HANDBOOK OF HYDROLOGY

David R. Maidment
Editor in Chief

Professor of Civil Engineering
University of Texas at Austin

McGRAW-HILL, INC.
New York San Francisco Washington, D.C. Auckland Bogotá
Caracas Lisbon London Madrid Mexico City
Milan Montreal New Delhi San Juan
Singapore Sydney Tokyo Toronto

CONTENTS

Part 2 Hydrologic Transport

Part 3 Hydrologic Statistics

Part 4 Hydrologic Technology

CHAPTER 19

ANALYSIS AND MODELING OF HYDROLOGIC TIME SERIES

Jose D. Salas
Engineering Research Center
Colorado State University
Fort Collins, Colorado

Time-series analysis has become a major tool in hydrology. It is used for building mathematical models to generate synthetic hydrologic records, to forecast hydrologic events, to detect trends and shifts in hydrologic records, and to fill in missing data and extend records. This chapter includes definitions related to the stochastic structure of hydrologic series; time-series analysis principles; stochastic models for single and multiple series, disaggregation models, Markov chains, and point process models; methods for filling in missing data and extension of records; data generation and simulation principles; and a summary of software available for analysis, modeling, and generation of hydrologic series. Applications of stochastic models to forecasting are included in Chap. 26.

19.1 STOCHASTIC STRUCTURE OF HYDROLOGIC TIME SERIES

19.1.1 Hydrologic Time Series

In general, hydrologic processes such as precipitation and runoff evolve on a continuous time scale. For instance, a recording gauging station in a stream provides a continuous record of water stage and discharge $y(t)$ through time. A plot of the flow hydrograph $y(t)$ versus time t constitutes a *stream-flow time series* in continuous time or a *continuous time series* (see Fig. 19.1.1*a* for illustration). However, most hydrologic processes of practical interest are defined in a discrete time scale. A *discrete time series* may be derived by sampling the continuous process $y(t)$ at discrete points in time, or by integrating the continuous time series over successive time intervals as shown in Fig. 19.1.1*a*. For example, a daily stream-flow series may be derived by sampling the flows of a stream once daily or by integrating the continuous-flow hydrograph on a daily basis. Most hydrologic series are defined on hourly, daily,

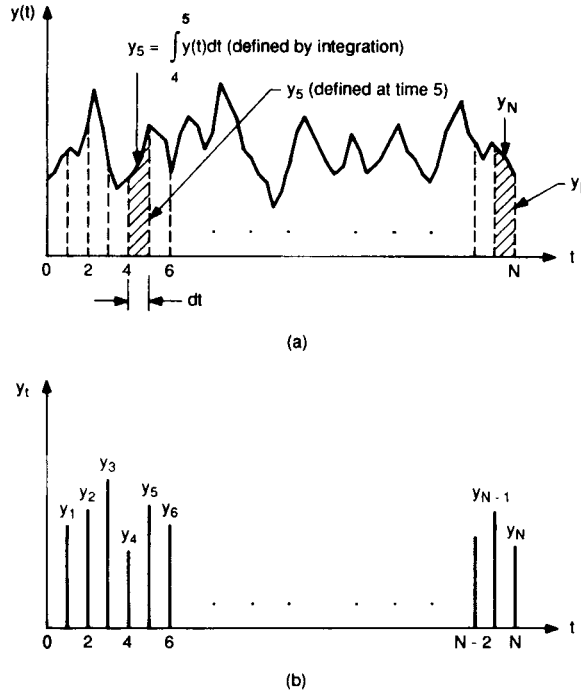

FIGURE 19.1.1 (*a*) A continuous time series $y(t)$. (*b*) A discrete time series y_t derived from the continuous series.

weekly, monthly, bimonthly, quarterly, and annual time intervals. The term *seasonal time series* is used for series defined at time intervals which are fractions of a year (usually multiples of a month). Figure 19.1.1*b* plots a discrete time series in the form of a bar or stick diagram; however, because of convenience or preference, the series is often plotted in the form of a continuous line by successively connecting the tops of the sticks. A continuous plot of a discrete time series should not be confused with a continuous time series.

Hydrologic time series may be classified into several categories depending on a number of factors. Each of these categories is defined below.

Single Time Series. A *single time series* (or *univariate series*) is simply a time series of one hydrologic variable at a given site. Consider a basin with five precipitation gauges and a stream network system with three stream-flow gauging stations. The precipitation time series measured at each site is a single time series. Likewise, the series resulting from the areal average of the five precipitation series is also a single time series. Similarly, the flow time series at any given site of the stream network system is a single time series.

Multiple Time Series. Consider the basin and flow network referred to above. The set of five precipitation time series of the basin represents a multivariate time series. Likewise, the set of three flow time series represents a multivariate series. In general, a

set of two or more time series constitutes a *multiple time series* or a *multivariate time series.* Furthermore, multiple time series may be a set of time series of different processes. For instance, the flow time series at sites 1, 2, and 3 and the corresponding precipitation time series at gauges 1 through 5 constitute a multiple time series. Additionally, a multiple time series may arise at a stream-flow gauging station when the station measures different variables such as discharge, flow depth, water temperature, and sediment transport or at a given weather station when it measures variables such as precipitation, air temperature, evaporation, and humidity.

Uncorrelated and Correlated Time Series. Figure 19.1.2 shows a single time series x_t. If the x's at time t depend (linearly) on the x's at time $t-k$, for $k = 1, 2, \ldots$, then the time series is called *autocorrelated, serially correlated,* or *correlated in time.* Otherwise, it is *uncorrelated.* An uncorrelated series is also called an *independent series.* Autocorrelation or *dependence* in some hydrologic time series such as stream flow usually arises from the effect of storage, such as surface, soil, and groundwater storages, which causes the water to remain in the system through subsequent time periods. For instance, basins with significant surface storage in the form of lakes, swamps, or glaciers produce stream-flow series showing significant autocorrelation. Likewise, subsurface storage, especially groundwater storage, produces significant autocorrelation in the stream-flow series derived from groundwater outflow. Conversely, time series of monthly or annual precipitation and time series of annual maximum flows (flood peaks) are usually uncorrelated, although in cases that a time series is *nonhomogeneous,* significant serial correlation may occur.[136]

Refer to the two series of Fig. 19.1.2. If the y's at time t depend (linearly) on the x's at time $t-k$, for $k = 0, 1, \ldots$ —then the two time series are *cross-correlated.* Several combinations of autocorrelation and cross-correlation exist. For instance, it is possible that both series y_t and x_t are uncorrelated in time, yet are cross-correlated with one another. Likewise, it is possible that each series can be autocorrelated, yet

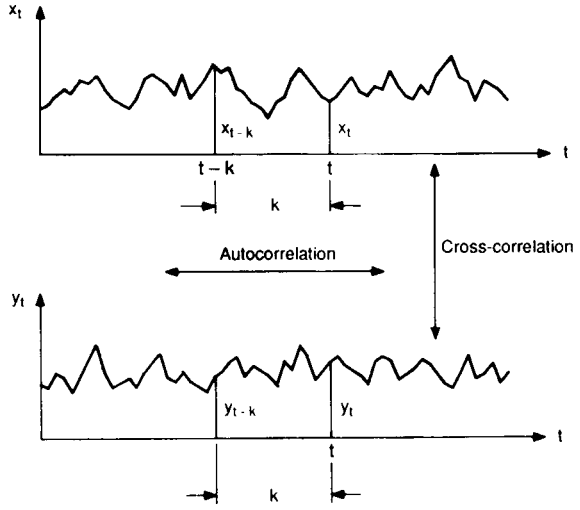

FIGURE 19.1.2 A pair of single time series x_t and y_t each having autocorrelation. Cross-correlation is between the two series.

there is no cross-correlation between them. Just as there are physical reasons why some hydrologic time series are autocorrelated, there are also physical reasons why two or more series are cross-correlated. Examples are precipitation series at two nearby sites and stream flow at two nearby gauging stations. In both cases, one would expect that the time series will be cross-correlated because the sites are relatively close to each other and therefore subject to similar climatic and hydrologic events. As the sites considered become father apart, their cross-correlation decreases. Likewise, one would expect a significant cross-correlation between stream-flow time series and the corresponding areal average precipitation time series over the same basin. One of the problems in hydrology is searching for significant correlation among time series.

Intermittent Time Series. Hydrologic time series are *intermittent* when the variable under consideration takes on nonzero and zero values throughout the length of record. For instance, the precipitation observed in a recording rain gauge is an intermittent continuous time series. Likewise, a discrete time series derived by integrating an intermittent continuous precipitation time series can be intermittent when the time interval of integration is relatively small. Thus, hourly, daily, and weekly rainfall are typically intermittent time series, while monthly and annual rainfall are usually nonintermittent. However, in semiarid and arid regions, even monthly and annual precipitation may be intermittent as well. Figure 19.1.3 shows a six-hourly rainfall series for a given gauging station and the corresponding stream flow series at two gauging stations. The rainfall series is intermittent, displaying a sequence of nonzero and zero rainfalls, while the stream-flow series is nonintermittent, with nonzero flows throughout the record. Stream-flow time series are often intermittent in semiarid and arid regions.

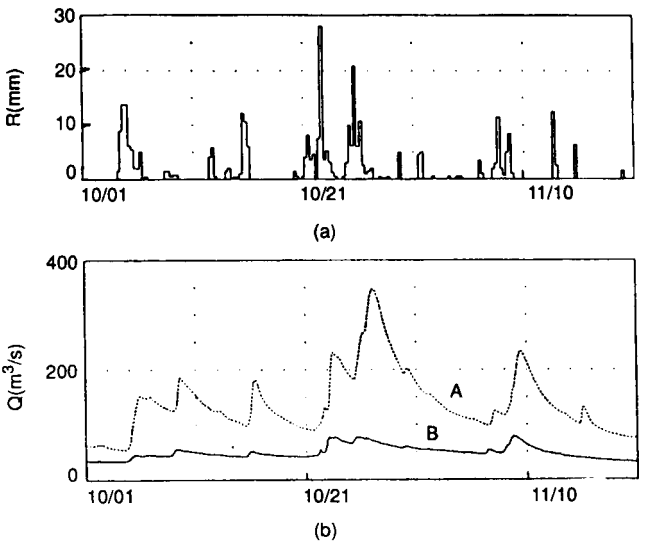

FIGURE 19.1.3 (*a*) Six-hour rainfall R, mm, during the period Oct. 1–Nov. 15, 1983, at site Te Haroto, Mohaka, New Zealand. (*b*) Corresponding stream flow Q, m^3/s, on the Mohaka River at Raupunga (A) and Glenfalls (B), New Zealand. *(Provided by S. M. Thompson.)*

Counting Time Series. The variable of interest may be the result of counting the occurrence of certain hydrologic events. An example is the count of rainy days for each month throughout the period of record. The resulting sequence of integer numbers $d_1, d_2, \ldots, d_n$ is a *counting time series.*

Regularly and Irregularly Spaced Time Series. Most time series are defined on a *regularly spaced* time interval; i.e., the value of the variable has been determined every hour, every day, or every week, etc., throughout the record. This is the case for most variables which are of interest in hydrology. However, in some cases, data may be collected at irregular time intervals. This is commonly true of water-quality measurements. Nearly all methods of time-series analysis require regularly spaced data, but some methods, such as the use of regression for trend analysis, can also be applied to irregularly spaced data. In this chapter it is assumed that the time series under consideration has been defined on a regular time scale.

Stationary and Nonstationary Time Series. A hydrologic time series is *stationary* if it is free of trends, shifts, or periodicity (cyclicity). This implies that the statistical parameters of the series, such as the mean and variance, remain constant through time. Otherwise, the time series is *nonstationary.* Generally, hydrologic time series defined on an annual time scale are stationary, although this assumption may be incorrect as a result of large-scale climatic variability, natural disruptions like a volcanic eruption, and human-induced changes such as the effect of reservoir construction on downstream flow. Hydrologic time series defined at time scales smaller than a year, such as monthly series, are typically nonstationary, mainly because of the annual cycle.

19.1.2 Partitioning of the Time-Series Structure

Hydrologic time series exhibit, in various degrees, trends, shifts or jumps, seasonality, autocorrelation, and nonnormality. These attributes of hydrologic time series are referred to as *components.* A time series can be *partitioned* or *decomposed* into its component series.

Trends and Shifts. In general, natural and human-induced factors may produce gradual and instantaneous trends and shifts (jumps) in hydrologic series. For example, a large forest fire in a river basin can immediately affect the runoff, producing a shift in the runoff series, whereas a gradual killing of a forest (for instance by an insect infestation that takes years for its population to build up) can result in gradual changes or trends in the runoff series (see Chap. 13 for further details). A large volcanic explosion such as the 1980 Mount St. Helens explosion, or a large landslide, can produce sudden changes in the sediment transport series of a stream. Trends in non-point-source water-quality series may be the result of long-term changes in agricultural practices and agricultural land development. Likewise, shifts in certain water-quality constituents may be caused by agricultural activities such as sudden changes in the use of certain types of pesticides. An important source of trends and shifts in stream-flow series arises from changes in land use and the development of reservoirs and diversion structures. The current concern about global warming and climatic changes is making hydrologists more aware of the occurrence of trends and shifts in hydrologic time series. Figure 19.1.4 shows the monthly series of water levels of Lake Victoria at Entebbe (Uganda), which has a significant upward shift.

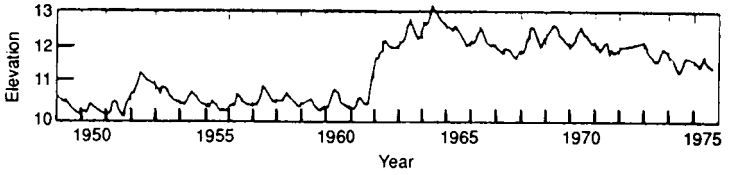

FIGURE 19.1.4 Time series of monthly levels of Lake Victoria at Entebbe, Uganda, for the period 1949–1975 showing an upward shift.[154]

Removing Trends. A hydrologic time series may exhibit shifts in one or more of its statistical characteristics. The most common ones are trends in the mean and in the variance. The partitioning of a time series with a simple trend is schematically shown in Fig. 19.1.5. A linear trend in the mean is shown in Fig. 19.1.5*a*. The trend $\bar{y}_t$ can be removed by the difference $y_t - \bar{y}_t$ as shown in Fig. 19.1.5*b*. The variance of such difference series, expressed by s_t^2, may be either a function of time (in which case there is a trend in the variance) or may be a constant, as shown graphically in Fig. 19.1.5*c*. The trend in the variance can be removed by $(y_t - \bar{y}_t)/s_t$ (the process of

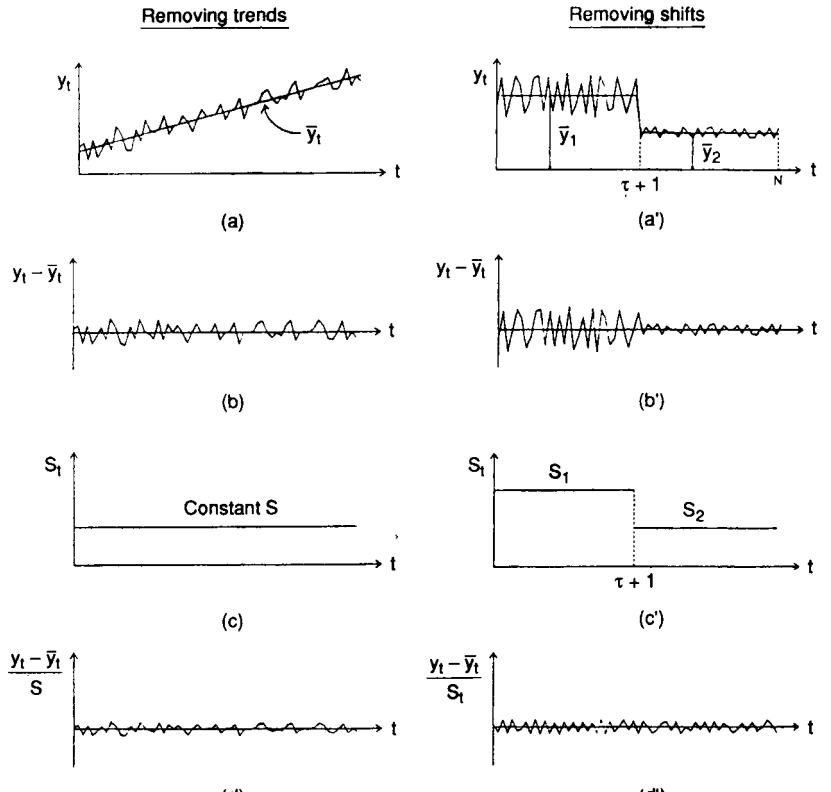

FIGURE 19.1.5 Partitioning an annual time series with trends and partitioning an ann... series with shifts.

constructing a new series by subtracting the mean and dividing by the standard deviation is called *standardization*). The residual series in Fig. 19.1.5*d* may still have other properties such as correlation structure which can be decomposed or removed.

Removing Shifts. A hydrologic series may also exhibit shifts in one or more of its statistical characteristics. Positive (upward) or negative (downward) shifts in the mean and variance are most commonly analyzed. Figure 19.1.5*a'* to *d'* gives examples of removing sudden downward shifts from an annual series. In Fig. 19.1.5*a'*, shifts in the mean and variance occur at time $\tau + 1$. The shift in the mean is removed by $y_t - \bar{y}_t$ as shown in Fig. 19.1.5*b'*, and the shift in the variance is removed by $(y_t - \bar{y}_t)/s_t$. The residual series plotted in Fig. 19.1.5*d'* now shows a series with mean zero and variance one and may further exhibit other properties such as autocorrelation.

Seasonality. Hydrologic series defined at time intervals smaller than a year (such as monthly series) generally exhibit distinct *seasonal* (or *periodic*) patterns. These result from the annual revolution of the earth around the sun which produces the annual cycle in most hydrologic processes. Some series of interest to hydrology, such as daily series of urban water use or daily series of hydropower generation, may also exhibit a *weekly pattern* due to variations of demands within a week. Likewise, hourly time series may have a distinct *diurnal pattern* due to the variations of demands within a day. Summer hourly rainfall series or certain water-quality constituents related to temperature may also exhibit distinct diurnal patterns due to the daily rotation of the earth which causes variations within the day of net radiation.

Seasonal or periodic patterns of hydrologic series translate into statistical characteristics which vary within the year (or within a week or a day as the case may be). Generally seasonal or periodic variations in the mean, variance, covariance, and skewness are important. Figure 19.1.6 shows how seasonal series are partitioned into basic components (the annual series is also shown for comparison). A part of the original time series y_t is plotted in Fig. 19.1.6*a'*, in which the seasonal (periodic) pattern is evident. It is a *periodic-stochastic* series since, in addition to the periodic pattern, a random pattern is also observed. This periodic-stochastic pattern repeats through time in a similar fashion. In contrast, the annual series in Fig. 19.1.6*a* does not show a periodic pattern; it simply varies about a constant mean $\bar{y}$ (see Fig. 19.1.6*b*). The fact that the series in Fig. 19.1.6*a'* behaves in a cyclic fashion implies that the mean of the series is also cyclic or periodic, as shown in Fig. 19.1.6*b'*. For instance, for monthly stream-flow series, each month will have its own mean $\bar{y}_t$ (refer to Sec. 19.2 for the definition of seasonal statistics).

Removing Seasonality in the Mean and Variance. Removing the seasonality in the mean is accomplished by taking the difference $y_t - \bar{y}_t$, where $\bar{y}_t$ is the monthly mean for January, February, . . . , if t is a monthly index. When this difference is plotted in Fig. 19.1.6*c'*, the series fluctuates about zero with a particular pattern. The variability of the series is initially small, then increases, and then decreases. This pattern repeats in the second year and subsequent years throughout the record. If such variability is measured by the variance s_t^2 in each time interval in the year (for instance, one variance for each month in the year for a monthly series) and s_t is plotted as in Fig. 19.1.6*d'*, it will exhibit a seasonal (periodic) pattern similar to that of the mean in Fig. 19.1.6*b'*. In contrast, for the annual series $y_t - \bar{y}_t$ in Fig. 19.1.6*c*, the variance s^2 is a constant. The seasonality in the variance can be removed by $(y_t - \bar{y}_t)/s_t$. This operation is also called *seasonal standardization* and often is referred to in literature as *deseasonalizing* the original series. Actually, this latter term

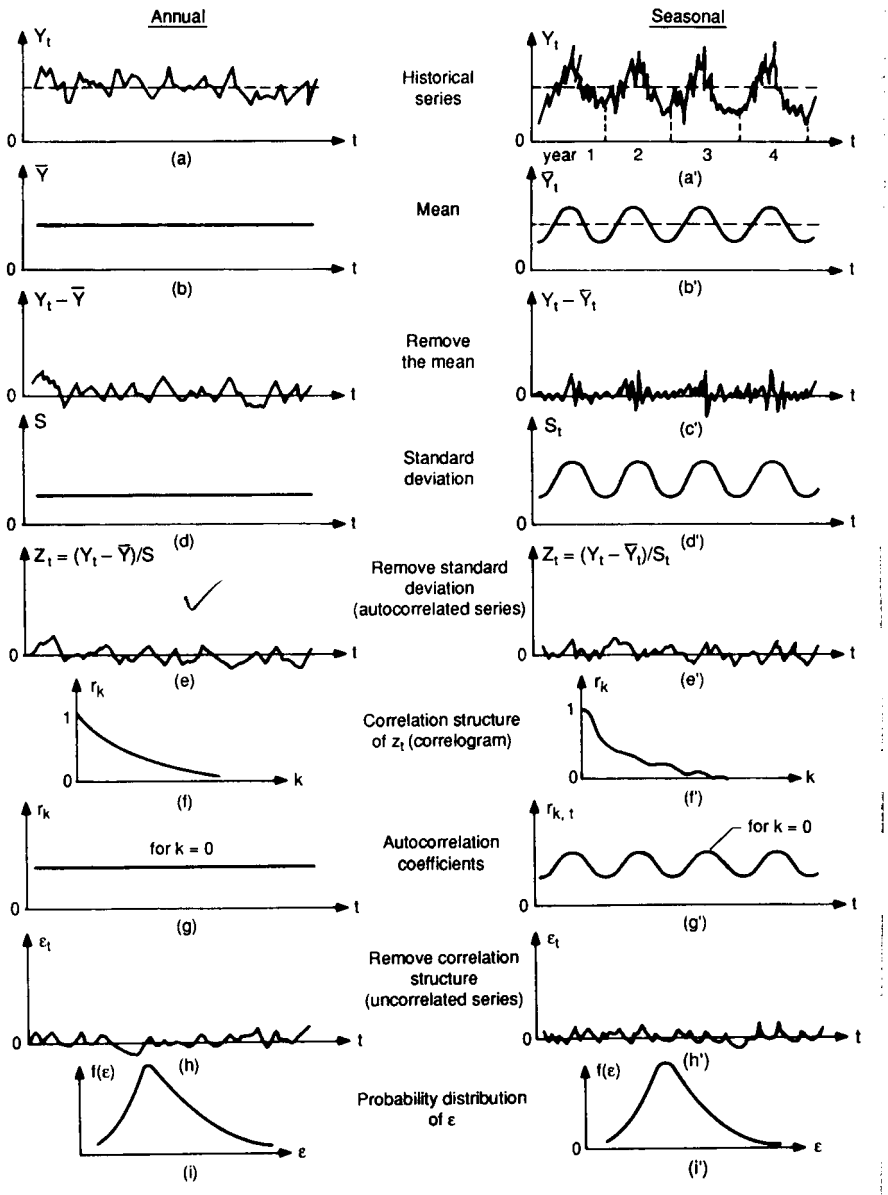

FIGURE 19.1.6 Partitioning an annual and a seasonal time series into their various components.

is a misnomer since it may imply that $z_t = (y_t - \bar{y}_t)/s_t$ is free of seasonality. However, other seasonalities may still be present in z_t.

Removing Seasonality in the Correlation. Unlike the seasonality in the mean and the variance, other remaining seasonalities are difficult to observe from the plot of the time series. Further analysis must be done to detect them. The series z_t in Fig. 19.1.6*e* and *e'* may have a dependence structure or autocorrelation. This is described by the *correlogram* r_k, which is a plot of the autocorrelation coefficient as a function of lag k, as shown in Fig. 19.1.6*f* and *f'*. However, a better way of analyzing the dependence structure of series z_t in Fig. 19.1.6*e'* is by determining the correlation on a season-by-season basis. For instance, for monthly series, one would correlate the February observations for all years of record with the corresponding January observations to obtain $r_{1,2}$, similarly for other months to obtain $r_{1,1}, \ldots, r_{1,12}$ and, in general, $r_{k,1}, \ldots, r_{k,12}$. The plot of $r_{k,t}$ for $k > 0$ may, depending on the hydrologic series under consideration, exhibit a seasonal or periodic pattern as in Fig. 19.1.6*g'*. In contrast, for annual series, the correlation coefficient r_k is assumed to remain constant, as depicted in Fig. 19.1.6*g*.

Regardless of whether the autocorrelation is constant or periodic, removing such correlation structure from the series requires a mathematical model to represent the correlation. A simple model that may be considered for series z_t in Fig. 19.1.6*e* is $z_t = r_1\ z_{t-1} + \varepsilon_t$, the *lag-one autoregressive process.* The residual series $\varepsilon_t = z_t - r_1 z_{t-1}$ is uncorrelated or free of autocorrelation. A similar operation, although it uses a different model, can be used to remove the periodic correlation structure of z_t in Fig. 19.1.6*e'*. Thus, the original seasonal series y_t has been partitioned into components, periodic mean, periodic variance, and periodic correlation, and these components have been removed from y_t, yielding a residual series ε_t free of periodicities. The residual series ε_t of Fig. 19.1.6*h* or *h'* is represented by a probability distribution function $f(\varepsilon)$ which may be normal or nonnormal.

19.2 TIME-SERIES ANALYSIS PRINCIPLES

This section addresses the estimation of a number of statistical properties of annual and seasonal hydrologic time series. In addition, the detection and estimation of trends, shifts, seasonality, and nonnormality are briefly discussed, as are some procedures for transforming a skewed series into a normally distributed series.

19.2.1 Statistical Properties of Time Series

Overall Sample Statistics. The *mean* and the *variance* of a time series y_t are estimated by

$$\bar{y} = \left(\frac{1}{N}\right) \sum_{t=1}^{N} y_t \tag{19.2.1}$$

$$s^2 = \frac{1}{N-1} \sum_{t=1}^{N} (y_t - \bar{y})^2 \tag{19.2.2}$$

respectively, where N = sample size. The square root of the variance is the *standard deviation* s, and $c_v = s/\bar{y}$ is the *coefficient of variation.* Likewise, the *skewness coeffi-*

cient is estimated by

$$g = \frac{N \sum_{t=1}^{N} (y_t - \bar{y})^3}{(N-1)(N-2)s^3} \tag{19.2.3}$$

in which $\bar{y}$ and s are as defined above. Since the sample statistics $\bar{y}$, s^2, and g are *estimators* of the population statistics μ, σ^2, and γ, respectively, sometimes the notations $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\gamma}$ are used to represent the sample statistics.

The *sample autocorrelation function* r_k (or autocorrelation coefficients) of a time series may be estimated by[80]

$$r_k = \frac{c_k}{c_0} \tag{19.2.4}$$

$$c_k = \left(\frac{1}{N}\right) \sum_{t=1}^{N-k} (y_{t+k} - \bar{y})(y_t - \bar{y}) \qquad k \geq 0 \tag{19.2.5}$$

The estimator r_k of Eq. (19.2.4) is an estimator of the population autocorrelation coefficient ρ_k. The plot of r_k versus k is the *correlogram*. Sometimes, the correlogram is used for choosing the type of stochastic model to represent a given time series. The lag-one serial correlation coefficient r_1 is a simple measure of the degree of time dependence of a series. When the correlogram decays rapidly to zero after a few lags, it may be an indication of small persistence or *short memory* in the series, while a slow decay of the correlogram may be an indication of large persistence or *long memory*. This short or long memory is shown schematically in Fig. 19.2.1.

Corrections for Bias. An estimator of a population statistic is *biased* when its mean value is different from the population statistic. The estimator r_k of Eq. (19.2.4) is a biased (downward) estimator of ρ_k; that is, the average value of r_k estimated by the formula is less than the true value ρ_k. Several procedures have been suggested to correct the values of r_k from limited samples.[195,206] Wallis and O'Connell[195] suggested the following correction to obtain an unbiased estimator of ρ_1

$$\hat{\rho}_1 = \frac{r_1 N + 1}{N - 4} \tag{19.2.6}$$

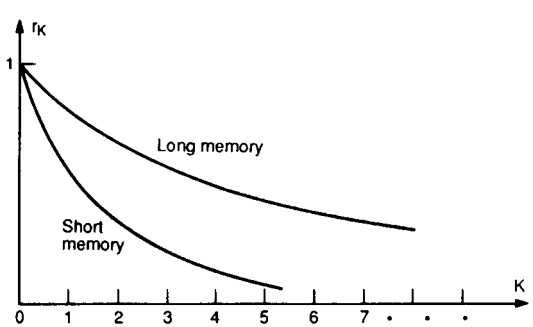

FIGURE 19.2.1 Schematic representation of a correlogram with short and long memory.

where N is the sample size. Likewise, s^2 of Eq. (19.2.2) is an unbiased estimator of σ^2 if the series is uncorrelated. When the series is autocorrelated, the effective sample size N is reduced. In this case, an unbiased estimator of the variance can be determined by[114,130]

$$\hat{\sigma}^2 = \frac{(N-1)\,s^2}{N-K} \tag{19.2.7}$$

where

$$K = \frac{[N(1-\hat{\rho}_1^2) - 2\,\hat{\rho}_1\,(1-\hat{\rho}_1^N)]}{[N(1-\hat{\rho}_1)^2]} \tag{19.2.8}$$

and s^2 and $\hat{\rho}_1$ are given by Eqs. (19.2.2) and (19.2.6), respectively.

The estimator g of Eq. (19.2.3) is an unbiased estimator of γ if the time series is uncorrelated and normally distributed. For nonnormal series, Bobee and Robitaille[9] proposed the unbiased estimator

$$\hat{\gamma}_0 = \frac{Lg[A + Bg^2(L^2/N)]}{\sqrt{N}} \tag{19.2.9a}$$

$$A = 1 + 6.5N^{-1} - 20.2N^{-2} \tag{19.2.9b}$$

$$B = 1.48N^{-1} + 6.77N^{-2} \tag{19.2.9c}$$

$$L = \frac{N-2}{\sqrt{(N-1)}} \tag{19.2.9d}$$

where $\pm L$ is the theoretical limit on skewness from a sample of size N.[88] Furthermore, for gamma correlated series, Fernandez and Salas[39] gave the unbiased estimator

$$\hat{\gamma} = \frac{\hat{\gamma}_0}{1 - 3.12\hat{\rho}_1^{3.7}N^{-0.49}} \tag{19.2.10}$$

in which $\hat{\gamma}_0$ and $\hat{\rho}_1$ are determined by Eqs. (19.2.9*a*) and (19.2.6), respectively.

Observed Annual Statistics. The overall sample statistics $\bar{y}$, s^2, c_v, g, and r_k are usually determined for annual hydrologic time series. Coefficients of variation c_v of annual flows are typically smaller than one, although they may be close to one or greater than one in streams in arid and semiarid regions. From an analysis of the annual flows of 126 rivers, McMahon and Mein[117] report a median value of c_v of 0.25. Coefficients of skewness g of annual flows are typically greater than zero. In some streams, small values of g are found, suggesting that annual flows are approximately normally distributed. On the other hand, in some streams of arid and semiarid regions, g can be greater than one. A range of g between -0.4 and about 2.0 and a median value of 0.40 has been reported.[117] Similarly, r_1 of annual flows are generally small but positive, although in some cases, because of sample variability, the r_1's are negative. It is quite typical to find values of r_1 in the range of $+0.0$ to 0.4 for annual stream-flow series. Yevjevich[202] found that, for a large number of rivers worldwide, the average value of r_1 was about 0.15, while McMahon and Mein[117] found a range of r_1 between -0.2 and 0.8 with a mean value of 0.23. Large values of r_1 for annual flows can be found for a number of reasons, including the effect of natural or manmade

surface storage such as lakes, reservoirs, or glaciers, the effect of slow groundwater storage response, and the effect of nonstationarities. Figure 19.2.2 shows a slow-decaying correlogram r_k for the annual flows of the White Nile River at Mongalla and a fast decaying r_k for the Blue Nile River at Khartoum, while the r_k for the Nile River at Aswan lies between the other two.

Seasonal Sample Statistics. Seasonal hydrologic time series, such as monthly flows, may be better described by considering statistics on a seasonal basis. Let the seasonal time series $y_{\nu,\tau}$, in which ν = year; τ = season; $\nu = 1, \ldots, N$; and $\tau = 1, \ldots, \omega$, with N and ω denoting the number of years of record and the number of seasons per year, respectively. The seasonal mean $\bar{y}_\tau$ is obtained by applying Eq. (19.2.1) for each season τ as

$$\bar{y}_\tau = \frac{1}{N}\sum_{\nu=1}^{N} y_{\nu,\tau} \qquad \tau = 1, \ldots, \omega \tag{19.2.11}$$

Likewise, Eqs. (19.2.2) and (19.2.3) can be applied on a seasonal basis to determine the seasonal variance s_τ^2 and the seasonal skewness coefficient g_τ, respectively, for $\tau = 1, \ldots, \omega$. Furthermore, the season-to-season correlation coefficient $r_{k,\tau}$ is determined by

$$r_{k,\tau} = \frac{c_{k,\tau}}{(c_{0,\tau}\, c_{0,\tau-k})^{1/2}} \tag{19.2.12}$$

$$c_{k,\tau} = \frac{1}{N}\sum_{\nu=1}^{N} (y_{\nu,\tau} - \bar{y}_\tau)(y_{\nu,\tau-k} - \bar{y}_{\tau-k}) \tag{19.2.13}$$

For instance, for monthly stream-flow time series, $r_{1,4}$ represents the correlation between the flows of the fourth month with those of the third month.

Description of Seasonal Statistics. Each of the statistics $\bar{y}_\tau$, s_τ, g_τ, and $r_{k,\tau}$ may be plotted versus time $\tau = 1, \ldots, \omega$ to observe whether they exhibit a seasonal pat-

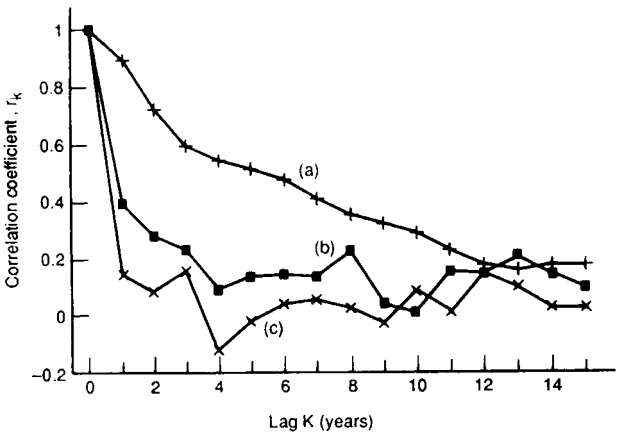

FIGURE 19.2.2 Correlogram of annual flows of (*a*) the White Nile River at Mongalla (1914–1983), (*b*) the Nile River at Aswan (1871–1989), and (*c*) the Blue Nile River at Khartoum (1912–1989).

tern (see Sec. 19.2.4 for testing for seasonality). These statistics may be fitted by Fourier series. This technique is especially effective with weekly and daily data.[155,206] Generally, for seasonal stream-flow series, $\bar{y}_\tau > s_\tau$, although for some streams $\bar{y}_\tau$ may be smaller than s_τ for the low-flow seasons. Furthermore, for intermittent stream-flow series, generally $\bar{y}_\tau < s_\tau$ throughout the year. Likewise, values of the skewness coefficient g_τ for the dry season are generally larger than those for the wet season, indicating that data in the dry season depart more from normality than those in the wet season. Values of the skewness for intermittent hydrologic series are usually larger than skewness for similar nonintermittent series. Seasonal correlations $r_{k,\tau}$ for stream flow during the dry season are generally larger than those for the wet season, and seasonal correlations $r_{k,\tau}$ for monthly precipitation are generally not significantly different from zero.[148] The data interval must be less than a month (typically, less than a week) to find significant autocorrelation in precipitation series.

Figure 19.2.3 shows $\bar{y}_\tau$, s_τ, g_τ, and $r_{k,\tau}$, $\tau = 1, \ldots, 12$ for the monthly flows of the Nile River at Aswan. A well-defined seasonality is shown in each statistic, meaning that the monthly statistics for the low-flow season are significantly different from those of the high-flow season. Furthermore, the seasonal patterns of $\bar{y}_\tau$ and s_τ are in phase while g_τ is out of phase relative to $\bar{y}_\tau$ and s_τ. Also, a well-defined seasonality in $r_{1,\tau}$ is evident with the larger values in dry months and smaller values in wet months. Most of the correlations $r_{12,\tau}$ are significant, with seasonality in phase with $r_{1,\tau}$. The correlations of Fig. 19.2.3 indicate the complex, long-term dependence (long-memory) structure of monthly flows of the Nile River. The significant correlations $r_{12,\tau}$ shown for this river are not typical for monthly stream-flow series. In fact, these correlations usually will be small or not significant. Conversely, rivers that exhibit long-term autocorrelation in seasonal flows will exhibit long-term autocorrelation in annual flows.

Sample Statistics for Multiple Series. For multiple time series, in addition to the above-defined sample statistics such as the mean, variance, skewness, and autocorrelation, the cross-correlation between each pair of time series can be determined. Consider the time series $y_t^{(i)}$ and $y_t^{(j)}$ at sites i and j, respectively. The sample lag-k cross-correlation coefficient is

$$r_k^{ij} = \frac{c_k^{ij}}{(c_0^{ii}\, c_0^{jj})^{1/2}} \tag{19.2.14}$$

$$c_k^{ij} = \left(\frac{1}{N}\right) \sum_{t=1}^{N-k} (y_{t+k}^{(i)} - \bar{y}^{(i)})\,(y_t^{(j)} - \bar{y}^{(j)}) \qquad k \geq 0 \tag{19.2.15}$$

and c_k^{ij} is the *cross-covariance* between the two series. The coefficient r_k^{ij} is an estimator of the population cross-correlation coefficient ρ_k^{ij}. The plot of r_k^{ij} versus k is the *cross-correlogram* or sample *cross-correlation function.* For instance, r_k^{ij} for annual flows of the Nile River at Aswan and the Blue Nile River at Khartoum is shown in Fig. 19.2.4. The lag-zero cross-correlation has a high value $r_0^{ij} = 0.85$. For n time series, the values r_k^{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, n$ are the elements of the cross-correlation matrix $\mathbf{M}_k$ (n by n matrix). The lag-zero cross-correlation matrix $\mathbf{M}_0$ is a symmetric matrix with 1's in the diagonal. The diagonal elements of the lag-k cross-correlation matrix $\mathbf{M}_k$ for $k > 0$ are the lag-k serial correlation coefficients for each site.

Likewise, for seasonal time series, the season-to-season cross-correlation between sites can be determined. Consider the seasonal time series $y_{\nu,\tau}^{(i)}$ and $y_{\nu,\tau}^{(j)}$ for sites i and j, respectively. The lag-k seasonal cross-correlation coefficient $r_{k,\tau}^{ij}$ between both series

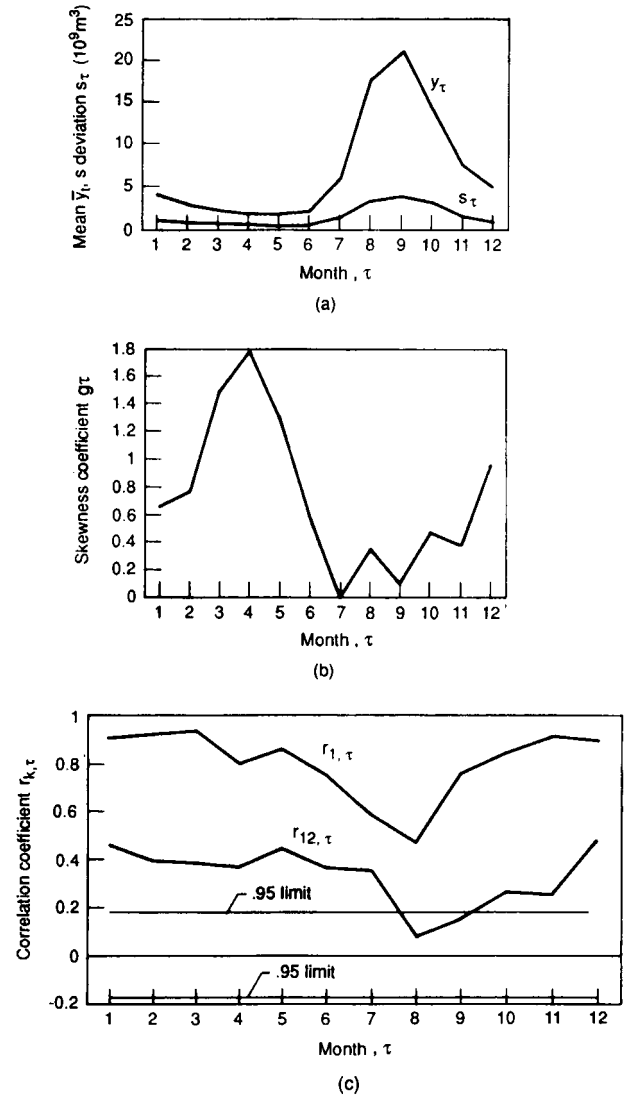

FIGURE 19.2.3 (*a*) Monthly mean $\bar{y}_\tau$ and standard deviation s_τ, (*b*) monthly skewness coefficient g_τ, and (*c*) lag-1 and lag-12 month-to-month correlations for monthly stream-flow series of the Nile River at Aswan (1871 – 1989).

can be determined by applying Eqs. (19.2.14) and (19.2.15) on a seasonal basis. For instance, for monthly stream flow, $r^{ij}_{1,4}$ is determined by correlating the stream-flow series of the fourth month of site i with those of the third month of site j.

Storage-Related Statistics. In modeling hydrologic time series for simulation studies of reservoir systems, storage-related statistics may be particularly important. Consider the hydrologic time series y_t, $t = 1, \ldots, N$ and a subsample $y_1, \ldots, y_n$

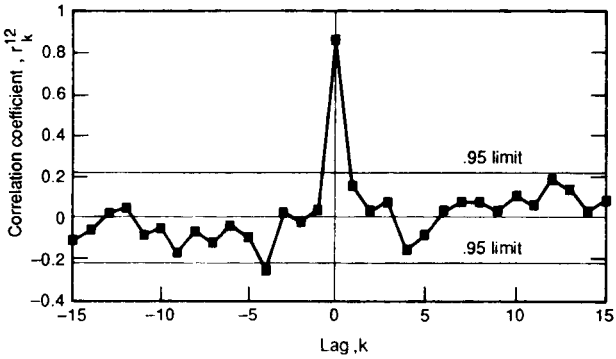

FIGURE 19.2.4 Cross-correlogram r_k^{12} of annual flows of the Nile River at Aswan (1) and the Blue Nile River at Khartoum (2), for the concurrent period 1912–1989.

with $n \leq N$. Form the sequence of partial sums S_i as

$$S_i = S_{i-1} + (y_i - \bar{y}_n) \qquad i = 1, \ldots, n \tag{19.2.16}$$

where $S_0 = 0$ and $\bar{y}_n$ is the sample mean. For instance, in the case of stream-flow time series, S_i represents the cumulative departure from the mean flow $\bar{y}_n$. The plot of S_i versus i, $i = 1, \ldots, n$ is the typical mass curve from which the minimum *storage capacity* D_n^* of a reservoir to deliver $\bar{y}_n$ through the time period n can be obtained.[41] For instance, Fig. 19.2.5 shows a mass curve for which $D_n^* = \max(D_1, \ldots, D_5)$. Related to D_n^* are the *range* R_n^* and the *rescaled range* R_n^{**} defined by

$$R_n^* = \max(S_0, S_1, \ldots, S_n) - \min(S_0, S_1, \ldots, S_n) \tag{19.2.17}$$

$$R_n^{**} = \frac{R_n^*}{s_n} \tag{19.2.18}$$

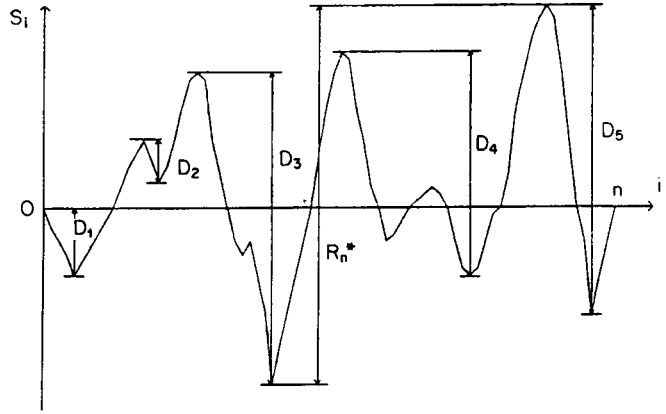

FIGURE 19.2.5 Mass curve S_i of Eq. (19.2.16), sequence of deficits $D_1, \ldots, D_5$, and range R_n^*.

in which s_n is the sample standard deviation. Thus, for a given sample of size N the ranges $R_2^*, \ldots, R_N^*$ and rescaled ranges $R_2^{**}, \ldots, R_N^{**}$ will result. Both R_n^* and R_n^{**} have been widely used in literature as measures of long-term dependence and for comparing alternative models of hydrologic series.[68,196,204]

In particular, Hurst[75] showed that for a large number of geophysical time series such as stream-flow, precipitation, temperature, and tree-ring series, the mean rescaled range $\bar{R}_n^{**}$ is proportional to n^h with $h > \frac{1}{2}$. The values of h obtained for different series gave a mean of about 0.73 and a standard deviation of 0.09. Theoretical results for normal, independent processes[37] and for autoregressive processes[110] indicated that asymptotically $h = \frac{1}{2}$ for these processes. The discrepancy between theoretical results stating that $h = \frac{1}{2}$ and Hurst's empirical finding that $h > \frac{1}{2}$ has become known as the *Hurst phenomenon.* Several estimators of h have been proposed and used in stochastic hydrology such as the original Hurst estimator K:[75]

$$K = \frac{\log (\bar{R}_n^{**})}{\log (n/2)} \tag{19.2.19}$$

It has been shown that the estimators of h are transient, meaning they depend on n, and, as $n \to \infty$, they generally converge to a limiting value, equal to $\frac{1}{2}$ for many time-series models.[153] One interpretation of the Hurst phenomenon has been to associate $h = \frac{1}{2}$ with short-memory models possessing short-term dependence structure and $h > \frac{1}{2}$ with long-memory models possessing long-term dependence. This is a valid interpretation in the asymptotic sense. However, such interpretation as a criterion for selecting stochastic models for hydrologic time-series simulation is not practical because a number of models, including ARMA processes (see Sec. 19.3 for definition of ARMA processes) can have long-term dependence structure, yet asymptotically their value of $h = \frac{1}{2}$. Nevertheless, estimates of h can be useful for comparing the performance of alternative modeling strategies and alternative estimation procedures.

Drought-Related Statistics. Drought-related statistics are also important in modeling hydrologic time series. Consider a hydrologic time series $y_t, t = 1, \ldots, N$ and a *demand level* d (also called a *crossing level*). Assume that y_t is an annual series and d is a constant (equal to the sample mean $\bar{y}$ or a fraction of $\bar{y}$) as shown in Fig. 19.2.6. A deficit occurs when $y_t < d$ consecutively during one or more years until $y_t > d$ again. Such a deficit can be defined by its duration L, by its magnitude M, and by its intensity $I = M/L$.[205] Since a number of deficits can occur in a given hydrologic

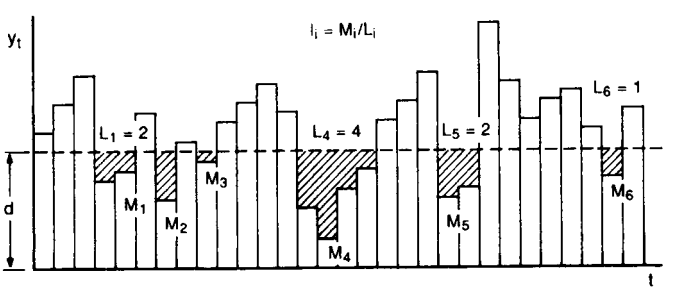

FIGURE 19.2.6 Definition of deficit length L_i, deficit magnitude M_i, and deficit intensity I_i.

sample (Fig. 19.2.6), the maximum deficit duration *(critical drought)* in a given sample $L^* = \max(L_1, \ldots, L_m)$ is often of most interest, where m is the number of deficits. Similar critical drought properties in relation to the magnitude and intensity are $M^* = \max(M_1, \ldots, M_m)$ and $I^* = \max(I_1, \ldots, I_m)$, respectively.

19.2.2 Determination and Testing of Trends

A number of parametric and nonparametric tests for trends have been suggested in the literature. This section includes one parametric and one nonparametric test. See also Sec. 17.2.3 of Chap. 17 for a discussion on the criteria for selecting parametric and nonparametric tests and Sec. 17.4 for trend analysis.

Detection and Estimation of Linear Trends. Assume that y_t, $t = 1, \ldots, N$ is an annual time series and N = sample size. A simple linear trend can be written as

$$y_t = a + bt \tag{19.2.20}$$

where a and b are the parameters of the regression model (see Sec. 17.4.2). Rejection of the hypothesis $b = 0$ can be considered as a detection of a linear trend. The hypothesis that $b = 0$ is rejected if

$$T_c = \left| \frac{\sqrt{N-2}}{r\sqrt{1-r^2}} \right| > T_{1-\alpha/2,\nu} \tag{19.2.21}$$

in which r is the cross-correlation coefficient between the sequences $y_1, \ldots, y_N$ and $1, \ldots, N$, and $T_{1-\alpha/2,\nu}$ is the $1 - \alpha/2$ quantile of the Student t distribution with $\nu = N - 2$ degrees of freedom. See Sec. 17.4.5 for the use of polynomial regression for nonlinear trends.

Mann-Kendall Test for Trends. This is a nonparametric test which tests for a trend in a time series without specifying whether the trend is linear or nonlinear. Consider the annual time series y_t, $t = 1, \ldots, N$. Each value $y_{t'}$, $t = 1, \ldots, N-1$ is compared with all subsequent values y_t, $t = t' + 1, t' + 2, \ldots, N$, and a new series z_k is generated by

$$\begin{aligned} z_k &= 1 \quad \text{if } y_t > y_{t'} \\ z_k &= 0 \quad \text{if } y_t = y_{t'} \\ z_k &= -1 \quad \text{if } y_t < y_{t'} \end{aligned} \tag{19.2.22}$$

in which $k = (t' - 1)(2N - t')/2 + (t - t')$. The Mann-Kendall statistic is given by the sum of the z_k series[71]

$$S = \sum_{t'=1}^{N-1} \sum_{t=t'+1}^{N} z_k \tag{19.2.23}$$

This statistic represents the number of positive differences minus the number of negative differences for all the differences considered.

The test statistic for $N > 40$ may be written as[71]

$$u_c = \frac{S + m}{\sqrt{V(S)}} \tag{19.2.24}$$

$$V(S) = \frac{1}{18}\left[N(N-1)(2N+5) - \sum_{i=1}^{n} e_i(e_i-1)(2e_i+5)\right] \qquad (19.2.25$$

where $m = 1$ if $S < 0$ and $m = -1$ if $S > 0$, n is the number of tied groups, and e_i the number of data in the ith (tied) group. The statistic u_c is assumed to be zero $S = 0$. Then the hypothesis of an upward or downward trend cannot be rejected the α significance level if $|u_c| > u_{1-\alpha/2}$, where $u_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of th standard normal distribution. Kendall[86] indicated that this test can be used even fo N as low as 10 if there are not too many tied values. Hirsch et al.[71] extended this test t seasonal time series.

Final Remarks. The Mann-Kendall test can be extended to test whether differen seasons exhibit trends in the same direction and of the same magnitude.[53,71] This tes has been applied to testing trends in water-quality time series.[190] Other nonparamet ric tests for trends useful for hydrologic time series have been suggested, such as th Hotelling-Pabst test[27] and the Sen test,[53] and there are also tests for seasonally corre lated data.[72] An excellent reference for nonparametric tests is Hollander and Wolfe.[7

19.2.3 Determination and Testing of Shifts (Jumps)

A number of parametric and nonparametric tests are available for testing and deter mining shifts in statistical properties of time series such as the mean and var iance.[53,164,173] One parametric and one nonparametric test are included here. Th reader is referred to Sec. 17.3.2 for further discussion concerning the assumption and applicability of tests for shifts.

t Test for Shift in the Mean. Suppose that $y_t, t = 1, \ldots, N$ is an annual hydrologi series which is uncorrelated and normally distributed with mean μ and standard deviation σ and N = sample size. The series is divided into two subseries of sizes N_1 and N_2 such that $N_1 + N_2 = N$. The first subseries $y_t, t = 1, 2, \ldots, N_1$, has mean μ_1 and standard deviation σ, and the second subseries $y_t, t = N_1 + 1, N_1 + 2, \ldots, N$ is assumed to have mean μ_2 and standard deviation σ. The simple t test can be used to test the hypothesis $\mu_1 = \mu_2$ when the two subseries have the same standard deviation σ. Rejection of the hypothesis can be considered as a detection of a shift. The tes statistic in this case is given by[106,173]

$$T_c = \frac{|\bar{y}_2 - \bar{y}_1|}{S\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}} \qquad (19.2.26$$

$$S = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N - 2}} \qquad (19.2.27$$

where $\bar{y}_1$ and $\bar{y}_2$ and s_1^2 and s_2^2 are the estimated means and variances of the first and the second subseries, respectively. The hypothesis $\mu_1 = \mu_2$ is rejected if $T_c > T_{1-\alpha/2,}$ where $T_{1-\alpha/2,\nu}$ is the $1 - \alpha/2$ quantile of the Student's t distribution with $\nu = N - 2$ degrees of freedom and α is the significance level of the test. Modifications of the tes are available when the variances in each group are different[173] and when the data exhibit some significant serial correlation.[106]

Mann-Whitney Test for Shift in the Mean. Suppose that y_t, $t = 1, \ldots, N$ is an annual hydrologic series that can be divided into two subseries $y_1, \ldots, y_{N_1}$ and $y_{N_1+1}, \ldots, y_N$ of sizes N_1 and N_2, respectively, such that $N_1 + N_2 = N$. A new series, z_t, $t = 1, \ldots, N$, is defined by rearranging the original data y_t in increasing order of magnitude. One can test the hypothesis that the mean of the first subseries is equal to the mean of the second subseries by using the statistic[173]

$$u_c = \frac{\sum_{t=1}^{N_1} R(y_t) - N_1(N_1 + N_2 + 1)/2}{[N_1 N_2 (N_1 + N_2 + 1)/12]^{1/2}} \tag{19.2.28}$$

where $R(y_t)$ is the rank of the observation y_t in ordered series z_t. The hypothesis of equal means of the two subseries is rejected if $|u_c| > u_{1-\alpha/2}$, where $u_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution and α is the significance level of the test. Equation (19.2.28) can be modified for the case of groups of values that are tied.[53]

Final Remarks. The foregoing tests are for a shift in the mean of a sample series when the point of change is known. However, when the point of change is not known, bayesian analysis can be used to detect the point of change and its amount.[105] Likewise, the foregoing tests assumed that the underlying series is uncorrelated. Although adjustments to the t test can be made for serial correlation,[106] intervention analysis is a more effective approach.[66] The F test has been used for changes in the variance. Furthermore, the tests included in this section assume a single series. For multiple series, double mass analysis can be used and the significance of a change in the slope can be tested. In the case of testing the mean of a series versus the mean of another series (or the mean of a group of series), the Mann-Whitney test can also be applied.[173] Furthermore, in the case of testing for equality in means for a group of series, one-way and two-way analysis of variance can be used.[58]

19.2.4 Testing for Seasonality

In most analysis and modeling of hydrologic time series, testing for seasonality in statistics is done by using simple procedures, mostly by observing the plot of the statistic under consideration versus the season τ. Figure 19.2.3 plots the mean $\bar{y}_\tau$ versus $\tau = 1, \ldots, 12$. The plot suggests that the $\bar{y}_\tau$'s during the low-flow season are quite different from those of the high-flow season. Thus, even though some of the $\bar{y}$'s in the low-flow season are similar to each other, one would conclude overall that $\bar{y}_\tau$ is a seasonal statistic. A similar argument can be made in relation to seasonality in other statistics such as s_τ, g_τ, and $r_{k,\tau}$.

More rigorous procedures for testing for seasonality can be made, although they are rarely applied. For instance, since $\bar{y}_\tau$ is an estimator of the population mean μ_τ, one could test the hypothesis that $\mu_\tau = \mu_{\tau'}$ (for any $\tau \neq \tau'$) versus $\mu_\tau \neq \mu_{\tau'}$. Under the assumption that the underlying variables are independent and normally distributed, one could apply the t test of Sec. 19.2.3. The independence assumption may not be quite true for seasonal stream-flow series, except if τ and τ' are far apart, while it is usually satisfied by seasonal precipitation series for any $\tau \neq \tau'$. This test can be extended to consider the hypothesis $\mu_1 = \cdots = \mu_\omega$ versus $\mu_1 \neq \cdots \neq \mu_\omega$ by applying *one-way analysis of variance.*[124] Likewise, one can test the hypothesis $\sigma_\tau^2 = \sigma_{\tau'}^2$ versus $\sigma_\tau^2 \neq \sigma_{\tau'}^2$ by using the two-tailed F test. Furthermore, extensions to test the hypothesis $\sigma_1^2 = \cdots = \sigma_\omega^2$ versus $\sigma_1^2 \neq \cdots \neq \sigma_\omega^2$ can be made by using an approximate *generalized likelihood-ratio test.*[124]

19.2.5 Testing for Normality

Several of the models and approaches included in this chapter assume that the variable under consideration is normally distributed. Therefore, it is usual practice to test the data for normality before further analysis. A widely used method for judging whether a certain data set is normally distributed is to plot the empirical frequency distribution of the data on normal probability paper (see also Sec. 17.2.2 of Chap. 17). A straight-line probability plot indicates that the data are normally distributed. Otherwise, the data are not normal and a transformation may be needed to make them normal. Given the availability of plotting packages for personal and workstation computers, the graphical test for normality is quite attractive in practice. More importantly, a powerful correlation plot test procedure is available.[94,95,193] However, other tests are also used, such as the chi-square test, the Kolmogorov-Smirnov test,[7,124] and procedures based on testing the hypothesis that the skewness coefficient is equal to zero or the kurtosis coefficient is equal to three.[173] See also Sec. 18.3.2 of Chap. 18 for details of normality tests.

19.2.6 Transformations to Normal

A widely used method for transforming data to normal is based on the *logarithmic transformation*. For instance, if x_t is the original series, $y_t = \log(x_t - c)$ is normally distributed provided that x_t is lognormal with lower-bound parameter c. Often the simple log transformation with $c = 0$ works approximately, even if the original variable x_t is not lognormal. Likewise, *power transformations* such as $y_t = (x_t - c)^b$ with $b < 1$ (usually ½ or ⅓) is an alternative. An equivalent method is the *Box-Cox transformation* given by[10]

$$y_t = \begin{cases} \dfrac{(x_t^\lambda - 1)}{\lambda} & \lambda \neq 0 \qquad (19.2.29a) \\ \ln(x_t) & \lambda = 0 \qquad (19.2.29b) \end{cases}$$

in which λ is a parameter which must be estimated so that the skewness of the y_t becomes zero.

19.3 TIME-SERIES MODELING

The concepts and principles discussed in the previous sections are used in this section for representing hydrologic time series by mathematical models. A number of *stochastic models* are presented here along with parameter estimation methods and model testing procedures. The models can be applicable for many hydrologic processes, particularly stream-flow and precipitation processes. The models included in this section belong to the class of autoregressive (AR), autoregressive with moving average terms (ARMA), and disaggregation modeling schemes. These models should be able to reproduce the most important statistical features of the hydrologic time series under consideration. Alternative models such as fractional gaussian noises,[111] broken line models,[31,118] and shifting level models[154] are not included in this section because AR and ARMA models can accommodate most typical cases. This is not to say that alternative models may not be useful. For instance, shifting-level models may be useful to capture the effect of climatic shifts in hydrologic series. Likewise modeling of short-term rainfall processes has been developing rapidly in recent years

The theory and modeling schemes included here simply introduce some basic concepts around which some of the recent rainfall models have been developing.

19.3.1 Modeling of Single Stationary Series

AR Models. A time series defined as[15,155]

$$y_t = \mu + \sum_{j=1}^{p} \phi_j (y_{t-j} - \mu) + \varepsilon_t \tag{19.3.1}$$

is called an *autoregressive model* of order p in which ε_t is an uncorrelated normal random variable (also referred to as *noise, innovation, error term,* or *series of shocks*) with mean zero and variance σ_ε^2; it is uncorrelated with $y_{t-1}, \ldots, y_{t-p}$. Since ε_t is normally distributed, then also y_t is normal. The parameters of the model are μ, $\phi_1, \ldots, \phi_p$, and σ_ε^2. The model, Eq. (19.3.1), is often denoted as the AR(p) model or simply the AR model. The AR(1) model takes the simple form

$$y_t = \mu + \phi_1 (y_{t-1} - \mu) + \varepsilon_t \tag{19.3.2}$$

Low-order AR models such as Eq. (19.3.2) have been widely used for modeling annual hydrologic time series (Refs. 42, 61, 114, 155, and 203, among others), and seasonal and daily series after seasonal standardization.[138,207]

The mean, variance, and autocorrelation function of the AR(p) process are[15,155]

$$E(y) = \mu \tag{19.3.3}$$

$$\text{Var}\,(y) = \sigma^2 = \frac{\sigma_\varepsilon^2}{\left(1 - \sum_{j=1}^{p} \phi_j \rho_j\right)} \tag{19.3.4}$$

$$\rho_k = \phi_1 \rho_{k-1} + \cdots + \phi_p \rho_{k-p} \tag{19.3.5}$$

respectively. The last expression is known as the *Yule-Walker equation.* The three foregoing equations are useful for determining the properties of a model, given the model parameters, and for estimating the parameters of the model given a set of observations $y_1, \ldots, y_N$. For the AR(1) model, Eqs. (19.3.4) and (19.3.5) give respectively

$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1 - \phi_1^2} \tag{19.3.6}$$

$$\rho_k = \phi_1 \rho_{k-1} = \phi_1^k \tag{19.3.7}$$

ARMA Models. A more versatile model than the AR is the *autoregressive moving average model*[12]

$$y_t = \mu + \sum_{j=1}^{p} \phi_j (y_{t-j} - \mu) + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{19.3.8}$$

with p autoregressive parameters $\phi_1, \ldots, \phi_p$, and q moving average parameters $\theta_1, \ldots, \theta_q$. The model is also denoted as an ARMA(p, q) model or simply as ARMA. Note that an ARMA(p, 0) model is the same as an AR(p) model and the ARMA(0, q) model is the same as the *moving average model* MA(q). The noise ε_t in

Eq. (19.3.8) is an uncorrelated normal process with mean zero and variance σ_ε^2 and is also uncorrelated with $y_{t-1}, \ldots, y_{t-p}$. A simple version of the ARMA(p, q) model is the ARMA(1, 1) as

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + \varepsilon_t - \theta_1\varepsilon_{t-1} \tag{19.3.9}$$

Low-order ARMA models such as Eq. (19.3.9) are useful for operational hydrology in general, especially for modeling annual series[23,67,68,107,115,129,155,156] and for seasonal series after seasonal standardization.[33,116]

ARMA models must fulfill the *stationarity* and *invertibility* conditions, which imply certain constraints of the parameters. These are specified by the solution of the *characteristic equations*[12]

$$u^p - \phi_1 u^{p-1} - \cdots - \phi_p = 0 \tag{19.3.10}$$

$$u^q - \theta_1 u^{q-1} - \cdots - \theta_q = 0 \tag{19.3.11}$$

whose roots in each case must lie within the unit circle. For instance, for the ARMA(1, 1) model, the constraints $-1 < \phi_1 < 1$ and $-1 < \theta_1 < 1$ arise from the foregoing equations.

The variance and the lag-1 autocorrelation coefficient of the ARMA(1, 1) model are

$$\sigma^2 = \frac{1 - 2\phi_1\theta_1 + \theta_1^2}{1 - \phi_1^2}\sigma_\varepsilon^2 \tag{19.3.12}$$

$$\rho_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 - 2\phi_1\theta_1 + \theta_1^2} \tag{19.3.13}$$

respectively. Furthermore, the autocorrelation function is

$$\rho_k = \phi_1\rho_{k-1} = \rho_1\phi_1^{k-1} \qquad k > 1 \tag{19.3.14}$$

Comparing Eqs. (19.3.7) and (19.3.14), one may observe that ρ_k of the AR(1) process is less flexible than that of the ARMA(1, 1) process, since the former depends on the sole parameter ϕ_1 while the latter depends on ϕ_1 and θ_1. Figure 19.3.1 gives some examples of typical correlograms for both the AR(1) and the ARMA(1, 1) processes. In relation to modeling hydrologic processes, one may say that AR processes are short-memory processes and ARMA processes are long-memory processes.[130,153]

GAR Models. Skewed hydrologic processes must be transformed into normal processes before AR and ARMA models are applied. However, a direct modeling approach which does not require a transformation may be a viable alternative. The *gamma autoregressive process* (GAR process) offers such an alternative. It is defined as[103]

$$y_t = \phi y_{t-1} + \varepsilon_t \tag{19.3.15}$$

where ϕ is the autoregressive coefficient, ε_t is a random component, and y_t has a three-parameter gamma marginal distribution. The noise ε_t can be obtained as a function of ϕ and the parameters of the gamma distribution λ, α, and β (the location, scale, and shape parameters, respectively) as

$$\varepsilon = \lambda(1 - \phi) + \eta \tag{19.3.16}$$

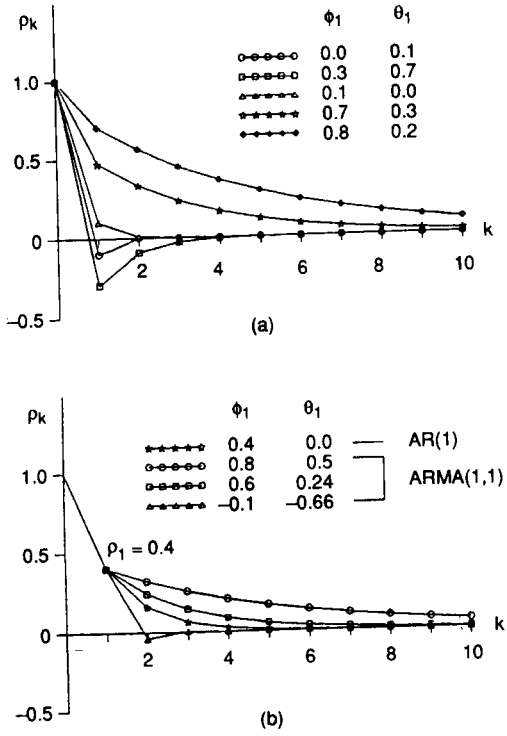

FIGURE 19.3.1 Correlograms ρ_k for (*a*) the ARMA(1, 1) process for various sets of parameters ϕ_1 and θ_1 and (*b*) the AR(1) and ARMA(1, 1) processes for which $\rho_1 = 0.4$.

$$\eta = 1 \qquad \text{if } M = 0$$
$$\eta = \sum_{j=1}^{M} E_j \phi^{U_j} \qquad \text{if } M > 0 \tag{19.3.17}$$

in which M is an integer random variable, Poisson-distributed, with mean $-\beta \ln(\phi)$, the set $U_1, \ldots, U_M$ are independent identically distributed (iid) random variables with uniform distribution (0, 1) and the set $E_1, \ldots, E_M$ are iid random variables exponentially distributed with mean $1/\alpha$. The GAR model has been applied to modeling annual stream-flow series.[39]

Product Models. While AR, ARMA, and GAR models are useful for modeling many hydrologic processes such as stream-flow processes in perennial rivers, they are impractical for *intermittent processes* such as stream flow in some ephemeral streams. Intermittent processes can be modeled as the product[79]

$$y_t = B_t z_t \tag{19.3.18}$$

where y_t is a nonnegative intermittent hydrologic variable, B_t is a discrete autocorrelated variable, z_t is a positive-valued continuous autocorrelated variable, and B_t and z_t are assumed to be mutually uncorrelated. B_t may be represented by a dependent (1, 0) *Bernoulli process* and z_t by an AR(1) process. Thus, the resulting product process y_t is intermittent and autoregressive. Product models have been applied for modeling short-term rainfall processes[17,24] and intermittent stream-flow processes.[159]

Parameter Estimation. Parameter estimation methods generally fall into three categories: *method of moments, method of maximum likelihood,* and *method of least squares.* The method of moments is based on taking as many moment equations as the number of parameters, substituting the population moments by the sample moments, and solving the equations simultaneously for the parameters. In the method of maximum likelihood, the likelihood function is first determined (this function is a function of the parameters given the observations), then the function (or its logarithm) is maximized and the parameters corresponding to such a maximum are the maximum likelihood estimators. In the method of least squares, the parameters which minimize the sum of square residuals $\Sigma\, \varepsilon_t^2$ are the least squares estimators; this is particularly useful for ARMA models. *Moment estimators* are generally available for all models. The method of maximum likelihood is the most efficient estimation method in a mean square error sense, although biases may be a problem, especially for small samples. The method of moments has been attractive in practice, since it is easier to apply and corrections for bias are available for some models. The method of least squares is an approximation to the method of maximum likelihood, and generally both methods require a numerical solution.

Moment estimators of the parameters of the AR(p) model can be obtained from Eqs. (19.3.3), (19.3.4), and (19.3.5). For instance, for the AR(1) model, the estimators $\hat{\mu}$, $\hat{\sigma}_\varepsilon^2$, and $\hat{\phi}_1$ are respectively

$$\hat{\mu} = \bar{y} \tag{19.3.19}$$

$$\hat{\sigma}_\varepsilon^2 = s^2 (1 - r_1^2) \tag{19.3.20}$$

$$\hat{\phi}_1 = r_1 \tag{19.3.21}$$

in which $\bar{y}$, s^2, and r_1 are the sample mean, variance, and lag-1 autocorrelation coefficient determined by Eqs. (19.2.1), (19.2.2), and (19.2.4), respectively. Corrections for bias may be made by using $\hat{\rho}_1$ of Eq. (19.2.6) instead of r_1 and $\hat{\sigma}^2$ of Eq. (19.2.7) instead of s^2.

Moment estimators for the ARMA(p, q) process can also be derived. For the ARMA(1, 1) model, the moment estimator of μ is also $\bar{y}$ and the estimators of σ_ε^2, ϕ_1, and θ_1 are respectively

$$\hat{\sigma}_\varepsilon^2 = \frac{s^2 (1 - \hat{\phi}_1^2)}{(1 - 2\,\hat{\phi}_1 \hat{\theta}_1 + \hat{\theta}_1^2)} \tag{19.3.22}$$

$$\hat{\phi}_1 = \frac{r_2}{r_1} \tag{19.3.23}$$

$$\hat{\theta}_1 = \frac{-b \pm \sqrt{b^2 - 4\,(r_1 - \hat{\phi}_1)^2}}{2\,(r_1 - \hat{\phi}_1)} \tag{19.3.24}$$

in which $b = 1 - 2\,\hat{\phi}_1 r_1 + \hat{\phi}_1^2$. O'Connell[130] provides a bias correction for s^2, and procedures for estimating ϕ_1 and θ_1 so that the Hurst slope is preserved.

Maximum likelihood estimators of ARMA models are generally found approximately by minimizing the sum $S(\underline{\phi}, \underline{\theta}) = \Sigma\, \varepsilon_t^2$, in which $\underline{\phi}$ and $\underline{\theta}$ represent the parameter sets $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$, respectively. The parameters $\underline{\phi}$ and $\underline{\theta}$ corresponding to the minimum sum will be the maximum likelihood estimators. Then, the estimator of σ_ε^2 is found by $\hat{\sigma}_\varepsilon^2 = (1/N)S(\underline{\phi}, \underline{\theta})$. A number of algorithms are available for minimizing the sum $S(\underline{\phi}, \underline{\theta})$.[12,15]

Fernandez and Salas[39] gave the moment estimators of the parameters of the GAR model as

$$\hat{\beta} = \frac{4}{\hat{\gamma}^2} \tag{19.3.25}$$

$$\hat{\alpha} = \frac{\sqrt{\hat{\beta}}}{\hat{\sigma}} \tag{19.3.26}$$

$$\hat{\lambda} = \hat{\mu} - \frac{\hat{\beta}}{\hat{\alpha}} \tag{19.3.27}$$

$$\hat{\phi} = \hat{\rho}_1 \tag{19.3.28}$$

in which $\hat{\rho}_1$, $\hat{\sigma}^2$, and $\hat{\gamma}$ are the unbiased estimators of Eqs. (19.2.6), (19.2.7), and (19.2.10), respectively and $\hat{\mu}$ is the unbiased sample mean of Eq. (19.2.1). Given that β, α, λ, and ϕ are determined, then the noise term ε of Eq. (19.3.15) is completely specified. Parameter estimation for product models such as Eq. (19.3.18) has been suggested.[17,24,159]

Model Testing and Selection. One of the basic tests for AR and ARMA models is in regard to the assumptions of normality and independence of the noise ε_t. Once the model parameters are determined, the residuals (noise) are found and tested for normality as in Sec. 19.2.5. A common test to determine if the ε's derived from an ARMA(p, q) model are independent is by the *Portemanteau lack of fit test.*[11] It uses the statistic

$$Q = N \sum_{k=1}^{L} r_k^2(\varepsilon) \tag{19.3.29}$$

where $r_k(\varepsilon)$ is the sample autocorrelation function of ε_t and L is the number of lags considered (for instance, $L = 0.25N$), Q is approximately chi-square distributed with $\nu = L - p - q$ degrees of freedom, and N is the sample size. If $Q < \chi^2_{1-\alpha,\nu}$, then ε_t is uncorrelated and the model from which the ε's were derived is judged to be an adequate model (α is the significance level of the test). Often, the bounds $\pm\, 1.96/\sqrt{N}$ are determined and one would like to see the $r_k(\varepsilon)$'s within the limits, especially for small k.

The adequacy of a time-series model is often examined by comparing the historical statistics with those derived from the model. The statistics considered are the mean, variance, skewness, and autocovariance, although more thorough comparisons include storage and drought-related statistics as well. If the method of moments is used for parameter estimation, certain statistics such as the mean and variance should be preserved or reproduced by the model. However, this may not necessarily be the case if transformations to normal are used in the modeling process or are due to biases of estimators. Matalas[114] showed that biases resulting from transformations may be important and suggested an estimation scheme to avoid such biases. Likewise, basic historical moments may not be reproduced by the model if

likelihood estimates are used. Sometimes, the comparison of the historical and model correlograms is used as the basis of judging the adequacy of a model. For instance, if the ARMA(1, 1) model is the model to be tested, the model correlogram ρ_k of Eq. (19.3.14) is compared with the historical or sample correlogram r_k. One would like to see that the model correlogram resembles the historical correlogram for the model to be adequate. Computer simulation is often used for comparing historical and model statistics. Stedinger and Taylor[176] offer a procedure to follow in using data generation studies for comparing historical and model-generated statistics.

Finally, comparison among models and model selection can be made by using the *Akaike information criteria* (AIC) suggested by Akaike.[1] A modified AIC called AICC has been suggested:[15]

$$\text{AICC}(p, q) = N \ln (\hat{\sigma}_\varepsilon^2) + \frac{2(p+q+1)N}{(N-p-q-2)} \tag{19.3.30}$$

in which $\hat{\sigma}_\varepsilon^2$ is the maximum likelihood estimator of the noise variance. The model which minimizes the AICC is selected.

19.3.2 Modeling of Single Periodic Series

PAR Models. Assume that a periodic hydrologic process is represented by $y_{\nu,\tau}$, in which ν defines the year and τ defines the season, such that $\tau = 1, \ldots, \omega$ and ω is the number of seasons in the year. Without loss of generality, τ could represent a day, week, month, or season. A time series defined as[155]

$$y_{\nu,\tau} = \mu_\tau + \sum_{j=1}^{p} \phi_{j,\tau}(y_{\nu,\tau-j} - \mu_{\tau-j}) + \varepsilon_{\nu,\tau} \tag{19.3.31}$$

is called a *periodic autoregressive model* of order p, in which $\varepsilon_{\nu,\tau}$ is an uncorrelated normal variable with mean zero and variance $\sigma_\tau^2(\varepsilon)$, and it is also uncorrelated with $y_{\nu,\tau-1}, \ldots, y_{\nu,\tau-p}$. The model parameters are μ_τ, $\phi_{1,\tau}, \ldots, \phi_{p,\tau}$ and $\sigma_\tau^2(\varepsilon)$ for $\tau = 1, \ldots, \omega$, and the model is often denoted as the PAR(p) or simply the PAR model. Note that in Eq. (19.3.31), if $\tau - j \le 0$, then $y_{\nu,\tau-j}$ becomes $y_{\nu-1,\omega+\tau-j}$ and $\mu_{\tau-j}$ becomes $\mu_{\omega+\tau-j}$. The PAR(1) model arises by making $p = 1$ in Eq. (19.3.31) as

$$y_{\nu,\tau} = \mu_\tau + \phi_{1,\tau}(y_{\nu,\tau-1} - \mu_{\tau-1}) + \varepsilon_{\nu,\tau} \tag{19.3.32}$$

Low-order PAR models such as Eq. (19.3.32) have been widely used in hydrology. For instance, the PAR(1) model was used by Hannan[64] for modeling monthly rainfall series and by Thomas and Fiering[181] for monthly stream-flow simulation. Likewise, PAR(1), PAR(2), and PAR(3) models have been used for simulation of seasonal hydrologic processes.[28,29,32,151]

PARMA Models. One may extend the PAR model [Eq. (19.3.31)] to include periodic moving average parameters. Such a model is the *periodic autoregressive moving average model* or PARMA(p, q) model. Low-order PARMA models are useful for modeling periodic hydrologic time series. For instance, the PARMA(1, 1) model is simply written as

$$y_{\nu,\tau} = \mu_\tau + \phi_{1,\tau}(y_{\nu,\tau-1} - \mu_{\tau-1}) + \varepsilon_{\nu,\tau} - \theta_{1,\tau}\varepsilon_{\nu,\tau-1} \tag{19.3.33}$$

This model has been applied to monthly stream-flow series.[69,155,180] Likewise, PARMA(2, 1) and PARMA(2, 2) models[5] and more complex *multiplicative PARMA*

models[121,160] may be needed for stream-flow modeling and simulation when preservation of both seasonal and annual statistics are desired.

Periodic GAR Model. Consider that $y_{\nu,\tau}$ is an autocorrelated variable with three-parameter gamma marginal distribution with location λ_τ, scale α_τ, and shape β_τ parameters varying with τ, $\tau = 1, \ldots, \omega$. The new variable $z_{\nu,\tau} = y_{\nu,\tau} - \lambda_\tau$ is a two-parameter gamma and can be represented by[38]

$$z_{\nu,\tau} = \phi_\tau z_{\nu,\tau-1} + (z_{\nu,\tau-1})^{\delta_\tau} w_{\nu,\tau} \tag{19.3.34}$$

where $z_{\nu,0} = z_{\nu-1,\omega}$, ϕ_τ is a periodic autoregressive coefficient, δ_τ is a periodic autoregressive exponent, and $w_{\nu,\tau}$ is the noise process. The *periodic GAR model* [Eq. (19.3.34)] has a periodic autocorrelation structure equivalent to that of the PAR(1) process. Refer to Fernandez and Salas[38] for properties and applications of the periodic GAR model.

Periodic Product Models. Intermittent hydrologic time series which are periodic and correlated can be modeled by[159]

$$y_{\nu,\tau} = B_{\nu,\tau} z_{\nu,\tau} \tag{19.3.35}$$

where $y_{\nu,\tau}$ is an intermittent periodic autocorrelated process, $B_{\nu,\tau}$ is a periodic autocorrelated Bernoulli (1, 0) process, and $z_{\nu,\tau}$ may be either an uncorrelated or correlated periodic process with a specified marginal distribution. Furthermore, the processes B and z are assumed to be mutually uncorrelated. Refer to Salas and Chebaane[159] and Chebaane et al.[25] for properties and applications of such *periodic product models* for stream-flow modeling and simulation.

Parameter Estimation. Estimation by method of moments and approximate maximum likelihood is available for the various periodic models included in this section. For instance, for the PAR(1) model of Eq. (19.3.32), the parameters μ_τ, ϕ_τ, and $\sigma_\tau^2(\varepsilon)$ may be estimated by

$$\hat{\mu}_\tau = \bar{y}_\tau \tag{19.3.36}$$

$$\hat{\phi}_{1,\tau} = \left(\frac{s_\tau}{s_{\tau-1}}\right) r_{1,\tau} \tag{19.3.37}$$

$$\hat{\sigma}_\tau^2(\varepsilon) = s_\tau^2 - s_{\tau-1}^2 r_{1,\tau}^2 \tag{19.3.38}$$

where $\bar{y}_\tau$, s_τ, and $r_{1,\tau}$ are the sample seasonal mean, seasonal standard deviation, and lag-1 season-to-season correlation coefficient, respectively.

Likewise, for the PARMA(1, 1) model of Eq. (19.3.33), the moment estimators are[157]

$$\hat{\phi}_{1,\tau} = \frac{c_{2,\tau}}{c_{1,\tau-1}} \tag{19.3.39}$$

$$\hat{\theta}_{1,\tau} = \hat{\phi}_{1,\tau} + \frac{(s_\tau^2 - \hat{\phi}_{1,\tau} c_{1,\tau})}{(\hat{\phi}_{1,\tau} s_{\tau-1}^2 - c_{1,\tau})} - \frac{(\hat{\phi}_{1,\tau+1} s_\tau^2 - c_{1,\tau+1})}{(\hat{\phi}_{1,\tau} s_{\tau-1}^2 - c_{1,\tau})\hat{\theta}_{1,\tau+1}} \tag{19.3.40}$$

$$\sigma_\tau^2(\varepsilon) = \frac{\hat{\phi}_{1,\tau+1} s_{\tau-1}^2 - c_{1,\tau+1}}{\hat{\theta}_{1,\tau+1}} \tag{19.3.41}$$

for $\tau = 1, \ldots, \omega$, where $c_{k,\tau}$ is determined by Eq. (19.2.13). Equation (19.3.40) is a system of equations that must be solved simultaneously to obtain $\theta_{1,\tau}, \tau = 1, \ldots, \omega$. Note that in Eqs. (19.3.37) through (19.3.41), $\tau - 1$ is to be interpreted as ω when $\tau = 1$ and as $\tau + 1$ when $\tau = \omega$. Moment estimation procedures for higher-order PAR and PARMA models are also available.[155,157]

In the method of maximum likelihood, the sum of the square residuals $S = \sum_{\nu=1}^{N} \sum_{\tau=1}^{\omega} \varepsilon_{\nu,\tau}^2$ in which $\varepsilon_{\nu,\tau}$, is the noise term of the PAR or PARMA model under consideration, is minimized to obtain the approximate maximum likelihood estimators $\hat{\phi}_{1,\tau}, \ldots, \hat{\phi}_{p,\tau}$ and $\hat{\theta}_{1,\tau}, \ldots, \hat{\theta}_{q,\tau}$. Then, the noise variance $\sigma_\tau^2(\varepsilon)$ can be obtained by $\hat{\sigma}_\tau^2(\varepsilon) = (1/N) \sum_{\nu=1}^{N} \varepsilon_{\nu,\tau}^2, \tau = 1, \ldots, \omega$ in which the noises are evaluated from the PAR or PARMA model equations, as the case may be.

Moment estimators of the parameters of the periodic GAR model of Eq. (19.3.34) are available.[38] Likewise, parameter estimation procedures for periodic product models are also available.[25,159]

Model Testing. Testing of PAR and PARMA models can be done by testing the basic assumptions of the models, i.e., that the noise $\varepsilon_{\nu,\tau}$ is uncorrelated and normal. After the residuals $\varepsilon_{\nu,\tau}$ are determined, the season-to--season correlations $r_{1,\tau}(\varepsilon)$, $\tau = 1, \ldots, \omega$ can be obtained from Eq. (19.2.12) and it can be verified that the r's fall within the bounds $\pm 1.96/\sqrt{N}$. Likewise, the residuals can be tested for normality on a seasonal basis. The model may be tested for adequacy to preserve certain historical statistics such as $\bar{y}_\tau$, s_τ^2, g_τ, *and* $r_{k,\tau}$ $(k = 1, \ldots)$. If the method of moments is used to estimate the parameters and the original data are approximately normally distributed, then the model must be able to reproduce such basic historical statistics. However, if transformations are used to make the original data normally distributed, then the model based on the method of moments generally will not reproduce the original statistics $\bar{y}_\tau$, s_τ^2, g_τ, and $r_{k,\tau}$, although it will reproduce similar statistics in the transformed domain. However, for the log-transformation and the PAR(1) model, it is possible to reproduce the basic statistics in the original domain by a procedure suggested by Burges.[18] Likewise, one may be interested to see the capability of a given model to reproduce statistics at aggregated time scales, typically at the annual time scale. Such statistics may be the annual correlation structure and storage- and drought-related statistics. In this case, data generation experiments are generally made.[155,176]

19.3.3 Modeling of Multiple Stationary Series

Analysis and modeling of multiple time series are widely needed in hydrology. For instance, one may like to model precipitation series at several sites in a river basin, stream-flow data recorded at several gauging stations in the stream network, or a mix of precipitation with stream-flow data recorded at various sites. Analysis and modeling procedures for multiple series are more involved than for a single series. In analyzing multiple series, vector and matrix notations are needed. However, the basic principles are similar, and we will still be referring to means, variances, and covariances, but in vector and matrix forms. We will start in this section with modeling of stationary series and the following section will deal with periodic series.

Multivariate AR and Multivariate ARMA Models. Consider a multiple time series $\mathbf{Y}_t$, a column vector with elements $y_t^{(1)}, \ldots, y_t^{(n)}$ in which n is the number of series

(number of sites or number of variables) under consideration. The *multivariate AR(1) model* suggested by Matalas[114] is defined as

$$\mathbf{Z}_t = \mathbf{A}_1\mathbf{Z}_{t-1} + \mathbf{B}\boldsymbol{\varepsilon}_t \tag{19.3.42}$$

in which $\mathbf{Z}_t = \mathbf{Y}_t - \boldsymbol{\mu}$, $\mathbf{A}_1$ and $\mathbf{B}$ are n- by n-parameter matrices and $\boldsymbol{\mu}$ is a column parameter vector with elements $\mu^{(1)}, \ldots, \mu^{(n)}$. The noise term $\boldsymbol{\varepsilon}_t$ is also a column vector of noises $\varepsilon_t^{(1)}, \ldots, \varepsilon_t^{(n)}$, each with zero mean such that $\mathbf{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^T) = \mathbf{I}$, where $\mathbf{T}$ denotes the transpose of the matrix and $\mathbf{I}$ is the identity matrix, and $\mathbf{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t-k}^T) = 0$ for $k \neq 0$. In addition, it is assumed that $\boldsymbol{\varepsilon}_t$ is uncorrelated with $\mathbf{Z}_{t-1}$ and $\boldsymbol{\varepsilon}_t$ is normally distributed. Model (19.3.42) has been widely used in operational hydrology. Higher-order multivariate AR models are also available.[134,155] Likewise, the *multivariate ARMA(1, 1) model* is written as[130,155]

$$\mathbf{Z}_t = \mathbf{A}_1\mathbf{Z}_{t-1} + \mathbf{B}\boldsymbol{\varepsilon}_t - \mathbf{C}_1\boldsymbol{\varepsilon}_{t-1} \tag{19.3.43}$$

in which $\mathbf{C}_1$ is an additional n- by n-parameter matrix.

Contemporaneous AR and ARMA Models. Using the full multivariate AR and multivariate ARMA models as described above often leads to complex parameter estimation, especially for the last model. Thus, model simplifications have been suggested. For instance, a simpler model will result from Eq. (19.3.42) if $\mathbf{A}_1$ is assumed to be a diagonal matrix.[114] In general, a *contemporaneous ARMA* (CARMA) model results if the matrices $\mathbf{A}_1$ and $\mathbf{C}_1$ of Eq. (19.3.43) are considered to be diagonal matrices.[22,155,158] In this case, model (19.3.43) implies a contemporaneous relationship in that only the dependence of concurrent values of the y's are considered important. Furthermore, the diagonalization of the parameter matrices allows model decoupling into component univariate models so that the model parameters do not have to be estimated jointly, and univariate modeling procedures can be employed.

To illustrate the foregoing concept, let a multivariate ARMA(p, q) process be

$$\mathbf{Z}_t = \sum_{j=1}^{p} \mathbf{A}_j\mathbf{Z}_{t-j} + \boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \mathbf{C}_j\boldsymbol{\varepsilon}_{t-j} \tag{19.3.44}$$

Assuming that matrices $\mathbf{A}_j$ and $\mathbf{C}_j$ are diagonals, model (19.3.44) can be decoupled into the model components

$$z_t^{(i)} = \sum_{j=1}^{p} a_j^{(i)} z_{t-j}^{(i)} + \varepsilon_t^{(i)} - \sum_{j=1}^{q} c_j^{(i)}\,\varepsilon_{t-j}^{(i)} \tag{19.3.45}$$

for $i = 1, \ldots, n$. Thus, the model components at each site are simply univariate ARMA(p, q) models where each $\varepsilon_t^{(i)}$, $i = 1, \ldots, n$, is uncorrelated, but are contemporaneously correlated with a variance-covariance matrix $\mathbf{G}$. Thus, the parameters a and c in each model, can be estimated by using univariate estimation procedures and the ε's can be modeled by

$$\boldsymbol{\varepsilon}_t = \mathbf{B}\boldsymbol{\zeta}_t \tag{19.3.46}$$

in which $\boldsymbol{\zeta}$ is normal, such that $\mathbf{E}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t^T) = \mathbf{I}$ and $\mathbf{E}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_{t-k}) = \mathbf{0}$ and $k \neq 0$. Note that one does not have to consider the same univariate ARMA(p, q) model for each site.

Parameter Estimation. Moment estimators for the multivariate AR(1) model (19.3.42) are[114]

$$\hat{\mathbf{A}}_1 = \hat{\mathbf{M}}_1\hat{\mathbf{M}}_0^{-1} \tag{19.3.47}$$

$$\hat{\mathbf{B}}\,\hat{\mathbf{B}}^T = \hat{\mathbf{M}}_0 - \hat{\mathbf{A}}_1\hat{\mathbf{M}}_1^T \tag{19.3.48}$$

in which $\hat{\mathbf{M}}_0$ and $\hat{\mathbf{M}}_1$ are the lag-zero and lag-one cross-covariance matrices of the multivariate series Z_t whose elements are determined by Eq. (19.2.15). Equation (19.3.48) gives estimates of the product $\hat{\mathbf{B}}\,\hat{\mathbf{B}}^T = \hat{\mathbf{D}}$ where $\hat{\mathbf{D}}$ is the right-hand side of (19.3.48). This matrix equation can be solved for $\hat{\mathbf{B}}$ by *principal component analysis*[114] or by *square root procedure*.[208] Matrices $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{B}}$ can be solved if $\hat{\mathbf{M}}_0$ and $\hat{\mathbf{M}}_1$ satisfy certain conditions. First of all, $\hat{\mathbf{M}}_0$ must be a positive definite matrix. This is generally satisfied when the sample sizes for all sites are the same. A technique that ensures that $\hat{\mathbf{M}}_0$ will be positive definite when the sample sizes are different is available.[30] In addition, matrix $\hat{\mathbf{B}}\,\hat{\mathbf{B}}^T = \hat{\mathbf{M}}_0 - \hat{\mathbf{A}}_1\hat{\mathbf{M}}_1^T$ must be positive definite. Estimators of $\hat{\mathbf{M}}_0$ and $\hat{\mathbf{M}}_1$ which ensure that $\hat{\mathbf{B}}\,\hat{\mathbf{B}}^T$ will be positive definite are also available.[30] With large matrices and with transformed data, numerical errors may still give an inconsistent $\mathbf{B}\,\mathbf{B}^T$ matrix. In such cases $\mathbf{B}\,\mathbf{B}^T$ can be adjusted.[13,119]

One can solve the matrix equation $\mathbf{B}\mathbf{B}^T = \mathbf{D}$ by using the square root method.[58] This method assumes that $\mathbf{B}$ is a *lower triangular matrix* (above diagonal elements are zero) and requires that $\mathbf{D}$ is a positive definite matrix. In such case the elements b^{ij} of $\mathbf{B}$ are

$$b^{ij} = \frac{d^{ji}}{b^{ij}} \qquad j = 1,\ i = 1,\ \ldots,\ n \tag{19.3.49a}$$

$$b^{ii} = \left[d^{ii} - \sum_{k=1}^{i-1} (b^{ik})^{1/2} \right]^{1/2} \qquad i = 2,\ \ldots,\ n \tag{19.3.49b}$$

$$b^{ij} = \frac{d^{ij} - \sum_{k=1}^{j-1} b^{ik}\, b^{jk}}{b^{ij}} \qquad j = 2,\ \ldots,\ n-1,\ i > j \tag{19.3.49c}$$

and $b^{ij} = 0$ for $i < j$, in which d^{ij}, $i, j = 1, \ldots, n$, are the elements of $\mathbf{D}$. In cases where $\mathbf{D}$ is a positive semidefinite matrix, a method based on principal components is needed.[13] An alternative method for solving $\mathbf{B}\mathbf{B}^T = \mathbf{D}$ which works for $\mathbf{D}$ positive definite or positive semidefinite is also available.[155]

The estimation difficulties normally encountered with the full multivariate ARMA models can be overcome by using CARMA models. Since the CARMA model (19.3.44) can be decoupled as in Eq. (19.3.45), univariate estimation procedures by the method of moments or maximum likelihood can be used to estimate the a's and c's. The elements of matrix $\hat{\mathbf{G}}$ for the CARMA(1, 1) model can be estimated as[178]

$$\hat{g}^{(ij)} = \frac{m_0^{(ij)}\,(1 - \hat{a}^{(i)}\,\hat{a}^{(j)})}{1 - \hat{a}^{(i)}\,\hat{c}^{(j)} - \hat{a}^{(j)}\,\hat{c}^{(i)} + \hat{c}^{(i)}\,\hat{c}^{(j)}} \tag{19.3.50}$$

for $i, j = 1, \ldots, n$, in which $m_0^{(ij)}$ is the ijth element of $\hat{\mathbf{M}}_0$, the lag-zero cross-covariance matrix of $\mathbf{Z}_t$. Finally, $\mathbf{B}$ of Eq. (19.3.46) will be estimated by solving $\hat{\mathbf{B}}\hat{\mathbf{B}}^{\mathrm{T}} = \hat{\mathbf{G}}$.

Model Testing. Model testing depends on the type of multivariate model considered. In the case of the full multivariate AR(p) and ARMA(p, q) models, one may test the assumption of normality of residuals $\boldsymbol{\varepsilon}_t$ and the assumptions $\mathbf{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^T) = \mathbf{I}$ and $\mathbf{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t-k}^T) = \mathbf{0}$ for $k \neq 0$. Thus, once the residuals $\varepsilon_t^{(i)}$, $i = 1, \ldots, n$, are found, one should test that $\hat{\mathbf{M}}_0(\varepsilon) = \mathbf{I}$ and that, at least, $\hat{\mathbf{M}}_1(\varepsilon) = \mathbf{0}$. The elements of these matrices

[except the diagonal of $\mathbf{M}_0(\varepsilon)$] must be within the limits of $\pm 1.96/\sqrt{N}$. Likewise, in the case of CARMA models, one can make similar tests as above for the residual variable ξ of Eq. (19.3.46). Testing whether a given model reproduces historical statistics is often performed by data generation procedures.[155]

19.3.4 Modeling of Multiple Periodic Series

Multivariate PAR and Multivariate PARMA Models. The *multivariate PAR(1) model* is given by[152]

$$\mathbf{Z}_{\nu,\tau} = \mathbf{A}_\tau \mathbf{Z}_{\nu,\tau-1} + \mathbf{B}_\tau \boldsymbol{\varepsilon}_{\nu,\tau} \tag{19.3.51}$$

in which $\mathbf{Z}_{\nu,\tau} = \mathbf{Y}_{\nu,\tau} - \boldsymbol{\mu}_\tau$, $\mathbf{A}_\tau$ and $\mathbf{B}_\tau$ are n- by n-parameter matrices, and $\boldsymbol{\mu}_\tau$ is a column parameter vector with elements $\mu_\tau^{(1)}, \ldots, \mu_\tau^{(n)}$. All parameters $\boldsymbol{\mu}_\tau$, $\mathbf{A}_\tau$, and $\mathbf{B}_\tau$ are periodic. The noise term $\boldsymbol{\varepsilon}_{\nu,\tau}$ is a column vector normally distributed with mean zero, and $\mathbf{E}(\boldsymbol{\varepsilon}_{\nu,\tau}\boldsymbol{\varepsilon}_{\nu,\tau}^T) = \mathbf{I}$ and $\mathbf{E}(\boldsymbol{\varepsilon}_{\nu,\tau}\boldsymbol{\varepsilon}_{\nu,\tau-k}^T) = \mathbf{0}$ for $k \neq 0$. In addition, it is assumed that $\boldsymbol{\varepsilon}_{\nu,\tau}$ is uncorrelated with $\mathbf{Z}_{\nu,\tau-1}$. This model has been widely used for generating seasonal hydrologic processes. Likewise, the *multivariate PARMA(1, 1) model* is written as[63,158]

$$\mathbf{Z}_{\nu,\tau} = \mathbf{A}_\tau \mathbf{Z}_{\nu,\tau-1} + \mathbf{B}_\tau \boldsymbol{\varepsilon}_{\nu,\tau} - \mathbf{C}_\tau \boldsymbol{\varepsilon}_{\nu,\tau-1} \tag{19.3.52}$$

in which $\mathbf{C}_\tau$ is an additional n by n periodic matrix parameter.

Contemporaneous PAR and PARMA Models. Simplifications of the foregoing models can be made. For instance, $\mathbf{A}_\tau$ of Eq. (19.3.51) can be made diagonal. In the case of the multivariate PARMA(1, 1) model of Eq. (19.3.52), it is more convenient to write the model as[4]

$$\mathbf{Z}_{\nu,\tau} = \mathbf{A}_\tau \mathbf{Z}_{\nu,\tau-1} + \boldsymbol{\varepsilon}_{\nu,\tau} - \mathbf{C}_\tau \boldsymbol{\varepsilon}_{\nu,\tau-1} \tag{19.3.53}$$

and consider that $\mathbf{A}_\tau$ and $\mathbf{C}_\tau$ are diagonal matrices. Then, the model can be decoupled into univariate models for each site. However, to maintain cross-correlation among sites, the vector $\boldsymbol{\varepsilon}_{\nu,\tau}$ will be assumed to have a variance-covariance matrix $\mathbf{G}_\tau$ or $\mathbf{E}(\boldsymbol{\varepsilon}_{\nu,\tau}\boldsymbol{\varepsilon}_{\nu,\tau}^T) = \mathbf{G}_\tau$ and $\mathbf{E}(\boldsymbol{\varepsilon}_{\nu,\tau}\,\boldsymbol{\varepsilon}_{\nu,\tau-k}^T) = 0$ for $k \neq 0$. Then, $\boldsymbol{\varepsilon}_{\nu,\tau}$ can be modeled as

$$\boldsymbol{\varepsilon}_{\nu,\tau} = \mathbf{B}_\tau \boldsymbol{\xi}_{\nu,\tau} \tag{19.3.54}$$

such that $\mathbf{E}(\boldsymbol{\xi}_{\nu,\tau}\boldsymbol{\xi}_{\nu,\tau}^T) = \mathbf{I}$ and $\mathbf{E}(\boldsymbol{\xi}_{\nu,\tau}\boldsymbol{\xi}_{\nu,\tau-k}^T) = 0$ for $k \neq 0$. The foregoing modeling scheme is a *contemporaneous PARMA(1, 1) model.* Similar simplifications can be made for higher-order models.

Parameter Estimation. Moment estimators of the parameters of the multivariate PAR(1) model (19.3.51) are[155]

$$\hat{\mathbf{A}}_\tau = \hat{\mathbf{M}}_{1,\tau}\,\hat{\mathbf{M}}_{0,\tau-1}^{-1} \tag{19.3.55}$$

$$\hat{\mathbf{B}}_\tau\hat{\mathbf{B}}_\tau^T = \hat{\mathbf{M}}_{0,\tau} - \hat{\mathbf{A}}_\tau\hat{\mathbf{M}}_{1,\tau}^T \tag{19.3.56}$$

in which $\hat{\mathbf{M}}_{k,\tau}$ ($k = 0, 1$) are sample season-to-season covariance matrices of $\mathbf{Z}_{\nu,\tau}$. Matrix $\hat{\mathbf{A}}_\tau$ is determined directly while $\hat{\mathbf{B}}_\tau$ must be found by the square root method. Estimation for the multivariate PARMA(1, 1) model is more complex.[63,158]

An alternative is to use the contemporaneous PARMA(1, 1) model of Eq. (19.3.53). Since it allows decoupling, univariate procedures can be used to estimate

the parameters $a_\tau^{(i)}$ and $c_\tau^{(i)}$, $i = 1, \ldots, n$ (elements of $\mathbf{A}_\tau$ and $\mathbf{C}_\tau$). Then, the elements g_τ^{ij} of matrix $\mathbf{G}_\tau$ may be obtained as[63]

$$g_\tau^{(ij)} = e_\tau^{(ij)} + f_\tau^{(ij)}\, g_{\tau-1}^{(ij)} \tag{19.3.57a}$$

$$e_\tau^{(ij)} = m_{0,\tau}^{(ij)} - a_\tau^{(i)}\, m_{0,\tau-1}^{(ij)}\, a_\tau^{(j)} \tag{19.3.57b}$$

$$f_\tau^{(ij)} = a_\tau^{(i)}\, c_\tau^{(j)} + a_\tau^{(i)}\, c_\tau^{(i)} - c_\tau^{(i)}\, c_\tau^{(j)} \tag{19.3.57c}$$

where $m_{0,\tau}^{(ij)}$, $i, j = 1, \ldots, n$ are the elements of $\hat{\mathbf{M}}_{0,\tau}$. Once matrix $\mathbf{G}_\tau$ is determined, then $\mathbf{B}_\tau$ of Eq. (19.3.54) can be determined from $\mathbf{B}_\tau \mathbf{B}_\tau^T = \mathbf{G}_\tau$.

Model Testing. Model testing for multivariate periodic models is similar to testing of multivariate stationary models except that periodicity must be considered. For instance, for full multivariate PAR or PARMA models, one can verify that the residuals $\boldsymbol{\varepsilon}_{\nu,\tau}$ are normally distributed and check the assumptions $\mathbf{E}(\boldsymbol{\varepsilon}_{\nu,\tau}\, \boldsymbol{\varepsilon}_{\nu,\tau}^T) = \mathbf{I}$ and $\mathbf{E}(\boldsymbol{\varepsilon}_{\nu,\tau}\, \boldsymbol{\varepsilon}_{\nu,\tau-k}^T) = \mathbf{0}$ for $k \neq 0$. Likewise, for contemporaneous PARMA models, one can make similar tests for the residuals $\boldsymbol{\xi}_{\nu,\tau}$ of Eq. (19.3.54). Additionally, one can do further testing by using data generation procedures.

19.3.5 Disaggregation of Annual to Seasonal Series

Generally, modeling of seasonal hydrologic time series is geared to preserving seasonal statistics only, while statistics at other levels of aggregation, such as annual statistics, may not be preserved. For instance, if the PAR(1) model is used to generate monthly flows, the historical monthly statistics are usually preserved, yet if such generated monthly flows are aggregated to obtain the corresponding annual flows, there is no assurance that the historical annual statistics will be preserved. *Disaggregation models* have been developed for reproducing statistics at more than one level of aggregation. Disaggregation models can be used for both temporal and spatial disaggregation; however, the models in this section are mostly described in terms of temporal disaggregation.

Traditional Valencia-Schaake Model. Assume that $\mathbf{X}$ and $\mathbf{Y}$ are normalized variables with mean zero. The basic form of the disaggregation model suggested by Valencia and Schaake is[189]

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{B}\boldsymbol{\varepsilon} \tag{19.3.58}$$

where $\mathbf{X}$ is an n vector of annual values at n sites, $\mathbf{Y}$ is an $n\omega$ vector of seasonal values in which ω is the number of seasons in the year, $\mathbf{A}$ and $\mathbf{B}$ are $n\omega$- by $n\omega$-parameter matrices, and $\boldsymbol{\varepsilon}$ is an $n\omega$ vector of independent standard normal variables. A desirable property of disaggregation models is additivity, i.e., the sum of the seasonal values must add up to the annual values. The parameters $\mathbf{A}$ and $\mathbf{B}$ may be estimated by[189]

$$\hat{\mathbf{A}} = \mathbf{S}_{YX}\mathbf{S}_{XX}^{-1} \tag{19.3.59}$$

$$\hat{\mathbf{B}}\hat{\mathbf{B}}^T = \mathbf{S}_{YY} - \mathbf{A}\mathbf{S}_{XY} \tag{19.3.60}$$

in which $\mathbf{S}_{UV}$ represents the sample covariance of the vectors $\mathbf{U}$ and $\mathbf{V}$. In the foregoing formulation, it is assumed that $\mathbf{X}$ (say the annual series) has been previously generated by a specified model such as the AR(1) or ARMA(1, 1) process.

Valencia-Schaake's model does not preserve the covariances of the first season of a year and any preceding season. To circumvent this, Eq. (19.3.58) is modified as[120]

$$\mathbf{Y} = \mathbf{AX} + \mathbf{B}\varepsilon + \mathbf{CZ} \tag{19.3.61}$$

where C is a new parameter matrix and Z is a vector of seasonal values from the previous year for each site. Usually, Z is a vector containing only the last season of the previous year in which case C is an $n\omega$ by n matrix. A, B, and C may be estimated by[99]

$$\hat{\mathbf{A}} = (\mathbf{S}_{YX} - \mathbf{S}^*_{YZ}\mathbf{S}^{-1}_{ZZ}\mathbf{S}^*_{ZX})\,(\mathbf{S}_{XX} - \mathbf{S}^*_{XZ}\mathbf{S}^{-1}_{ZZ}\mathbf{S}^*_{ZX})^{-1} \tag{19.3.62a}$$

$$\hat{\mathbf{C}} = (\mathbf{S}^*_{YZ} - \hat{\mathbf{A}}\mathbf{S}^*_{XZ})\mathbf{S}^{-1}_{ZZ} \tag{19.3.62b}$$

$$\hat{\mathbf{B}}\hat{\mathbf{B}}^T = \mathbf{S}_{YY} - \hat{\mathbf{A}}\mathbf{S}_{XY} - \hat{\mathbf{C}}\mathbf{S}^*_{ZY} \tag{19.3.62c}$$

where $\mathbf{S}^*_{XZ} = \mathbf{S}_{XX'}\mathbf{S}^{-1}_{XX}\mathbf{S}_{X'Z}$, $\mathbf{S}^*_{YZ} = \mathbf{S}_{YZ} + \mathbf{S}_{YX}\mathbf{S}^{-1}_{XX}(\mathbf{S}^*_{XZ} - \mathbf{S}_{XZ})$ and $\mathbf{X}'$ is the vector of the previous year. With A, B, and C thus estimated, model (19.3.61) preserves the covariances $\mathbf{S}_{YY}$ and $\mathbf{S}_{YX}$ as well as the additivity property. However, Eqs. (19.3.62) assume an annual model which reproduces $\mathbf{S}_{XX}$ and $\mathbf{S}_{XX'}$.

On the other hand, a scheme which does not depend on the annual model's structure can be formulated as[177]

$$\mathbf{Y}_\tau = \mathbf{AX}_t + \boldsymbol{\varepsilon}_t \tag{19.3.63a}$$

$$\boldsymbol{\varepsilon}_t = \mathbf{C}\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\xi}_t \tag{19.3.63b}$$

in which the subscript t is now introduced, $\boldsymbol{\varepsilon}_t$ is independent of $\mathbf{X}_t$, and $\boldsymbol{\xi}_t$ is a random component with covariance matrix parameter $\mathbf{S}_{\xi\xi}$. The parameters may be estimated by

$$\hat{\mathbf{A}} = \mathbf{S}_{YX}\mathbf{S}^{-1}_{XX} \tag{19.3.64a}$$

$$\hat{\mathbf{C}} = \mathbf{S}_{\varepsilon\varepsilon'}\mathbf{S}^{-1}_{\varepsilon\varepsilon} \tag{19.3.64b}$$

$$\mathbf{S}_{\xi\xi} = \mathbf{S}_{\varepsilon\varepsilon} - \hat{\mathbf{C}}\mathbf{S}^T_{\varepsilon\varepsilon}\hat{\mathbf{C}}^T \tag{19.3.64c}$$

in which $\mathbf{S}_{\varepsilon\varepsilon} = \mathbf{S}_{YY} - \hat{\mathbf{A}}\mathbf{S}_{XY}$ and $\boldsymbol{\varepsilon}' = \boldsymbol{\varepsilon}_{t-1}$. The model scheme reproduces the moments $\mathbf{S}_{YY}$, $\mathbf{S}_{YX}$, and $\mathbf{S}_{XX}$.

Lane's Condensed Model. The foregoing disaggregation models have too many parameters, a problem which may be significant, especially when the number of sites is large and the sample size is small. Lane[98] sets to zero some parameters of model (19.3.61). Thus

$$\mathbf{Y}_\tau = \mathbf{A}_\tau\mathbf{X} + \mathbf{B}_\tau\boldsymbol{\varepsilon} + \mathbf{C}_\tau\mathbf{Y}_{\tau-1} \qquad \tau = 1, \ldots, \omega \tag{19.3.65}$$

is a model in which the number of parameters is reduced considerably (ω sets of parameters $\mathbf{A}_\tau$, $\mathbf{B}_\tau$, and $\mathbf{C}_\tau$). However, a shortcoming is that the additivity property is lost because the model is applied separately to each season. This shortcoming can be avoided by adjusting the seasonal values so that they add exactly to the annual values at each site. The estimation of model parameters and approximate adjustments can be found in Lane.[98]

Santos-Salas Step Disaggregation Model. In some of the disaggregation approaches, it is necessary to solve $\hat{\mathbf{B}}\hat{\mathbf{B}}^T = \hat{\mathbf{D}}$ for $\hat{\mathbf{B}}$. Since $\hat{\mathbf{B}}\hat{\mathbf{B}}^T$ should be positive semidefinite (because of the additivity property), the principal components tech-

nique is usually followed. However, when the matrices involved are large (which is typical for multisite seasonal disaggregation), the solution for $\hat{\mathbf{B}}$ usually deteriorates, and large computer storage capacity is required.[13] The disaggregation problem can be made computationally more amenable if it can be done in steps (stages or cascades); thus the size of the matrices involved will decrease and, consequently, so will the number of parameters. For instance, Fig. 19.3.2*a* schematically shows that annual flows are disaggregated into monthly flows directly in one step (this is the usual approach), while Fig. 19.3.2*b* shows such disaggregation is performed in two steps, into quarterly flows in the first step, then each quarterly flow is further disaggregated into monthly flows in the second step. However, even in the latter approach, large matrices will result when the number of seasons and the number of sites are large.

Santos and Salas[165,166] proposed a step disaggregation in which, at each step, the disaggregation is always into two parts or two seasons. For instance, Fig. 19.3.3 shows that the yearly value $\mathbf{X}_\nu$ is disaggregated into 12 monthly values by first disaggregating the year into the first month $\mathbf{Y}_{\nu,1}$ and the sum of the remaining 11 months $\sum_{\tau=2}^{12} \mathbf{Y}_{\nu,\tau}$. Then, this latter sum is disaggregated into the second month $\mathbf{Y}_{\nu,2}$ and the sum of the remaining 10 months $\sum_{\tau=3}^{12} \mathbf{Y}_{\nu,\tau}$ and so on until the months $\mathbf{Y}_{\nu,11}$ and $\mathbf{Y}_{\nu,12}$ are obtained in the eleventh disaggregation step. This stepwise disaggregation scheme leads to a maximum parameter matrix size of 2×2 for single site disaggregation and $2n \times 2n$ for multisite disaggregation. Note that the seasonal covariance between flows is preserved if model (19.3.61) is applied in each step. The step disaggregation model as

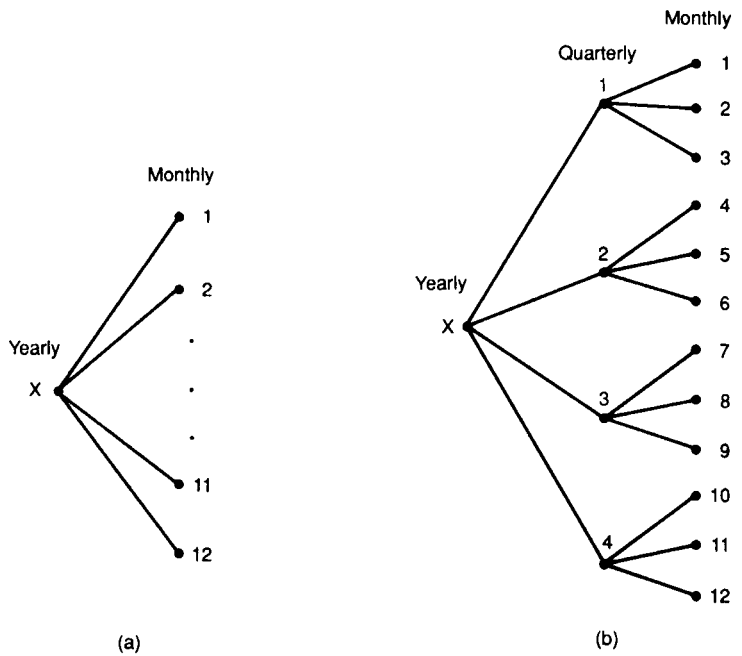

FIGURE 19.3.2 Disaggregation of annual flows X into monthly flows (*a*) in one step, and (*b*) in two steps, first into quarterly flows, then into monthly flows.

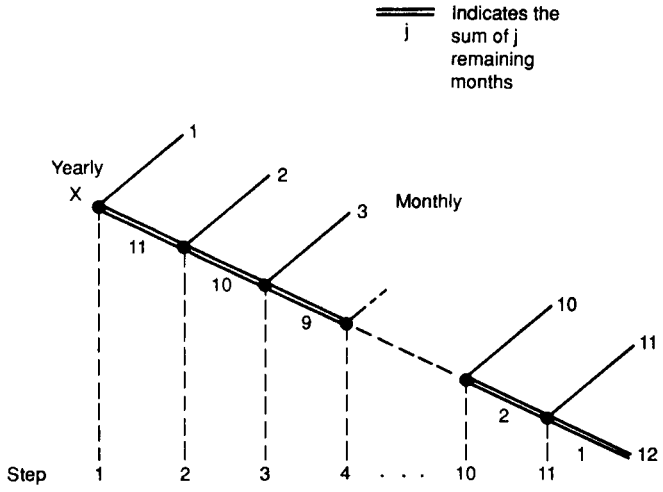

FIGURE 19.3.3 Step disaggregation of annual flows X into monthly flows in 11 steps. First step disaggregates into month 1 and the sum of months 2–12, the second step pertains to month 2 and sum of months 3–12, and so on.[166]

suggested above has the advantage over the previous models in that it has the minimum size of matrices involved, thus the smallest number of parameters, while keeping the additivity property. However, it has the same drawback of all previous models when the underlying variables are not normally distributed. However, the approach proposed by Todini[182] can be implemented in conjunction with the step disaggregation model for skewed variables.

Stedinger, Pei, and Cohn's Stagewise Models. Stedinger et al.[179] suggested a single site disaggregation model which reproduces seasonal statistics and the covariance of seasonal flows with annual flows, assuming lognormally distributed seasonal flows and lognormally distributed annual flows. Consider that $y_{v,\tau} = \log(Q_{v,\tau} - q_\tau)$ is normally distributed seasonal flows, $Q_{v,\tau}$ is the original seasonal flows, and q_τ is the lower bound for season τ. In addition, consider that $x_v = \log(Q_v - q)$ is also normally distributed, Q_v is the original annual flows, and q is the lower bound. The model to generate $y_{v,\tau}$ may be written as[179]

$$y_{v,1} = a_1 + b_1 x_v + \varepsilon_{v,1} \tag{19.3.66a}$$

$$y_{v,2} = a_2 + b_2 x_v + d_2 y_{v,1} + \varepsilon_{v,2} \tag{19.3.66b}$$

$$y_{v,\tau} = a_\tau + b_\tau x_v + c_\tau y_{v,\tau-1} + d_\tau \sum_{j=1}^{\tau-1} w_\tau y_{v,j} + \varepsilon_{v,\tau} \quad \tau = 3, \ldots, \omega \tag{19.3.66c}$$

where $\varepsilon_{v,\tau}$ are uncorrelated zero-mean normal random variables. The term $\Sigma\, w_\tau y_{v,j}$ allows for reproduction of the variance of the first-order approximation of Q_v, the actual lognormal annual flow, and $w_\tau = \exp(\mu_\tau + 0.5\,\sigma_\tau^2)$, in which μ_τ and σ_τ^2 are the mean and variance of $y_{v,\tau}$. The seasonal variables $y_{v,\tau}$ from Eq. (19.3.66) will produce a generated annual flow equal to $Q_v^* = \sum_{\tau=1}^{\omega} [q_\tau + \exp(y_{v,\tau})]$ which will not be equal to

the original generated annual flow Q_v. Then, the seasonal flows can be adjusted as

$$Q_{v,\tau} = \left(\frac{Q_v}{Q_v^*}\right)[q_\tau + \exp(y_{v,\tau})] \tag{19.3.67}$$

in which case the sum of $Q_{v,1}, \ldots, Q_{v,\omega}$ will be exactly equal to Q_v.

Model (19.3.66) essentially amounts to adding an extra term to Lane's model (19.3.65). Likewise, model (19.3.66), if applied to real space flows, is similar to Santos-Salas' step model. Model (19.3.66) relates each monthly flow with the yearly flow explicitly, and the additivity property is preserved by including an extra term which represents the sum of the previously generated monthly flows. On the other hand, in Santos and Salas' step model, the relationship between each monthly flow and the annual flow is not explicit. However, if such a step model is written in regression form, then a term representing the sum of the remaining flows in the year will appear. Thus, the models are different, but accomplish essentially the same thing in the real space flows. The advantage of model (19.3.66) is that it can generate lognormally distributed seasonal flows which, with some adjustment, will add up to a lognormally distributed annual flow. A shortcoming of the model is that there is no provision for preserving the covariance of flows of the first season of this year with flows of the last season of the previous year. The multisite version of model (19.3.66) is also available.[59,61]

19.3.6 Markov Chains

The models included in the previous sections are applicable for continuous variables. However, various processes in hydrology can be formulated as discrete-valued processes or continuous processes can be discretized for computational convenience. In these cases, the theory of *Markov chains* may be applicable. Markov chains have been used in hydrology for modeling processes such as precipitation, stream flow, soil moisture, and water storage in reservoirs. Gabriel and Neumann[50] developed a Markov chain model for the occurrence of dry and wet days in daily rainfall. Many others have used Markov chains for modeling precipitation processes[3,26,40,62,122,149,172,184] and water storage processes.[51,56,109,126,210]

Definition and Properties. Consider $X(t)$ to be a discrete-valued process which started at time 0 and developed through time. The values that the $X(t)$ process takes on are denoted by x_t, $t = 0, 1, \ldots$. Then

$$P[X(t) = x_t | X(0) = x_0, X(1) = x_1, \ldots, X(t-1) = x_{t-1}] \tag{19.3.68}$$

is the probability of the process being equal to x_t at time t, given its entire history. If the foregoing probability simplifies to

$$P[X(t) = x_t | X(t-1) = x_{t-1}] \tag{19.3.69}$$

it means that the outcome of the process at time t can be defined by using only the outcome at time $t-1$. A process which has this property is a *first-order Markov chain* or a *simple Markov chain.* Higher-order Markov chains can be formulated; however, only simple Markov chains will be considered here. Furthermore, the notation $X(t) = j, j = 1, \ldots, r$ will be used instead of x_t, which means that $X(t)$ is at *state* j, and r is the number of states. For instance, in modeling daily rainfall, one may consider two states, $j = 1$ for a dry day (no rain) and $j = 2$ for a wet day. Figure 19.3.4 shows schematically the definition of states.

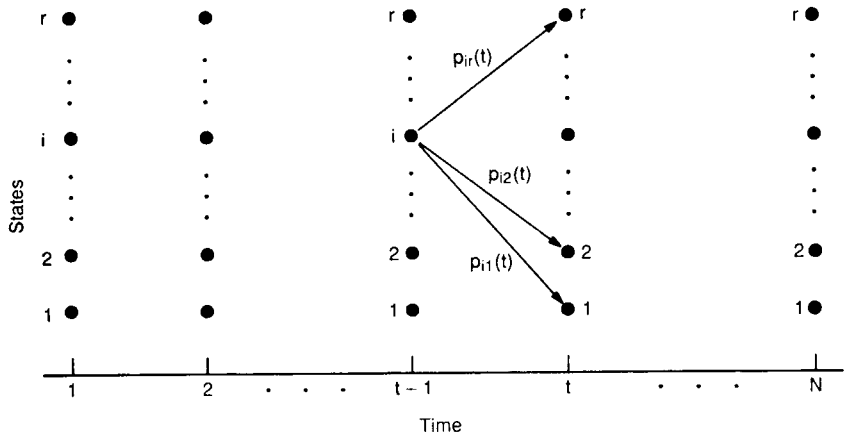

FIGURE 19.3.4 Definition of states for a Markov chain $X(t)$ and the corresponding transition probabilities $p_{ij}(t), j = 1, \ldots, r$.

Transition Probability Matrix. A simple Markov chain is defined by its *transition probability matrix* $\mathbf{P}(t)$, which is a square matrix with elements $p_{ij}(t)$ given by

$$p_{ij}(t) = P[X(t) = j | X(t-1) = i] \tag{19.3.70}$$

for all i, j pairs. Figure 19.3.4 shows that the chain may go from state i at time $t-1$ to states $1, \ldots, r$ at time t, with corresponding transition probabilities $p_{i1}(t), \ldots, p_{ir}(t)$. Then,

$$\sum_{j=1}^{r} p_{ij}(t) = 1 \qquad i = 1, \ldots, r$$

Furthermore, if the transition probability matrix $\mathbf{P}(t)$ does not depend on time, the Markov chain is a *homogeneous chain* or a *stationary chain.* In this case, the notations P and p_{ij} are used. For the rest of this section, a homogeneous Markov chain is assumed.

n-Step Probability. Assume that the chain is now in state i and after n time steps it is in state j. The transition probability from i to j in n steps, denoted by $p_{ij}^{(n)}$, is given by[133]

$$p_{ij}^{(n)} = \sum_{k=1}^{r} p_{ik}^{(n-1)} p_{kj} \qquad n > 1 \tag{19.3.71}$$

and $p_{ij}^{(1)} = p_{ij}$. Thus, $p_{ij}^{(n)}, i, j = 1, \ldots, r$ are elements of the n-step *transition probability matrix* $\mathbf{P}^{(n)}$. It may be shown that $\mathbf{P}^{(n)}$ can be found by multiplying the one-step transition probability matrix $\mathbf{P}$ by itself n times.

Marginal Distribution. The probability distribution of the chain being at any state j at time t, denoted by $q_j(t) = P[X(t) = j], j = 1, \ldots, r$, is called the *marginal distribution* of the process. Thus, $q_j(0)$ is the distribution of the initial states. The marginal

state probability $q_j(t)$, given that $q_j(0)$ is known, may be determined as[133]

$$q_j(t) = \sum_{i=1}^{r} q_i(0) p_{ij}^{(t)} \tag{19.3.72}$$

Also, $\mathbf{q}(t) = \mathbf{q}(0)\,\mathbf{P}^t$, in which $\mathbf{q}(t)$ denotes the (row) vector of marginal state probabilities.

Steady-State Probabilities. The steady-state probability vector $\mathbf{q}^*$ with elements $q_1^*, \ldots, q_r^*$ represents the average fraction of time the chain is in states $1, \ldots, r$, respectively. It can be found by estimating $\mathbf{P}^{(t)}$ for large t until it converges. Also, the elements q_i^*, $i = 1, \ldots, r$ can be found by solving the system of equations

$$q_i^* = \sum_{k=1}^{r} q_k^* p_{ki} \qquad i = 1, \ldots, r \tag{19.3.73a}$$

$$\sum_{i=1}^{r} q_i^* = 1 \tag{19.3.73b}$$

Example. Assume that daily rainfall for a given site is represented by a simple Markov chain with two states, $j = 1$ for dry and $j = 2$ for wet, and a transition probability matrix $\mathbf{P}$ with elements $p_{11} = 0.6$, $p_{12} = 0.4$, $p_{21} = 0.3$, and $p_{22} = 0.7$. Assume also that initially the day is dry or $j = 1$ at $t = 0$. This also means that the initial marginal state probability vector is $\mathbf{q}(0) = [1, 0]$. Find: (1) the probability that the next day will be a dry day, (2) the probability that after 2 days, the day will be wet, (3) the probability of states dry and wet after 3 days, and (4) the probability of states dry and wet at any given day (regardless of the initial state). Since initially the day is dry, then $p_{11}^{(1)} = p_{11} = 0.60$ and Eq. (19.3.71) gives $p_{12}^{(2)} = p_{11}p_{12} + p_{12}p_{22} = 0.6 \times 0.4 + 0.4 \times 0.7 = 0.52$. The probabilities of states dry and wet after 3 days are determined by

$$\mathbf{q}(3) = \mathbf{q}(0)\begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}^3 = [1 \quad 0]\begin{bmatrix} 0.444 & 0.556 \\ 0.417 & 0.583 \end{bmatrix} = [0.444 \quad 0.556]$$

Finally, the probabilities of states dry and wet regardless of the initial state (long-run probabilities) are obtained by solving the system of Eqs. (19.3.73). Alternatively, it may be obtained from $\mathbf{P}^{(t)}$ where t is large. For example, for $t = 8$, it may be shown

$$\mathbf{P}^{(8)} = \begin{bmatrix} 0.429 & 0.571 \\ 0.429 & 0.571 \end{bmatrix}$$

Therefore, $\mathbf{q}^* = [0.429 \quad 0.571]$ with approximation to the third decimal figure.

Estimation and Testing. Estimation for a simple Markov chain amounts to estimating the elements p_{ij} of the transition probability matrix. For instance, consider modeling weekly rainfall for a period of ω weeks during the summer and denote by state 1 when it does not rain (dry week) and by state 2 when it rains (wet week). Thus, a sequence of rainfall states for a given summer may appear as 2112211222212 in which $\omega = 13$ (weeks). Thus, for a homogeneous simple Markov chain, the probabilities p_{11}, p_{12}, p_{22}, and p_{22} will be estimated. Assume that the total sample, considering N years of data, is ωN. Then n_{11} = number of times a dry week is followed by another dry week, $\mathbf{n}_{12}$ = number of times a dry week is followed by a wet week, n_{21} = number

of times a wet week is followed by a dry week, and n_{22} = number of times a wet week is followed by a wet week. Furthermore, denote $n_1 = n_{11} + n_{12}$ and $n_2 = n_{21} + n_{22}$. Then, $\hat{p}_{ij} = n_{ij}/n_i$, $i = 1, 2$ and $j = 1, 2$. This algorithm can be extended for r states.

To test whether a simple Markov chain is an adequate model to describe the process under consideration, one can check some of the assumptions of the model and check whether it is able to reproduce some relevant properties of the process. For instance, one should check whether Eq. (19.3.68) simplifies to Eq. (19.3.69). Statistical methods for such tests are available.[40] In addition, one can compare the n-step transition probability of Eq. (19.3.71) with that obtained from the observed data. For $t = 2$ and $r = 2$, one can compare $p_{21}^{(2)} = p_{21}p_{11} + p_{22}p_{21}$ with $\hat{p}_{21}^{(2)}$ obtained from the data. Likewise, for a two-state Markov chain, the probability that the chain remains in state 1 during k steps is $P(L_1 = k) = p_{11}^{k-1}p_{12}$ and the probability that the chain stays in state 2 during k steps is $P(L_2 = k) = p_{22}^{k-1}p_{21}$. These probabilities can be compared with the corresponding probabilities obtained from the historical data.

Remarks. The emphasis of this section has been on homogeneous, first-order (simple) Markov chains. Although in some cases this model may be adequate to model the hydrologic process under consideration, often more complex models may be necessary. For instance, in modeling daily rainfall processes, the parameters of the Markov chain are often assumed to vary with time across the year.[40,149,201] Higher-order Markov chains may be necessary in other cases.[26] In addition, maximum likelihood estimation of parameters has been suggested.[149,201] Furthermore, selection of the order of Markov chain models can be based on the Akaike information criteria.[26,52,149,185]

19.3.7 Point Process Modeling

Simple Point Process. The theory of *point processes* has also been suggested for modeling time series of short-term rainfall since Le Cam[104] and Todorovic and Yevjevich[183] suggested that the occurrence of rainfall showers can be modeled by a *Poisson process.* Assume that storm arrivals are governed by a Poisson process. This means that the number of storms $N(t)$ in a time interval $(0, t)$ is Poisson-distributed with parameter λt, or

$$P[N(t) = n] = \frac{(\lambda t)^n}{n!} \exp(-\lambda t) \qquad n = 0, 1, \ldots \tag{19.3.74}$$

in which λ represents the storm arrival rate (per unit time). Referring to Fig. 19.3.5*a*, n storms arrived in the interval $(0, t)$ at times $t_1, \ldots, t_n$. The number of storms in any time interval T is also Poisson-distributed with parameter λT. The second assumption is that a *white noise* (random) rainfall amount R is associated with a storm arrival. For instance, R can be gamma-distributed with scale parameter μ and shape parameter δ. In addition, $N(t)$ and R are assumed to be independent. Thus, in Fig. 19.3.5*a*, rainfall amounts $r_1, \ldots, r_n$ are associated with storms occurring at times $t_1, \ldots, t_n$. Such a rainfall-generating process, called *Poisson white noise* (PWN), is a simple example of a point process.[35]

Likewise, the cumulative rainfall (mass curve process) in the interval $(0, t)$ is given by

$$Z(t) = \sum_{j=1}^{N(t)} R_j \tag{19.3.75}$$

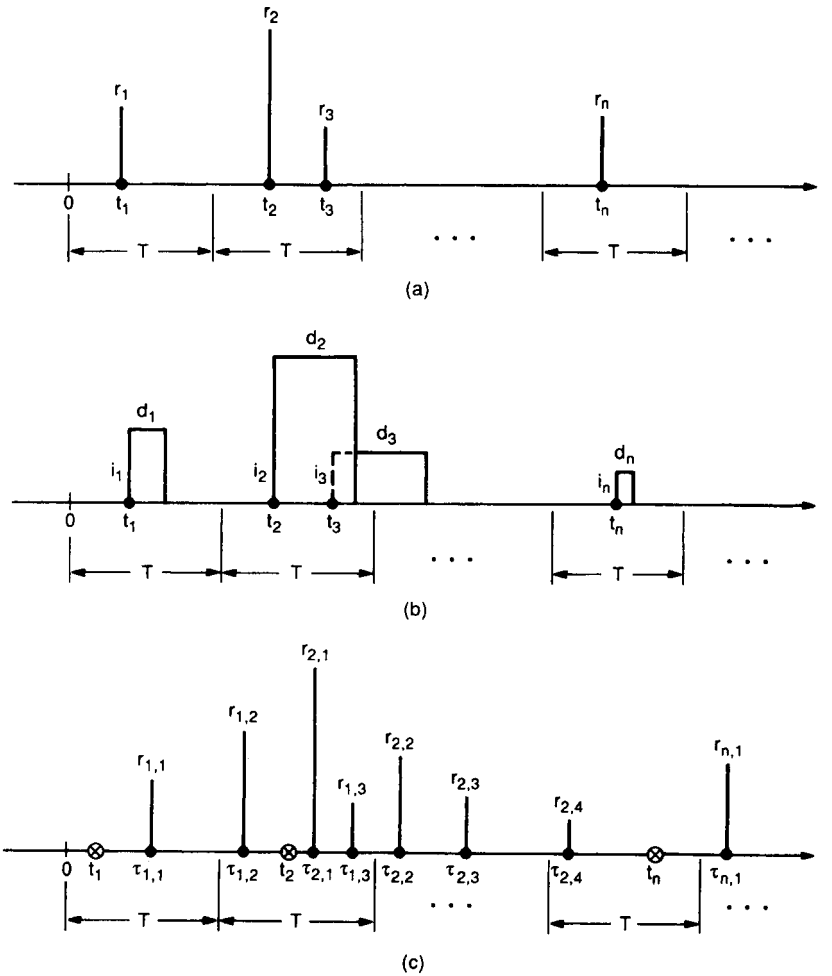

FIGURE 19.3.5 Schematic representation of (*a*) Poisson white noise, (*b*) Poisson rectangular pulse, and (*c*) Neyman-Scott white noise processes.

Such cumulative rainfall $Z(t)$ is called a *compound Poisson process.*[133] While the PWN model and $Z(t)$ describe the rainfall process in continuous time, a related process is the cumulative rainfall Y_i over nonoverlapping time intervals T as depicted in Fig. 19.3.5 (for instance, $T = 1$ h). Thus, the rainfall process Y_i in discrete time is defined by

$$Y_i = Z(iT) - Z(iT - T) \qquad i = 1, \ldots \tag{19.3.76}$$

The basic statistical properties of this process have been studied.[21,35,133] However, the rainfall process Y_i does not reproduce some of the important features of observed rainfall patterns. For instance, the lag-one serial correlation coefficients for hourly

and daily precipitation at Denver Airport station for the months of June and December based on the 1948–1983 record are 0.446 and 0.172, respectively, while the process Y_i derived from the PWN model has zero autocorrelation.[133] Despite this shortcoming, even such a simple model can produce useful results for predicting the distribution of annual precipitation[35] and the distribution of extreme precipitation events.[20]

In the PWN model, the rainfall is assumed to occur instantaneously, so the storms have zero duration, which is unrealistic. Instead, one may consider that rainfalls occur with finite durations (rectangular pulses) as schematically shown in Fig. 19.3.5*b*. Each rainfall occurrence has a random intensity I and a random duration D. This is called the *Poisson rectangular pulse* (PRP) model. I and D can be assumed independent, and each exponentially distributed. Figure 19.3.5*b* shows a PRP process with n storms in the interval $(0, t)$ occurring at times $t_1, \ldots, t_n$ and associated with them are storms of intensities and durations $(i_1, d_1), \ldots, (i_n, d_n)$. In this case, storms can overlap, and, as a result, the aggregated process Y_i will be correlated. Properties and estimation procedures for such PRP process are available.[2,128,144,147] The PRP model is better conceptualized than the PWN model, but it is still limited when applied to rainfall data. Thus, more complex models have been suggested such as those based on the concept of clusters.

Cluster Processes. The concept of *clusters* was originally suggested by Neyman and Scott[127] in modeling the spatial distribution of galaxies. Le Cam[104] and Kavvas and Delleur[84,85] applied this concept of space clustering to model daily rainfall. Further developments have been made.[43,77,141,142,144–147] Here the concept of clusters as applied to modeling rainfall processes at a point is briefly described. The *Neyman-Scott cluster process* can be described as a two-level mechanism for modeling rainfall. First, storm-generating mechanisms (systems), or simply storms, arrive governed by a Poisson process with parameter λt. With reference to Fig. 19.3.5*c*, assume that in the period $(0, t)$, n storms arrive at points $t_1, \ldots t_n$. Then, associated with each storm, there are M precipitation bursts which are Poisson or geometrically distributed with parameter ν. In Fig. 19.3.5*c* there are three precipitation bursts associated with the storm that arrived at time t_1, four precipitation bursts associated with the storm that arrived at time t_2, and, in general, m_j precipitation bursts associated with a storm that arrived at time t_j. In addition, the time of occurrence, τ, of bursts relative to the storm origin t_j may be assumed to be exponentially distributed. For instance, the three bursts arising from the first storm are located at times $\tau_{1,1}$, $\tau_{1,2}$, and $\tau_{1,3}$ relative to time t_1. Finally, if the precipitation burst is described by an instantaneous random precipitation of depth R, then the resulting precipitation process is known as *Neyman-Scott white noise* (NSWN), while if the precipitation burst is described by a rectangular pulse of random intensity I and random duration D, then the precipitation process is known as the *Neyman-Scott rectangular pulse* (NSRP).

Properties and estimation of parameters for Neyman-Scott models are available.[36,43,45,77,84,85,128,141,142,147] The usual estimation approach has been the method of moments, although maximum likelihood has been suggested.[43] An apparent major problem is that parameters estimated based on one level of aggregation, say hourly data, are inconsistent with those estimated from another level of aggregation, say daily.[43,128,144] The problem seems to be that, as data are aggregated, information is lost and corresponding second-order statistics do not have enough information to give reliable estimates of the parameters of the generating process (model). As a consequence, the parameters of the generating process become significantly biased and have large variance. Estimation based on weighted moments of various time scales in a least squares fashion is an alternative.[19,36,77] Physical considerations may be useful

in setting up constraints in some of the parameters and for initializing the estimates to be based on statistical considerations.[77,78] Koepsell and Valdes[92] applied this concept, based on the space-time cluster model suggested by Waymire et al.[200] for modeling rainfall in Texas, and pointed out the difficulty in estimating the parameters even when using physical reasoning as well as statistical analysis.

Remarks. The models included in this section essentially involved concepts of point processes in one dimension as applied to modeling precipitation processes at a point or a site within the class of Poisson processes and Neyman-Scott cluster processes. However, other alternative (somewhat related) temporal precipitation models have been suggested such as those based on Cox processes,[168,169] renewal processes,[44,57,143] and Barlett-Lewis processes.[146] Likewise, alternative space-time multidimensional precipitation models have been suggested.[77,96,170,171] For both temporal and space-time categories, which class of models is best is still open to question. Even small differences in modeling a certain component of a Neyman-Scott process can lead to significant differences in inferring the rest of the model parameters from actual data.[43] Furthermore, all precipitation models proposed to date are limited in some respects; for example, they do not include the daily periodicity observed in actual convective rainfall processes.[128]

19.4 FILLING IN MISSING OBSERVATIONS AND EXTENSION OF RECORDS

Records of hydrologic processes such as precipitation and stream flow are usually short and often have missing observations. Therefore, one of the first steps in any hydrologic data analysis is to *fill in missing values* and to *extend short records.* Observations may be missing for a number of reasons such as interruption of measurements because of equipment (mechanical, electrical, or electronic) failure, effects of extreme natural phenomena such as hurricanes or landslides or of human-induced factors such as wars and civil disturbances, mishandling of observed records by field personnel, or accidental loss of data files in the computer system. Likewise, hydrologic records are generally short; however, no matter the length of the record at a given site, if there are nearby sites with longer records, it may be possible to extend shorter records.

This section includes a number of information transfer techniques which can be applicable to both filling in missing observations and extending records. Some classical methods for filling in missing hydrologic data, such as the normal ratio method and the weighted distance interpolation method, are not included here. The methods included here are based on linear regression and time-series analysis.

19.4.1 Methods Based on Simple Linear Regression

Simple linear regression is most commonly applied for transferring hydrologic information between two gauging stations. Consider a short and a long sequence of a pair of hydrologic random variables represented by y_t and x_t, respectively. For instance, one variable can represent flows, and the other can also represent flows or rainfall. Assume that N_1 is the length of the short sequence and $(N_1 + N_2)$ is the length of the long sequence. The length N_1 also denotes the concurrent period of record. Without

loss of generality, the sequences y and x may be represented as

$$y_1, y_2, \ldots, y_{N_1}$$
$$x_1, x_2, \ldots, x_{N_1}, x_{N_1+1}, \ldots, x_{N_1+N_2}$$

Then, a simple linear regression model may be established to extend the short sequence y_t.

Mathematical Model. A *simple linear regression model* between variables y_t and x_t may be generally represented as[113]

$$y_t = a + bx_t + \alpha\theta(1-\rho^2)^{1/2}\sigma_y\varepsilon_t \tag{19.4.1}$$

where y_t = dependent variable (short record)
x_t = independent variable (long record)
a, b = population parameters of the regression
α = coefficient
$\theta = 1$ when the noise ε_t is added; $\theta = 0$ when ε_t is not added
ρ = population cross-correlation coefficient between y_t and x_t
σ_y = population standard deviation of y_t
ε_t = normal uncorrelated variable with mean zero and variance one which is uncorrelated with x_t

Estimation of Parameters. The estimators of a and b are given by[113]

$$\hat{a} = \bar{y}_1 - \hat{b}\bar{x}_1 \tag{19.4.2}$$

$$\hat{b} = \hat{r}\frac{s_1(y)}{s_1(x)} \tag{19.4.3}$$

where $\bar{y}_1$ and $\bar{x}_1$ are estimated means of the variables y_t and x_t, respectively, based on the concurrent record of size N_1, $s_1(y)$ and $s_1(x)$ are the corresponding estimated standard deviations of y_t and x_t, and $\hat{r}$ is the sample cross-correlation coefficient (refer to Sec. 19.2.1 for estimation of these statistics). In addition, σ_y of Eq. (19.4.1) is estimated by $s_1(y)$ and α is given by

$$\alpha = \left[\frac{N_2(N_1-4)(N_1-1)}{(N_2-1)(N_1-3)(N_1-2)}\right]^{1/2} \tag{19.4.4}$$

Model (19.4.1) can be used to extend the short sequence y_t; i.e., the values $\hat{y}_{N_1+1}, \ldots, \hat{y}_{N_1+N_2}$ are estimated from the concurrent values $x_{N_1+1}, \ldots, x_{N_1+N_2}$. Therefore, the new mean $\bar{y}$ and the new variance $s^2(y)$ of the extended sequence $y_1, \ldots, y_{N_1}, \hat{y}_{N_1+1}, \ldots, \hat{y}_{N_1+N_2}$ are[113]

$$\bar{y} = \bar{y}_1 + \frac{N_2}{(N_1+N_2)}\hat{b}(\bar{x}_2 - \bar{x}_1) \tag{19.4.5}$$

$$s^2(y) = \frac{1}{(N_1+N_2-1)}\left[(N_1-1)s_1^2(y) + (N_2-1)\hat{b}^2 s_2^2(x) + \frac{N_1N_2}{(N_1+N_2)}\hat{b}^2(\bar{x}_2-\bar{x}_1)^2 + (N_2-1)\theta^2\alpha^2(1-\hat{r}^2)s_1^2(y)\right] \tag{19.4.6}$$

Model (19.4.1) without noise ε_t ($\theta = 0$) may be used for filling in missing data only

when just a few records are missing. For a significant number of missing records or for extension of short records in general, model (19.4.1) with $\theta = 0$ causes the variance of the extended record to be underestimated.[113] While this problem is eliminated by considering model (19.4.1) with noise ($\theta = 1$), this requires generating random numbers, which does not lead to a unique extended sequence. Hirsch[70] proposed a method known as *maintenance of variance extension* in which model (19.4.1) with $\theta = 0$ is considered in such a way that the mean $\bar{y}$ and the variance $s^2(y)$ of Eqs. (19.4.5) and (19.4.6), respectively, are maintained. Section 17.4.10 in Chap. 17 discusses some of the principles behind this method. Following Hirsch, model (19.4.1) with $\theta = 0$ may be used for record extension in which the parameters a and b are determined by[192]

$$\hat{a}_1 = \frac{(N_1 + N_2)\,\bar{y} - N_1\,\bar{y}_1}{N_2} - \hat{b}_1 \bar{x}_2 \tag{19.4.7}$$

$$\hat{b}_1 = \left[\frac{(N_1 + N_2 - 1)s^2(y) - (N_1 - 1)s_1^2(y) - N_1(\bar{y}_1 - \bar{y})^2 - N_2\,(\hat{a} - \bar{y})^2}{(N_2 - 1)s_2^2(x)} \right]^{1/2} \tag{19.4.8}$$

where $\bar{y}$ and $s^2(y)$ are estimated by Eqs. (19.4.5) and (19.4.6), respectively; $\bar{y}_1$ and $s_1^2(y)$ are the sample mean and the variance of the original short record; and $\bar{x}_2$ and $s_2^2(x)$ are the sample mean and variance of the longer record $x_{N_1+1}, \ldots, x_{N_1+N_2}$. Thus, Eqs. (19.4.7) and (19.4.8), when used for estimating $\hat{y}_{N_1+1}, \ldots, \hat{y}_{N_1+N_2}$, will produce a record $y_1, \ldots, y_{N_1}, \hat{y}_{N_1+1}, \ldots, \hat{y}_{N_1+N_2}$ which will have mean $\bar{y}$ and variance $s^2(y)$.

Criteria for Improving Estimators of Parameters. In using correlation analysis to extend the short record y_t on the basis of a longer record x_t, a question arises whether the combined record of y_t consisting of N_1 recorded values and N_2 estimated values improves the estimates of the parameters such as the mean and variance. The criteria to be briefly described here assumes model (19.4.1) with noise $\theta = 1$. The variance of the mean $\bar{y}$ based on the longer record is[113]

$$\text{Var}\,(\bar{y}) = \frac{\sigma_y^2}{N_1}\left[1 - \frac{N_2}{N_1 + N_2}\left(r^2 - \frac{1 - r^2}{N_1 - 3}\right)\right] \tag{19.4.9}$$

For $\bar{y}$ (extended record) to be a better estimator of the population mean μ_y than $\bar{y}_1$ (short record), Var $(\bar{y})$ must be smaller than Var $(\bar{y}_1)$. This occurs if

$$|r| > \left(\frac{1}{N_1 - 2}\right)^{1/2} \tag{19.4.10}$$

The right side of Eq. (19.4.10) is called the critical minimum correlation coefficient for improving the estimate of the mean. Such critical correlation is shown in Table 19.4.1 under $m = 1$, column (1) for various values of N_1.

Following the same concept described above, critical minimum correlation coefficients for improving the estimator of the variance can be found. Table 19.4.1, under $m = 1$, column (2), gives such critical correlations for $\theta = 1$, N_1 varying from 8 to 60 and $N_2 = 60$. For all practical purposes, critical values in the table can be applied for any value of N_2.

TABLE 19.4.1 Critical Correlations for Improving the Estimates of the Mean (1) and the Variance (2) for $m = 1, 2, 3, 4, 5$, Values of N_1 From 8 Through 60, and $N_2 = 60$

The assumed models are Eqs. (19.4.1) and (19.4.11) with $\theta = 1$ (m = number of concurrent records used)

	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$	
N_1	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
8	0.408	0.720	0.577	0.835	0.707	0.914	0.816		0.913	
10	0.354	0.650	0.500	0.763	0.612	0.841	0.707	0.799	0.791	0.945
12	0.316	0.597	0.447	0.707	0.548	0.785	0.632	0.845	0.707	0.892
14	0.289	0.556	0.408	0.661	0.500	0.739	0.577	0.199	0.645	0.847
16	0.267	0.522	0.380	0.624	0.463	0.700	0.535	0.759	0.598	0.808
18	0.250	0.494	0.354	0.592	0.433	0.666	0.500	0.725	0.559	0.774
20	0.236	0.469	0.333	0.565	0.408	0.637	0.471	0.695	0.527	0.744
25	0.209	0.422	0.295	0.510	0.361	0.578	0.417	0.634	0.466	0.681
30	0.189	0.386	0.267	0.469	0.327	0.533	0.378	0.587	0.423	0.632
35	0.174	0.359	0.246	0.436	0.302	0.498	0.348	0.548	0.389	0.592
40	0.162	0.336	0.229	0.410	0.281	0.468	0.324	0.517	0.363	0.559
45	0.152	0.317	0.216	0.387	0.264	0.443	0.305	0.490	0.341	0.531
50	0.144	0.301	0.204	0.368	0.250	0.422	0.289	0.467	0.323	0.506
55	0.137	0.288	0.194	0.352	0.238	0.404	0.275	0.447	0.307	0.485
60	0.131	0.276	0.186	0.338	0.227	0.387	0.263	0.429	0.294	0.466

Note: Values of the critical correlation for improving the estimate of the variance are for $N_2 = 60$, but they can be used for any N_2, since the critical correlation does not vary significantly as N_2 varies.

19.4.2 Methods Based on Multiple Linear Regression

Multiple linear regression for transferring information to a site with a short record may be applied when two or more nearby sites with longer records are available. Assume that the short record of size N_1 is represented by y_t and the m longer records of size $N_1 + N_2$ are represented by the vector $\mathbf{x}_t$ as[54]

$$
\begin{array}{llll}
y_1, & y_2, & \ldots, & y_{N_1} \\
x_1^{(1)}, & x_2^{(1)}, & \ldots, & x_{N_1}^{(1)}, x_{N_1+1}^{(1)}, \ldots, x_{N_1+N_2}^{(1)} \\
 & & \vdots & \\
x_1^{(m)}, & x_2^{(m)}, & \ldots, & x_{N_1}^{(m)}, x_{N_1+1}^{(m)}, \ldots, x_{N_1+N_2}^{(m)}
\end{array}
$$

It is also assumed that the concurrent observations are drawn from a multivariate normal population with parameters $\mu_x^{(i)}$, μ_y, $\sigma_x^{2(i)}$, σ_y^2, and R, where $\mu_x^{(i)}$ and $\sigma_x^{2(i)}$ denote the population mean and variance of $x_t^{(i)}$, respectively, for sites $i = 1, \ldots, m$; μ_y and σ_y^2 are the population mean and variance of y_t, respectively; and R is the population *multiple correlation coefficient*. In addition, it is assumed that at each site the observations are serially uncorrelated. The problem is to transfer information from the m sites with records of length $N_1 + N_2$ to the site with a short record.

Mathematical Model. The short record y_t of length N_1 may be related to the m records x_t by the multiple linear regression model[54]

$$y_t = a + \sum_{i=1}^{m} b_i x_t^{(i)} + \alpha\theta(1 - R^2)^{1/2}\,\sigma_y\,\varepsilon_t \tag{19.4.11}$$

where $\theta = 1$ if noise is added, otherwise $\theta = 0$; a and b are estimated by

$$\hat{a} = \bar{y}_1 - \sum_{i=1}^{m} \hat{b}_i \bar{x}_1^{(i)} \tag{19.4.12}$$

$$\hat{b}_i = \sum_{j=1}^{m} \hat{d}_1^{(ij)}\,\hat{c}_1^{(j)} \qquad i = 1, \ldots, m \tag{19.4.13}$$

with $\bar{y}_1$ and $\bar{x}_1^{(i)}$, $i = 1, \ldots, m$, the sample means of y_t and $x_t^{(i)}$, respectively, based on the sample of size N_1; $\hat{d}_1^{(ij)}$ are the elements of the inverse of the matrix whose elements are the lag-zero cross-covariances between $x_t^{(i)}$ and $x_t^{(j)}$, $i, j = 1, \ldots, m$; and $\hat{c}_1^{(j)}$ are the lag-zero cross-covariances between $x_t^{(j)}$ and y_t, $j = 1, \ldots, m$ [these cross-covariances may be determined from Eq. (19.2.15)]. The multiple-correlation coefficient R is estimated from the N_1 concurrent observations as

$$\hat{R} = \left[\frac{N_1 \sum_{i=1}^{m} \hat{b}_i \hat{c}_1^{(i)}}{\sum_{t=1}^{N_1} (y_t - \bar{y}_1)^2}\right]^{1/2} \tag{19.4.14}$$

and the coefficient α is given by

$$\alpha = \left[\frac{N_2(N_1 - 2m - 2)(N_1 - 1)}{(N_2 - 1)(N_1 - m - 2)(N_1 - m - 1)}\right]^{1/2} \tag{19.4.15}$$

Then, the new estimators of the mean μ_y and of the variance σ_y^2 are[54]

$$\bar{y} = \bar{y}_1 + \frac{N_2}{N_1 + N_2} \sum_{i=1}^{m} \hat{b}_i [\bar{x}_2^{(i)} - \bar{x}_1^{(i)}] \tag{19.4.16}$$

$$s^2(y) = \frac{1}{N_1 + N_2 - 1}\left\{(N_1 - 1)\,s_1^2(y) + N_2 \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{b}_i \hat{b}_j \hat{g}_2^{(ij)} + \frac{N_1 N_2}{N_1 + N_2}\left[\sum_{i=1}^{m} \hat{b}_i (\bar{x}_2^{(i)} - \bar{x}_1^{(i)})\right]^2 + (N_2 - 1)\,\alpha^2\theta_2\,(1 - \hat{R}^2)\,s_1^2(y)\right\} \tag{19.4.17}$$

where $\bar{y}_1$ and $s_1^2(y)$ are the sample mean and variance of the short series y_t; $\bar{x}_1^{(i)}$ and

$\bar{x}_2^{(i)}$ are the sample means of the short sample $x_1^{(i)}, \ldots, x_N^{(i)}$ and the additional sample $x_{N_1+1}^{(i)}, \ldots, x_{N_1+N_2}^{(i)}$, respectively; and $\hat{\rho}_2^{(ij)}$, $i, j = 1, \ldots, m$, are the lag-zero cross-covariances between $x_t^{(i)}$ and $x_t^{(j)}$ for the additional samples of size N_2. The maintenance-of-variance extension method described in Sec. 19.4.1 under simple linear regression has been extended to the multivariate case.[60]

Criteria for Improving Estimators of Parameters. The criterion for improving estimates of parameters of the short record for the multiple linear regression model is based on comparing the variances of the original and new estimators. The mean of the short sample y_t is improved if[54]

$$|R| > \left(\frac{m}{N_1 - 2}\right)^{1/2} \tag{19.4.18}$$

Table 19.4.1, under columns (1), gives the critical minimum multiple correlation coefficient for various values of N_1 and m. Likewise, the variance of the short sample is improved if the estimated multiple correlation coefficient R is larger than a critical value.[54,125] Such critical values are given in Table 19.4.1 under columns (2) for $m = 1$ through 5 and $N_1 = 8$ through 60.

19.4.3 Methods Based on Time-Series Models

Missing observations can be filled in and records can be extended by using many of the time-series models described in Sec. 19.3.

Use of AR(1) and PAR(1) Models. These models can be used to fill in missing observations for a given site when no other nearby sites with concurrent information are available. Assume the AR(1) model (19.3.2):

$$y_t = \mu + \phi(y_{t-1} - \mu) + \varepsilon_t \tag{19.4.19}$$

where the model parameters can be estimated from available data. If an observation at time t is missing, but y_{t-1} is known, then y_t can be determined from Eq. (19.4.19) using $\varepsilon_t = 0$ (the mean of ε_t). Since model (19.4.19) assumes stationarity, nonstationary data must be made stationary before the model is applied. A more convenient model for seasonal data is the PAR(1) model (19.3.32):

$$y_{\nu,\tau} = \mu_\tau + \phi_{1,\tau}(y_{\nu,\tau-1} - \mu_\tau) + \varepsilon_{\nu,\tau} \tag{19.4.20}$$

which can be used to fill in missing seasonal data such as monthly observations. The foregoing models should not be used to fill in successive missing observations.

Use of Multivariate Models. Multivariate models can be used for filling in missing data and extension of records. Suppose that y_t is the site with missing records and that data at sites $x_t^{(1)}$ and $x_t^{(2)}$ are available. Then, a multivariate model can be formulated as

$$y_t = a + \sum_{j=1}^{p} b_j y_{t-j} + \sum_{j=0}^{p_1} b_j^{(1)} x_{t-j}^{(1)} + \sum_{j=0}^{p_2} b_j^{(2)} x_{t-j}^{(2)} + \varepsilon_t \tag{19.4.21}$$

in which a, b_j, $j = 1, \ldots, p$; $b_j^{(1)}$, $j = 0, \ldots, p_1$ and $b_j^{(2)}$, $j = 0, \ldots, p_2$ are the parameters to be estimated from the data. Estimation of parameters and testing can

be made by least squares.[34] Usually, p, p_1, and p_2 are small—of the order of 1 or 2. Note that model (19.4.21) falls in the category of *ARMAX* (ARMA with exogenous variables) and *transfer function models.* Applications of multivariate models can be found in Beauchamp et al.[6] and Kottegoda and Elgy.[93]

19.5 MONTE CARLO SIMULATION*

19.5.1 Introduction

Basic Concepts of Monte Carlo Simulation. Consider a *hydrologic system,* in which I represents the *input* and O represents the *output.* The system can be simple or complex, and the input and output of each can be either a single variable or a vector of several variables or any combination of these. For instance, the system may be a watershed system in which the input is simply the average precipitation series over the basin and the output is the stream-flow series at the outlet of the basin. In general, *Monte Carlo simulation* is a method for obtaining the probability distribution of output O given the probability distribution of the input I. Thus, in Monte Carlo simulation studies, three steps are usually required, namely, determining the input, transforming the input into the output, and then analyzing the output.

The input to be used in Monte Carlo simulation studies may be the historical hydrologic records, or artificially or synthetically generated records. In fact, this is one of the purposes of the stochastic models that are included in Sec. 19.3. Further discussion on how to generate synthetic records based on such models is given in Secs. 19.5.2 and 19.5.3 below. The transformation of the input into the output is made by means of a mathematical model which represents the behavior of the physical system under study. In the case of a reservoir system in which the input is the set of inflows to the reservoir and the output is the set of reservoir outflows, the inflows are transformed into the outflows by operating the reservoir according to the reservoir operating rule, a set of constraints, and the mass balance equation of the reservoir. Thus, the transformation process involves routing the input through the system to obtain the output. Put another way, the system input drives the system, which transforms it into the system output. Finally, the system output is analyzed statistically so that it can be used for decision making. The analysis of the output may consist of determining basic output statistics, such as the mean and variance, box plots to observe the variance of the output graphically, and the overall frequency distribution of the output variables under consideration. Details of such analysis can be found in Secs. 17.2 and 17.3 of Chap. 17.

Applications of Monte Carlo Simulation Studies. Some examples of typical applications of Monte Carlo simulation studies in hydrology are included here for illustration. The applications selected are purposely presented in a simplified schematic manner in order to illustrate the underlying concepts only. These concepts can then be extended and applied to more complex cases.

Design the Capacity of a Reservoir. Assume that the capacity of a reservoir for water supply will be determined so that a given water demand d will be delivered from the reservoir throughout a specified planning horizon. N years of historical stream-flow record (inflows to the reservoir) are available; in this case, for simplicity, it is assumed that N coincides with the planning horizon. One possible solution may be to deter-

*Part of the material in this section was contributed by Fidel Saenz de Ormijana.

mine the needed reservoir capacity assuming that the historical record will be identically repeated in the future; this capacity is the output of the system shown as $O(h)$ in Fig. 19.5.1. However, such identical realization of inflows is unlikely to occur in the future. An alternative approach for determining the reservoir capacity is Monte Carlo simulation. Thus, a mathematical model of the inflows x_t is determined and then used to generate a large number of possible sequences that may occur in the future. Then, for each sequence $x_t(i)$, also denoted as $I(i)$ in Fig. 19.5.1, the reservoir capacity $O(i)$ is determined, yielding the set $O(1), \ldots, O(m)$ where m is the number of realizations considered (usually large). This also means that the uncertain occurrence of future inflows is translated into an uncertain reservoir capacity. The set $O(1), \ldots, O(m)$ can be analyzed statistically to provide the hydrologist with additional information to make a decision on what reservoir capacity to use. This use of Monte Carlo analysis for reservoir design has been widely suggested in the literature.[41,107,108]

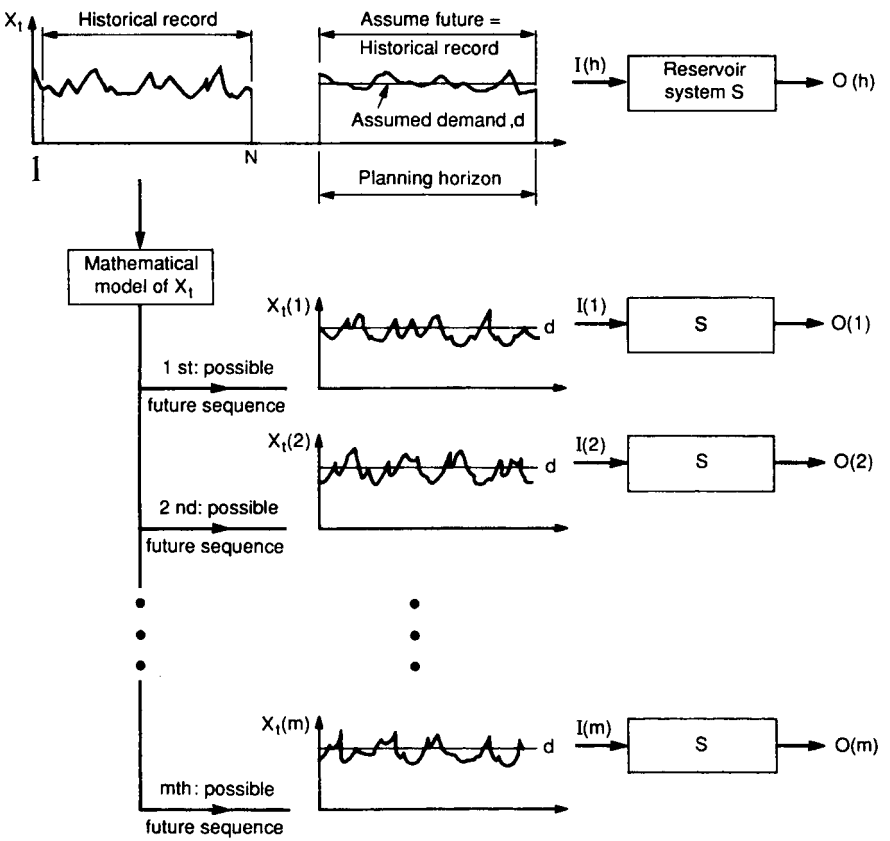

FIGURE 19.5.1 Schematic representation of a hydrologic system in which historical records $I(h)$ and synthetic records $I(i)$, $i = 1, \ldots, m$ are used as inputs for determining historical output $O(h)$ and simulated outputs $O(i)$, $i = 1, \ldots, m$, respectively.

Evaluating the Performance of a Reservoir of Given Capacity. The performance of a reservoir of specified capacity, operating rule, and projected water demands can be evaluated to determine, for instance, the reliability of the system to meet target demands, the total shortages or how often the reservoir may run dry, how often it may spill, and the duration and magnitude of those episodes. This problem can be approached by Monte Carlo simulation studies in which the set of inflows to the reservoir is the system input I and the system output O can be any desired set of performance measures of the reservoir system under consideration. As shown in Fig. 19.5.1, the simulation study can be made by using the historical record alone or by using synthetically generated inflow records. For reservoirs with seasonal regulation, and short operation planning horizon as compared to the historical record, the use of historical records alone may be sufficient. However, since these requirements are rarely met, synthetic records are typically used.

Evaluating the Performance of a Reservoir of Specified Operating Rules. In this case, the capacity of a reservoir and projected water demands are known, but the effect of an operating rule in the performance of the reservoir is being evaluated. A reservoir may have been operated historically under a certain operating rule and an alternative operating rule may need to be considered. In this case, Monte Carlo simulation studies are conducted in which the same inputs (synthetic records) are used for each operating rule. Then the two sets of outputs (performance measures) are compared statistically to determine whether the alternative operating rule gave significantly different results than the original rule.

Evaluating the Performance of Irrigation Water Delivery Systems. The assessment of alternatives to improve irrigation water delivery systems can be made by considering the system's performance under uncertainty. The objective of water delivery systems may be defined in terms of the desire to best meet the water requirements at the farm level. Performance measures related to such objectives can be based on a number of system state variables, for instance, the amounts of water required at a number of diversion points in the system and the amounts of water actually delivered to the diversion points, all variables defined throughout the irrigation season.[65] Monte Carlo simulation studies of system inputs, such as reservoir inflows, effective precipitation, and crop evapotranspiration, can be generated by multivariate models (such as those referred to in Sec. 19.3) for a length equal to the irrigation season. These inputs can then be routed through the irrigation system simulation model to obtain the system's state variables. This process is repeated a large number of times to provide an array of performance measures which can be analyzed statistically.

Determining the Dependable Capacity of Hydropower Systems. A common approach in the analysis of hydropower systems is to use the critical period of record for determining project dependable capacity. However, this approach usually underestimates the power capacity actually available for marketing purposes. An alternative approach is to use Monte Carlo simulation by which inflows to a reservoir system can be synthetically generated for a length equal to the operational horizon (usually several years long), then routed through the system to obtain the hydropower output (power capacity and energy). The process can be repeated several times to obtain an array of hydropower outputs $O(i)$, $i = 1, \ldots, m$ as indicated in Fig. 19.5.1. The output is then subject to statistical analysis to assist in defining a marketable hydropower output.[97]

Determining the Drought Properties of Water-Supply Systems. Drought properties of various return periods are needed to assess the degree to which a water-supply

system will be able to cope with future droughts and, accordingly, to plan alternative water-supply strategies. The estimation of long-term droughts, for instance, the drought length, magnitude, and intensity as defined in Sec. 19.2, based on the historical record alone, are not reliable when the return periods are of the order of the historical record or they cannot be determined when the historical record is smaller than the return period. In these cases, drought properties can be determined by synthetically generating inflows at key points in the water-supply system under consideration, of length equal to the return period of interest, routing such inflows through the system to obtain the flows at the diversion point of the water-supply system. These inflows in turn can be analyzed to obtain the drought property of interest, denoted by $O(i)$ in Fig. 19.5.1. The process is repeated several times to find the array of drought properties $O(1), \ldots, O(m)$ whose average is an estimate of the T-year drought.[48,49]

Other Applications of Monte Carlo Simulation. A number of other applications of Monte Carlo simulation can be found in literature. For instance, for water systems in which a short record of stream flow and a longer record of precipitation are available, one may use the concurrent precipitation–stream flow records to calibrate a watershed model and generate a number of sequences of synthetic precipitation records which then are routed through the watershed model to generate synthetic streamflow records. In turn, these records can be used for design, evaluation of operational rules, or for any other purpose at hand. In some cases, Monte Carlo simulation has been used for estimating floods of long return periods.[46] Hourly rainfall and daily pan evaporation can be synthetically generated and routed through a previously calibrated watershed model to produce a long record of hourly stream flow. Annual flood peaks and volumes can be obtained from such synthetic flow traces, and frequency analysis will provide the flood estimates for the desired return periods.

Monte Carlo simulation has been used for studying the impacts of global climate change on the operation of water resources projects[102] and for determining the variability of the system's output as a result of the uncertainty in the parameters of the system's model. For instance, consider a watershed model used for transforming precipitation into runoff which involves a set of parameters, one of which (for the sake of simplicity) is assumed to be uncertain. Based on previous applications of the model in similar watersheds, one may estimate or assume the distribution of such a parameter. One can randomly sample values of the parameter and for each value run the watershed model to find the runoff sequence corresponding to a given precipitation input sequence. Thus, m values of the parameter will produce m sets of runoff sequences which can be analyzed statistically to determine the effect of parameter uncertainty on the simulated runoff.

Monte Carlo simulation studies are also used for many problems in groundwater hydrology, typically for deriving the distribution of the underlying output variable of a groundwater flow equation, given the distribution of the parameters and boundary conditions. For instance, one can use a steady-state groundwater flow equation for a two-dimensional isotropic nonuniform medium and study the effect of spatial variability of hydraulic conductivity K on the hydraulic head (system output). The variability of K can be modeled by a given probability distribution function and covariance function. Thus, a large number m of realizations of random spatially correlated parameters K can be generated and the groundwater flow equation solved for each realization to find the set of m hydraulic heads at various points in space, from which one can find the distribution of hydraulic heads.[47,55,83] One can extend the foregoing concept to study the effect of variability of other parameters such as porosity, the effect of variability on boundary conditions, and the effect of variability of model inputs such as groundwater recharge. Finally, an important application of Monte

Carlo simulation is to establish the uncertainty in travel time and spread of pollutants in porous media as a function of the uncertainty in the parameters of the groundwater contamination transport model. For this purpose, one can follow the random sampling and simulation approach described previously.

19.5.2 Generation of Random Inputs

In the early days of Monte Carlo simulation, random inputs were generated by randomly drawing cards from a stack, by sequentially reading values from tables of random numbers, or by using devices such as coins or dice. However, with the advent of digital computers and the development of mathematical and statistical techniques, the use of these procedures has become obsolete. Instead, mathematically based approaches for generating random inputs for Monte Carlo simulation studies have become the state of the art. This section includes some basic models for generating uniform and normal random numbers.

Generation of Uniform Random Numbers. Random numbers from a uniform distribution between the bounds 0 and 1 are considered here. These random numbers, referred to as $u(0, 1)$ are widely used for generating random numbers from other distributions. The most popular generators are called *linear congruential* generators, which are integer algorithms of the type $x_i = (ax_{i-1} + b) \bmod c$ where a, b, and c are integers and the mod c notation indicates that x_i is the remainder after dividing $(ax_{i-1} + b)$ by c. This procedure may also be expressed as

$$x_i = ax_{i-1} + b - c \operatorname{Int}\left(\frac{ax_{i-1} + b}{c}\right) \tag{19.5.1}$$

For example, if $a = 13$, $b = 3$, $c = 11$, and $x_0 = 3$, then the first five random numbers generated are 9, 10, 1, 5, and 2. This algorithm produces integer numbers between 0 and $c - 1$. The values of the generated numbers depend on the constants a, b, and c and on the initial value x_0 called the *seed number*. Linear congruential generators always have a finite cycle length; i.e., the sequence repeats itself exactly after generating a certain number of values. There are conditions for a, b, and c, under which the generator yields all the values between 0 and $c - 1$ before repeating.[91]

To obtain numbers uniformly distributed between zero and one, simply let $u_i = x_i/c$. Since u must not be zero or one exactly, c is always a very large value. For certain choices of the parameters a, b, and c, the variable u is approximately uniformly distributed between zero and one. For instance, a common choice of parameters, supported by the IMSL[76] subroutine library, is $a = 16807$, $b = 0$, and $c = 2^{31} - 1$, which may be used with 32-bit-word computers. For computers with smaller word size, a and c should be chosen so that their product does not exceed the machine word size, otherwise an integer overflow will occur. Likewise, the seed number x_0 must be provided in order to start the generation. It may be obtained through some random mechanism, for instance, by reading the computer clock. However, debugging a computer code and checking the output is easier if repeated simulations give the same results. In these cases, it is convenient to use the same starting seed. Table 19.5.1 gives 30 starting seeds for generating uniform random numbers by the linear congruential generator supported by IMSL, as noted above. Each seed value x_0 can be used to generate a set of 131,072 uniform random numbers and all sets will be independent of each other. Additional choices of a, b, and c and alternative generators are available.[14,137]

TABLE 19.5.1 Starting Seeds x_0 for Generating Uniform Random Numbers Based on Eq. (19.5.1) with $a = 16807$, $b = 0$, and $c = 2^{31} - 1$

748932582	250756106	431442774
1985072130	1025663860	1659181395
1631331038	186056398	400219676
67377721	522237216	1904711401
366304404	213453332	263704907
1094585182	1651217741	350425820
1767585417	909094944	873344587
1980520317	2095891343	1416387147
392682216	203905359	1881263549
64298628	2001697019	1456845529

Source: From Bratley et al.[14]

Generally, random number generators such as the linear congruential generator described above produce *pseudo-random numbers* because they follow a deterministic sequence even though the successive values of that sequence are uncorrelated. There are ways to improve the randomness of generated random numbers. One can do *shuffling*, i.e., temporarily store an array of random numbers, using each new generated random number to obtain one from the stored array, then replacing the numbers in the array by newly generated numbers as they are used. Shuffling can do no harm (it has no effect on a string of perfectly generated random numbers) and it improves significantly the sequences produced by a good generator. Press et al.[137] give FORTRAN, Pascal, and C codes for shuffling. *Combining* two or more generators to generate a set of random numbers is another approach.[14] Finally, one can use shuffling and combining simultaneously.[112]

Generation of Normal Random Numbers. A fast and simple method for generating standard normal random numbers ε (i.e., with zero mean and unit standard deviation) is the approximation based on the lambda distribution[82,132,139,140,167,186] as $\epsilon = 4.91\,[u^{0.14} - (1-u)^{0.14}]$, in which u is a uniform (0, 1) random number. The accuracy (the difference between the true standard normal variable and that obtained by the approximation) of this approximation is 0.0032 for $|\epsilon| < 2$, and 0.0038 for $2 < |\epsilon| < 3$, and increases rapidly for $|\epsilon| > 3$. The variable range is limited to $|\epsilon| < 4.91$ (the probability of ϵ being outside this interval is less than 10^{-6}). The variance and kurtosis of the lambda approximation are 0.997 and 2.972, close to the theoretical values 1 and 3, respectively, for a normal distribution. If greater accuracy is desired or the simulation focuses on tail behavior (such as for extreme value analysis), a better approximation may be used.

Several approximations for the normal distribution based on polynomial equations can be found.[209] A rational approximation with relative accuracy 10^{-6} is[131]

$$\epsilon(u) = t + \frac{p_0 + p_1 t + p_2 t^2 + p_3 t^3 + p_4 t^4}{q_0 + q_1 t + q_2 t^2 + q_3 t^3 + q_4 t^4} \qquad 0.5 \le u < 1 \qquad (19.5.2a)$$

$$\epsilon(u) = -\epsilon(1-u) \qquad 0 < u < 0.5 \qquad (19.5.2b)$$

where $t = [-2 \ln (1-u)]^{1/2}$, u is a uniform (0, 1) random number, and the coefficients p_i and q_i, $i = 0, 1, \ldots, 4$ are given by $p_0 = -0.322232431088$, $p_1 = -1$,

$p_2 = -0.342242088547$, $p_3 = -0.0204231210245$, $p_4 = -0.0000453642210148$, $q_0 = 0.099348462606$, $q_1 = 0.588581570495$, $q_2 = 0.531103462366$, $q_3 = 0.103537752285$, and $q_4 = 0.0038560700634$.

Two other methods are often used to generate standard normal random numbers. The sum of uniform random numbers $\epsilon = \sum_{i=1}^{n} u_i - n/2$, where n is commonly 12, yields an approximate normal number with the same mean, variance, and skewness as the standard normal, but kurtosis is 2.9, against a kurtosis of 3.0 for the normal distribution. The range of ϵ is limited to $|\epsilon| < 6$, and the difference between the actual standard normal value and the value obtained by the above approximation is less than 0.009 for $|\epsilon| < 2$, but it can be as high as 0.9 for $2 < |\epsilon| < 3$. Likewise, the Box-Muller method gives two standard normal numbers based on $\epsilon_1 = (-2 \ln u_1)^{1/2} \cos(2\pi u_2)$ and $\epsilon_2 = (-2 \ln u_1)^{1/2} \sin(2\pi u_2)$ in which u_1 and u_2 are two uniform (0, 1) random numbers. This method is slightly faster than the above rational approximation, but the random numbers produced are not independent when the required uniform random numbers are generated by a linear congruential generator.

Remarks. In general, one can generate random numbers from any continuous distribution with cumulative distribution function $F(x)$ by the *inversion method.* The cumulative distribution function $F(x)$ varies between 0 and 1, so if a uniform (0, 1) random number u is generated, this value is made equal to $F(x)$ and the distribution function inverted to find the value of x, then $x = F^{-1}(u)$ is a random number from the distribution $F(x)$. This method can be extended to generate discrete random numbers as well.

19.5.3 Generation of Correlated Inputs

Autocorrelated and cross-correlated inputs are often needed for Monte Carlo simulation studies. For this purpose one can use a number of alternatives including procedures based on the available historical time series and on mathematical models such as those presented in Sec. 19.3. Some of these methods are discussed here in some detail for illustration.

Use of Historical Data. Using historical data is a widespread method of assessing alternative inputs when the available historic series of the input variables are long enough to define a sufficient approximation to the input distribution. For example, in simulating reservoir performance, if the historic stream-flow series is long compared to the simulation horizon, it may be possible to break down the historic stream-flow series into a number of subseries of length equal to the length of the simulation horizon, simulate the system operation with each subseries, and obtain the corresponding output statistics (such as supply reliability) over the simulation horizon. Yet another alternative in using the historic series is to work with the N *wrapped-around series* $\{x_2, x_3, \ldots, x_{N-1}, x_N, x_1\}$, etc., obtained by circular permutation of the historic values. This procedure is also known as *index-sequential.* A major drawback with this procedure is that the resulting set of N input series yields N outputs which are not independent and, as a consequence, the outputs have less precision. However, this approach has been used in some cases.[74,87,97]

Use of Univariate Time-Series Models. Suppose that we need to generate N consecutive observations of an ARMA(p, q) model with known or estimated parameters.

The ARMA(p, q) model defined by Eq. (19.3.8) is rewritten here as

$$z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \qquad (19.5.3)$$

where $z_t = y_t - \mu$ represents the deviations of the process from its mean, $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are autoregressive and moving average coefficients, respectively, and ϵ_t is normally distributed with mean zero and standard deviation σ_ϵ. To generate the z_t's, one can simply generate ϵ_t as needed by the procedures described in Sec. 19.5.2. The problem is that to generate the first value z_1, one needs to known in advance the values $z_{-p+1}, \ldots, z_0$. A number of procedures can be used to get around this problem. A convenient procedure is to use a *warm-up period* in which $r + N$ values $z_{-r+1}, \ldots, z_0, z_1, \ldots, z_N$ are generated setting $z_t = 0$ (its mean) for $t \le 0$, the first r values are deleted, and only the remaining N are used. If r is large enough, this will remove the bias introduced by taking $z_t = 0$ for $t \le 0$ and the effect on $z_1, \ldots, z_N$ will be negligible. A value of r around 50 has been recommended[42] for generations based on low-order ARMA models. In addition, for some low-order ARMA models, the initialization of the generation can be made directly without using the warm-up period.[107,161]

Finally, the ARMA(p, q) normal series are generated by $y_t = \mu + z_t$. If the original data x_t has been previously transformed by using $y_t = g(x_t)$, in which $g(\cdot)$ is a transformation function [for instance, $y_t = \log(x_t)$], then the generated series in the real or original domain will be $x_t = g^{-1}(y_t)$, the inverse transformation function [for instance, $x_t = \exp(y_t)$ if the logarithmic transformation was used]. Likewise, the generation of periodic AR and periodic ARMA time series can be made by following similar procedures except that the periodic parameters must be considered in the generation algorithm.

Use of Multivariate Time-Series Models. Generally the principles involved in univariate generation can be extended to multivariate generation. For instance, one can generate a set of n cross-correlated normal variables $\boldsymbol{\epsilon}_t$ with variance-covariance matrix $\mathbf{G} = \mathbf{B}\mathbf{B}^T$ by using the simple stationary multivariate model (19.3.46) $\boldsymbol{\epsilon}_t = \mathbf{B}\boldsymbol{\xi}_t$ in which $\mathbf{B}$ is an n by n lower triangular matrix and $\boldsymbol{\xi}$ is an n by 1 vector of uncorrelated standard normal variables. First generate the set $\xi_1^{(1)}, \ldots, \xi_1^{(n)}$ of uncorrelated standard normal random numbers. Then, generate successively the set of cross-correlated numbers by expanding the multivariate model as

$$\begin{aligned}
\epsilon_1^{(1)} &= b^{11}\xi_1^{(1)} \\
\epsilon_1^{(2)} &= b^{21}\xi_1^{(1)} + b^{22}\xi_1^{(2)} \\
&\vdots \\
\epsilon_1^{(n)} &= b^{n1}\xi_1^{(1)} + \cdots + b^{nn}\xi_1^{(n)}
\end{aligned}$$

Finally, other sets of cross-correlated numbers $\boldsymbol{\epsilon}_t$ for $t = 2, \ldots$ can be generated by repeating the same procedure.

Likewise, one can generate series which are simultaneously autocorrelated and cross-correlated by using, for instance, multivariate AR(1) model (19.3.42), multivariate ARMA(1, 1) model (19.3.43), or the CARMA model (19.3.44). In any case, initial values are required to start the generation. The initialization can be made by using the warm-up period approach as in the univariate generation; however, a direct method is available for generation based on the multivariate ARMA(1, 1) model.[178] Finally, the foregoing generation procedures can be extended to the case of multivariate autocorrelated and cross-correlated periodic series by using the PAR(1) model (19.3.51), and PARMA(1, 1) model (19.3.52), or the contemporaneous PAR or

PARMA model (19.3.53). The generation procedure is similar to the cases of stationary models except that periodic parameters are used.[155,161]

Use of Disaggregation Models. Consider here as an illustration a generation based on the traditional Valencia-Schaake model. For ease of reference, rewrite model (19.3.58) as $\mathbf{Y}' = \mathbf{A}\mathbf{X}' + \mathbf{B}\boldsymbol{\epsilon}$, where $\mathbf{Y}'$ is the vector of the series being generated (for instance, seasonal series), $\mathbf{X}'$ is the series to be disaggregated (for instance, annual series), $\boldsymbol{\epsilon}$ represents a vector of uncorrelated standard normal variables, and $\mathbf{A}$ and $\mathbf{B}$ are the parameters of the model. It is assumed that all random variables are normally distributed with mean zero, i.e., $\mathbf{Y}' = \mathbf{Y} - \bar{Y}$ and $\mathbf{X}' = \mathbf{X} - \bar{X}$. The parameter matrices obtained for a case of disaggregating annual data at one site into three seasons are

$$\mathbf{A} = \begin{bmatrix} 0.4821 \\ 0.4837 \\ 0.0342 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 17.6242 & 0 & 0 \\ -14.9866 & 4.2467 & 0 \\ -2.6376 & -4.2467 & 0 \end{bmatrix}$$

Note that the column in matrix A adds to unity and all the columns in matrix **B** add to zero in order to preserve additivity of the seasonal flows to form the correct annual flow. In addition, the analysis of annual and seasonal data gives $\bar{x} = 461.04$, $\bar{y}_1 = 168.68$, $\bar{y}_2 = 269.00$, and $\bar{y}_3 = 23.36$. Likewise, the model for annual data is an AR(1) as $X_t' = 0.3X_{t-1} + \xi_t$ in which ξ_t is normal with mean zero and standard deviation 209.5, and $X_t = X_t' + 461.04$, where X_t is the annual flow value. Thus, seasonal series are generated by

$$\begin{bmatrix} y'_{\nu,1} \\ y'_{\nu,2} \\ y'_{\nu,3} \end{bmatrix} = \begin{bmatrix} 0.4821 \\ 0.4837 \\ 0.0342 \end{bmatrix} X'_\nu + \begin{bmatrix} 17.6242 & 0 & 0 \\ -14.9866 & 4.2467 & 0 \\ -2.6376 & -4.2467 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{\nu,1} \\ \epsilon_{\nu,2} \\ \epsilon_{\nu,3} \end{bmatrix} \tag{19.5.5}$$

where ν denotes the year.

The step-by-step generating procedure is

1. Generate a sequence of annual values. For instance, three generated values are $X_1' = 262.9$, $X_2' = -287.5$, and $X_3' = 90.9$.
2. Disaggregate the generated annual values into seasonal values. For the first year, generate three normal (0, 1) values such as $\epsilon_{1,1} = -0.319$, $\epsilon_{1,2} = 0.994$, and $\epsilon_{1,3} = 0.662$. Then Eq. (19.5.5) gives:

$$y'_{1,1} = 0.4821 \times 262.9 + 17.6242 \times (-0.319) = 121.12$$

$$y'_{1,2} = 0.4837 \times 262.9 - 14.9866 \times (-0.319) + 4.2467 \times 0.994 = 136.16$$

$$y'_{1,3} = 0.0342 \times 262.9 - 2.6376 \times (-0.319) - 4.2467 \times 0.994 = 5.61$$

   In a similar manner, the rest of the seasonal values can be computed from the other annual values.
3. Add the seasonal means to the generated seasonal series as $y_{1,1} = 168.68 + 121.12 = 289.80$, $y_{1,2} = 269.00 + 136.16 = 405.16$, and $y_{1,3} = 23.36 + 5.61 = 28.97$.

19.5.4 Length and Number of Simulations

In any analysis involving Monte Carlo simulation, the questions arise of how long the sample size of the generated sequences should be and how many samples should be

generated. Answers to such questions vary from one analysis to another depending on the problem under consideration. This section attempts to give some practical guides for some typical problems in hydrology.

Length of Generated Samples. The length of the generated sample can be the same, shorter, or longer than the length of the historical sample, depending on the particular situation. For instance, consider the case of evaluating a stochastic model in regard to its ability to reproduce historical statistics. The model can be verified theoretically to some extent, especially in relation to statistics such as the mean, standard deviation, and correlations. But the interest may be in regard to more complex statistics such as storage- and drought-related statistics which must be verified by simulation. In any case, the generated sample length should be equal to the length of the historical sample. Now, consider that data generation is required for design of a reservoir. Clearly, in this case, the length of the generated sample must be equal to the planning horizon or economic life of the reservoir being designed.

Likewise, reservoir operational studies require similar considerations. If the purpose of the generation is to test alternative operating rules, the length of the simulation should be equal to the length of the operational planning horizon. In the case that a reservoir is operated under a well-established rule and the purpose is to determine, for instance, the hydropower output, the length of the generation should also coincide with the operational planning horizon (the Western Area Power Administration of the U.S. Department of Energy typically uses 5 to 10 years for this purpose[17]). Finally, if the purpose of the simulation is to determine extreme droughts of a specified return period T, the length of the generation must coincide with T.

Number of Samples to Generate. Enough samples should be generated so that the required output statistics are estimated accurately. To a large extent, the number of samples depends on what statistic of the Monte Carlo simulation output is of interest for the problem at hand. For example, to determine the mean of the output O with a given accuracy, one can use the normal approximation to establish the $1-\alpha$ confidence limits on the population mean $\mu(O)$, from which one can write $P[-u_{1-\alpha/2}\ \sigma(O)/\sqrt{m} < \overline{O} - \mu(O) < u_{1-\alpha/2}\ \sigma(O)/\sqrt{m}] = 1-\alpha$ where $\overline{O}$ = sample mean of the output, $\sigma(O)$ = population variance, $u_{1-\alpha/2} = 1-\alpha/2$ quantile of the normal distribution with mean zero and variance one, $1-\alpha$ = confidence level, and m = sample size. Thus, if $\overline{O}$ must be within $0.1\ \sigma(O)$ of $\mu(O)$ with a probability $1-\alpha = 0.95$, then $u_{0.975} = 1.96$ and the sample size required is given by $1.96\ \sigma(O)/\sqrt{m} = 0.1\sigma(O)$, which gives $m = 384$. Likewise, for an accuracy of $0.2\sigma(O)$, $m = 96$. A better approximation may be generally obtained by using instead the confidence limits based on the t distribution. In this case, the number of samples is obtained, for instance, by solving $t_{1-\alpha/2,m-1} = 0.1\sqrt{m}$ for m [for $0.1\sigma(O)$ accuracy], where $t_{1-\alpha/2,m-1}$ is the $1-\alpha/2$ quantile of the t distribution with $m-1$ degrees of freedom.

Likewise, to determine the standard deviation of the output O with a given accuracy, one can establish the confidence limits on the population variance $\sigma^2(O)$. Thus, one can write

$$P\left(\frac{\sqrt{m-1}}{\sqrt{\chi^2_{1-\alpha/2,m-1}}} < \frac{\sigma(O)}{S(O)} < \frac{\sqrt{m-1}}{\sqrt{\chi^2_{\alpha/2,m-1}}}\right) = 1-\alpha$$

in which $S(O)$ = sample standard deviation and $\chi^2_{\beta,m-1} = \beta$-quantile of the chi-square distribution with $m-1$ degrees of freedom. For example, for $m = 384$ and $1-\alpha = 0.95$, $P[0.934 < \sigma(O)/S(O) < 1.077] = 0.95$, which means that with a sam-

ple size of 384 one can determine the sample standard deviation of the output such that its ratio with the population standard deviation is within about 15 percent.

Furthermore, one may be interested in determining the distribution of the output O with a prescribed accuracy. In this case, one can use the confidence limits on the true distribution F based on the Kolomogorov theorem.[68] Thus, $P(\max|\hat{F} - F| < 1.36/\sqrt{m}) = 0.95$ states that the maximum absolute difference between the sample distribution $\hat{F}$ and the population distribution F is less than $1.36/\sqrt{m}$ with probability 0.95. For instance, for $m = 5000$, the error in estimating the distribution is less than 0.019 with probability 0.95.

In addition, some practical guides have been offered. For example, when data generation is required for designing a reservoir, if annual data are used, as many as 1000 samples may be needed to accurately define the probability distribution of the maximum storage required.[108] On the other hand, if monthly flows are used, fewer samples may be adequate. In general, the number of samples varies from about 300 for streams that exhibit low variability to the order of 1000 for streams with high variability.[108] Obviously, these practical guides are less precise than the statistical criteria given above.

19.5.5 Model and Parameter Uncertainty

Much of the material of the previous sections was presented without making explicit reference to model uncertainty and parameter uncertainty. However, the importance of such uncertainties for many applications of Monte Carlo simulation must be recognized.[81,89,90,194] Model uncertainty arises because hydrologic processes are inherently complex, so alternative mathematical formulations have been proposed to reproduce the historical record in a statistical sense. Thus, for a particular problem at hand, the hydrologist is faced with the problem of selecting a model among the several alternatives. For instance, in the case of stream-flow modeling, a number of alternative short-memory and long-memory models have been proposed. On the other hand, parameter uncertainty arises because the model parameters are estimated from historical samples which usually are small. Naturally, both model uncertainty and parameter uncertainty are closely related.[90]

Model uncertainty can be alleviated in a number of ways. One can use physical arguments in deriving the model structure. For instance, in the case of stream-flow modeling, conceptual arguments have been proposed to justify the use of ARMA and PARMA processes.[5,41,155] Likewise, in the case of modeling precipitation processes, conceptual arguments have been proposed to justify the family of cluster processes.[197-199] Given that a family of models has been selected, the problem is still to define the type of model within the family. Statistical considerations may be used to assist in identifying the type of model. For instance, in the case of ARMA models, the type of model is selected by using diagnostic checks and other criteria such as the Akaike information criteria.[1] Likewise, model selection can be made by Monte Carlo simulation experiments to compare the performance of competing models in reproducing historical statistics, especially those statistics which have not been used in parameter estimation and those which are pertinent to the study at hand. Furthermore, in complex systems involving several sites, the issue of model selection is really more in terms of selecting among model strategies involving an array of univariate, multivariate, and disaggregation models. In these cases experience will likely provide the best solution.

TABLE 19.6.1 Summary of General-Purpose Programs for Time-Series Analysis, Modeling, and Forecasting

Name	Brief description	Reference
1. BMDP	Time-series analysis, modeling, and forecasting, estimation of missing data, analysis of variance, and nonparametric tests. Time-series modules include univariate and bivariate spectral analysis, intervention analysis, and ARMA and transfer-function models. Estimation of missing data includes modules based on regression analysis. Nonparametric analysis modules include several tests for detection of shifts.	BMDP[8] BMDP Statistical Software 1440 Sepulveda Blvd., Suite 316 Los Angeles, CA 90025 U.S.A.
2. IMSL	Time-series analysis including modeling simulation and forecasting, analysis of variance, and nonparametric tests. Modeling, simulation, and forecasting are based on ARMA models. Forecasting includes Kalman filtering. Programs for nonparametric tests include tests for detection of trends and shifts, tests of randomness, and tests of goodness of fit.	IMSL[76] IMSL 14141 Southwest Freeway, Suite 3000 Sugarland, TX 77478-3498 U.S.A.
3. ITSM	Time-series analysis and modeling and forecasting based on ARMA models.	Brockwell and Davis[16] ITSM Statistical Dept. Colorado State University Fort Collins, CO 80523 U.S.A.
4. MINITAB	Time-series analysis and modeling based on ARMA models. It includes programs for analysis of variance and for nonparametric tests to detect trends and shifts and test for randomness.	Minitab[123] Minitab Inc. 3081 Enterprise Dr. State College, PA 16801-3008 U.S.A.
5. SAS/ETS	Time-series analysis and modeling and forecasting based on ARMA models.	SAS/ETS[150] SAS Institute Inc. SAS Campus Dr. Cary, NC 27513 U.S.A.
6. SPSS	Time-series analysis, modeling and forecasting, analysis of variance and nonparametric tests. Modeling and forecasting are based on ARMA models. Programs for nonparametric tests include tests for detection of shifts, tests of randomness and tests of goodness of fit.	SPSS[174] SPSS Inc. 444 N. Michigan Ave. Chicago, IL 60611-3962 U.S.A.

TABLE 19.6.1 Summary of General-Purpose Programs for Time-Series Analysis, Modeling, and Forecasting *(Continued)*

Name	Brief description	Reference
7. Statgraphics	Time-series analysis and modeling, analysis of variance, and nonparametric tests. Time-series modeling is based on ARMA models. Modules on nonparametric methods include tests for detection of shifts, randomness, and goodness of fit.	Statgraphics[175] STSC Inc. 2115 E. Jefferson St. Rockville, MD 20852 U.S.A.

Parameter uncertainty is a major issue in stochastic modeling and simulation studies.[90,91] Procedures have been developed to deal with parameter uncertainty,[116,155,179,188,191] although in practice they are rarely applied. For instance, consider the AR(1) model of Eq. (19.3.2). To include parameter uncertainty in data generation studies based on this model, one needs to know the distribution of the parameters μ, ϕ_1, and σ_ϵ^2 given a historical sample of size N. In this case the sample distributions are[116,155] $\mu \sim N[\hat{\mu}, \hat{\sigma}_\epsilon^2/[(1-\hat{\phi}_1)^2 N]]$, $\phi_1 \sim N[\hat{\phi}_1, (1-\hat{\phi}_1^2)/(N-1)]$ and $\sigma_\epsilon^2 \sim N[\hat{\sigma}_\epsilon^2, 2\hat{\sigma}_\epsilon^4/N]$, where $\hat{\mu}$, $\hat{\phi}_1$, and $\hat{\sigma}_\epsilon^2$ denote the estimates of μ, ϕ_1, and σ_ϵ^2 obtained from the historical sample and $\sim$ denotes *distributed as.* The procedure to simulate m sequences of length n which includes parameter uncertainty is:[116,155] (1) follow steps 2 and 3 for $i = 1, \ldots, m$; (2) for sequence i, generate the parameter set $\mu(i)$, $\phi_1(i)$, and $\sigma_\epsilon^2(i)$ from the above normal distribution, respectively; and (3) using model (19.3.2) and the parameters $\mu(i)$, $\phi_1(i)$, and $\sigma_\epsilon^2(i)$, generate the sequence $y_1(i)$, $y_2(i)$, $\ldots$, $y_n(i)$ following the approach suggested in Sec. 19.5.3. Similar procedures for univariate ARMA and PAR models are also available.[116,155,179] For multivariate models the problem is more complex; however, bayesian procedures are available.[188] In complex modeling and simulation studies involving seasonal data and multiple sites, the consideration of parameter uncertainty is generally complex. However, since such complex systems usually involve an array of models, it is always possible to include parameter uncertainty at least in some parts of the modeling process.[155,179]

19.6 COMPUTER PROGRAMS FOR TIME-SERIES ANALYSIS

Analysis, modeling, and simulation of hydrologic time series can be done effectively with the aid of computers. Several alternative software packages are available. Some of them are general-purpose programs for analysis of any kind of time series and others are specifically oriented for hydrologic time series. General-purpose program packages, such as those included in Table 19.6.1, are attractive because usually they are well-documented, have more statistical features, and are accompanied by good

TABLE 19.6.2 Summary of Special Programs for Hydrologic Time-Series Analysis, Modeling, and Simulation

Name	Brief description	Reference
1. HEC-4	Analysis, modeling, and simulation of multiple time series of monthly flows based on multiple linear regression models. It can also generate synthetic monthly flows at ungauged sites.	U.S. Army Corps of Engineers[187] Dept. of the Army Hydrologic Engineering Center 609 2d St. Davis, CA 95616-4687 U.S.A.
2. LAST	Analysis, modeling, and simulation of multiple annual and seasonal streamflow data. The main features of the approach are (1) preservation of annual serial correlation and annual cross-correlations, (2) generation of "key" stations annual flows and disaggregation of these values into component substations on an annual basis, and (3) disaggregation of annual flows into seasonal flows preserving both season-to-season correlations and cross-correlation between sites. Modeling and simulation are based on univariate and multivariate AR(1) and AR(2) models and disaggregation models.	Lane,[98] Lane and Frevert[100, 101] LAST Bureau of Reclamation, D-5077 Earth Sciences Div. P.O. Box 25007 Denver, CO 80225-007 U.S.A.
3. SPIGOT	Three modeling and generation schemes are used: (1) aggregated annual flow for the entire basin is generated by a univariate model, then is disaggregated into basin (aggregated) monthly flows, which in turn are disaggregated into key site monthly values; (2) aggregated annual flow for the entire basin is generated by a univariate model, then is disaggregated into monthly flows at key sites in a single step; and (3) annual flows at key sites are generated by a multivariate model and are disaggregated into monthly flows by a multivariate disaggregation model. In all schemes, modeling and simulation of annual flows are based on either univariate or multivariate AR(0) or AR(1) models, as the case may be.	Grygier and Stedinger[61] Dr. J. R. Stedinger School of Civil Engineering Hollister Hall Cornell University Ithaca, NY 14853-3501 U.S.A.

TABLE 19.6.2 Summary of Special Programs for Hydrologic Time-Series Analysis, Modeling, and Simulation *(Continued)*

Name	Brief description	Reference
4. CSUPAC1	Consists of programs CSU001 and CSU002 for modeling and generation of single-site hydrologic series and programs CSU003 and CSU004 for modeling and generation of multisite series. Univariate modeling is based on PAR(0), PAR(1), PAR(2), and PARMA(1, 1) models, and multivariate modeling is based on low-order contemporaneous PAR or contemporaneous PARMA models. Options for alternative transformations and Fourier series analysis are included.	Dr. J. D. Salas Engineering Research Center Colorado State University Fort Collins, CO 80523 U.S.A.
5. WASIM	Consists of programs WASIM1 and WASIM2 for modeling and generation of hydrologic time series based on stationary ARMA models, and program WASIM3, which includes parameter uncertainty in the generation.	McLeod and Hipel[116] Dr. Angus I. McLeod Statistics and Actuarial Science Group University of Western Ontario London, Ontario Canada N6A5B9

graphical display capabilities. However, generally, they are expensive. Furthermore, most general-purpose packages, for the most part, do not consider the model structures that one normally finds in hydrologic time series (an exception is the family of stationary ARMA models); estimation based on short samples, which is a typical case in hydrology; or periodicity in the covariance structure, which is also typical in seasonal hydrologic time series. Nor do they consider aggregation and disaggregation schemes on direct approaches to deal with nonnormal series.

On the other hand, programs developed specifically for hydrologic time series, such as those included in Table 19.6.2, have the advantage that the underlying models and estimation procedures involve features that are unique to hydrologic time series. However, no single package can handle all cases that may arise in practice, yet most packages can be applicable to the typical cases. Likewise, most packages have been developed for modeling seasonal and annual hydrologic series, such as stream-flow series, while packages for modeling and generation of short-term processes such as hourly rainfall are not readily available.

Generally, any computer package must be used with care. This is especially true in programs that do estimation and generation all at once. In this section, general-purpose programs and special programs for hydrologic time-series analysis, modeling and generation are presented in summarized form in Tables 19.6.1 and 19.6.2. The intent here is to make hydrologists aware of what computer packages are available without attempting detailed comparisons among them.

REFERENCES

1. Akaike, H., "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1924.
2. Bacchi, B., P. Burlando, and R. Rosso, "Extreme Value Analysis of Stochastic Models of Point Rainfall," 3d Scientific Assembly of the IAHS, Poster Session, Baltimore, May 10–19, 1989.
3. Bardossy, A., and E. J. Plate, "Modeling Daily Rainfall Using a Semi-Markov Representation of Circulation Pattern Occurrence," *J. Hydrol.*, vol. 122, pp. 33–47, 1991.
4. Bartolini, P., and J. D. Salas, "Properties of Multivariate Periodic ARMA(1, 1) Models," Int. Symp. Multivariate Analysis in Hydrologic Processes, Colorado State University, Fort Collins, Colo., July 15–17, 1985.
5. Bartolini, P., and J. D. Salas, "Modeling of Streamflow Processes at Different Time Scales," submitted to *Water Resour. Res.*, 1992.
6. Beauchamp, J. J., D. J. Downing, and S. F. Railsback, "Comparison of Regression and Time Series Methods for Synthesizing Missing Streamflow Records," *Water Resour. Bull.*, vol. 25, no. 5, pp. 961–975, 1989.
7. Benjamin, J. R., and C. A. Cornell. "Probability, Statistics and Decisions for Civil Engineers," McGraw-Hill, New York, 1970.
8. BMDP, *BMDP Statistical Software.* University of California Press, 1981.
9. Bobee, B., and R. Robitaille, "Correction of Bias in the Estimation of the Coefficient of Skewness," *Water Resour. Res.*, vol. 11, no. 6, pp. 851–854, 1975.
10. Box, G. E. P., and D. R. Cox, "An Analysis of Transformations," *J. Royal Stat. Soc.*, Series B, vol. 26, pp. 211–252, 1964.
11. Box, G. E. P., and Pierce, D. A., "Distribution of Autocorrelations in Autoregressive Integrated Moving Average Time Series Models," *J. Am. Stat. Assoc.*, vol. 65, pp. 1509–1526, 1970.
12. Box, G. E. P., and G. M. Jenkins, "Time Series Analysis Forecasting and Control," Holden-Day, San Francisco, 1976.
13. Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology,* Addison-Wesley, Reading, Mass., 1985.
14. Bratley, P., B. L. Fox, and L. E. Schrage, *A Guide to Simulation* (2d ed.), Springer-Verlag, New York, 1987.
15. Brockwell, P. J., and R. A. Davis, *Time Series: Theory and Methods* (2d ed.), Springer-Verlag, New York, 1991*a*.
16. Brockwell, P. J., and R. A. Davis, *ITSM: An Interactive Time Series Modeling Package for the PC,* Springer-Verlag, New York, 1991*b*.
17. Buishand, T. A., "Stochastic Modeling of Daily Rainfall Sequences," Dept. of Mathematics and Dept. of Land Water Use, Agricultural University, Wageningen, The Netherlands, 1977.
18. Burges, S. J., "Some Problems with Lognormal Markov Runoff Models," *J. Hydraul. Div., Am. Soc. Cir. Eng.* (now *J. Hydraul. Eng.*), vol. 98, no. HY9, pp. 1487–1496, 1972.
19. Burlando, P., "Modelli Stochastici per la Simulazione e la Previsione della Precipitazione nel Tempo," Ph.D. thesis, Institute of Hydraulics, Politecnico di Milano, Italy, 1989.
20. Burlando, P., and R. Rosso, "Stochastic Models of Temporal Rainfall, Reproducibility, Estimation and Prediction of Extreme Events," NATO Advanced Study Institute on Stochastic Hydrology in Water Resources Systems: Simulation and Optimization, Peñiscola, Spain, Sept. 18–29, 1989.
21. Cadavid, L. G., J. D. Salas, and D. C. Boes, "Disaggregation of Short-Term Precipitation