth quantile is

$$x_p = \exp[m + s z_p] \tag{7.101}$$

wherein z_p is the p^{th} quantile of the standard normal distribution. To estimate sample means and other statistics one can fill in the missing observations with

$$x(j) = \exp\{y(j)\} \quad \text{for} \quad j = 1, \ldots, (n - r) \tag{7.102}$$

where

$$y(j) = m + sF^{-1}\{P_0[(j - a)/(n - r + 1 - 2a)]\} \tag{7.103}$$

Once a complete sample is constructed, standard estimators of the sample mean and variance can be calculated, as can medians and ranges. By filling in the missing small observations, and then using complete-sample estimators of statistics of interest, the procedure is relatively insensitive to the assumption that the observations actually have a lognormal distribution.

Maximum likelihood estimators are quite flexible, and are more efficient than plotting-position methods when the values of the observations are not recorded because they are below or above the perception threshold (Kroll and Stedinger, 1996). Maximum likelihood methods allow the observations to be represented by exact values, ranges and various thresholds that either were or were not exceeded at various times. This can be particularly important with historical flood data sets because the magnitudes of many historical floods are not recorded precisely, and it may be known that a threshold was never crossed or was crossed at most once or twice in a long period (Stedinger and Cohn, 1986; Stedinger, 2000; O'Connell et al., 2002). Unfortunately, maximum likelihood estimators for the LP3 distribution have proven to be problematic. However, recently developed expected moment estimators seem to do as well as maximum likelihood estimators with the LP3 distribution (Cohn et al., 1997, 2001; Griffs et al., 2004).

While often a computational challenge, maximum likelihood estimators for complete samples, and samples with some observations censored, pose no conceptual challenge. One need only write the maximum likelihood function for the data and proceed to seek the parameter values that maximize that function. Thus if $F(x \mid \boldsymbol{\theta})$ and $f(x \mid \boldsymbol{\theta})$ are the cumulative distribution and probability density functions that should describe the data, and $\boldsymbol{\theta}$ is the vector of parameters of the distribution, then for the case described above wherein $x_1, \ldots, x_r$ are r of n observations that exceeded a threshold T, the likelihood function would be (Stedinger and Cohn, 1986):

$$\begin{aligned} &\mathrm{L}(\boldsymbol{\theta} \mid r, n, x_1, \ldots, x_r) \\ &\quad = F(T \mid \boldsymbol{\theta})^{(n-r)} f(x_1 \mid \boldsymbol{\theta}) f(x_2 \mid \boldsymbol{\theta}) \cdots f(x_r \mid \boldsymbol{\theta}) \end{aligned} \qquad (7.104)$$

Here, $(n - r)$ observations were below the threshold T, and the probability an observation is below T is $F(T \mid \boldsymbol{\theta})$ which then appears in Equation 7.104 to represent that observation. In addition, the specific values of the r observations $x_1, \ldots, x_r$ are available, where the probability an observation is in a small interval of width δ around x_i is δ $f(x_i \mid \boldsymbol{\theta})$. Thus strictly speaking the likelihood function also includes a term δ^r. Here what is known of the magnitude of all of the n observations is included in the likelihood function in the appropriate way. If all that were known of some observation was that it exceeded a threshold M, then that value should be represented by a term $[1 - F(M \mid \boldsymbol{\theta})]$ in the likelihood function. Similarly, if all that were known was that the value was between L and M, then a term $[F(M \mid \boldsymbol{\theta}) - F(L \mid \boldsymbol{\theta})]$ should be included in the likelihood function. Different thresholds can be used to describe different observations corresponding to changes in the quality of measurement procedures. Numerical methods can be used to identify the parameter vector that maximizes the likelihood function for the data available.

5. Regionalization and Index-Flood Method

Research has demonstrated the potential advantages of 'index flood' procedures (Lettenmaier et al., 1987; Stedinger and Lu, 1995; Hosking and Wallis, 1997; Madsen, and Rosbjerg, 1997a). The idea behind the index-flood approach is to use the data from many hydrologically 'similar' basins to estimate a dimensionless flood distribution (Wallis, 1980). Thus this method 'substitutes space for time' by using regional information to compensate for having relatively short records at each site. The concept underlying the index-flood method is that the distributions of floods at different sites in a 'region' are the same except for a scale or index-flood parameter that reflects the size, rainfall and runoff characteristics of each watershed. Research is revealing when this assumption may be reasonable. Often a more sophisticated multi-scaling model is appropriate (Gupta and Dawdy, 1995a; Robinson and Sivapalan, 1997).

Generally the mean is employed as the index flood. The problem of estimating the p^{th} quantile x_p is then reduced to estimation of the mean, μ_x, for a site and the ratio x_p/μ_x of the p^{th} quantile to the mean. The mean can often be estimated adequately with the record available at a site, even if that record is short. The indicated ratio is estimated using regional information. The British Flood Studies Report (NERC, 1975) calls these normalized flood distributions "growth curves".

Key to the success of the index-flood approach is identification of sets of basins that have similar coefficients of variation and skew. Basins can be grouped geographically, as well as by physiographic characteristics including drainage area and elevation. Regions need not be geographically contiguous. Each site can potentially be assigned its own unique region consisting of sites with which it is particularly similar (Zrinji and Burn, 1994), or regional regression equations can be derived to compute normalized regional quantiles as a function of a site's physiographic characteristics and other statistics (Fill and Stedinger, 1998).

Clearly the next step for regionalization procedures, such as the index-flood method, is to move away from estimates of regional parameters that do not depend upon basin size and other physiographic parameters. Gupta et al. (1994) argue that the basic premise of the index-flood method – that the coefficient of variation of floods is relatively constant – is inconsistent with the known relationships between the coefficient of variation (CV) and drainage area (see also Robinson and Sivapalan, 1997). Recently, Fill and Stedinger (1998) built such a relationship into an index-flood procedure by using a regression model to explain variations in the normalized quantiles. Tasker and Stedinger (1986) illustrated how one might relate log-space skew to physiographic basin characteristics (see also Gupta and Dawdy, 1995b). Madsen and Rosbjerg (1997b) did the same for a regional model of κ for the GEV distribution. In both studies, only a binary variable representing 'region' was found useful in explaining variations in these two shape parameters.

Once a regional model of alternative shape parameters is derived, there may be some advantage to combining such regional estimators with at-site estimators employing an empirical Bayesian framework or some other weighting schemes. For example, Bulletin 17B recommends weigh at-site and regional skewness estimators, but almost certainly places too much weight on the at-site values (Tasker and Stedinger, 1986). Examples of empirical Bayesian procedures are provided by Kuczera (1982), Madsen and Rosbjerg (1997b) and Fill and Stedinger (1998). Madsen and Rosbjerg's (1997b) computation of a κ-model with a New Zealand data set demonstrates how important it can be to do the regional analysis carefully, taking into account the cross-correlation among concurrent flood records.

When one has relatively few data at a site, the index-flood method is an effective strategy for deriving flood frequency estimates. However, as the length of the available record increases, it becomes increasingly advantageous to also use the at-site data to estimate the coefficient of variation as well. Stedinger and Lu (1995) found that the L-moment/GEV index-flood method did quite well for 'humid regions' ($CV \approx 0.5$) when $n < 25$, and for semi-arid regions ($CV \approx 1.0$) for $n < 60$, if reasonable care is taken in selecting the stations to be included in a regional analysis. However, with longer records, it became advantageous to use the at-site mean and L-CV with a regional estimator of the shape parameter for a GEV distribution. In many cases this would be roughly equivalent to fitting a Gumbel distribution corresponding to a shape parameter $\kappa = 0$. Gabriele and Arnell (1991) develop the idea of having regions of different sizes for different parameters. For realistic hydrological regions, these and other studies illustrate the value of regionalizing estimators of the shape, and often the coefficient of variation of a distribution.

6. Partial Duration Series

Two general approaches are available for modelling flood and precipitation series (Langbein, 1949). An annual maximum series considers only the largest event in each year. A partial duration series (PDS) or peaks-over-threshold (POT) approach includes all 'independent' peaks above a truncation or threshold level. An objection to using annual maximum series is that it employs only the largest event in each year, regardless of whether the second-largest event in a year exceeds the largest events of other years. Moreover, the largest annual flood flow in a dry year in some arid or semi-arid regions may be zero, or so small that calling them floods is misleading. When considering rainfall series or pollutant discharge events, one may be interested in modelling all events that occur within a year that exceed some threshold of interest.

Use of a partial duration series framework avoids such problems by considering all independent peaks that exceed a specified threshold. Furthermore, one can estimate annual exceedance probabilities from the analysis of partial duration series. Arguments in favour of partial duration series are that relatively long and reliable records are often available, and if the arrival rate for peaks over the threshold is large enough (1.65 events/year for the Poisson-arrival with exponential-exceedance model), partial duration series analyses should yield more accurate estimates of extreme quantiles than the corresponding annual-maximum frequency analyses (NERC, 1975; Rosbjerg, 1985). However, when fitting a three-parameter distribution, there seems to be little advantage from the use of a partial duration series approach over an annual maximum approach. This is true even when the partial duration series includes many more peaks than the maximum series because both contain the same largest events (Martins and Stedinger, 2001a).

A drawback of partial duration series analyses is that one must have criteria to identify only independent peaks (and not multiple peaks corresponding to the same event). Thus, such analysis can be more complicated than analyses using annual maxima. Partial duration models, perhaps with parameters that vary by season, are often used to estimate expected damages from hydrological events when more than one damage-causing event can occur in a season or within a year (North, 1980).

A model of a partial duration series has at least two components: first, one must model the arrival rate of events larger than the threshold level; second, one must model the magnitudes of those events. For example, a Poisson distribution has often been used to model the arrival of events, and an exponential distribution to describe the magnitudes of peaks that exceed the threshold.

There are several general relationships between the probability distribution for annual maximum and the frequency of events in a partial duration series. For a partial duration series model, let λ be the average arrival rate of flood peaks greater than the threshold x_0 and let $G(x)$ be the probability that flood peaks, when they occur, are less than $x > x_0$, and thus those peaks fall in the range $[x_0, x]$. The annual exceedance probability for a flood, denoted $1/T_a$, corresponding to an annual return period T_a, is related to the corresponding exceedance probability $q_e = [1 - G(x)]$ for level x in the partial duration series by

$$1/T_a = 1 - \exp\{-\lambda q_e\} = 1 - \exp\{-1/T_p\} \qquad (7.105)$$

where $T_p = 1/(\lambda q_e)$ is the average return period for level x in the partial duration series.

Many different choices for $G(x)$ may be reasonable. In particular, the generalized Pareto distribution (GPD) is a simple distribution useful for describing floods that exceed a specified lower bound. The cumulative distribution function for the generalized three-parameter Pareto distribution is:

$$F_X(x) = 1 - [1 - \kappa(x - \xi)/\alpha]^{1/\kappa} \qquad (7.106)$$

with mean and variance

$$\mu_X = \xi + \alpha/(1+\kappa)\kappa$$

$$\sigma_X^2 = \alpha^2/[(1+\kappa)^2(1+2\kappa)] \qquad (7.107)$$

where for $\kappa < 0$, $\xi < x < \infty$, whereas for $\kappa > 0$, $\xi < x < \xi + \alpha/\kappa$ (Hosking and Wallis, 1987). A special case of the GPD is the two-parameter exponential distribution with $\kappa = 0$. Method of moment estimators work relatively well (Rosbjerg et al., 1992).

Use of a generalized Pareto distribution for $G(x)$ with a Poisson arrival model yields a GEV distribution for the annual maximum series greater than x_0 (Smith, 1984; Stedinger et al., 1993; Madsen et al., 1997a). The Poisson-Pareto and Poisson-GPD models provide very reasonable descriptions of flood risk (Rosbjerg et al., 1992). They have the advantage that they focus on the distribution of the larger flood events, and regional estimates of the GEV distribution's shape parameter κ from annual maximum and partial duration series analyses can be used interchangeably.

Madsen and Rosbjerg (1997a) use a Poisson-GPD model as the basis of a partial duration series index-flood procedure. Madsen et al. (1997b) show that the estimators are fairly efficient. They pooled information from many sites to estimate the single shape parameter κ and the arrival rate where the threshold was a specified percentile of the daily flow duration curve at each site. Then at-site information was used to estimate the mean above-threshold flood. Alternatively, one could use the at-site data to estimate the arrival rate as well.

7. Stochastic Processes and Time Series

Many important random variables in water resources are functions whose values change with time. Historical records of rainfall or streamflow at a particular site are a sequence of observations called a *time series*. In a time series, the observations are ordered by time, and it is generally the case that the observed value of the random variable at one time influences one's assessment of the distribution of the random variable at later times. This means that the observations are not independent. Time series are conceptualized as being a single observation of a *stochastic process,* which is a generalization of the concept of a random variable.

This section has three parts. The first presents the concept of *stationarity* and the basic statistics generally used to describe the properties of a stationary stochastic process. The second presents the definition of a Markov process and the Markov chain model. Markov chains are a convenient model for describing many phenomena, and are often used in synthetic flow and rainfall generation and optimization models. The third part discusses the sampling properties of statistics used to describe the characteristics of many time series.

7.1. Describing Stochastic Processes

A random variable whose value changes through time according to probabilistic laws is called a *stochastic process*. An observed *time series* is considered to be one realization of a stochastic process, just as a single observation of a random variable is one possible value the random variable may assume. In the development here, a stochastic process is a sequence of random variables $\{X(t)\}$ ordered by a discrete time variable $t = 1, 2, 3, \dots$

The properties of a stochastic process must generally be determined from a single time series or realization. To do this, several assumptions are usually made. First, one generally assumes that the process is stationary. This means that the probability distribution of the process is not changing over time. In addition, if a process is strictly stationary, the joint distribution of the random variables $X(t_1), \dots, X(t_n)$ is identical to the joint distribution of $X(t_1 + t), \dots, X(t_n + t)$ for any t; the joint distribution depends only on the differences $t_i - t_j$ between the times of occurrence of the events.

For a stationary stochastic process, one can write the mean and variance as

$$\mu_X = \mathrm{E}[X(t)] \tag{7.108}$$

and

$$\sigma^2 = \mathrm{Var}[X(t)] \tag{7.109}$$

Both are independent of time t. The *autocorrelations*, the correlation of X with itself, are given by

$$\rho_X(k) = \mathrm{Cov}[X(t), X(t+k)]/\sigma_X^2 \tag{7.110}$$

for any positive integer time lag k. These are the statistics most often used to describe stationary stochastic processes.

When one has available only a single time series, it is necessary to estimate the values of μ_X, σ_X^2, and $\rho_X(k)$ from values of the random variable that one has observed. The mean and variance are generally estimated essentially as they were in Equation 7.39.

$$\hat{\mu}_X = \bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t \tag{7.111}$$

$$\sigma_X^2 = \frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2 \tag{7.112}$$

while the autocorrelations $\rho_X(k)$ for any time lag k can be estimated as (Jenkins and Watts, 1968)

$$\hat{\rho}_X(k) = r_k = \frac{\sum_{t=1}^{T-k}(x_{t+k} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \tag{7.113}$$

The sampling distribution of these estimators depends on the correlation structure of the stochastic process giving rise to the time series. In particular, when the observations are positively correlated, as is usually the case in natural streamflows or annual benefits in a river basin simulation, the variances of the estimated $\bar{x}$ and $\hat{\sigma}_X^2$ are larger than would be the case if the observations were independent. It is sometimes wise to take this inflation into account. Section 7.3 discusses the sampling distribution of these statistics.

All of this analysis depends on the assumption of stationarity, for only then do the quantities defined in Equations 7.108 to 7.110 have the intended meaning. Stochastic processes are not always stationary. Agricultural and urban development, deforestation, climatic variability and changes in regional resource management can alter the distribution of rainfall, streamflows, pollutant concentrations, sediment loads and groundwater levels over time. If a stochastic process is not essentially stationary over the time span in question, then statistical techniques that rely on the stationary assumption do not apply and the problem generally becomes much more difficult.

7.2. Markov Processes and Markov Chains

A common assumption in many stochastic water resources models is that the stochastic process $X(t)$ is a *Markov process*. A first-order Markov process has the property that the dependence of future values of the process on past values depends only on the current value and not on previous values or observations. In symbols for $k > 0$,

$$F_X[X(t+k) \mid X(t), X(t-1), X(t-2), \ldots] = F_X[X(t+k) \mid X(t)] \qquad (7.114)$$

For Markov processes, the current value summarizes the state of the processes. As a consequence, the current value of the process is often referred to as the *state*. This makes physical sense as well when one refers to the state or level of an aquifer or reservoir.

A special kind of Markov process is one whose state $X(t)$ can take on only discrete values. Such a process is called a *Markov chain*. Often in water resources planning, continuous stochastic processes are approximated by Markov chains. This is done to facilitate the construction of simple stochastic models. This section presents the basic notation and properties of Markov chains.

Consider a stream whose annual flow is to be represented by a discrete random variable. Assume that the distribution of streamflows is stationary. In the following development, the continuous random variable representing the annual streamflows (or some other process) is approximated by a random variable Q_y in year y, which takes on only n discrete values q_i (each value representing a continuous range or interval of possible streamflows) with unconditional probabilities p_i where

$$\sum_{i=1}^{n} p_i = 1 \qquad (7.115)$$

It is frequently the case that the value of Q_{y+1} is not independent of Q_y. A Markov chain can model such dependence. This requires specification of the *transition probabilities* p_{ij},

$$p_{ij} = \Pr[Q_{y+1} = q_j \mid Q_y = q_i] \qquad (7.116)$$

A transition probability is the conditional probability that the next state is q_j, given that the current state is q_i. The transition probabilities must satisfy

$$\sum_{j=1}^{n} p_{ij} = 1 \quad \text{for all } i \qquad (7.117)$$

Figure 7.11 shows a possible set of transition probabilities in a matrix. Each element p_{ij} in the matrix is the probability of a transition from streamflow q_i in one year to streamflow q_j in the next. In this example, a low flow tends to be followed by a low flow, rather than a high flow, and vice versa.

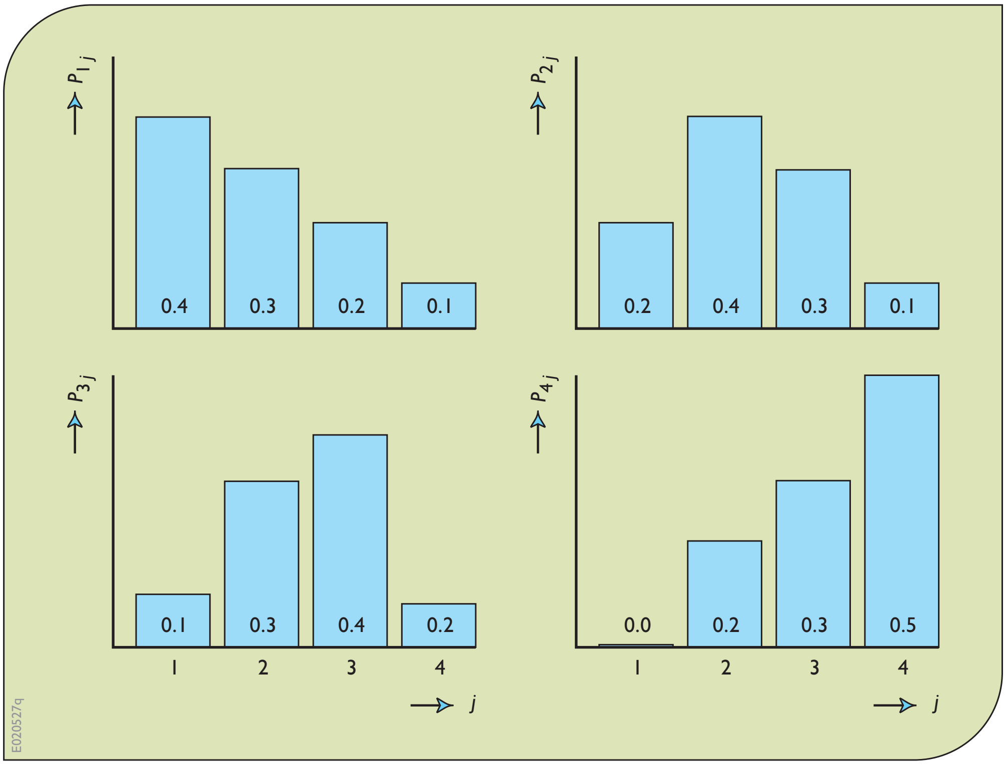

Figure 7.11. Matrix (above) and histograms (below) of streamflow transition probabilities showing probability of streamflow q_j (represented by index j) in year $y+1$ given streamflow q_i (represented by index i) in year y.

year	P_1^y	P_2^y	P_3^y	P_4^y
y	0.000	1.000	0.000	0.000
y + 1	0.200	0.400	0.300	0.100
y + 2	0.190	0.330	0.310	0.170
y + 3	0.173	0.316	0.312	0.199
y + 4	0.163	0.312	0.314	0.211
y + 5	0.159	0.310	0.315	0.216
y + 6	0.157	0.309	0.316	0.218
y + 7	0.156	0.309	0.316	0.219
y + 8	0.156	0.309	0.316	0.219
y + 9	0.156	0.309	0.316	0.219

Table 7.8. Successive streamflow probabilities based on transition probabilities in Figure 7.11.

Let $\boldsymbol{P}$ be the transition matrix whose elements are p_{ij}. For a Markov chain, the transition matrix contains all the information necessary to describe the behaviour of the process. Let p_i^y be the probability that the streamflow Q_y is q_i (in state i) in year y. Then the probability that $Q_{y+1} = q_j$ is the sum of the products of the probabilities p_i^y that $Q_y = q_i$ times the probability p_{ij} that the next state Q_{y+1} is q_j given that $Q_y = q_i$. In symbols, this relationship is written:

$$p_j^{y+1} = p_1^y p_{1j} + p_2^y p_{2j} + \cdots + p_n^y p_{nj} = \sum_{i=1}^{n} p_i^y p_{ij} \qquad (7.118)$$

Letting $\boldsymbol{p}^y$ be the row vector of state resident probabilities $(p_i^y, \ldots, p_n^y)$, this relationship may be written

$$\boldsymbol{p}^{(y+1)} = \boldsymbol{p}^{(y)}\boldsymbol{P} \qquad (7.119)$$

To calculate the probabilities of each streamflow state in year $y + 2$, one can use $\boldsymbol{p}^{(y+1)}$ in Equation 7.119 to obtain

$$\boldsymbol{p}^{(y+2)} = \boldsymbol{p}^{(y+1)}\boldsymbol{P} \text{ or } \boldsymbol{p}^{(y+2)} = \boldsymbol{p}^y\boldsymbol{P}^2$$

Continuing in this manner, it is possible to compute the probabilities of each possible streamflow state for years $y + 1$, $y + 2$, $y + 3$, … , $y + k$, … as

$$\boldsymbol{p}^{(y+k)} = \boldsymbol{p}^y\boldsymbol{P}^k \qquad (7.120)$$

Returning to the four-state example in Figure 7.11, assume that the flow in year y is in the interval represented by q_2. Hence in year y the unconditional streamflow probabilities p_i^y are (0, 1, 0, 0). Knowing each p_i^y, the probabilities p_j^{y+1}corresponding to each of the four streamflow states can be determined. From Figure 7.11, the probabilities p_j^{y+1} are 0.2, 0.4, 0.3 and 0.1 for $j = 1, 2, 3$ and 4, respectively. The probability vectors for nine future years are listed in Table 7.8.

As time progresses, the probabilities generally reach limiting values. These are the *unconditional* or *steady-state* probabilities. The quantity p_i has been defined as the unconditional probability of q_i. These are the steady-state probabilities which $\boldsymbol{p}^{(y+k)}$ approaches for large k. It is clear from Table 7.8 that as k becomes larger, Equation 7.118 becomes

$$p_j = \sum_{i=1}^{n} p_i p_{ij} \qquad (7.121)$$

or in vector notation, Equation 7.119 becomes

$$\boldsymbol{p} = \boldsymbol{pP} \qquad (7.122)$$

where $\boldsymbol{p}$ is the row vector of unconditional probabilities $(p_1, \ldots, p_n)$. For the example in Table 7.8, the probability vector $\boldsymbol{p}$ equals (0.156, 0.309, 0.316, 0.219).

The steady-state probabilities for any Markov chain can be found by solving simultaneous Equation 7.122 for all but one of the states j together with the constraint

$$\sum_{i=1}^{n} p_i = 1 \qquad (7.123)$$

Annual streamflows are seldom as highly correlated as the flows in this example. However, monthly, weekly and especially daily streamflows generally have high serial correlations. Assuming that the unconditional steady-state probability distributions for monthly streamflows are stationary, a Markov chain can be defined for each month's streamflow. Since there are twelve months in a year, there would be twelve transition matrices, the elements of which could be denoted as p_{ij}^t. Each defines the probability of a streamflow q_j^{t+1} in month $t + 1$, given a streamflow q_i^t in month t. The steady-state stationary probability vectors for each month can be found by the procedure outlined above, except that now all twelve matrices are used to calculate all twelve steady-state probability vectors. However, once the steady-state vector $\boldsymbol{p}$ is found for one month, the others are easily computed using Equation 7.120 with t replacing y.

7.3. Properties of Time-Series Statistics

The statistics most frequently used to describe the distribution of a continuous-state stationary stochastic process are the sample mean, variance and various autocorrelations. Statistical dependence among the observations, as is frequently found in time series, can have a marked effect on the distribution of these statistics. This part of Section 7 reviews the sampling properties of these statistics when the observations are a realization of a stochastic process.

The sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7.124}$$

when viewed as a random variable is an unbiased estimate of the mean of the process μ_X, because

$$\mathrm{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] = \mu_X \tag{7.125}$$

However, correlation among the X_i's, so that $\rho_X(k) \neq 0$ for $k > 0$, affects the variance of the estimated mean $\bar{X}$.

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{E}[(\bar{X} - \mu_X)^2]\\ &= \frac{1}{n^2}\mathrm{E}\left\{\sum_{t=1}^{n}\sum_{s=1}^{n}(X_t - \mu_X)(X_s - \mu_X)\right\}\\ &= \frac{\sigma_X^2}{n}\left\{1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_X(k)\right\}\end{aligned} \tag{7.126}$$

The variance of $\bar{X}$, equal to σ_x^2/n for independent observations, is inflated by the factor within the brackets. For $\rho_X(k) \geq 0$, as is often the case, this factor is a nondecreasing function of n, so that the variance of $\bar{X}$ is inflated by a factor whose importance does not decrease with increasing sample size. This is an important observation, because it means that the average of a correlated time series will be less precise than the average of a sequence of independent random variables of the same length with the same variance.

A common model of stochastic series has

$$\rho_X(k) = [\rho_X(1)]^k = \rho^k \tag{7.127}$$

This correlation structure arises from the autoregressive Markov model discussed at length in Section 8. For this correlation structure

sample size *n*	correlation of consecutive observations p = 0.0	p = 0.3	p = 0.6
25	0.050	0.067	0.096
50	0.035	0.048	0.069
100	0.025	0.034	0.050

Table 7.9. Standard error of $\bar{X}$ when $\sigma_x = 0.25$ and $\rho_X(k) = \rho^k$.

$$\mathrm{Var}(\bar{X}) = \frac{\sigma_X^2}{n}\left\{1 + \frac{2\rho}{n}\frac{[n(1-\rho) - (1-\rho^n)]}{(1-\rho)^2}\right\} \tag{7.128}$$

Substitution of the sample estimates for σ_X^2 and $\rho_X(1)$ in the equation above often yields a more realistic estimate of the variance of $\bar{X}$ than does the estimate s_X^2/n if the correlation structure $\rho_X(k) = \rho^k$ is reasonable; otherwise, Equation 7.126 may be employed. Table 7.9 illustrates the effect of correlation among the X_t values on the standard error of their mean, equal to the square root of the variance in Equation 7.126.

The properties of the estimate of the variance of X,

$$\hat{\sigma}_X^2 = v_X^2 = \frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X})^2 \tag{7.129}$$

are also affected by correlation among the X_t's. Here v rather than s is used to denote the variance estimator, because n is employed in the denominator rather than $n - 1$. The expected value of v_x^2 becomes

$$\mathrm{E}[v_X^2] = \sigma_X^2\left\{1 - \frac{1}{n} - \frac{2}{n}\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_X(k)\right\} \tag{7.130}$$

The bias in v_X^2 depends on terms involving $\rho_X(1)$ through $\rho_X(n - 1)$. Fortunately, the bias in v_X^2 decreases with n and is generally unimportant when compared to its variance.

Correlation among the X_t's also affects the variance of v_X^2. Assuming that X has a normal distribution (here the variance of v_X^2 depends on the fourth moment of X), the variance of v_X^2 for large n is approximately (Kendall and Stuart, 1966, Sec. 48.1).

$$\mathrm{Var}(v_X^2) \cong 2\frac{\sigma_X^4}{n}\left\{1 + 2\sum_{k=1}^{\infty}\rho_X^2(k)\right\} \tag{7.131}$$

where for $\rho_X(k) = \rho^k$, Equation 7.131 becomes

$$\mathrm{Var}(v_X^2) \cong 2\frac{\sigma_X^4}{n}\left(\frac{1+\rho^2}{1-\rho^2}\right) \qquad (7.132)$$

Like the variance of $\overline{X}$, the variance of v_X^2 is inflated by a factor whose importance does not decrease with n. This is illustrated by Table 7.10 which gives the standard deviation of v_X^2 divided by the true variance σ_X^2 as a function of n and ρ when the observations have a normal distribution and $\rho_X(k) = \rho^k$. This would be the coefficient of variation of v_X^2 were it not biased.

A fundamental problem of time-series analyses is the estimation or description of the relationship among the random variable values at different times. The statistics used to describe this relationship are the autocorrelations. Several estimates of the autocorrelations have been suggested. A simple and satisfactory estimate recommended by Jenkins and Watts (1968) is:

$$\hat{\rho}_X(k) = r_k = \frac{\sum_{t=1}^{n-k}(x_t-\overline{x})(x_{t+k}-\overline{x})}{\sum_{t=1}^{n}(x_t-\overline{x})^2} \qquad (7.133)$$

Here r_k is the ratio of two sums where the numerator contains $n - k$ terms and the denominator contains n terms. The estimate r_k is biased, but unbiased estimates frequently have larger mean square errors (Jenkins and Watts, 1968). A comparison of the bias and variance of r_1 is provided by the case when the X_t's are independent normal variates. Then (Kendall and Stuart, 1966)

$$E[r_1] = -\frac{1}{n} \qquad (7.134a)$$

and

$$\mathrm{Var}(r_1) = \frac{(n-2)^2}{n^2(n-1)} \cong \frac{1}{n} \qquad (7.134b)$$

For $n = 25$, the expected value of r_1 is -0.04 rather than the true value of zero; its standard deviation is 0.19. This results in a mean square error of $(\mathrm{E}[r_1])^2 + \mathrm{Var}(r_1) = 0.0016 + 0.0353 = 0.0369$. Clearly, the variance of r_1 is the dominant term.

For X_t values that are not independent, exact expressions for the variance of r_k generally are not available.

sample size n	correlation of consecutive observations $\rho = 0.0$	 $\rho = 0.3$	 $\rho = 0.6$
25	0.28	0.31	0.41
50	0.20	0.22	0.29
100	0.14	0.15	0.21

E021101k

Table 7.10. Standard deviation of $[v_X^2/\sigma_X^2]$ when observations have a normal distribution and $\rho_X(k) = \rho^k$.

However, for normally distributed X_t and large n (Kendall and Stuart, 1966),

$$\mathrm{Var}(r_k) \cong \frac{1}{n}\sum_{l=-\infty}^{+\infty}\left[\rho_X^2(l) + \rho_X(l+k)\rho_X(l-k) - 4\rho_X(k)\rho_X(l)\rho_X(k-l) + 2\rho_X^2(k)\rho_X^2(l)\right] \qquad (7.135)$$

If $\rho_X(k)$ is essentially zero for $k > q$, then the simpler expression (Box et al., 1994)

$$\mathrm{Var}(r_k) \cong \frac{1}{n}\left[1 + 2\sum_{t=1}^{q}\rho_X^2(l)\right] \qquad (7.136)$$

is valid for r_k corresponding to $k > q$. Thus for large n, $\mathrm{Var}(r_k) \geq l/n$ and values of r_k will frequently be outside the range of $\pm 1.65/\sqrt{n}$, even though $\rho_X(k)$ may be zero.

If $\rho_X(k) = \rho^k$, Equation 7.136 reduces to

$$\mathrm{Var}(r_k) \cong \frac{1}{n}\left[\frac{(l+\rho^2)(l-\rho^{2k})}{l-\rho^2} - 2k\rho^{2k}\right] \qquad (7.137)$$

In particular for r_1, this gives

$$\mathrm{Var}(r_1) \cong \frac{1}{n}(1-\rho^2) \qquad (7.138)$$

Approximate values of the standard deviation of r_1 for different values of n and ρ are given in Table 7.11.

The estimates of r_k and r_{k+j} are highly correlated for small j; this causes plots of r_k versus k to exhibit slowly varying cycles when the true values of $\rho_X(k)$ may be zero. This increases the difficulty of interpreting the sample autocorrelations.

sample size n	correlation of consecutive observations		
	$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.6$
25	0.20	0.19	0.16
50	0.14	0.13	0.11
100	0.10	0.095	0.080

Table 7.11. Approximate standard deviation of r_1 when observations have a normal distribution and $\rho_X(k) = \rho^k$.

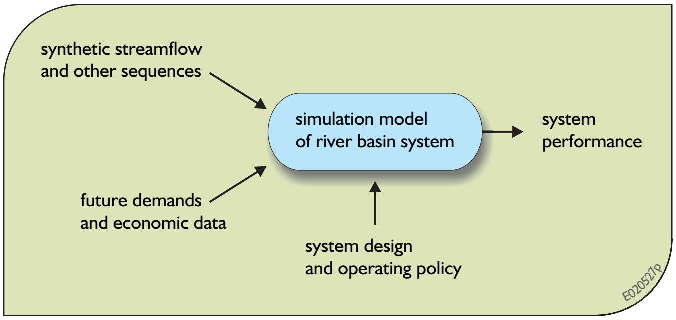
synthetic streamflow and other sequences
simulation model of river basin system
system performance
future demands and economic data
system design and operating policy
E020527p

Figure 7.12. Structure of a simulation study, indicating the transformation of a synthetic streamflow sequence, future demands and a system design and operating policy into system performance statistics.

8. Synthetic Streamflow Generation

8.1. Introduction

This section is concerned primarily with ways of generating sample data such as streamflows, temperatures and rainfall that are used in water resource systems simulation studies (e.g., as introduced in the next section). The models and techniques discussed in this section can be used to generate any number of quantities used as inputs to simulation studies. For example Wilks (1998, 2002) discusses the generation of wet and dry days, rainfall depths on wet days and associated daily temperatures. The discussion here is directed toward the generation of streamflows because of the historical development and frequent use of these models in that context (Matalas and Wallis, 1976). In addition, they are relatively simple compared to more complete daily weather generators and many other applications. Generated streamflows have been called *synthetic* to distinguish them from historical observations (Fiering, 1967). The activity has been called stochastic hydrological modelling. More detailed presentations can be found in Marco et al. (1989) and Salas (1993).

River basin simulation studies can use many sets of streamflow, rainfall, evaporation and/or temperature sequences to evaluate the statistical properties of the performance of alternative water resources systems. For this purpose, synthetic flows and other generated quantities should resemble, statistically, those sequences that are likely to be experienced during the planning period. Figure 7.12 illustrates how synthetic streamflow, rainfall and other stochastic sequences are used in conjunction with projections of future demands and other economic data to determine how different system designs and operating policies might perform.

Use of only the historical flow or rainfall record in water resource studies does not allow for the testing of alternative designs and policies against the range of sequences that are likely to occur in the future. We can be very confident that the future historical sequence of flows will not be the historical one, yet there is important information in that historical record. That information is not fully used if only the historical sequence is simulated. Fitting continuous distributions to the set of historical flows and then using those distributions to generate other sequences of flows, all of which are statistically similar and equally weighted, gives one a broader range of inputs to simulation models. Testing designs and policies against that broader range of flow sequences that could occur more clearly identifies the variability and range of possible future performance indicator values. This in turn should lead to the selection of more robust system designs and policies.

The use of synthetic streamflows is particularly useful for water resources systems having large amounts of over-year storage. Use of only the historical hydrological record in system simulation yields only one time history of how the system would operate from year to year. In water resources systems with relatively little storage, so that reservoirs and/or groundwater aquifers refill almost every year, synthetic hydrological sequences may not be needed if historical sequences of a reasonable length are available. In this second case, a twenty-five-year historical record provides twenty-five descriptions of the possible within-year operation of the system. This may be sufficient for many studies.

Generally, use of stochastic sequences is thought to improve the precision with which water resources system performance indices can be estimated, and some studies have shown this to be the case (Vogel and Shallcross, 1996; Vogel and Stedinger, 1988). In particular, if system operation performance indices have thresholds and sharp breaks, then the coarse descriptions provided by historical series are likely to provide relative inaccurate estimates of the expected values of such statistics. For example, suppose that shortages only invoke non-linear penalties on average one year in twenty. Then in a sixty-year simulation there is a 19% probability that the penalty will be invoked at most once, and an 18% probability it will be invoked five or more times. Thus the calculation of the annual average value of the penalty would be highly unreliable unless some smoothing of the input distributions is allowed, associated with a long simulation analysis.

On the other hand, if one is only interested in the mean flow, or average benefits that are mostly a linear function of flows, then use of stochastic sequences will probably add little information to what is obtained simply by simulating the historical record. After all, the fitted models are ultimately based on the information provided in the historical record, and their use does not produce new information about the hydrology of the basin.

If in a general sense one has available *n* years of record, the statistics of that record can be used to build a stochastic model for generating thousands of years of flow. These synthetic data can then be used to estimate more exactly the system performance, assuming, of course, that the flow-generating model accurately represents nature. But the initial uncertainty in the model parameters resulting from having only *n* years of record would still remain (Schaake and Vicens, 1980).

An alternative is to run the historical record (if it is sufficiently complete at every site and contains no gaps of missing data) through the simulation model to generate *n* years of output. That output series can be processed to produce estimates of system performance. So the question is the following: Is it better to generate multiple input series based on uncertain parameter values and use those to determine average system performance with great precision, or is it sufficient just to model the *n*-year output series that results from simulation of the historical series?

The answer seems to depend upon how well behaved the input and output series are. If the simulation model is linear, it does not make much difference. If the simulation model were highly non-linear, then modelling the input series would appear to be advisable. Or if one is developing reservoir operating policies, there is a tendency to make a policy sufficiently complex to deal very well with the few droughts in the historical record, giving a false sense of security and likely misrepresenting the probability of system performance failures.

Another situation where stochastic data generating models are useful is when one wants to understand the impact on system performance estimates of the parameter uncertainty stemming from short historical records. In that case, parameter uncertainty can be incorporated into streamflow generating models so that the generated sequences reflect both the variability that one would expect in flows over time as well as the uncertainty of the parameter values of the models that describe that variability (Valdes et al., 1977; Stedinger and Taylor, 1982a,b; Stedinger et al., 1985; Vogel and Stedinger, 1988).

If one decides to use a stochastic data generator, the challenge is to use a model that appropriately describes the important relationships, but does not attempt to reproduce more relationships than are justified or can be estimated with available data sets.

Two basic techniques are used for streamflow generation. If the streamflow population can be described by a stationary stochastic process (a process whose parameters do not change over time), and if a long historical streamflow record exists, then a stationary stochastic streamflow model may be fit to the historical flows. This statistical model can then generate synthetic sequences that describe selected characteristics of the historical flows. Several such models are discussed below.

The assumption of stationarity is not always plausible, particularly in river basins that have experienced marked changes in runoff characteristics due to changes in land cover, land use, climate or the use of groundwater during the period of flow record. Similarly, if the physical characteristics of a basin change substantially in the future, the historical streamflow record may not provide reliable estimates of the distribution of future unregulated flows. In the absence of the stationarity of streamflows or a representative historical record, an alternative scheme is to assume that precipitation is a stationary stochastic process and to route either historical or synthetic precipitation sequences through an appropriate rainfall–runoff model of the river basin.

8.2. Streamflow Generation Models

The first step in the construction of a statistical streamflow generating model is to extract from the historical streamflow record the fundamental information about the joint distribution of flows at different sites and at different times. A streamflow model should ideally capture what is judged to be the fundamental characteristics of the joint distribution of the flows. The specification of what characteristics are fundamental is of primary importance.

One may want to model as closely as possible the true marginal distribution of seasonal flows and/or the marginal distribution of annual flows. These describe both how much water may be available at different times and also how variable is that water supply. Also, modelling the joint distribution of flows at a single site in different months, seasons and years may be appropriate. The persistence of high flows and of low flows, often described by their correlation, affects the reliability with which a reservoir of a given size can provide a given yield (Fiering, 1967; Lettenmaier and Burges, 1977a, 1977b; Thyer and Kuczera, 2000). For multi-component reservoir systems, reproduction of the joint distribution of flows at different sites and at different times will also be important.

Sometimes, a streamflow model is said to resemble statistically the historical flows if it produces flows with the same mean, variance, skew coefficient, autocorrelations and/or cross-correlations as were observed in the historical series. This definition of statistical resemblance is attractive because it is operational and only requires an analyst to only find a model that can reproduce the observed statistics. The drawback of this approach is that it shifts the modelling emphasis away from trying to find a good model of marginal distributions of the observed flows and their joint distribution over time and over space, given the available data, to just reproducing arbitrarily selected statistics. Defining statistical resemblance in terms of moments may also be faulted for specifying that the parameters of the fitted model should be determined using the observed sample moments, or their unbiased counterparts. Other parameter estimation techniques, such as maximum likelihood estimators, are often more efficient.

Definition of resemblance in terms of moments can also lead to confusion over whether the population parameters should equal the sample moments, or whether the fitted model should generate flow sequences whose sample moments equal the historical values. The two concepts are different because of the biases (as discussed in Section 7) in many of the estimators of variances and correlations (Matalas and Wallis, 1976; Stedinger, 1980, 1981; Stedinger and Taylor, 1982a).

For any particular river basin study, one must determine what streamflow characteristics need to be modelled. The decision should depend on what characteristics are important to the operation of the system being studied, the available data, and how much time can be spared to build and test a stochastic model. If time permits, it is good practice to see if the simulation results are in fact sensitive to the generation model and its parameter values by using an alternative model and set of parameter values. If the model's results are sensitive to changes, then, as always, one must exercise judgement in selecting the appropriate model and parameter values to use.

This section presents a range of statistical models for the generation of synthetic data. The necessary sophistication of a data-generating model depends on the intended use of the generated data. Section 8.3 below presents the simple autoregressive Markov model for generating annual flow sequences. This model alone is too simple for many practical studies, but is useful for illustrating the fundamentals of the more complex models that follow. It seems, therefore, worth some time exploring the properties of this basic model.

Subsequent sections discuss how flows with any marginal distribution can be produced, and present models for generating sequences of flows that can reproduce the persistence of historical flow sequences. Other parts of this section present models for generating concurrent flows at several sites and for generating seasonal or monthly flows that preserve the characteristics of annual flows. More detailed discussions for those wishing to study synthetic streamflow models in greater depth can be found in Marco et al. (1989) and Salas (1993).

8.3. A Simple Autoregressive Model

A simple model of annual streamflows is the autoregressive Markov model. The historical annual flows q_y are thought of as particular values of a stationary stochastic process Q_y. The generation of annual streamflows and other variables would be a simple matter if annual flows were independently distributed. In general, this is not the case and a generating model for many phenomena should capture the relationship between values in different years or in other time periods. A common and reasonable assumption is that annual flows are the result of a first-order Markov process (as discussed in Section 7.2).

Assume for now that annual streamflows are normally distributed. In some areas the distribution of annual flows is in fact nearly normal. Streamflow models that produce non-normal streamflows are discussed as an extension of this simple model.

The joint normal density function of two streamflows Q_y and Q_w in years y and w having mean μ, variance σ^2, and year-to-year correlation ρ between flows is

$$f(q_y, q_w) = \frac{1}{2\pi\sigma^2(1-\rho^2)^{0.5}}$$
$$\exp\left[\frac{(q_y-\mu)^2 - 2\rho(q_y-\mu)(q_w-\mu)+(q_w-\mu)^2}{2\sigma^2(1-\rho^2)}\right] \tag{7.139}$$

The joint normal distribution for two random variables with the same mean and variance depend only on their common mean μ, variance σ^2, and the correlation ρ between the two (or equivalently the covariance $\rho\sigma^2$).

The sequential generation of synthetic streamflows requires the conditional distribution of the flow in one year given the value of the flows in previous years. However, if the streamflows are a first-order (lag 1) Markov process, then the distribution of the flow in year $y + 1$ depends entirely on the value of the flow in year y. In addition, if the annual streamflows have a multivariate normal distribution, then the conditional distribution of Q_{y+1} is normal with mean and variance

$$E[Q_{y+1} \mid Q_y = q_y] = \mu + \rho(q_y - \mu) \tag{7.140}$$
$$\text{Var}(Q_{y+1} \mid Q_y = q_y) = \sigma^2(1 - \rho^2)$$

where q_y is the value of the random variable Q_y in year y. This relationship is illustrated in Figure 7.13. Notice that the larger the absolute value of the correlation ρ between the flows, the smaller the conditional variance of Q_{y+1}, which in this case does not depend at all on the value q_y.

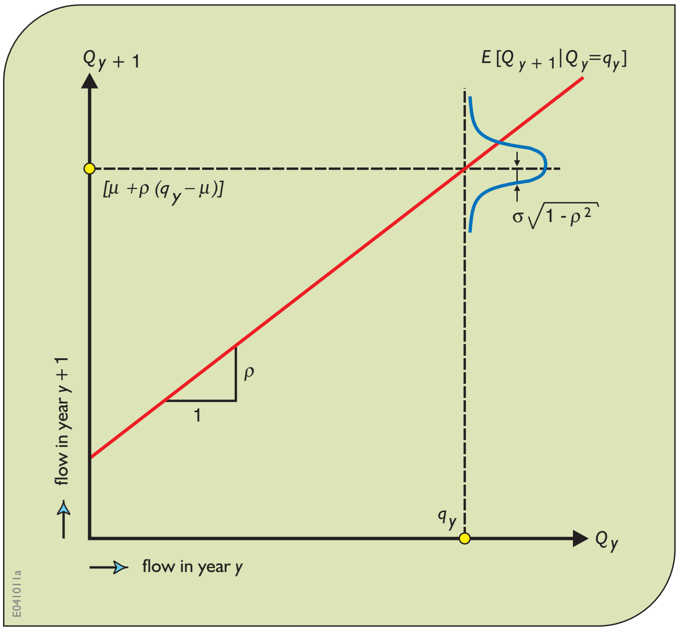

Figure 7.13. Conditional distribution of Q_{y+1} given $Q_y = q_y$ for two normal random variables.

Synthetic normally distributed streamflows that have mean μ, variance σ^2, and year-to-year correlation ρ, are produced by the model

$$Q_{y+1} = \mu + \rho(Q_y - \mu) + V_y\sigma\sqrt{1-\rho^2} \tag{7.141}$$

where V_y is a standard normal random variable, meaning that it has zero mean, $E[V_y] = 0$, and unit variance, $E\left[V_y^2\right] = 1$. The random variable V_y is added here to provide the variability in Q_{y+1} that remains even after Q_y is known. By construction, each V_y is independent of past flows Q_w where $w \le y$, and V_y is independent of V_w for $w \ne y$. These restrictions imply that

$$E[V_wV_y] = 0 \quad \text{for} \quad w \ne y \tag{7.142}$$

and

$$E[(Q_w - \mu)V_y] = 0 \quad \text{for} \quad w \le y \tag{7.143}$$

Clearly, Q_{y+1} will be normally distributed if both Q_y and V_y are normally distributed because sums of independent normally distributed random variables are normally distributed.

It is a straightforward procedure to show that this basic model indeed produces streamflows with the specified moments, as demonstrated below.

Using the fact that $E[V_y] = 0$, the conditional mean of Q_{y+1} given that Q_y equals q_y is

$$\begin{aligned} E[Q_{y+1} \mid q_y] &= E[\mu + \rho(q_y - \mu) + V_y \sigma\sqrt{1-\rho^2}] \\ &= \mu + \rho(q_y - \mu) \end{aligned} \qquad (7.144)$$

Since $E[V_y^2] = \text{Var}[V_y] = 1$, the conditional variance of Q_{y+1} is

$$\begin{aligned} \text{Var}[Q_{y+1} \mid q_y] &= E[\{Q_{y+1} - E[Q_{y+1} \mid q_y]\}^2 | q_y] \\ &= E[\{\mu + \rho(q_y - \mu) \\ &\quad + V_y\sigma\sqrt{1-\rho^2} - [\mu + \rho(q_y - \mu)]\}^2 \\ &= E[V_y\sigma\sqrt{1-\rho^2}\,]^2 = \sigma^2(1-\rho^2) \end{aligned} \qquad (7.145)$$

Thus, this model produces flows with the correct conditional mean and variance.

To compute the unconditional mean of Q_{y+1} one first takes the expectation of both sides of Equation 7.141 to obtain

$$E[Q_{y+1}] = \mu + \rho(E[Q_y] - \mu) + E[V_y]\sigma\sqrt{1-\rho^2} \qquad (7.146)$$

where $E[V_y] = 0$. If the distribution of streamflows is independent of time so that for all y, $E[Q_{y+1}] = E[Q_y] = E[Q]$, it is clear that $(1 - \rho)\, E[Q] = (1 - \rho)\, \mu$ or

$$E[Q] = \mu \qquad (7.147)$$

Alternatively, if Q_y for $y = 1$ has mean μ, then Equation 7.146 indicates that Q_2 will have mean μ. Thus repeated application of the Equation 7.146 would demonstrate that all Q_y for $y > 1$ have mean μ.

The unconditional variance of the annual flows can be derived by squaring both sides of Equation 7.141 to obtain

$$\begin{aligned} E[(Q_{y+1} - \mu)^2] &= E[\{\rho(Q_y - \mu) + V_y\sigma\sqrt{1-\rho^2}\}^2] \\ &= \rho^2 E[(Q_y - \mu)^2] + 2\rho\sigma\sqrt{1-\rho^2} \\ &\quad \times E[(Q_y - \mu)V_y] + \sigma^2(1-\rho^2)E\left[V_y^2\right] \end{aligned} \qquad (7.148)$$

Because V_y is independent of Q_y (Equation 7.143), the second term on the right-hand side of Equation 7.148 vanishes. Hence the unconditional variance of Q satisfies

$$E[(Q_{y+1} - \mu)^2] = \rho^2 E[(Q_y - \mu)^2] + \sigma^2(1 - \rho^2) \qquad (7.149)$$

Assuming that Q_{y+1} and Q_y have the same variance yields

$$E[(Q - \mu)^2] = \sigma^2 \qquad (7.150)$$

so that the unconditional variance is σ^2, as required.

Again, if one does not want to assume that Q_{y+1} and Q_y have the same variance, a recursive argument can be adopted to demonstrate that if Q_1 has variance σ^2, then Q_y for $y \geq 1$ has variance σ^2.

The covariance of consecutive flows is another important issue. After all, the whole idea of building these time-series models is to describe the year-to-year correlation of the flows. Using Equation 7.141 one can show that the covariance of consecutive flows must be

$$\begin{aligned} E[(Q_{y+1} - \mu)(Q_y - \mu)] &= E\{[\rho(Q_y - \mu) \\ &\quad + V_y\sigma\sqrt{1-\rho^2}\,](Q_y - \mu)\} \\ &= \rho E[(Q_y - \mu)^2] = \rho\sigma^2 \end{aligned} \qquad (7.151)$$

where $E[(Q_y - \mu)V_y] = 0$ because V_y and Q_y are independent (Equation 7.143).

Over a longer time scale, another property of this model is that the covariance of flows in year y and $y + k$ is

$$E[(Q_{y+k} - \mu)(Q_y - \mu)] = \rho^k\sigma^2 \qquad (7.152)$$

This equality can be proven by induction. It has already been shown for $k = 0$ and 1. If it is true for $k = j - 1$, then

$$\begin{aligned} E[(Q_{y+j} - \mu)(Q_y - \mu)] &= E\{[\rho(Q_{y+j-1} - \mu) \\ &\quad + V_{y+j-1}\sigma\sqrt{1-\rho^2}\,](Q_y - \mu)\} \\ &= \rho E[(Q_y - \mu)](Q_{y+j-1} - \mu)] \\ &= \rho[\rho^{j-1}\sigma^2] = \rho^j\sigma^2 \end{aligned} \qquad (7.153)$$

where $E[(Q_y - \mu)V_{y+j-1}] = 0$ for $j \geq 1$. Hence Equation 7.152 is true for any value of k.

It is important to note that the results in Equations 7.144 to 7.152 do not depend on the assumption that the random variables Q_y and V_y are normally distributed. These relationships apply to all autoregressive Markov processes of the form in Equation 7.141 regardless of the distributions of Q_y and V_y. However, if the flow Q_y in year $y = 1$ is normally distributed with mean μ and variance σ^2, and if the V_y are independent normally distributed random variables with mean zero and unit variance, then the generated Q_y for $y \geq 1$ will also be normally distributed with mean μ and variance σ^2. The next section considers how this and other models can be used to generate streamflows that have other than a normal distribution.

8.4. Reproducing the Marginal Distribution

Most models for generating stochastic processes deal directly with normally distributed random variables. Unfortunately, flows are not always adequately described by the normal distribution. In fact, streamflows and many other hydrological data cannot really be normally distributed because of the impossibility of negative values. In general, distributions of hydrological data are positively skewed, having a lower bound near zero and, for practical purposes, an unbounded right-hand tail. Thus they look like the gamma or lognormal distribution illustrated in Figures 7.3 and 7.4.

The asymmetry of a distribution is often measured by its coefficient of skewness. In some streamflow models, the skew of the random elements V_y is adjusted so that the models generate flows with the desired mean, variance and skew coefficient. For the autoregressive Markov model for annual flows

$$\begin{aligned} E[(Q_{y+1} - \mu)^3] &= E\left[\rho(Q_y - \mu) + V_y\sigma\sqrt{1-\rho^2}\,\right]^3 \\ &= \rho^3 E[(Q_y - \mu)^3] \\ &\quad + \sigma^3(1-\rho^2)^{3/2}\, E\left[V_y^3\right] \end{aligned} \tag{7.154}$$

so that

$$\gamma_Q = \frac{E[(Q-\mu)^3]}{\sigma^3} = \frac{(1-\rho^2)^{3/2}}{1-\rho^3}\, E\left[V_y^3\right] \tag{7.155}$$

By appropriate choice of the skew of V_y, $E[V_y^3]$, the desired skew coefficient of the annual flows can be produced. This method has often been used to generate flows that have approximately a gamma distribution by using V_y's with a gamma distribution and the required skew. The resulting approximation is not always adequate (Lettenmaier and Burges, 1977a).

The alternative and generally preferred method is to generate normal random variables and then transform these variates to streamflows with the desired marginal distribution. Common choices for the distribution of streamflows are the two-parameter and three-parameter lognormal distributions or a gamma distribution. If Q_y is a lognormally distributed random variable, then

$$Q_y = \tau + \exp(X_y) \tag{7.156}$$

where X_y is a normal random variable. When the lower bound τ is zero, Q_y has a two-parameter lognormal distribution. Equation 7.156 transforms the normal variates X_y into lognormally distributed streamflows. The transformation is easily inverted to obtain

$$X_y = \ln(Q_y - \tau) \quad \text{for} \quad Q_y > \tau \tag{7.157}$$

where Q_y must be greater than its lower bound τ.

The mean, variance, skewness of X_y and Q_y are related by the formulas (Matalas, 1967)

$$\begin{aligned} \mu_Q &= \tau + \exp\left(\mu_X + \frac{1}{2}\sigma_X^2\right) \\ \sigma_Q^2 &= \exp\left(2\mu_X + \sigma_X^2\right)\left[\exp\left(\sigma_X^2\right) - 1\right] \\ \gamma_Q &= \frac{\exp\left(3\sigma_X^2\right) - 3\exp\left(\sigma_X^2\right) + 2}{\left[\exp\left(\sigma_X^2\right) - 1\right]^{3/2}} \\ &= 3\phi + \phi^3 \quad \text{where } \phi = \left[\exp\left(\sigma_X^2\right) - 1\right]^{1/2} \end{aligned} \tag{7.158}$$

If normal variates X_s^y and X_y^u are used to generate lognormally distributed streamflows Q_s^y and Q_y^u at sites s and u, then the lag-k correlation of the Q_y's, denoted $\rho_Q(k;\, s,\, u)$, is determined by the lag-k correlation of the X variables, denoted $\rho_X(k;\, s,\, u)$, and their variances $\sigma_x^2(s)$ and $\sigma_x^2(u)$, where

$$\rho_Q(k;s,u) = \frac{\exp[\rho_X(k;\, s,\, u)\sigma_X(s)\sigma_X(u)] - 1}{\left\{\exp\left[\sigma_X^2(s)\right] - 1\right\}^{1/2}\ \left\{\exp\left[\sigma_X^2(u)\right] - 1\right\}^{1/2}} \tag{7.159}$$

The correlations of the X_y^s can be adjusted, at least in theory, to produce the observed correlations among the Q_y^s variates. However, more efficient estimates of the true correlation of the Q_y^s values are generally obtained by transforming the historical flows q_y^s into their normal equivalent $x_y^s = \ln(q_y^s - \tau)$ and using the historical correlations of these X_y^s values as estimators of $\rho_X(k;\, s,\, u)$ (Stedinger, 1981).

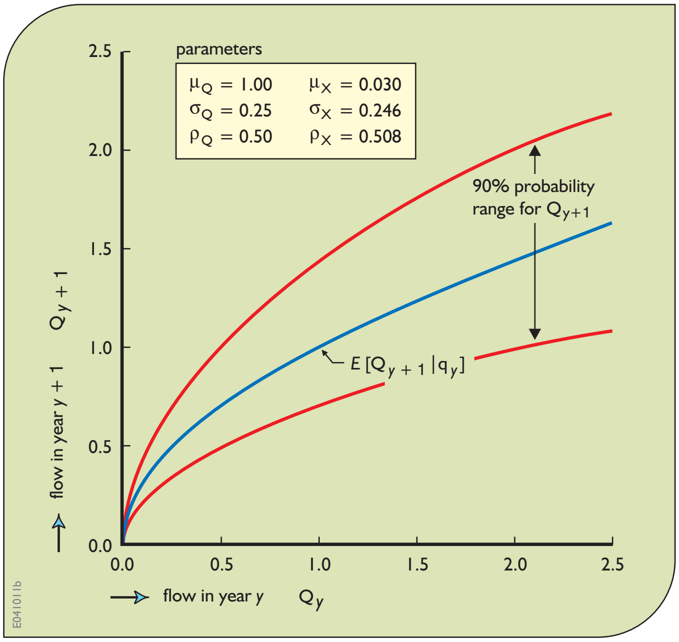

Figure 7.14. Conditional mean of Q_{y+1} given $Q_y = q_y$ and 90% probability range for the value of Q_{y+1}.

Some insight into the effect of this logarithmic transformation can be gained by considering the resulting model for annual flows at a single site. If the normal variates follow the simple autoregressive Markov model

$$X_{y+1} - \mu = \rho_X(X_y - \mu) + V_y\sigma_X\sqrt{1-\rho_X^2} \qquad (7.160)$$

then the corresponding Q_y follow the model (Matalas, 1967)

$$Q_{y+1} = \tau + D_y\{\exp[\mu_x(1 - \rho_x)]\}(Q_y - \tau)^{\rho x} \qquad (7.161)$$

where

$$D_y = \exp[(1 - \rho_X^2)^{1/2}\,\sigma_X V_y] \qquad (7.162)$$

The conditional mean and standard deviation of Q_{y+1} given that $Q_y = q_y$ now depends on $(q_y - \tau)^{\rho x}$. Because the conditional mean of Q_{y+1} is no longer a linear function of q_y, (as shown in Figure 7.14), the streamflows are said to exhibit differential persistence: low flows are now more likely to follow low flows than high flows are to follow high flows. This is a property often attributed to real streamflow distributions. Models can be constructed to capture the relative persistence of wet and dry periods (Matalas and Wallis, 1976; Salas, 1993; Thyer and Kuczera, 2000). Many weather generators for precipitation and temperature include such tendencies by employing a Markov chain description of the occurrence of wet and dry days (Wilks, 1998).

8.5. Multivariate Models

If long concurrent streamflow records can be constructed at the several sites at which synthetic streamflows are desired, then ideally a general multi-site streamflow model could be employed. O'Connell (1977), Ledolter (1978), Salas et al. (1980) and Salas (1993) discuss multivariate models and parameter estimation. Unfortunately, identification of the most appropriate model structure is very difficult for general multivariate models.

This section illustrates how the basic univariate annual-flow model in Section 8.3 can be generalized to the multivariate case. This exercise reveals how multivariate models can be constructed to reproduce specified variances and covariances of the flow vectors of interest, or some transformation of those values. This multi-site generalization of the annual AR(1) or autoregressive Markov model follows the approach taken by Matalas and Wallis (1976). This general approach can be further extended to generate multi-site/multi-season modelling procedures, as is done in Section 8.6, employing what have been called disaggregation models. However, while the size of the model matrices and vectors increases, the model is fundamentally the same from a mathematical viewpoint. Hence this section starts with the simpler case of an annual flow model.

For simplicity of presentation and clarity, vector notation is employed. Let $\mathbf{Z}_y = \left(Z_y^1, \ldots, Z_y^n\right)^T$ be the column vector of transformed zero-mean annual flows at sites $s = 1, 2, \ldots, n$, so that

$$E[Z_y^s] = 0 \qquad (7.163)$$

In addition, let $\mathbf{V}_y = \left(V_y^1, \ldots, V_y^n\right)^T$ be a column vector of standard-normal random variables, where V_y^s is independent of V_w^r for $(r, w) \neq (s, y)$ and independent of past flows Z_w^r where $y \geq w$. The assumption that the variables have zero mean implicitly suggests that the mean value has already been subtracted from all the variables. This makes the notation simpler and eliminates the need to include a constant term in the models. With all the random variables having zero mean, one can focus on reproducing the variances and covariances of the vectors included in a model.

A sequence of synthetic flows can be generated by the model

$$\mathbf{Z}_{y+1} = \mathrm{A}\mathbf{Z}_y + \mathrm{B}\mathbf{V}_y \qquad (7.164)$$

where A and B are $(n \times n)$ matrices whose elements are chosen to reproduce the lag 0 and lag 1 cross-covariances of the flows at each site. The lag 0 and lag 1 covariances and cross-covariances can most economically be manipulated by use of the two matrices S_0 and S_1. The lag-zero covariance matrix, denoted S_0, is defined as

$$\mathrm{S}_0 = \mathrm{E}[\mathbf{Z}_y\mathbf{Z}_y^{\mathrm{T}}] \qquad (7.165)$$

and has elements

$$\mathrm{S}_0(i, j) = \mathrm{E}[\mathbf{Z}_y^i\mathbf{Z}_y^j] \qquad (7.166)$$

The lag-one covariance matrix, denoted S_1, is defined as

$$\mathrm{S}_1 = \mathrm{E}[\mathbf{Z}_{y+1}\,\mathbf{Z}_y^{\mathrm{T}}] \qquad (7.167)$$

and has elements

$$\mathrm{S}_1(i, j) = \mathrm{E}[\mathbf{Z}_{y+1}^i\,\mathbf{Z}_y^j] \qquad (7.168)$$

The covariances do not depend on y because the streamflows are assumed to be stationary.

Matrix S_1 contains the lag 1 covariances and lag 1 cross-covariances. S_0 is symmetric because the cross-covariance $\mathrm{S}_0(i, j)$ equals $\mathrm{S}_0(j, i)$. In general, S_1 is not symmetric.

The variance–covariance equations that define the values of A and B in terms of S_0 and S_1 are obtained by manipulations of Equation 7.164. Multiplying both sides of that equation by $\mathbf{Z}_y^{\mathrm{T}}$ and taking expectations yields

$$\mathrm{E}[\mathbf{Z}_{y+1}\mathbf{Z}_y^{\mathrm{T}}] = \mathrm{E}[\mathrm{A}\mathbf{Z}_y\mathbf{Z}_y^{\mathrm{T}}] + \mathrm{E}[\mathrm{B}\mathbf{V}_y\mathbf{Z}_y^{\mathrm{T}}] \qquad (7.169)$$

The second term on the right-hand side vanishes because the components of $\mathbf{Z}_y$ and $\mathbf{V}_y$ are independent. Now the first term in Equation 7.169, $\mathrm{E}[\mathrm{A}\mathbf{Z}_y\,\mathbf{Z}_y^{\mathrm{T}}]$, is a matrix whose (i, j)th element equals

$$\mathrm{E}\left[\sum_{k=1}^{n} a_{ik}\mathbf{Z}_y^k\mathbf{Z}_y^j\right] = \sum_{k=1}^{n} a_{ik}\mathrm{E}[\mathbf{Z}_y^k\mathbf{Z}_y^j] \qquad (7.170)$$

The matrix with these elements is the same as the matrix $\mathrm{AE}\left[\mathbf{Z}_y\,\mathbf{Z}_y^{\mathrm{T}}\right]$.

Hence, A – the matrix of constants – can be pulled through the expectation operator just as is done in the scalar case where $\mathrm{E}[a\mathbf{Z}_y + b] = a\mathrm{E}[\mathbf{Z}_y] + b$ for fixed constants a and b.

Substituting S_0 and S_1 for the appropriate expectations in Equation 7.169 yields

$$\mathrm{S}_1 = \mathrm{AS}_0 \text{ or } \mathrm{A} = \mathrm{S}_1\mathrm{S}_0^{-1} \qquad (7.171)$$

A relationship to determine the matrix B is obtained by multiplying both sides of Equation 7.164 by its own transpose (this is equivalent to squaring both sides of the scalar equation $a = b$) and taking expectations to obtain

$$\mathrm{E}[\mathbf{Z}_{y+1}\,\mathbf{Z}_{y+1}^{\mathrm{T}}] = \mathrm{E}[\mathrm{A}\mathbf{Z}_y\mathbf{Z}_y^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}] + \mathrm{E}[\mathrm{A}\mathbf{Z}_y\mathbf{V}_y^{\mathrm{T}}\mathrm{B}^{\mathrm{T}}] + \mathrm{E}[\mathrm{B}\mathbf{V}_y\mathbf{Z}_y\mathrm{A}^{\mathrm{T}}] + \mathrm{E}[\mathrm{B}\mathbf{V}_y\mathbf{V}_y^{\mathrm{T}}\mathrm{B}^{\mathrm{T}}] \qquad (7.172)$$

The second and third terms on the right-hand side of Equation 7.172 vanish because the components of $\mathbf{Z}_y$ and $\mathbf{V}_y$ are independent and have zero mean. $\mathrm{E}[\mathbf{V}_y\mathbf{V}_y^{\mathrm{T}}]$ equals the identity matrix because the components of $\mathbf{V}_y$ are independently distributed with unit variance. Thus

$$\mathrm{S}_0 = \mathrm{AS}_0\mathrm{A}^{\mathrm{T}} + \mathrm{BB}^{\mathrm{T}} \qquad (7.173)$$

Solving for the B matrix, one finds that it should satisfy the quadratic equation

$$\mathrm{BB}^{\mathrm{T}} = \mathrm{S}_0 - \mathrm{AS}_0\mathrm{A}^{\mathrm{T}} = \mathrm{S}_0 - \mathrm{S}_1\,\mathrm{S}_0^{-1}\mathrm{S}_1^{\mathrm{T}} \qquad (7.174)$$

The last equation results from substitution of the relationship for A given in Equation 7.171 and the fact that S_0 is symmetric; hence, S_0^{-1} is symmetric.

It should not be too surprising that the elements of B are not uniquely determined by Equation 7.174. The components of the random vector $\mathbf{V}_y$ may be combined in many ways to produce the desired covariances as long as B satisfies Equation 7.174. A lower triangular matrix that satisfies Equation 7.174 can be calculated by Cholesky decomposition (Young, 1968; Press et al., 1986).

Matalas and Wallis (1976) call Equation 7.164 the *lag-1 model*. They do not call it the Markov model because the streamflows at individual sites do not have the covariances of an autoregressive Markov process given in Equation 7.152. They suggest an alternative model for what they call the *Markov model*. It has the same structure as the lag-1 model except it does not preserve the lag-1 cross-covariances. By relaxing this requirement, they obtain a simpler model with fewer parameters that generates flows that have covariances of an autoregressive Markov process at each site. In their Markov model, the new A matrix is simply a diagonal matrix, whose diagonal elements are the lag-1 correlations of flows at each site:

$$A = \mathrm{diag}[\rho(1; i, i)] \qquad (7.175)$$

where $\rho(1; i, i)$ is the lag-one correlation of flows at site i.

The corresponding B matrix depends on the new A matrix and S_0, where as before

$$BB^T = S_0 - AS_0A^T \qquad (7.176)$$

The idea of fitting time-series models to each site separately and then correlating the innovations in those separate models to reproduce the cross-correlation between the series is a very general and useful modelling idea that has seen a number of applications with different time-series models (Matalas and Wallis, 1976; Stedinger et al., 1985; Camacho et al., 1985; Salas, 1993).

8.6. Multi-Season, Multi-Site Models

In most studies of surface water systems it is necessary to consider the variations of flows within each year. Streamflows in most areas have within-year variations, exhibiting wet and dry periods. Similarly, water demands for irrigation, municipal and industrial uses also vary, and the variations in demand are generally out of phase with the variation in within-year flows; more water is usually desired when streamflows are low, and less is desired when flows are high. This increases the stress on water delivery systems and makes it all the more important that time-series models of streamflows, precipitation and other hydrological variables correctly reproduce the seasonality of hydrological processes.

This section discusses two approaches to generating within-year flows. The first approach is based on the disaggregation of annual flows produced by an annual flow generator to seasonal flows. Thus the method allows for reproduction of both the annual and seasonal characteristics of streamflow series. The second approach generates seasonal flows in a sequential manner, as was done for the generation of annual flows. Thus the models are a direct generalization of the annual flow models already discussed.

8.6.1. Disaggregation Models

The disaggregation model proposed by Valencia and Schaake (1973) and extended by Mejia and Rousselle (1976) and Tao and Delleur (1976) allows for the generation of synthetic flows that reproduce statistics both at the annual and at the seasonal level. Subsequent improvements and variations are described by Stedinger and Vogel (1984), Maheepala and Perera (1996), Koutsoyiannis and Manetas (1996) and Tarboton et al. (1998).

Disaggregation models can be used for either multi-season single-site or multi-site streamflow generation. They represent a very flexible modelling framework for dealing with different time or spatial scales. Annual flows for the several sites in question or the aggregate total annual flow at several sites can be the input to the model (Grygier and Stedinger, 1988). These must be generated by another model, such as those discussed in the previous sections. These annual flows or aggregated annual flows are then disaggregated to seasonal values.

Let $\mathbf{Z}_y = (Z_y^1, \ldots, Z_y^N)^T$ be the column vector of N transformed normally distributed annual or aggregate annual flows for N separate sites or basins. Next, let $\mathbf{X}_y = (X_{1y}^1, \ldots, X_{Ty}^1, X_{1y}^2, \ldots, X_{Ty}^2, \ldots, X_{1y}^n, \ldots, X_{Ty}^n)^T$ be the column vector of nT transformed normally distributed seasonal flows X_{ty}^s for season t, year y, and site $s = 1, \ldots, n$.

Assuming that the annual and seasonal series, Z_y^s and X_{ty}^s, have zero mean (after the appropriate transformation), the basic disaggregation model is

$$\mathbf{X}_y = A\mathbf{Z}_y + B\mathbf{V}_y \qquad (7.177)$$

where $\mathbf{V}_y$ is a vector of nT independent standard normal random variables, and A and B are, respectively, $nT \times N$ and $nT \times nT$ matrices. One selects values of the elements of A and B to reproduce the observed correlations among the elements of $\mathbf{X}_y$ and between the elements of $\mathbf{X}_y$ and $\mathbf{Z}_y$. Alternatively, one could attempt to reproduce the observed correlations of the untransformed flows as opposed to the transformed flows, although this is not always possible (Hoshi et al., 1978) and often produces poorer estimates of the actual correlations of the flows (Stedinger, 1981).

The values of A and B are determined using the matrices $S_{zz} = E[\mathbf{Z}_y\mathbf{Z}_y^T]$, $S_{xx} = E[\mathbf{X}_y\mathbf{X}_y^T]$, $S_{xz} = E[\mathbf{X}_y\mathbf{Z}_y^T]$, and $S_{zx} = E[\mathbf{Z}_y\mathbf{X}_y^T]$ where S_{zz} was called S_0 earlier. Clearly, $S_{xz}^T = S_{zx}$. If S_{xz} is to be reproduced, then by multiplying Equation 7.177 on the right by $\mathbf{Z}_y^T$ and taking expectations, one sees that A must satisfy

$$E[\mathbf{X}_y\mathbf{Z}_y^T] = E[A\mathbf{Z}_y\mathbf{Z}_y^T] \qquad (7.178)$$

or

$$S_{xz} = AS_{zz} \tag{7.179}$$

Solving for the coefficient matrix A one obtains

$$A = S_{xz}S_{zz}^{-1} \tag{7.180}$$

To obtain an equation that determines the required values in the matrix B, one can multiply both sides of Equation 7.177 by their transpose and take expectations to obtain

$$S_{xx} = AS_{zz}A^T + BB^T \tag{7.181}$$

Thus, to reproduce the covariance matrix S_{xx}, the B matrix must satisfy

$$BB^T = S_{xx} - AS_{zz}A^T \tag{7.182}$$

Equations 7.180 and 7.182 for determining A and B are completely analogous to Equations 7.171 and 7.174 for the A and B matrices of the lag 1 models developed earlier. However, for the disaggregation model as formulated, BB^T, and hence the matrix B, can actually be singular or nearly so (Valencia and Schaake, 1973). This occurs because the real seasonal flows sum to the observed annual flows. Thus given the annual flow at a site and all but one $(T - 1)$ of the seasonal flows, the value of the unspecified seasonal flow can be determined by subtraction.

If the seasonal variables X_{ty}^s correspond to non-linear transformations of the actual flows Q_{ty}^s, then BB^T is generally sufficiently non-singular that a B matrix can be obtained by Cholesky decomposition. On the other hand, when the model is used to generate values of X_{ty}^s to be transformed into synthetic flows Q_{ty}^s, the constraint that these seasonal flows should sum to the given value of the annual flow is lost. Thus the generated annual flows (equal to the sums of the seasonal flows) will deviate from the values that were to have been the annual flows. Some distortion of the specified distribution of the annual flows results. This small distortion can be ignored, or each year's seasonal flows can be scaled so that their sum equals the specified value of the annual flow (Grygier and Stedinger, 1988). The latter approach eliminates the distortion in the distribution of the generated annual flows by distorting the distribution of the generated seasonal flows. Koutsoyiannis and Manetas (1996) improve upon the simple scaling algorithm by including a step that rejects candidate vectors $\mathbf{X}_y$ if the required adjustment is too large, and instead generates another vector $\mathbf{X}_y$. This reduces the distortion in the monthly flows that results from the adjustment step.

The disaggregation model has substantial data requirements. When the dimension of $\mathbf{Z}_y$ is n and the dimension of the generated vector $\mathbf{X}_y$ is m, the A matrix has mn elements. The lower diagonal B matrix and the symmetric S_{xx} matrix, upon which it depends, each have $m(m + 1)/2$ nonzero or non-redundant elements. For example, when disaggregating two aggregate annual flow series to monthly flows at five sites, $n = 2$ and $m = 12 \times 5 = 60$; thus, A has 120 elements while B and S_{xx} each have 1,830 nonzero or non-redundant parameters. As the number of sites included in the disaggregation increases, the size of S_{xx} and B increases rapidly. Very quickly the model can become overly parameterized, and there will be insufficient data to estimate all parameters (Grygier and Stedinger, 1988).

In particular, one can think of Equation 7.177 as a series of linear models generating each monthly flow X_{ty}^k for $k = 1$, $t = 1, \ldots, 12$; $k = 2$, $t = 1, \ldots, 12$ up to $k = n$, $t = 1, \ldots, 12$ that reproduces the correlation of each X_{ty}^k with all n annual flows, Z_y^k, and all previously generated monthly flows. Then when one gets to the last flow in the last month, the model will be attempting to reproduce $n + (12n - 1) = 13n - 1$ annual to monthly and cross-correlations. Because the model implicitly includes a constant, this means one needs $k^* = 13n$ years of data to obtain a unique solution for this critical equation. For $n = 3$, $k^* = 39$. One could say that with a record length of forty years, there would be only one degree of freedom left in the residual model error variance described by B. That would be unsatisfactory.

When flows at many sites or in many seasons are required, the size of the disaggregation model can be reduced by disaggregation of the flows in stages. Such condensed models do not explicitly reproduce every season-to-season correlation (Lane, 1979; Stedinger and Vogel, 1984; Gryier and Stedinger, 1988; Koutsoyiannis and Manetas, 1996), Nor do they attempt to reproduce the cross-correlations among all the flow variates at the same site within a year (Lane, 1979; Stedinger et al., 1985). Contemporaneous models, like the Markov model developed earlier in Section 8.5, are models developed for individual sites whose innovation vectors $\mathbf{V}_y$ have the needed cross-correlations to reproduce the cross-correlations of the concurrent flows (Camacho et al., 1985), as was done in Equation 7.176. Grygier and

Stedinger (1991) describe how this can be done for a condensed disaggregation model without generating inconsistencies.

8.6.2. Aggregation Models

One can start with annual or seasonal flows, and break them down into flows in shorter periods representing months or weeks. Alternatively one can start with a model that describes the shortest time step flows. This latter approach has been referred to as aggregation to distinguish it from disaggregation.

One method for generating multi-season flow sequences is to convert the time series of seasonal flows Q_{ty} into a homogeneous sequence of normally distributed zero-mean unit-variance random variables Z_{ty}. These can then be modelled by an extension of the annual flow generators that have already been discussed. This transformation can be accomplished by fitting a reasonable marginal distribution to the flows in each season so as to be able to convert the observed flows q^s_{ty} into their transformed counterparts Z^s_{ty}, and vice versa. Particularly when shorter streamflow records are available, these simple approaches may yield a reasonable model of some streams for some studies. However, they implicitly assume that the standardized series is stationary, in the sense that the season-to-season correlations of the flows do not depend on the seasons in question. This assumption seems highly questionable.

This theoretical difficulty with the standardized series can be overcome by introducing a separate streamflow model for each month. For example, the classic Thomas–Fiering model (Thomas and Fiering, 1970) of monthly flows may be written

$$Z_{t+1,y} = \beta_t Z_{ty} + \sqrt{1 - \beta_t^2 V_{ty}} \tag{7.183}$$

where the Z_{ty}'s are standard normal random variables corresponding to the streamflow in season t of year y, β_t is the season-to-season correlation of the standardized flows, and V_{ty} are independent standard normal random variables. The problem with this model is that it often fails to reproduce the correlation among non-consecutive months during a year and thus misrepresents the risk of multi-month and multi-year droughts (Hoshi et al., 1978).

For an aggregation approach to be attractive, it is necessary to use a model with greater persistence than the Thomas–Fiering model. A general class of time-series models that allow reproduction of different correlation structures are the Box–Jenkins Autoregressive-Moving average models (Box et al., 1994). These models are presented by the notation ARMA(p,q) for a model which depends on p previous flows, and q extra innovations V_{ty}. For example, Equation 7.141 would be called an AR(1) or AR(1,0) model. A simple ARMA(1,1) model is

$$Z_{t+1} = \phi_1 \cdot Z_t + V_{t+1} - \theta_1 \cdot V_t \tag{7.184}$$

The correlations of this model have the values

$$\rho_1 = (1 - \theta_1\phi_1)(\phi_1 - \theta_1)/(1 + \theta_1^2 - 2\phi_1\theta_1) \tag{7.185}$$

for the first lag. For $i > 1$

$$\rho_i = \phi^{i-1}\rho_1 \tag{7.186}$$

For ϕ values near and $0 < \theta_1 < \phi_1$, the autocorrelations ρ_k can decay much slower than those of the standard AR(1) model.

The correlation function ρ_k of general ARMA(p,q) model,

$$Z_{t+1} = \sum_{i=1}^{p} \phi_i \cdot Z_{t+1-i} + V_{t+1} - \sum_{j=1}^{q} \theta_j \cdot V_{t+1-j} \tag{7.187}$$

is a complex function that must satisfy a number of conditions to ensure the resultant model is stationary and invertible (Box et al., 1994).

ARMA(p,q) models have been extended to describe seasonal flow series by having their coefficients depend upon the season – these are called periodic autoregressive-moving average models, or PARMA. Salas and Obeysekera (1992), Salas and Fernandez (1993), and Claps et al., (1993) discuss the conceptual basis of such stochastic streamflow models. For example, Salas and Obeysekera (1992) found that low-order PARMA models, such as a PARMA(2,1), arise from reasonable conceptual representations of persistence in rainfall, runoff and groundwater recharge and release. Claps et al. (1993) observe that the PARMA(2,2) model, which may be needed if one wants to preserve year-to-year correlation, poses a parameter estimation challenge that is almost unmanageable (see also Rasmussen et al., 1996). The PARMA (1,1) model is more practical and easy to extend to the multivariate case (Hirsch, 1979; Stedinger et al., 1985; Salas, 1993; Rasmussen et al., 1996). Experience has shown that PARMA(1,1) models do a better job of reproducing the correlation of seasonal flows beyond lag 1 than does a Thomas–Fiering PAR(1,0) model (see for example, Bartolini and Salas, 1993).

9. Stochastic Simulation

This section introduces stochastic simulation. Much more detail on simulation is contained in later chapters. As discussed in Chapter 3, simulation is a flexible and widely used tool for the analysis of complex water resources systems. Simulation is trial and error. One must define the system being simulated, both its design and operating policy, and then simulate it to see how it works. If the purpose is to find the best design and operating policy, many such alternatives must be simulated and their results must be compared. When the number of alternatives to simulate becomes too large for the time and money available for such analyses, some kind of preliminary screening, perhaps using optimization models, may be justified. This use of optimization for preliminary screening – that is, for eliminating alternatives prior to a more detailed simulation – is discussed in Chapters 3, 4 and later chapters.

As with optimization models, simulation models may be deterministic or stochastic. One of the most useful tools in water resources systems planning is stochastic simulation. While optimization can be used to help define reasonable design and operating policy alternatives to be simulated, simulations can better reveal how each such alternative will perform. Stochastic simulation of complex water resources systems on digital computers provides planners with a way to define the probability distributions of multiple performance indices of those systems.

When simulating any system, the modeller designs an experiment. Initial flow, storage and water quality conditions must be specified if these are being simulated. For example, reservoirs can start full, empty or at random representative conditions. The modeller also determines what data are to be collected on system performance and operation, and how they are to be summarized. The length of time the simulation is to be run must be specified and, in the case of stochastic simulations, the number of runs to be made must also be determined. These considerations are discussed in more detail by Fishman (2001) and in other books on simulation. The use of stochastic simulation and the analysis of the output of such models are introduced here primarily in the context of an example to illustrate what goes into a stochastic simulation model and how one can deal with the information that is generated.

9.1. Generating Random Variables

Included in any stochastic simulation model is some provision for the generation of sequences of random numbers that represent particular values of events such as rainfall, streamflows or floods. To generate a sequence of values for a random variable, the probability distribution for the random variable must be specified. Historical data and an understanding of the physical processes are used to select appropriate distributions and to estimate their parameters (as discussed in Section 7.2).

Most computers have algorithms for generating random numbers uniformly distributed (equally likely) between zero and one. This uniform distribution of random numbers is defined by its cdf and pdf;

$$F_U(u) = 0 \quad \text{for} \quad u \le 0,$$

$$= u \quad \text{for} \quad 0 \le u \le 1$$

and

$$= 1 \quad \text{if} \quad u \ge 1 \qquad (7.188)$$

so that

$$f_U(u) = 1 \quad \text{if} \quad 0 \le u \le 1 \text{ and } 0 \text{ otherwise} \qquad (7.189)$$

These uniform random variables can then be transformed into random variables with any desired distribution. If $F_Q(q_t)$ is the cumulative distribution function of a random variable Q_t in period t, then Q_t can be generated using the inverse of the distribution.

$$Q_t = F_Q^{-1}[U_t] \qquad (7.190)$$

Here U_t is the uniform random number used to generate Q_t. This is illustrated in Figure 7.15.

Analytical expressions for the inverse of many distributions, such as the normal distribution, are not known, so special algorithms are employed to efficiently generate deviates with these distributions (Fishman, 2001).

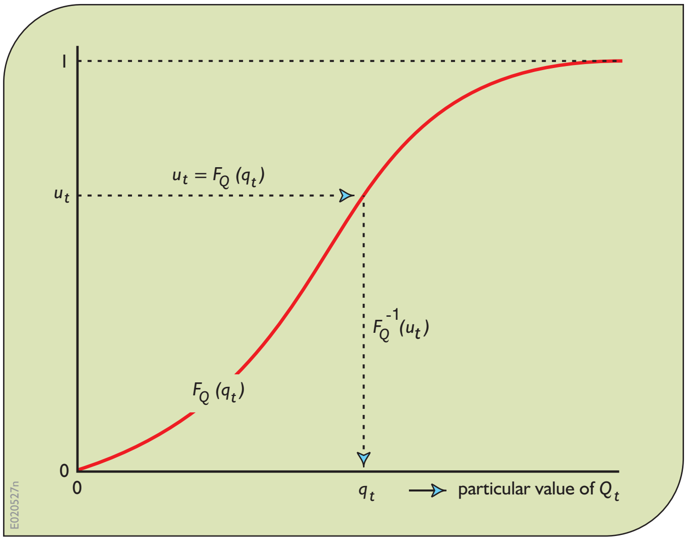

Figure 7.15. The probability distribution of a random variable can be inverted to produce values of the random variable.

9.2. River Basin Simulation

An example will demonstrate the use of stochastic simulation in the design and analysis of water resources systems. Assume that farmers in a particular valley have been plagued by frequent shortages of irrigation water. They currently draw water from an unregulated river to which they have water rights. A government agency has proposed construction of a moderate-size dam on the river upstream of the points where the farmers withdraw water. The dam would be used to increase the quantity and reliability of irrigation water available to the farmers during the summer growing season.

After preliminary planning, a reservoir with an active capacity of 4×10^7 m^3 has been proposed for a natural dam site. It is anticipated that, because of the increased reliability and availability of irrigation water, the quantity of water desired will grow from an initial level of 3×10^7 m^3/yr after construction of the dam to 4×10^7 m^3/yr within six years. After that, demand will grow more slowly to 4.5×10^7 m^3/yr – the estimated maximum reliable yield. The projected demand for summer irrigation water is shown in Table 7.12.

A simulation study can evaluate how the system will be expected to perform over a twenty-year planning period. Table 7.13 contains statistics that describe the hydrology at the dam site. The estimated moments are computed from the forty-five-year historic record.

year	water demand (x 10^7 m^3 /yr)
1	3.0
2	3.2
3	3.4
4	3.6
5	3.8
6	4.0
7	4.1
8	4.2
9	4.3
10	4.3
11	4.4
12	4.4
13	4.4
14	4.4
15	4.5
16	4.5
17	4.5
18	4.5
19	4.5
20	4.5

Table 7.12. Projected water demand for irrigation water.

Using the techniques discussed in the previous section, a Thomas–Fiering model is used to generate twenty-five lognormally distributed synthetic streamflow sequences. The statistical characteristics of the synthetic flows are those listed in Table 7.14. Use of only the forty-five-year historic flow sequence would not allow examination of the system's performance over the large range of streamflow sequences, which could occur during the twenty-year planning period. Jointly, the synthetic sequences should be a description of the range of inflows that the system might experience. A larger number of sequences could be generated.

Table 7.13. Characteristics of the river flow.

	winter	summer	annual	
mean flow	4.0	2.5	6.5	x $10^7 m^3$
standard deviation	1.5	1.0	2.3	x $10^7 m^3$
correlation of flows:				
winter with following summer		0.65		
summer with following winter		0.60		

E02110p

9.3. The Simulation Model

The simulation model is composed primarily of continuity constraints and the proposed operating policy. The volume of water stored in the reservoir at the beginning of seasons 1 (winter) and 2 (summer) in year y are denoted by S_{1y} and S_{2y}. The reservoir's winter operating policy is to store as much of the winter's inflow Q_{1y} as possible. The winter release R_{1y} is determined by the rule

$$R_{1y} = \begin{cases} S_{1y} + Q_{1y} - K & \text{if} \quad S_{1y} + Q_{1y} - R_{\min} > K \\ R_{\min} & \text{if} \quad K \geq S_{1y} + Q_{1y} - R_{\min} \geq 0 \\ S_{1y} + Q_{1y} & \text{otherwise} \end{cases} \tag{7.191}$$

where K is the reservoir capacity of $4 \times 10^7\,m^3$ and $R_{\min}$ is $0.50 \times 10^7\,m^3$, the minimum release to be made if possible. The volume of water in storage at the beginning of the year's summer season is

$$S_{2y} = S_{1y} + Q_{1y} - R_{1y} \tag{7.192}$$

The summer release policy is to meet each year's projected demand or target release D_y, if possible, so that

$$\begin{aligned} R_{2y} &= S_{2y} + Q_{2y} - K && \text{if} \quad S_{2y} + Q_{2y} - D_y > K \\ &= D_y && \text{if} \quad 0 \leq S_{2y} + Q_{2y} - D_y \leq K \\ &= S_{2y} + Q_{2y} && \text{otherwise} \end{aligned} \tag{7.193}$$

This operating policy is illustrated in Figure 7.16.

The volume of water in storage at the beginning of the next winter season is

$$S_{1,y+1} = S_{2y} + Q_{2y} - R_{2y} \tag{7.194}$$

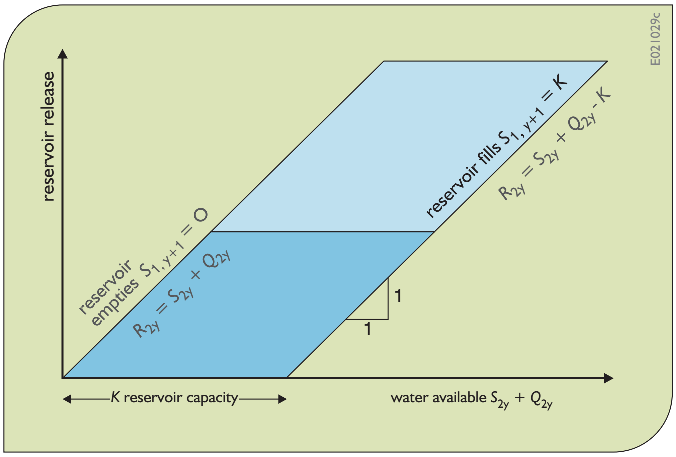
reservoir release
reservoir empties $S_{1,y+1} = 0$
$R_{2y} = S_{2y} + Q_{2y}$
reservoir fills $S_{1,y+1} = K$
$R_{2y} = S_{2y} + Q_{2y} - K$
1
1
K reservoir capacity
water available $S_{2y} + Q_{2y}$
E021029c

Figure 7.16. Summer reservoir operating policy. The shaded area denotes the feasible region of reservoir releases.

9.4. Simulation of the Basin

The question to be addressed by this simulation study is how well the reservoir will meet the farmers' water requirements. Three steps are involved in answering this question. First, one must define the performance criteria or indices to be used to describe the system's performance. The appropriate indices will, of course, depend on the problem at hand and the specific concerns of the users and managers of a water resources system. In this example of a reservoir-irrigation system, several indices will be used relating to the reliability with which target releases are met and the severity of any shortages.

The second step is to simulate the proposed system to evaluate the specified indices. For our reservoir-irrigation system, the reservoir's operation was simulated twenty-five times using the twenty-five synthetic streamflow sequences, each twenty years in length. Each of the twenty simulated years consisted of first a winter and then a summer season. At the beginning of the first winter season, the reservoir was taken to be empty ($S_{1y} = 0$ for $y = 1$) because construction would just have been completed. The target release or demand for water in each year is given in Table 7.13.

The third and final step, after simulating the system, is to interpret the resulting information so as to gain an understanding of how the system might perform both with the proposed design and operating policy and with modifications in either the system's design or its operating policy. To see how this may be done, consider the operation of our example reservoir-irrigation system.

The reliability p_y of the target release in year y is the probability that the target release D_y is met or exceeded in that year:

$$p_y = \Pr[R_{2y} \geq D_y] \tag{7.195}$$

The system's reliability is a function of the target release D_y, the hydrology of the river, the size of the reservoir and the operating policy of the system. In this example, the reliability also depends on the year in question. Figure 7.17 shows the total number of failures that occurred in each year of the twenty-five simulations. In three of these, the reservoir did not contain sufficient water after the initial winter season to meet the demand the first summer. After year 1, few failures occur in years 2 through 9 because of the low demand. Surprisingly few failures occur in years 10 and 13, when demand has reached its peak; this is because the reservoir was normally full at the beginning of this period as a result of lower demand in the earlier years. Starting in years 14 and after, failures occurred more frequently because of the higher demand placed on the system. Thus one has a sense of how the reliability of the target releases changes over time.

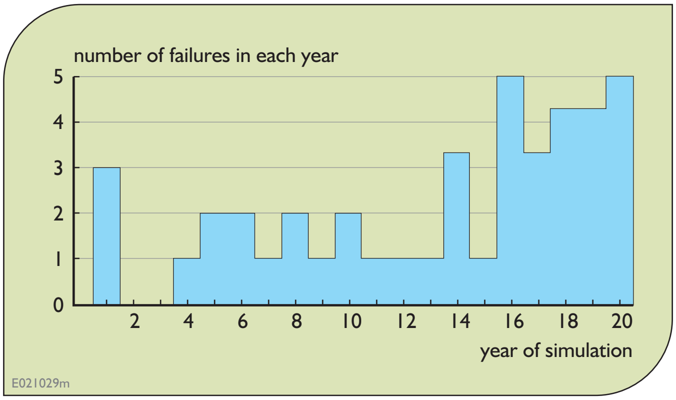

Figure 7.17. Number of failures in each year of twenty-five twenty-year simulations.

9.5. Interpreting Simulation Output

Table 7.14 contains several summary statistics of the twenty-five simulations. Column 2 of the table contains the average failure frequency in each simulation, which equals the number of years the target release was not met divided by twenty, the number of years simulated. At the bottom of column 2 and the other columns are several statistics that summarize the twenty-five values of the different performance indices. The sample estimates of the mean and variance of each index are given as one way of summarizing the distribution of the observations. Another approach is specification of the sample median, the approximate inter-quartile range $x_{(6)}-x_{(20)}$, and/or the range $x_{(1)}-x_{(25)}$ of the observations, where $x_{(i)}$ is the ith largest observation. Either set of statistics could be used to describe the centre and spread of each index's distribution.

Suppose that one is interested in the distribution of the system's failure frequency or, equivalently, the reliability with which the target can be met. Table 7.14 reports that the mean failure rate for the twenty-five simulations is 0.084, implying that the average reliability over the twenty-year period is $1 - 0.084 = 0.916$, or 92%. The median failure rate is 0.05, implying a median reliability of 95%. These are both reasonable estimates of the centre of the distribution of the failure frequency. Note that the actual failure frequency ranged from 0 (seven times) to 0.30. Thus the system's reliability ranged from 100% to as low as 70%, 75% and 80% in runs 17, 8, and 11, respectively. Obviously, the farmers are interested not only in knowing the mean failure frequency but also the range of failure frequencies they are likely to experience.

If one knew the form of the distribution function of the failure frequency, one could use the mean and standard deviation of the observations to determine an interval within which the observations would fall with some pre-specified probability. For example, if the observations are normally distributed, there is a 90% probability that the index falls within the interval $\mu_x \pm 1.65\sigma_x$. Thus, if the simulated failure rates are normally distributed, then there is about a 90% probability

Table 7.14. Results of 25 twenty-year simulations.

E021101q

	frequency of failure to meet:		total shortage TS $x10^7 m^3$	average deficit, AD
simulation number, *i*	target	80% of target		
1	0.10	0.0	1.25	0.14
2	0.15	0.05	1.97	0.17
3	0.10	0.05	1.79	0.20
4	0.10	0.05	1.67	0.22
5	0.05	0.0	0.21	0.05
6	0.0	0.0	0.00	0.00
7	0.15	0.05	1.29	0.10
8	0.25	0.10	4.75	0.21
9	0.0	0.0	0.00	0.00
10	0.10	0.0	0.34	0.04
11	0.20	0.0	1.80	0.11
12	0.05	0.05	1.28	0.43
13	0.05	0.0	0.53	0.12
14	0.10	0.0	0.88	0.11
15	0.15	0.05	1.99	0.15
16	0.05	0.0	0.23	0.05
17	0.30	0.05	2.68	0.10
18	0.10	0.0	0.76	0.08
19	0.0	0.0	0.00	0.00
20	0.0	0.0	0.00	0.00
21	0.0	0.0	0.00	0.00
22	0.05	0.05	1.47	0.33
23	0.0	0.0	0.00	0.00
24	0.0	0.0	0.00	0.00
25	0.05	0.0	0.19	0.04
mean $\bar{x}$	0.084	0.020	1.00	0.106
standard deviation of values; s_x	0.081	0.029	1.13	0.110
median	0.05	0.00	0.76	0.10
approximate interquartile range; $x_{(6)}$ -$x_{(20)}$	0.0 - 0.15	0.0 - 0.05	0.0 - 1.79	0.0 - 0.17
range; $x_{(1)}$ -$x_{(25)}$	0.0 - 0.30	0.0 - 0.10	0.0 - 4.75	0.0 - 0.43

that the actual failure rate observed in any simulation is within the interval $\bar{x} \pm 1.65s_x$. In our case this interval would be $[0.084 - 1.65(0.081), 0.084 + 1.65(0.081)] = [-0.050, 0.218]$. Clearly, the failure rate cannot be less than zero, so this interval makes little sense in our example.

A more reasonable approach to describing the distribution of a performance index whose probability distribution function is not known is to use the observations themselves. If the observations are of a continuous random variable, the interval $x_{(i)} - x_{(n+1-i)}$ provides a reasonable estimate of an interval within which the random variable falls with probability

$$P = \frac{n+1-i}{n+1} - \frac{i}{n+1} = \frac{n+1-2i}{n+1} \qquad (7.196)$$

In our example, the range $x_{(1)} - x_{(25)}$ of the twenty-five observations is an estimate of an interval in which a continuous random variable falls with probability $(25 + 1 - 2)/(25 + 1) = 92\%$, while $x_{(6)} - x_{(20)}$ corresponds to probability $(25 + 1 - 2 \times 6)/(25 + 1) = 54\%$.

Table 7.14 reports that, for the failure frequency, $x_{(1)}-x_{(25)}$ equals 0–0.30, while $x_{(6)}-x_{(20)}$ equals 0–0.15. Reflection on how the failure frequencies are calculated reminds us that the failure frequency can only take on the discrete, non-negative values 0, 1/20, 2/20, … , 20/20. Thus, the random variable X cannot be less than zero. Hence, if the lower endpoint of an interval is zero, as is the case here, then $0 - x_{(k)}$ is an estimate of an interval within which the random variable falls with a probability of at least $k/(n + 1)$. For k equal to 20 and 25, the corresponding probabilities are 77% and 96%.

Often, the analysis of a simulated system's performance centres on the average value of performance indices, such as the failure rate. It is important to know the accuracy with which the mean value of an index approximates the true mean. This is done by the construction of *confidence intervals*. A confidence interval is an interval that will contain the unknown value of a parameter with a specified probability. Confidence intervals for a mean are constructed using the t statistic,

$$t = \frac{\bar{x} - \mu_x}{s_x/\sqrt{n}} \qquad (7.197)$$

which, for large n, has approximately a standard normal distribution. Certainly, $n = 25$ is not very large, but the approximation to a normal distribution may be sufficiently good to obtain a rough estimate of how close the average frequency of failure $\bar{x}$ is likely to be to μ_x. A $100(1 - 2\alpha)\%$ confidence interval for μ_x is, approximately,

$$\bar{x} - t_\alpha \frac{s_x}{\sqrt{n}} \le \mu_x \le \bar{x} + t_\alpha \frac{s_x}{\sqrt{n}}$$

or

$$0.084 - t_\alpha \left(\frac{0.081}{\sqrt{25}}\right) \le \mu_x \le 0.084 + t_\alpha \left(\frac{0.081}{\sqrt{25}}\right) \qquad (7.198)$$

If $\alpha = 0.05$, then using a normal distribution $t_\alpha = 1.65$ and Equation 7.118 becomes $0.057 \le \mu_x \le 0.11$.

Hence, based on the simulation output, one can be about 90% sure that the true mean failure frequency lies between 5.7% and 11%. This corresponds to a reliability of between 89% and 94%. By performing additional simulations to increase the size of n, the width of this confidence interval can be decreased. However, this increase in accuracy may be an illusion because the uncertainty in the parameters of the streamflow model has not been incorporated into the analysis.

Failure frequency or system reliability describes only one dimension of the system's performance. Table 7.14 contains additional information on the system's performance related to the severity of shortages. Column 3 lists the frequencies with which the shortage exceeded 20% of that year's demand. This occurred in approximately 2% of the years, or in 24% of the years in which a failure occurred. Taking another point of view, failures in excess of 20% of demand occurred in nine out of twenty-five, or in 36% of the simulation runs.

Columns 4 and 5 of Table 7.14 contain two other indices that pertain to the severity of the failures. The total shortfall, TS, in Column 4 is calculated as the sum of the positive differences between the demand and the release in the summer season over the twenty-year period.

$$TS = \sum_y [D_{2y} - R_{2y}]^+$$

where

$$[Q]^+ = Q \quad \text{if } Q > 0; \quad 0 \text{ otherwise} \qquad (7.199)$$

The total shortfall equals the total amount by which the target release is not met in years in which shortages occur.

Related to the total shortfall is the average deficit. The deficit is defined as the shortfall in any year divided by the target release in that year. The average deficit, AD, is

$$AD = \frac{1}{m} \sum_{y=1}^{20} \frac{\left[D_{2y} - R_{2y}\right]}{D_{2y}} \tag{7.200}$$

where m is the number of failures (deficits) or nonzero terms in the sum.

Both the total shortfall and the average deficit measure the severity of shortages. The mean total shortfall $\overline{TS}$, equal to 1.00 for the twenty-five simulation runs, is a difficult number to interpret. While no shortage occurred in seven runs, the total shortage was 4.7 in run 8, in which the shortfall in two different years exceeded 20% of the target. The median of the total shortage values, equal to 0.76, is an easier number to interpret in that one knows that half the time the total shortage was greater and half the time less than this value.

The mean average deficit $\overline{AD}$ is 0.106, or 11%. However, this mean includes an average deficit of zero in the seven runs in which no shortages occurred. The average deficit in the eighteen runs in which shortages occurred is (11%)(25/18) = 15%. The average deficit in individual simulations in which shortages occurred ranges from 4% to 43%, with a median of 11.5%.

After examining the results reported in Table 7.14, the farmers might determine that the probability of a shortage exceeding 20% of a year's target release is higher than they would like. They can deal with more frequent minor shortages, not exceeding 20% of the target, with little economic hardship, particularly if they are warned at the beginning of the growing season that less than the targeted quantity of water will be delivered. Then they can curtail their planting or plant crops requiring less water.

In an attempt to find out how better to meet the farmers' needs, the simulation program was re-run with the same streamflow sequences and a new operating policy in which only 80% of the growing season's target release is provided (if possible) if the reservoir is less than 80% full at the end of the previous winter season. This gives the farmers time to adjust their planting schedules and may increase the quantity of water stored in the reservoir to be used the following year if the drought persists.

As the simulation results with the new policy in Table 7.15 demonstrate, this new operating policy appears to have the expected effect on the system's operation. With the new policy, only six severe shortages in excess of 20% of demand occur in the twenty-five twenty-year simulations, as opposed to ten such shortages with the original policy. In addition, these severe shortages are all less severe than the corresponding shortages that occur with the same streamflow sequence when the original policy is followed.

The decrease in the severity of shortages is obtained at a price. The overall failure frequency has increased from 8.4% to 14.2%. However, the latter value is misleading because in fourteen of the twenty-five simulations, a failure occurs in the first simulation year with the new policy, whereas only three failures occur with the original policy. Of course, these first-year failures occur because the reservoir starts empty at the beginning of the first winter and often does not fill that season. Ignoring these first-year failures, the failure rates with the two policies over the subsequent nineteen years are 8.2% and 12.0%. Thus the frequency of failures in excess of 20% of demand is decreased from 2.0% to 1.2% by increasing the frequency of all failures after the first year from 8.2% to 12.0%. Reliability decreases, but so does vulnerability. If the farmers are willing to put up with more frequent minor shortages, then it appears that they can reduce their risk of experiencing shortages of greater severity.

The preceding discussion has ignored the statistical issue of whether the differences between the indices obtained in the two simulation experiments are of sufficient statistical reliability to support the analysis. If care is not taken, observed changes in a performance index from one simulation experiment to another may be due to sampling fluctuations rather than to modifications of the water resource system's design or operating policy.

As an example, consider the change that occurred in the frequency of shortages. Let X_{1i} and X_{2i} be the simulated failure rates using the ith streamflow sequence with the original and modified operating policies. The random variables $Y_i = X_{1i} - X_{2i}$ for i equal 1 through 25 are independent of each other if the streamflow sequences are generated independently, as they were.

One would like to confirm that the random variable Y tends to be negative more often than it is positive, and hence, that policy 2 indeed results in more failures overall. A direct test of this theory is provided by the sign test. Of the twenty-five paired simulation runs, $y_i < 0$ in twenty-one cases and $y_i = 0$ in four cases. We can ignore the times

Table 7.15. Results of 25 twenty-year simulations with modified operating policy to avoid severe shortages.

E021101r

simulation number, *i*	frequency of failure to meet:		total shortage TS $x10^7 m^3$	average deficit, AD
	target	80% of target		
1	0.10	0.0	1.80	0.20
2	0.30	0.0	4.70	0.20
3	0.25	0.0	3.90	0.20
4	0.20	0.05	3.46	0.21
5	0.10	0.0	1.48	0.20
6	0.05	0.0	0.60	0.20
7	0.20	0.0	3.30	0.20
8	0.25	0.10	5.45	0.26
9	0.05	0.0	0.60	0.20
10	0.20	0.0	3.24	0.20
11	0.25	0.0	3.88	0.20
12	0.10	0.05	1.92	0.31
13	0.10	0.0	1.50	0.20
14	0.15	0.0	2.52	0.20
15	0.25	0.05	3.76	0.18
16	0.10	0.0	1.80	0.20
17	0.30	0.0	5.10	0.20
18	0.15	0.0	2.40	0.20
19	0.0	0.0	0.0	0.0
20	0.05	0.0	0.76	0.20
21	0.10	0.0	1.80	0.20
22	0.10	0.05	2.37	0.26
23	0.05	0.0	0.90	0.20
24	0.05	0.0	0.90	0.20
25	0.10	0.0	1.50	0.20
mean $\bar{x}$	0.142	0.012	2.39	0.201
standard deviation of values; s_x	0.087	0.026	1.50	0.050
median	0.10	0.00	1.92	0.20
approximate interquartile range; $x_{(6)}$ - $x_{(20)}$	0.05 - 0.25	0.0 - 0.0	0.90 - 3.76	0.20 - 0.20
range; $x_{(1)}$ - $x_{(25)}$	0.0 - 0.30	0.0 - 0.10	0.0 - 5.45	0.0 - 0.31

when $y_i = 0$. Note that if $y_i < 0$ and $y_i > 0$ were equally likely, then the probability of observing $y_i < 0$ in all twenty-one cases when $y_i \neq 0$ is 2^{-21} or 5×10^{-7}. This is exceptionally strong proof that the new policy has increased the failure frequency.

A similar analysis can be made of the frequency with which the release is less than 80% of the target. Failure frequencies differ in the two policies in only four of the twenty-five simulation runs. However, in all four cases where they differ, the new policy resulted in fewer severe failures. The probability of such a lopsided result, were it equally likely that either policy would result in a lower frequency of failures in excess of 20% of the target, is $2^{-4} = 0.0625$. This is fairly strong evidence that the new policy indeed decreases the frequency of severe failures.

Another approach to this problem is to ask if the difference between the average failure rates $\bar{x}_1$ and $\bar{x}_2$ is statistically significant; that is, can the difference between X_1 and X_2 be attributed to the fluctuations that occur in the average of any finite set of random variables? In this example the significance of the difference between the two means can be tested using the random variable Y_i defined as $X_{1i} - X_{2i}$ for i equal 1 through 25. The mean of the observed y_i's is

$$\begin{aligned}\bar{y} &= \frac{1}{25}\sum_{i-1}^{25}(x_{1i} - x_{2i}) = \bar{x}_1 - \bar{x}_2 \\ &= 0.084 - 0.142 = -0.058\end{aligned} \tag{7.201}$$

and their variance is

$$s_y^2 = \frac{1}{25}\sum_{i=1}^{25}(x_{1i} - x_{2i} - \bar{y})^2 = (0.0400)^2 \tag{7.202}$$

Now, if the sample size n, equal to 25 here, is sufficiently large, then t defined by

$$t = \frac{\bar{y} - \mu_Y}{s_Y/\sqrt{n}} \tag{7.203}$$

has approximately a standard normal distribution. The closer the distribution of Y is to that of the normal distribution, the faster the convergence of the distribution of t is to the standard normal distribution with increasing n. If $X_{1i} - X_{2i}$ is normally distributed, which is not the case here, then each Y_i has a normal distribution and t has Student's t-distribution.

If $E[X_{1i}] = E[X_{2i}]$, then μ_Y equals zero, and upon substituting the observed values of $\bar{y}$ and s_Y^2 into Equation 7.123, one obtains

$$t = \frac{-0.0580}{0.0400/\sqrt{25}} = -7.25 \tag{7.204}$$

The probability of observing a value of t equal to -7.25 or smaller is less than 0.1% if n is sufficiently large that t is normally distributed. Hence it appears very improbable that μ_Y equals zero.

This example provides an illustration of the advantage of using the same streamflow sequences when simulating both policies. Suppose that different streamflow sequences were used in all the simulations. Then the expected value of Y would not change, but its variance would be given by

$$\begin{aligned}\mathrm{Var}(Y) &= E[X_1 - X_2 - (\mu_1 - \mu_2)]^2 \\ &= E[(X_1 - \mu_1)^2] - 2E[(X_1 - \mu_1) \\ &\quad \times (X_2 - \mu_2)] + E[(X_2 - \mu_2)^2] \\ &= \sigma_{x_1}^2 - 2\mathrm{Cov}(X_1, X_2) + \sigma_{x_2}^2\end{aligned} \tag{7.205}$$

where $\mathrm{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$ and is the *covariance* of the two random variables. The covariance between X_1 and X_2 will be zero if they are independently distributed, as they would be if different randomly generated streamflow sequences were used in each simulation. Estimating $\sigma_{x_1}^2$ and $\sigma_{x_2}^2$ by their sample estimates, an estimate of what the variance of Y would be if $\mathrm{Cov}(X_1, X_2)$ were zero is

$$\hat{\sigma}_Y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 = (0.081)^2 + (0.087)^2 = (0.119)^2 \tag{7.206}$$

The actual sample estimate σ_Y equals 0.040; if independent streamflow sequences are used in all simulations, σ_Y will take a value near 0.119 rather than 0.040 (Equation 7.202). A standard deviation of 0.119, with $\mu_y = 0$, yields a value of the t test statistic

$$t = \frac{\bar{y} - \mu_Y}{0.119/\sqrt{25}} = -2.44 \tag{7.207}$$

If t is normally distributed, the probability of observing a value less than -2.44 is about 0.8%. This illustrates that use of the same streamflow sequences in the simulation of both policies allows one to better distinguish the differences in the policies' performance. By using the same streamflow sequences, or other random inputs, one can construct a simulation experiment in which variations in performance caused by different random inputs are confused as little as possible with the differences in performance caused by changes in the system's design or operating policy.

10. Conclusions

This chapter has introduced statistical concepts that analysts use to describe the randomness or uncertainty of their data. Most of the data used by water resources systems analysts is uncertain. This uncertainty comes from not understanding as well as we would like how our water resources systems (including their ecosystems) function as well as not being able to forecast, perfectly, the future. It is that simple. We do not know the exact amounts, qualities and their distributions over space and time of either the supplies of water we manage or the water demands we try to meet. We also do not know the benefits and costs, however measured, of any actions we take to manage both water supply and water demand.

The chapter began with an introduction to probability concepts and methods for describing random variables and parameters of their distributions. It then reviewed some of the commonly used probability distributions and how to determine the distributions of sample data, how to work with censored and partial duration series data, methods of regionalization, stochastic processes and time-series analyses.

The chapter concluded with an introduction to a range of univariate and multivariate stochastic models that are used to generate stochastic streamflow, precipitation depths, temperatures and evaporation. These methods are used to generate temporal and spatial stochastic process that serve as inputs to stochastic simulation models for system design, for system operations studies, and for evaluation of the reliability and precision of different estimation algorithms. The final section of this chapter provides an example of stochastic simulation, and the use of statistical methods to summarize the results of such simulations.

This is merely an introduction to some of the statistical tools available for use when dealing with uncertain data. Many of the concepts introduced in this chapter will be used in the chapters that follow on constructing and implementing various types of optimization, simulation and statistical models. The references cited in the reference section provide additional and more detailed information.

Although many of the methods presented in this and in some of the following chapters can describe many of the characteristics and consequences of uncertainty, it is unclear as to whether or not society knows exactly what to do with such information. Nevertheless, there seems to be an increasing demand from stakeholders involved in planning processes for information related to the uncertainty associated with the impacts predicted by models. The challenge is not only to quantify that uncertainty, but also to communicate it in effective ways that inform, and not confuse, the decision-making process.

11. References

AYYUB, B.M. and MCCUEN, R.H. 2002. *Probability, statistics, and reliability for engineers and scientists*. Boca Raton, Chapman and Hill, CRC Press.

BARTOLINI, P. and SALAS, J. 1993. Modelling of streamflow processes at different time scales. *Water Resources Research*, Vol. 29, No. 8, pp. 2573–87.

BATES, B.C. and CAMPBELL, E.P. 2001. A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall–runoff modelling. *Water Resources Research*, Vol. 37, No. 4, pp. 937–47.

BEARD, L.R. 1960. Probability estimates based on small normal-distribution samples. *Journal of Geophysical Research*, Vol. 65, No. 7, pp. 2143–8.

BEARD, L.R. 1997. Estimating flood frequency and average annual damages. *Journal of Water Resources Planning and Management*, Vol. 123, No. 2, pp. 84–8.

BENJAMIN, J.R. and CORNELL, C.A. 1970. *Probability, statistics and decisions for civil engineers*. New York, McGraw-Hill.

BICKEL, P.J. and DOKSUM, K.A. 1977. *Mathematical statistics: basic ideas and selected topics*. San Francisco, Holden-Day.

BOBÉE, B. 1975. The log Pearson type 3 distribution and its applications in hydrology. *Water Resources Research*, Vol. 14, No. 2, pp. 365–9.

BOBÉE, B. and ASHKAR, F. 1991. *The gamma distribution and derived distributions applied in hydrology*. Littleton Colo., Water Resources Press.

BOBÉE, B. and ROBITAILLE, R. 1977. The use of the Pearson type 3 distribution and log Pearson type 3 distribution revisited. *Water Resources Research*, Vol. 13, No. 2, pp. 427–43.

BOX, G.E.P.; JENKINS, G.M. and RISINSEL, G.C. 1994. *Times series analysis: forecasting and control*, 3rd Edition. New Jersey, Prentice-Hall.

CAMACHO, F.; MCLEOD, A.I. and HIPEL, K.W. 1985. Contemporaneous autoregressive-moving average (CARMA) modelling in water resources. *Water Resources Bulletin*, Vol. 21, No. 4, pp. 709–20.

CARLIN, B.P. and LOUIS, T.A. 2000. *Bayes and empirical Bayes methods for data analysis*, 2nd Edition. New York, Chapman and Hall, CRC.

CHOWDHURY, J.U. and STEDINGER, J.R. 1991. Confidence intervals for design floods with estimated skew coefficient. *Journal of Hydraulic Engineering*, Vol. 117, No. 7, pp. 811–31.

CHOWDHURY, J.U.; STEDINGER, J.R. and LU, L.H. 1991. Goodness-of-fit tests for regional GEV flood distributions. *Water Resources Research*, Vol. 27, No. 7, pp. 1765–76.

CLAPS, P. 1993. Conceptual basis for stochastic models of monthly streamflows. In: J.B. Marco, R. Harboe and J.D. Salas (eds). *Stochastic hydrology and its use in water resources systems simulation and optimization*, Dordrecht, Kluwer Academic, pp. 229–35.

CLAPS, P.; ROSSI, F. and VITALE, C. 1993. Conceptual-stochastic modelling of seasonal runoff using autoregressive moving average models and different scales of aggregation. *Water Resources Research*, Vol. 29, No. 8, pp. 2545–59.

COHN, T.A.; LANE, W.L. and BAIER, W.G. 1997. An algorithm for computing moments-based flood quantile estimates when historical flood information is available. *Water Resources* Research, Vol. 33, No. 9, pp. 2089–96.

COHN, C.A.; LANE, W.L. and STEDINGER, J.R. 2001. Confidence intervals for EMA flood quantile estimates. *Water Resources Research*, Vol. 37, No. 6, pp. 1695–1706.

CRAINICEANU, C.M.; RUPPERT, D.; STEDINGER, J.R. and BEHR, C.T. 2002. *Improving MCMC mixing for a GLMM describing pathogen concentrations in water supplies, in case studies in Bayesian analysis*. New York, Springer-Verlag.

D'AGOSTINO, R.B. and STEPHENS, M.A. 1986. *Goodness-of-fit procedures*. New York, Marcel Dekker.

DAVID, H.A. 1981. *Order statistics,* 2nd edition. New York, Wiley.

FIERING, M.B. 1967. *Streamflow synthesis*. Cambridge, Mass., Harvard University Press.

FILL, H. and STEDINGER, J. 1995. L-moment and PPCC goodness-of-fit tests for the Gumbel distribution and effect of autocorrelation. *Water Resources Research,* Vol. 31, No. 1, pp. 225–29.

FILL, H. and STEDINGER, J. 1998. Using regional regression within index flood procedures and an empirical Bayesian estimator. *Journal of Hydrology*, Vol. 210, Nos 1–4, pp. 128–45.

FILLIBEN, J.J. 1975. The probability plot correlation test for normality. *Technometrics*, Vol. 17, No. 1, pp. 111–17.

FISHMAN, G.S. 2001. *Discrete-event simulation: modelling, programming, and analysis*. Berlin, Springer-Verlag.

GABRIELE, S. and ARNELL, N. 1991. A hierarchical approach to regional flood frequency analysis. *Water Resources Research*, Vol. 27, No. 6, pp. 1281–9.

GELMAN, A.; CARLIN, J.B.; STERN, H.S. and RUBIN, D.B. 1995. *Bayesian data analysis*. Boca Raton, Chapman and Hall, CRC.

GILKS, W.R.; RICHARDSON, S. and SPIEGELHALTER, D.J. (eds). 1996. *Markov chain Monte Carlo in practice*. London and New York, Chapman and Hall.

GIESBRECHT, F. and KEMPTHORNE, O. 1976. Maximum likelihood estimation in the three-parameter log normal distribution. *Journal of the Royal Statistical Society B*, Vol. 38, No. 3, pp. 257–64.

GREENWOOD, J.A. and DURAND, D. 1960. Aids for fitting the gamma distribution by maximum likelihood. *Technometrics*, Vol. 2, No. 1, pp. 55–65.

GRIFFS, V.W.; STEDINGER, J.R. and COHN, T.A. 2004. LP3 quantile estimators with regional skew information and low outlier adjustments. *Water Resources Research*, Vol. 40, forthcoming.

GRYGIER, J.C. and STEDINGER, J.R. 1988. Condensed disaggregation procedures and conservation corrections. *Water Resources Research*, Vol. 24, No. 10, pp. 1574–84.

GRYGIER, J.C. and STEDINGER, J.R. 1991. *SPIGOT: a synthetic flow generation software package, user's manual and technical description, version 2.6*. Ithaca, N.Y., School of Civil and Environmental Engineering, Cornell University.

GUMBEL, E.J. 1958. *Statistics of extremes*. New York, Columbia University Press.

GUPTA, V.K. and DAWDY, D.R. 1995a. Physical interpretation of regional variations in the scaling exponents of flood quantiles. *Hydrological Processes*, Vol. 9, Nos. 3–4, pp. 347–61.

GUPTA, V.K. and DAWDY, D.R. 1995b. Multiscaling and skew separation in regional floods. *Water Resources Research*, Vol. 31, No. 11, pp. 2761–76.

GUPTA, V.K.; MESA, O.J. and DAWDY, D.R. 1994. Multiscaling theory of flood peaks: regional quantile analysis. *Water Resources Research*, Vol. 30, No. 12, pp. 3405–12.

HAAN, C.T. 1977. *Statistical methods in hydrology*. Ames, Iowa, Iowa State University Press.

HAAS, C.N. and SCHEFF, P.A. 1990. Estimation of averages in truncated samples. *Environmental Science and Technolo*gy, Vol. 24, No. 6, pp. 912–19.

HELSEL, D.R. 1990. Less than obvious: statistical treatment of data below the detection limit. *Environ. Sci. and Technol.*, Vol. 24, No. 12, pp. 1767–74.

HELSEL, D.R. and COHN, T.A. 1988. Estimation of descriptive statistics for multiple censored water quality data. *Water Resources Research*, Vol. 24, No. 12, pp. 1997–2004.

HIRSCH, R.M. 1979. Synthetic hydrology and water supply reliability. *Water Resources Research*, Vol. 15, No. 6, pp. 1603–15.

HIRSCH, R.M. and STEDINGER, J.R. 1987. Plotting positions for historical floods and their precision. *Water Resources Research*, Vol. 23, No. 4, pp. 715–27.

HOSHI, K.; BURGES, S.J. and YAMAOKA, I. 1978. Reservoir design capacities for various seasonal operational hydrology models. *Proceedings of the Japanese Society of Civil Engineers*, No. 273, pp. 121–34.

HOSHI, K.; STEDINGER, J.R. and BURGES, S. 1984. Estimation of log normal quantiles: Monte Carlo results and first-order approximations. *Journal of Hydrology*, Vol. 71, Nos 1–2, pp. 1–30.

HOSKING, J.R.M. 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of Royal Statistical Society*, *B*, Vol. 52, No. 2, pp. 105–24.

HOSKING, J.R.M. and WALLIS, J.R. 1987. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, Vol. 29, No. 3, pp. 339–49.

HOSKING, J.R.M. and WALLIS, J.R. 1995. A comparison of unbiased and plotting-position estimators of L-moments. Water Resources Research, Vol. 31, No. 8, pp. 2019–25.

HOSKING, J.R.M. and WALLIS, J.R. 1997. *Regional frequency analysis: an approach based on L-moments*. Cambridge, Cambridge University Press.

HOSKING, J.R.M.; WALLIS, J.R. and WOOD, E.F. 1985. Estimation of the generalized extreme-value distribution by the method of probability weighted moments. *Technometrics*, Vol. 27, No. 3, pp. 251–61.

IACWD (Interagency Advisory Committee on Water Data). 1982. *Guidelines for determining flood flow frequency*, Bulletin 17B. Reston, Va., US Department of the Interior, US Geological Survey, Office of Water Data Coordination.

JENKINS, G.M. and WATTS, D.G. 1968. *Spectral Analysis and its Applications*. San Francisco, Holden-Day.

KENDALL, M.G. and STUART, A. 1966. *The advanced theory of statistics*, Vol. 3. New York, Hafner.

KIRBY, W. 1974. Algebraic boundness of sample statistics. *Water Resources Research*, Vol. 10, No. 2, pp. 220–2.

KIRBY, W. 1972. Computer oriented Wilson–Hilferty transformation that preserves the first 3 moments and lower bound of the Pearson Type 3 distribution. *Water Resources Research*, Vol. 8, No. 5, pp. 1251–4.

KITE, G.W. 1988. *Frequency and risk analysis in hydrology*. Littleton, Colo. Water Resources Publications.

KOTTEGODA, M. and ROSSO, R. 1997. *Statistics, probability, and reliability for civil and environmental engineers*. New York, McGraw-Hill.

KOUTSOYIANNIS, D. and MANETAS, A. 1996. Simple disaggregation by accurate adjusting procedures. *Water Resources Research*, Vol. 32, No. 7, pp. 2105–17.

KROLL, K. and STEDINGER, J.R. 1996. Estimation of moments and quantiles with censored data. *Water Resources Research,* Vol. 32, No. 4, pp. 1005–12.

KUCZERA, G. 1999. Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. Water *Resources Research,* Vol. 35, No. 5, pp. 1551–7.

KUCZERA, G. 1983. Effects of sampling uncertainty and spatial correlation on an empirical Bayes procedure for combining site and regional information. *Journal of Hydrology*, Vol. 65, No. 4, pp. 373–98.

KUCZERA, G. 1982. Combining site-specific and regional information: an empirical Bayesian approach. *Water Resources Research*, Vol. 18, No. 2, pp. 306–14.

LANDWEHR, J.M.; MATALAS, N.C. and WALLIS, J.R. 1978. Some comparisons of flood statistics in real and log space. *Water Resources Research*, Vol. 14, No. 5, pp. 902–20.

LANDWEHR, J.M.; MATALAS, N.C. and WALLIS, J.R. 1979. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research*, Vol. 15, No. 5, pp. 1055–64.

LANE, W. 1979. *Applied stochastic techniques (users manual).* Denver, Colo., Bureau of Reclamation, Engineering and Research Center, December.

LANGBEIN, W.B. 1949. Annual floods and the partial duration flood series, EOS. *Transactions of the American Geophysical Union*, Vol. 30, No. 6, pp. 879–81.

LEDOLTER, J. 1978. The analysis of multivariate time series applied to problems in hydrology. *Journal of Hydrology*, Vol. 36, No. 3–4, pp. 327–52.

LETTENMAIER, D.P. and BURGES, S.J. 1977a. Operational assessment of hydrological models of long-term persistence. *Water Resources Research*, Vol. 13, No. 1, pp. 113–24.

LETTENMAIER, D.P. and BURGES, S.J. 1977b. An operational approach to preserving skew in hydrological models of long-term persistence. *Water Resources Research*, Vol. 13, No. 2, pp. 281–90.

LETTENMAIER, D.P.; WALLIS, J.R. and WOOD, E.F. 1987. Effect of regional heterogeneity on flood frequency estimation. *Water Resources Research*, Vol. 23, No. 2, pp. 313–24.

MACBERTHOUEX, P. and BROWN, L.C. 2002. *Statistics for environmental engineers*, 2nd Edition. Boca Raton, Fla., Lewis, CRC Press.

MADSEN, H.; PEARSON, C.P.; RASMUSSEN, P.F. and ROSBJERG, D. 1997a. Comparison of annual maximum series and partial duration series methods for modelling extreme hydrological events 1: at-site modelling. *Water Resources Research*, Vol. 33, No. 4, pp. 747–58.

MADSEN, H.; PEARSON, C.P. and ROSBJERG, D. 1997b. Comparison of annual maximum series and partial duration series methods for modelling extreme hydrological events 2: regional modelling. *Water Resources Research,* Vol. 33, No. 4, pp. 759–70.

MADSEN, H. and ROSBJERG, D. 1997a. The partial duration series method in regional index flood modelling. *Water Resources Research*, Vol. 33, No. 4, pp. 737–46.

MADSEN, H. and ROSBJERG, D. 1997b. Generalized least squares and empirical Bayesian estimation in regional partial duration series index-flood modelling. *Water Resources Research*, Vol. 33, No. 4, pp. 771–82.

MAHEEPALA, S. and PERERA, B.J.C. 1996. Monthly hydrological data generation by disaggregation. *Journal of Hydrology*, No. 178, 277–91.

MARCO, J.B.; HARBOE, R. and SALAS, J.D. (eds). 1989. Stochastic hydrology and its use in water resources systems simulation and optimization. NATO ASI Series. Dordrecht, Kluwer Academic.

MARTINS, E.S. and STEDINGER, J.R. 2000. Generalized maximum likelihood GEV quantile estimators for hydrological data. *Water Resources Research*, Vol. 36, No. 3, pp. 737–44.

MARTINS, E.S. and STEDINGER, J.R. 2001a. Historical information in a GMLE-GEV framework with partial duration and annual maximum series. *Water Resources Research*, Vol. 37, No. 10, pp. 2551–57.

MARTINS, E.S. and STEDINGER, J.R. 2001b. Generalized maximum likelihood Pareto-Poisson flood risk analysis for partial duration series. *Water Resources Research*, Vol. 37, No. 10, pp. 2559–67.

MATALAS, N.C. 1967. Mathematical assessment of synthetic hydrology. *Water Resources Research*, Vol. 3, No. 4, pp. 937–45.

MATALAS, N.C. and WALLIS, J.R. 1973. Eureka! It fits a Pearson type 3 distribution. *Water Resources Research*, Vol. 9, No. 3, pp. 281–9.

MATALAS, N.C. and WALLIS, J.R. 1976. Generation of synthetic flow sequences. In: A.K. Biswas (ed.), *Systems approach to water management*. New York, McGraw-Hill.

MEJIA, J.M. and ROUSSELLE, J. 1976. Disaggregation models in hydrology revisited. *Water Resources Research*, Vol. 12, No. 2, pp. 185–6.

NERC (Natural Environmental Research Council) 1975. Flood studies report, Vol. 1: hydrological studies. London.

NORTH, M. 1980. Time-dependent stochastic model of floods. *Journal of the Hydraulics Division*, ASCE, Vol. 106, No. HY5, pp. 649–65.

O'CONNELL, D.R.H.; OSTENAA, D.A.; LEVISH, D.R. and KLINGER, R.E. 2002. Bayesian flood frequency analysis with paleohydrological bound data. *Water Resources Research*, Vol. 38, No. 5, pp. 161–64.

O'CONNELL, P.E. 1977. ARMA models in synthetic hydrology. In: T.A. Ciriani, V. Maione and J.R. Wallis (eds), *Mathematical models for surface water hydrology*. New York, Wiley.

PRESS, W.H.; FLANNERY, B.P.; TEUKOLSKY, S.A. and VETTERLING, W.T. 1986. *Numerical recipes: the art of scientific computing*. Cambridge, UK, Cambridge University Press.

RAIFFA, H. and SCHLAIFER, R. 1961. *Applied statistical decision theory*. Cambridge, Mass., MIT Press.

RASMUSSEN, P.F. and ROSBJERG, D. 1989. Risk estimation in partial duration series. *Water Resources Research*, Vol. 25, No. 11, pp. 2319–30.

RASMUSSEN, P.F. and ROSBJERG, D. 1991a. Evaluation of risk concepts in partial duration series. *Stochastic Hydrology and Hydraulics*, Vol. 5, No. 1, pp. 1–16.

RASMUSSEN, P.F. and ROSBJERG, D. 1991b. Prediction uncertainty in seasonal partial duration series. *Water Resources Research*, Vol. 27, No. 11, pp. 2875–83.

RASMUSSEN, R.F.; SALAS, J.D.; FAGHERAZZI, L.; RASSAM.; J.C. and BOBÉE, R. 1996. Estimation and validation of contemporaneous PARMA models for streamflow simulation. *Water Resources Research*, Vol. 32, No. 10, pp. 3151–60.

ROBINSON, J.S. and SIVAPALAN, M. 1997. Temporal scales and hydrological regimes: implications for flood frequency scaling. *Water Resources Research*, Vol. 33, No. 12, pp. 2981–99.

ROSBJERG, D. 1985. Estimation in Partial duration series with independent and dependent peak values. *Journal of Hydrology*, *No.* 76, pp. 183–95.

ROSBJERG, D. and MADSEN, H. 1998. Design with uncertain design values, In: H. Wheater and C. Kirby (eds), *Hydrology in a changing environment*, New York, Wiley. Vol. 3, pp. 155–63.

ROSBJERG, D.; MADSEN, H. and RASMUSSEN, P.F. 1992. Prediction in partial duration series with generalized Pareto-distributed exceedances. *Water Resources Research*, Vol. 28, No. 11, pp. 3001–10.

SALAS, J.D. 1993. Analysis and modelling of hydrological time series. In: D. Maidment (ed.), *Handbook of hydrology*, Chapter 17. New York, McGraw-Hill.

SALAS, J.D.; DELLEUR, J.W.; YEJEVICH, V. and LANE, W.L. 1980. *Applied modelling of hydrological time series*. Littleton, Colo., Water Resources Press Publications.

SALAS, J.D. and FERNANDEZ, B. 1993. Models for data generation in hydrology: univariate techniques. In: J.B. Marco, R. Harboe and J.D. Salas (eds), *Stochastic hydrology and its use in water resources systems simulation and optimization*, Dordrecht, Kluwer Academic. pp. 76–95.

SALAS, J.D. and OBEYSEKERA, J.T.B. 1992. Conceptual basis of seasonal streamflow time series. *Journal of Hydraulic Engineering*, Vol. 118, No. 8, pp. 1186–94.

SCHAAKE, J.C. and VICENS, G.J. 1980. Design length of water resource simulation experiments. *Journal of Water Resources Planning and Management*, Vol. 106, No. 1, pp. 333–50.

SLACK, J.R.; WALLIS, J.R. and MATALAS, N.C. 1975. On the value of information in flood frequency analysis. *Water Resources Research*, Vol. 11, No. 5, pp. 629–48.

SMITH, R.L. 1984. Threshold methods for sample extremes. In: J. Tiago de Oliveira (ed.), *Statistical extremes and applications*, Dordrecht, D. Reidel. pp. 621–38.

STEDINGER, J.R. 1980. Fitting log normal distributions to hydrological data. *Water Resources Research*, Vol. 16, No. 3, pp. 481–90.

STEDINGER, J.R. 1981. Estimating correlations in multivariate streamflow models. *Water Resources Research*, Vol. 17, No. 1, pp. 200–08.

STEDINGER, J.R. 1983. Estimating a regional flood frequency distribution. *Water Resources Research*, Vol. 19, No. 2, pp. 503–10.

STEDINGER, J.R. 1997. Expected probability and annual damage estimators. *Journal of Water Resources Planning and Management*, Vol. 123, No. 2, pp. 125–35. [With discussion, Leo R. Beard, 1998. *Journal of Water Resources Planning and Management*, Vol. 124, No. 6, pp. 365–66.]

STEDINGER, J.R. 2000. Flood frequency analysis and statistical estimation of flood risk. In: E.E. Wohl (ed.), *Inland flood hazards: human, riparian and aquatic communities*, Chapter 12. Stanford, UK, Cambridge University Press.

STEDINGER, J.R. and BAKER, V.R. 1987. Surface water hydrology: historical and paleoflood information. *Reviews of Geophysics*, Vol. 25, No. 2, pp. 119–24.

STEDINGER, J.R. and COHN, T.A. 1986. Flood Frequency analysis with historical and paleoflood information. *Water Resources Research*, Vol. 22, No. 5, pp. 785–93.

STEDINGER, J.R. and LU, L. 1995. Appraisal of regional and index flood quantile estimators. *Stochastic Hydrology and Hydraulics*, Vol. 9, No. 1, pp. 49–75.

STEDINGER, J.R.; PEI, D. and COHN, T.A. 1985. A disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations. *Water Resources Research*, Vol. 21, No. 5, pp. 665–7.

STEDINGER, J.R. and TAYLOR, M.R. 1982a. Synthetic streamflow generation. Part 1: model verification and validation. *Water Resources Research*, Vol. 18, No. 4, pp. 919–24.

STEDINGER, J.R. and TAYLOR, M.R. 1982b. Synthetic streamflow generation. Part 2: effect of parameter uncertainty. *Water Resources Research*, Vol. 18, No. 4, pp. 919–24.

STEDINGER, J.R. and VOGEL, R. 1984. Disaggregation procedures for the generation of serially correlated flow vectors. *Water Resources Research*, Vol. 20, No. 1, pp. 47–56.

STEDINGER, J.R.; VOGEL, R.M. and FOUFOULA-GEORGIOU, E. 1993. Frequency analysis of extreme events, In: D. Maidment (ed.), *Handbook of hydrology*, Chapter 18. New York, McGraw-Hill.

STEPHENS, M. 1974. Statistics for Goodness of Fit, *Journal of the American Statistical Association*, Vol. 69, pp. 730–37.

TAO, P.C. and DELLEUR, J.W. 1976. Multistation, multi-year synthesis of hydrological time series by disaggregation. *Water Resources Research*, Vol. 12, No. 6, pp. 1303–11.

TARBOTON, D.G.; SHARMA, A. and LALL, U. 1998. Disaggregation procedures for stochastic hydrology based on nonparametric density estimation. *Water Resources Research*, Vol. 34, No. 1, pp. 107–19.

TASKER, G.D. and STEDINGER, J.R. 1986. Estimating generalized skew with weighted least squares regression. *Journal of Water Resources Planning and Management*, Vol. 112, No. 2, pp. 225–37.

THOM, H.C.S. 1958. A note on the gamma distribution. *Monthly Weather Review*, Vol. 86, No. 4, 1958. pp. 117–22.

THOMAS, H.A., JR. and FIERING, M.B. 1962. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In: A. Maass, M.M. Hufschmidt, R. Dorfman, H.A. Thomas, Jr., S.A. Marglin and G.M. Fair (eds), *Design of water resources systems*. Cambridge, Mass., Harvard University Press.

THYER, M. and KUCZERA, G. 2000. Modelling long-term persistence in hydroclimatic time series using a hidden state Markov model. *Water Resources Research*, Vol. 36, No. 11, pp. 3301–10.

VALDES, J.R.; RODRIGUEZ, I. and VICENS, G. 1977. Bayesian generation of synthetic streamflows 2: the multivariate case. *Water Resources Research*, Vol. 13, No. 2, pp. 291–95.

VALENCIA, R. and SCHAAKE, J.C., Jr. 1973. Disaggregation processes in stochastic hydrology. *Water Resources Research*, Vol. 9, No. 3, pp. 580–5.

VICENS, G.J.; RODRÍGUEZ-ITURBE, I. and SCHAAKE, J.C., Jr. 1975. A Bayesian framework for the use of regional information in hydrology. *Water Resources Research,* Vol. 11, No. 3, pp. 405–14.

VOGEL, R.M. 1987. The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional hypotheses. *Water Resources Research,* Vol. 22, No. 4, pp. 587–90.

VOGEL, R.M. and FENNESSEY, N.M. 1993. L-moment diagrams should replace product moment diagrams. *Water Resources Research*, Vol. 29, No. 6, pp. 1745–52.

VOGEL, R.M. and MCMARTIN, D.E. 1991. Probability plot goodness-of-fit and skewness estimation procedures for the Pearson type III distribution. *Water Resources Research*, Vol. 27, No. 12, pp. 3149–58.

VOGEL, R.M. and SHALLCROSS, A.L. 1996. The moving blocks bootstrap versus parametric time series models. *Water Resources Research*, Vol. 32, No. 6, pp. 1875–82.

VOGEL, R.M. and STEDINGER, J.R. 1988. The value of stochastic streamflow models in over-year reservoir design applications. *Water Resources Research*, Vol. 24, No. 9, pp. 1483–90.

WALLIS, J.R. 1980. Risk and uncertainties in the evaluation of flood events for the design of hydraulic structures. In: E. Guggino, G. Rossi and E. Todini (eds), *Piene e Siccita,* Catania, Italy, Fondazione Politecnica del Mediterraneo. pp. 3–36.

WALLIS, J.R.; MATALAS, N.C. and SLACK, J.R. 1974a. Just a Moment! *Water Resources Research*, Vol. 10, No. 2, pp. 211–19.

WALLIS, J.R.; MATALAS, N.C. and SLACK, J.R. 1974b. *Just a moment! Appendix*. Springfield, Va., National Technical Information Service, PB-231 816.

WANG, Q. 2001. A. Bayesian joint probability approach for flood record augmentation. *Water Resources Research*, Vol. 37, No. 6, pp. 1707–12.

WANG, Q.J. 1997. LH moments for statistical analysis of extreme events. *Water Resources Research*, Vol. 33, No. 12, pp. 2841–48.

WILK, M.B. and GNANADESIKAN, R. 1968. Probability plotting methods for the analysis of data. *Biometrika*, Vol. 55, No. 1, pp. 1–17.

WILKS, D.S. 2002. Realizations of daily weather in forecast seasonal climate. *Journal of Hydrometeorology*, No. 3, pp. 195–207.

WILKS, D.S. 1998. Multi-site generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, No. 210, pp. 178–91.

YOUNG, G.K. 1968. Discussion of 'Mathematical assessment of synthetic hydrology' by N.C. Matalas and reply. *Water Resources Research*, Vol. 4, No. 3, pp. 681–3.

ZELLNER, A. 1971. An introduction to Bayesian inference in econometrics, New York, Wiley.

ZRINJI, Z. and BURN, D.H. 1994. Flood Frequency analysis for ungauged sites using a region of influence approach. *Journal of Hydrology*, No. 13, pp. 1–21.