STRUCTURAL ANALYSIS OF HYDROLOGIC TIME SERIES

by

VUJICA YEVJEVICH

November 1972

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

56

STRUCTURAL ANALYSIS OF HYDROLOGIC TIME SERIES

by

Vujica Yevjevich*

Hydrology Papers
Colorado State University
Fort Collins, Colorado 80521

November 1972

No. 56

*Professor of Civil Engineering, Colorado State University, Fort Collins, Colorado.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

ACKNOWLEDGMENT

The results of investigation presented in this paper are obtained within the activities of the research projects "Stochastic Processes in Hydrology" sponsored by U. S. National Science Foundation, Grant No. GK-11444, with the writer as the principal investigator. This support is gratefully acknowledged. Several graduate research assistants have worked closely with the writer on this research project, studying various aspects of hydrologic stochastic time processes, either in their special studies or in M.S. and Ph.D. dissertations. Some of these works are referred to in the text.

This investigation of hydrologic time series has been helped by a cooperative effort between Colorado State University and the U. S. Bureau of Reclamation, in developing a practical method, which is application oriented, in the structural analysis and mathematical description of time series for the purpose of generating new samples. This work is condensed in the report "Mathematical Model for Obtaining New Samples from the Available Hydrologic Time Series". The report, based on findings in this paper, is not included in this text, awaiting tests by various practical applications to hydrologic time series of sufficiently diverse properties. The writer of this paper, specialized personnel in computer applications, and graduate research assistants, represented Colorado State University. Albert E. Gibbs and Eugene A. Cristafano, both of the Division of Planning Coordination, and hydrologists of Region 7, represented the U. S. Bureau of Reclamation, Denver, Colorado, in this endeavor. This cooperation is acknowledged.

ABSTRACT

Structural analysis and mathematical description of hydrologic time series are based on a set of well defined hypotheses, with the assumption that a development cannot be better than hypotheses underlying it. Techniques are presented on how to infer the existence of periodic deterministic component parameters in time series. The unavailability of exact inference techniques is replaced by approximations, whenever the complexity of hydrologic time series does not justify the use of existing statistical inference techniques. Once periodicities are inferred, Fourier analysis is used to mathematically describe periodicities in parameters by a minimum number of low frequency harmonics and their estimated coefficients.

Inferred dependence models for the stationary stochastic components of a given order of stationarity, after all periodicities in parameters are removed, are basically of the autoregressive linear type. The assumption is that the autoregressive coefficients may be both periodic and nonperiodic. Several misinterpretations of autoregressive linear models are discussed. These misinterpretations are due to doubts often advanced by some superficial arguments on the applicability of these models in hydrology.

The frequency distribution curve of the independent stochastic stationary component, derived from the inferred dependence model, is approximated by best fit as one among various probability distribution functions studied. They are more or less simple. The small interval time series (say 1-day, 2-day, 3-day, and 7-day discrete time series) require less simple probability distribution functions to fit these frequency curves, while for the large interval time series (say 15-day, monthly, or bimonthly time series) simple probability functions produce good fits.

Biases in time series which should not be reproduced or perpetuated by structural analysis, mathematical description and generation of new samples, are outlined and discussed.

Chapter 1

INTRODUCTION

To place this study in context with the writer's continuous efforts regarding the structural analysis of hydrologic time series, the problems related to this analysis and the methods of attacking them are first defined and outlined. Basic concepts are sketched to provide the reader insight into the philosophy of the approach used in this study.

1.1 Previous Work

It has been shown in previous Hydrology Papers [1, 2, 3, 4]** and in an article [5] that hydrologic time series of precipitation and runoff have definite structural patterns. To define the various components of a hydrologic time series, the general results of these patterns are briefly summarized.

Sequences of annual precipitation, annual effective precipitation (precipitation minus evaporation) and natural annual runoff from river basins are approximately stationary time series, or with series properties independent of the absolute time. Series of annual precipitation are nearly independent time series, and the series of annual effective precipitation and annual runoff are either independent or dependent stochastic variables. In this latter case the variables are usually of a simple linear dependence, approximately of the first- and the second-order autoregressive (or Markov) linear models [1,2].

Series of monthly precipitation, monthly effective precipitation, monthly runoff, as well as monthly series of many other hydrologic variables, have periodic components of 12 months in both monthly means and monthly standard deviations. When these periodic components are removed from a monthly series, the remaining part or component may be considered approximately an independent stationary stochastic process for monthly precipitation, and approximately a linearly dependent stationary stochastic process for monthly runoff [3]. Similar patterns appear in series of daily river flows [4, 5]. The autocovariances (or autocorrelation coefficients) as well as the higher-order parameters may also be periodic for some time series of hydrologic random variables. Runoff series (say daily flow series) are among the most complex geophysical time processes, however.

**Reference numbers in the text refer to the references at the end of the paper.

1.2 Objective of this Study

The objective of this study is to present a detailed analysis of the structure of hydrologic time series beyond the previous work. With this in mind inquires into the structure of hydrologic time series are made in several ways: (1) separation of a discrete time series (of time intervals less than a year) into periodic and stochastic components; (2) investigation of whether various parameters of time series are periodic or nonperiodic; (3) determination of significant harmonics (amplitudes are significantly greater than for nonperiodic series) in the periodic parameters; (4) analysis of whether stochastic components are dependent or independent; (5) fitting of adequate mathematical dependence models and the computation of independent stochastic variables from dependent stochastic components; (6) fitting probability distribution functions to independent stochastic components and selecting the function of best fit; (7) derivation of the structure as a final mathematical model of a time series; (8) description of various biases in time series that influence both the structural analysis and the final mathematical models; (9) selection of statistical inference techniques for an objective but practical structural analysis of time series; (10) eventual physical explanation of various structural properties of hydrologic time series; (11) separation of mathematical models into deterministic and stochastic parts, with the parameters of deterministic components subject to sampling errors; (12) computation of explained variances of time series by various components, and similar inquiries.

1.3 Significance of the Analysis

Two basic results are significant from this study: (1) an improved understanding and mathematical description of hydrologic stochastic processes by a better analysis of time series structure, and (2) developments of an improved methodology for the generation of samples of hydrologic time series by the experimental statistical (the Monte Carlo) method.

To explore these results, an understanding of the composition of the structure of a hydrologic time series is needed. This composition should result from the analysis of the structure of any hydrologic time series. First, there may be either a trend or a long-term persistent movement appearing as a smooth broad motion extending over years, and/or as slippages (positive and negative jumps) and other transient deterministic components created either by nonhomogeneity and inconsistency in data or by sampling variations. Second, a periodic movement with a basic cycle of the year is nearly always present in time series of daily and monthly values. Third, when the deterministic parts in the form of trends, slippages, and periodic components in statistical parameters are removed from a time series by using the algebraic composition of its structure, only the stochastic component remains, usually as a stationary process. Thus, to reach the ultimate in analysis of time series structure, its decomposition into components is necessary. Because a time series may be represented by an algebraic equation between various components does not mean that these components may be used as separate series for the solution of various water resource problems, though in some cases this approach may be feasible.

Considering various components as being produced by unrelated causative factors may lead to wrong conclusions. It is necessary to remember that the worthwhile objective of a detailed analysis of a hydrologic stochastic process is an explanation of the time series properties by physical causative factors. To study the potential physical causative factors of a component separately, it must be isolated. In periodic-stochastic processes this is done by separating the deterministic processes from the stochastic process. In other words, deterministic components should be identified, proven, and separated from the remaining part of the series, thereby isolating the stochastic component.

This analysis of the structure of a hydrologic series and the physical explanation for its components may be compared, only along general lines, to the analysis of a communication time series. The signal components of a communication time series are equivalent to deterministic (periodic, transient) components of hydrologic time series. Its noise (or stochasticity) is comparable to the stochastic component of hydrologic series. The difference is that physical description and explanation of signal and noise components, and their connection, is often much simpler in communications engineering than in hydrology; the components of hydrologic series are still far from being well understood particularly the interaction between periodic and stochastic components. Although periodicities are explained by astronomic cycles, and consequently by the periodicity in the energy supply from the sun over various areas of the earth's surface, and further interactions and responses of various earth's environments, the complexity of periodic components needs a much better physical analysis than is presently available. Similarly, though the stochastic component may be explained by various random processes in the air, over oceans and at the continental surfaces, that is over various geophysical environments, many more efforts are needed to improve its physical understanding, explanation and description. Analysis of the structure of hydrologic time series is considered here as a necessary initial step for a comprehensive physical explanation of composition of time and space hydrologic stochastic processes.

1.4 Physical Background of Hydrologic Stochastic Processes

Basic characteristics of hydrologic time series, such as eventual long-term trends and other persistencies, the periodicities of the year and of the day, and the randomness and time dependence of stochastic variations, may be physically or statistically explained in the following ways.

(a) Long-range trends and other eventual long-range persistencies. Inconsistency (systematic errors) and nonhomogeneity (changes in nature by man-made or by natural processes) are mainly responsible for the long- range trends or sudden changes. They must be identified and removed if they are not expected either to continue or to continue in a different manner. However, causative factors produced by the historical study of operating gaging stations and environmental changes in river basins should support the statistical detection of trends and positive and negative jumps.

Trends and cyclicity are often the results of sampling fluctuations for short time series. When periodicity is only a result of sampling variation, it is called in this paper pseudo or sampling cyclicity. The main problem is to determine for these trends and pseudo-cycles not to be statistically significant. For example, a series of N years is divided into four

subseries each of the size N/4. The four means of these subseries are ranked as $\bar{x}_1 = \bar{x} + \epsilon_1$, $\bar{x}_2 = \bar{x} + \epsilon_2$, $\bar{x}_3 = \bar{x} + \epsilon_3$ and $\bar{x}_4 = \bar{x} + \epsilon_4$, with $\epsilon_1 < \epsilon_2 < 0 < \epsilon_3 < \epsilon_4$, and $\bar{x}$ is the mean of N values. These four independent values can be permutated in 24 sequences. The probability to obtain the upward trend of these means, $\bar{x}_1 \rightarrow \bar{x}_2 \rightarrow \bar{x}_3 \rightarrow \bar{x}_4$, is 1/24, and the downward trend, $\bar{x}_4 \rightarrow \bar{x}_3 \rightarrow \bar{x}_2 \rightarrow \bar{x}_1$, is also 1/24. To have a pseudo-cycle, like $\bar{x}_1 \rightarrow \bar{x}_3 \rightarrow \bar{x}_2 \rightarrow \bar{x}_4$, or $\bar{x}_2 \rightarrow \bar{x}_1 \rightarrow \bar{x}_4 \rightarrow \bar{x}_3$, or similar combinations of small-large-small-large, there are four combinations with the total probability of 1/6. Therefore, to show a weak long-range trend or pseudo-cyclicity only by the sampling variation the probability is 5/12, if four subseries are used as an example.

These sampling fluctuations in small samples should not be overlooked in any study of the structure of hydrologic time series. Regional studies should decide whether there is any significant trend or significant pseudo-cyclicity to be assigned to a particular series in the area. However, the set of series in a region for this investigation should not come from stations too near each other, because of the high correlation among hydrologic series of neighboring stations.

In conclusion, the apparent long-range trends and pseudo-cyclicity should not be considered as a permanent property of any series of annual values of a hydrologic variable (after the known non-homogeneity and inconsistency are removed), if it is not confirmed by regional studies. If a regional study shows that the phenomenon follows a stationary process of annual values, the sampling fluctuation with trends and pseudo-cyclicity at some stations inside this region should not be considered as significant, and should not be perpetuated in structural analysis and mathematical description of time series, nor should it be perpetuated in generating new samples by the Monte Carlo method.

(b) Sources of within-the-year periodicity, randomness and time dependence. Astronomic cycles cause periodicity ("signals") in various hydrologic time series. Turbulence, large-scale vorticities, heat transfer, and similar sources of randomness of fluid mechanics, air opacity for radiation waves, thermodynamic and other processes, are responsible for randomness or "noise" in these series. Storage of various quantitites in hydrologic environments and the resulting smoothing effects are factors that attenuate the periodic process and create or increase the time dependence in stochastic variation. Inputs to hydrologic environments are mainly a composition of periodic and random parts which often mutually interact. These environments respond in three ways: (1) by smoothing or magnifying the inputs; (2) by adding, amplifying, or dampening the periodicity if the environments have some periodic aspects in their responses, and (3) by adding or modifying randomness due to various factors in the environment which change and/or react with chance elements in them. In summary, the analysis of hydrologic time series should show components and their properties as described.

The basic approach in this investigation is the assumption that periodic components are deterministic properties of various time series parameters. Further, a stationary stochastic process is superposed on them in a given manner and is described by algebraic equations of time series composition. Therefore, hydrologic time series are nonstationary processes. By the described fundamental assumptions a nonstationary process can be decomposed into deterministic components and a stationary stochastic process. This approach requires analysis of periodicity in as many parameters as necessary to obtain a given order of stationarity of the stochastic component.

Experience indicates periodicity components are mainly deterministic processes. It is difficult to find physical factors in various hydrologic environments that would change the basic astronomic periodicities. These environments can modify the amplitudes and cause the phases of various subharmonics of the basic periodicity to change. If this can not be accepted in reality, the structural analysis of hydrologic time series can not use all the advantages of mathematical techniques developed for stationary processes, so new techniques must be developed for hydrologic processes based exclusively on the nonstationarity of these processes. Many present-day approaches in hydrology for the analysis of time series and the generation of new samples are based on a nonstationary approach to the treatment of stochastic aspects of these series.

1.5 Definition of the Independent Stochastic Component of Discrete Hydrologic Time Series

If m_τ and s_τ are designated as monthly or daily means, and monthly or daily standard deviations, respectively, (or for any other time

interval in which a year is divided), with τ designating the discrete positions inside the year, then the standardization of a variable x_i gives

$$\epsilon_i = \frac{x_i - m_\tau}{s_\tau} \quad , \tag{1.1}$$

in which x_i are discrete values of a series and ϵ_i is the new reduced variable which may or may not be the first approximation of the stationary stochastic component, independent or dependent. Further analysis may show periodicities in other properties, such as in the autocorrelation coefficients of the ϵ_i series, in the higher-order moments and particularly in the skewness and the excess coefficients of the ϵ_i distribution at each position τ (say, in each of the 12 months, or in each of the 365 days). These periodicities can be also removed by appropriate mathematical models to single out the stationary stochastic component of a series.

Stationary stochastic components often come out to be approximately linearly dependent, with mathematical autoregressive dependence models such as

$$\epsilon_i = \rho_1 \epsilon_{i-1} + \sqrt{1 - \rho_1^2} \quad \xi_i \quad , \tag{1.2}$$

or

$$\epsilon_i = \alpha_1 \epsilon_{i-1} + \alpha_2 \epsilon_{i-2} + \\ + \sqrt{1 - \alpha_1^2 - \alpha_2^2 - 2\alpha_1 \alpha_2 \rho_1} \quad \xi_i \quad , \tag{1.3}$$

or by the higher-order linear autoregressive models. Equations 1.2 and 1.3 represent the two simplest autoregressive linear models, of the first-order and the second-order, respectively. The estimate of ρ_1 in Equation 1.2 is usually by r_1, the first serial correlation coefficient of the sample, and the estimate of α_1 and α_2 in Equation 1.3 is usually by a_1 and a_2 of the sample, which depend on r_1 and r_2 (the first and the second serial correlation coefficients of the sample), though r_1, or r_1 and r_2 are not the unbiased estimates of ρ_1, or ρ_1 and ρ_2. Then by using Equations 1.2 and 1.3 the independent stationary stochastic component can be determined either by

$$\xi_i = \frac{\epsilon_i - r_1 \epsilon_{i-1}}{\sqrt{1 - r_1^2}} \tag{1.4}$$

or by

$$\xi_i = \frac{\epsilon_i - a_1 \epsilon_{i-1} - a_2 \epsilon_{i-2}}{\sqrt{1 - a_1^2 - a_2^2 - 2a_1 a_2 r_1}} \quad , \tag{1.5}$$

for Equations 1.2 and 1.3 respectively, and similarly for the higher-order linear models. If parameters that are functions of the higher-order moments are shown to be periodic, these periodicities can be similarly mathematically described and removed.

The independent stochastic component, designated in this text as the ξ_i random variable and assumed to be identically distributed at all τ positions of the period ω, should be as nearly stationary and independent time process as the analysis of available data and statistical inference permit or justify. This definition of an independent stochastic component is used throughout the following text.

1.6 Time Series Measures

Experience shows that many hydrologic time processes follow the astronomic periodicities of the day and the year. When these variables are integrated over 24 hours as the average or total daily values, the cycle of the day is no longer present in the discrete series of daily values. Similarly, an integration of the continuous process over 365 days, or a summation of discrete values of a series from 1 to ω, where ω is the number of values in any year of this discrete process resulting in average or total annual values, the cycle of the year is no longer present in the discrete series of annual values.

For a continuous time series, x_t, a time interval, Δt, is selected to sum or average the process inside the consecutive and non-overlapping sequence of these intervals. This procedure creates a new discrete series with Δt defined here as the time series measure. Usually, in hydrology, Δt are multiples of either hour, day or month, with the intervals themselves of the hour, the day, the month, and the year included. For a majority of hydrologic time series, hours and fractions or multiples of hours are used to derive a new series when the short time series measures are relevant. When the opposite is the case, the day or month or their multiples are used. Many hydrologic services publish data as hourly, daily, monthly and annual series. Further analysis is mainly concerned with time series measures of the day and multiples of the day, and the month, though

derivations are valid for any Δt. For this study only those values of Δt are relevant that avoid the daily periodicity, ($\Delta t \geq 1$ day), but keep the annual periodicity, or 1 day $\leq \Delta t <$ one year, with Δt a fraction of the year, so that this annual periodicity remains in the new discrete time series. Daily and monthly series are often taken here as the two examples of the Δt selection. The values of Δt of 1-day, 2-day, 3-day, 7-day, 13-day, 14-day, 15-day, 1-month, 2-month, 3-month, 6-month, and similar interval lengths, by which a year can be divided into approximately ω equal intervals, fit the patterns to be investigated in this paper. If hourly values are studied, or values of intervals that are fractions of an hour or multiples of an hour but fractions of the day, then the two periodicities of the day and the year would show in the discrete series for many hydrologic variables. The methodology to be outlined in the subsequent text can be applied to any time measure of similar properties.

If the annual cycle is denoted by T (the year), then $\omega = T/\Delta t$ is the number of discrete intervals, or there are ω discrete values x_i of the random variable within any year. The sequence of these values are denoted by τ, with $\tau = 1, 2, \ldots, \omega$. For monthly values $\omega = 12$, for weekly values $\omega = 52$, and for the daily values $\omega = 365$.

1.7 Condensation of Hydrologic Information

Information from hydrologic observed data can be presented as tables, graphs, and mathematical models. Usually the best presentation is as condensed mathematical models. The information in a sample of 100 years of observations of monthly precipitation or monthly runoff, or daily discharge and intermittent precipitation events, or of any other hydrologic variable, may be condensed into a few mathematical models containing the necessary estimation of parameters. Basically, the following mathematical models are appropriate.

(a) Algebraic structural models. These describe the connection between the periodic-deterministic and the stochastic components. The simplest example is given by Equation 1.1. Sometimes complex rather than simple models are likely to fit these connections in various practical cases.

(b) Models for periodicities. These describe various periodic-deterministic components. Primarily, they are sets of trigonometric functions in the Fourier analysis of periodic components. The complete description and removal of periodic components should reduce the series to stationary stochastic components, provided they do not contain trends and jumps.

(c) Models of dependence of stationary stochastic components. These models are a further description of the time series structure. The independent stationary stochastic variable should be well defined and identified by these dependence models. These dependence mathematical models are deterministic functions relating the random variables.

(d) Models of univariate probability distribution functions. These are descriptive mathematical models of the distribution of independent stationary stochastic components, or of independent random components identically distributed over all τ positions inside the period ω.

As an example, in the simplest form, the assumption is made that Equation 1.1 describes the first relation between the deterministic components, m_τ and s_τ, and the stochastic component, ϵ_i. A further improvement may be made by using the periodic functions μ_τ and σ_τ as fitted to m_τ and s_τ values, of the periodic movements in the mean and standard deviation; if they are proportional, then $\eta_\tau = \sigma_\tau/\mu_\tau$ is a constant to be estimated by the constant coefficient of variation, C_v. If m_τ follows a simple periodic movement and can be described by m harmonics in the Fourier series analysis, then 2m+1 is the number of parameters to be estimated, two for each harmonic, and the general mean, μ_x estimated by the sample mean, m_x. If this is the case, Equation 1.1 becomes

$$y_i = (x_i - \mu_\tau)/\eta_x\mu_\tau = x_i/\mu_\tau\,\eta_x - 1/\eta_x. \qquad (1.6)$$

It has now one additional parameter, η_x, to be estimated by C_v, for the periodicity in s_τ. Assume further that ϵ_i is a stationary variable and follows the second-order linear autoregressive model of Equation 1.3, as the third equation. It adds another two parameters, α_1 and α_2, to be estimated by the sample statistics, a_1 and a_2. If ξ_i of Equation 1.5 is identically distributed over all τ (12 months, or 365 days) and if it follows the log-normal probability distribution with three parameters μ_n, σ_n, and γ (to be estimated by the mean of logarithms m_n,

the standard deviation of logarithms s_n, and the lower boundary g, respectively), then it gives the fourth equation with three additional parameters to estimate.

In conclusion, the above example gives four mathematical equations, with $\nu = 2m + 7$ parameters. For $m = 1$ (only a 12-month periodic function for m_τ), $\nu = 9$; for $m = 2$, $\nu = 11$, and for $m = 6$ (the maximum number of harmonics of monthly time series for the 12-month periodicity), $\nu = 19$. It is clear that there are seven basic parameters μ_x, η_x, $\alpha_1, \alpha_2, \mu_n, \sigma_n$, and γ and as many pairs of Fourier coefficients as there are harmonics in the second equation, all to be estimated from the sample data. The proper analysis of the structure of hydrologic time series should lessen the total number of parameters to be estimated from data.

If four mathematical equations are given and ν parameters estimated from data, with all necessary statistical inferences performed in developing these four equations, it can be rightfully claimed that all available information in a large amount of observational data has been extracted in the form of descriptive mathematical models. If additional data become available statistical inference should be performed again for these models, and their parameters re-estimated from an increased sample. In the future, instead of publishing books of tables and/or long series of graphs, four or more equations in general forms and a list of parameters may contain the extracted information.

One objection to the hypothesis of extracting information by mathematical models and their parameters, is that the condensed information does not refer to various random functions (new derived variables) of the basic process, like extremes, runs, ranges, and similar variables. The basic postulate of probability theory is that properties of any function of a random variable may be determined from the stochastic process of that variable, if this process is known and properly mathematically described. Thus, this objection can be overcome either analytically by developing characteristics of functions of the basic random process, which approach is difficult with complex hydrologic series, or through the experimental statistical or data generation method by generating new samples of the basic process, with the derivation from the generated new samples of a sample of any new random variable which is a function of the basic process. Tests may be designed to demonstrate that the properties of various variables derived directly from the historic sample are statistically indistinguishable from the properties of variables derived either analytically or by the experimental statistical method.

The current cases of hydrologic series may be more complex than the above example, particularly since μ_τ and σ_τ may not be proportional; there may be a cycle in α_1 and α_2 of Equation 1.3. The stochastic components of Equations 1.4 and 1.5 may not be identically distributed over τ, or the periodicity may exist in the skewness coefficient, β_τ, or in the excess coefficient and in other higher-moments parameters. However, it is still feasible to determine the independent stochastic component identically distributed over all τ positions by deriving additional mathematical models for other periodic components.

An advantage to identifying a stationary independent stochastic component is that the sample size for the estimation of its parameters may be long. If a series has the periodicity ω, with $\omega = T/\Delta t$, which is the number of discrete values in a year, and if n is the number of years, then $N = n\omega$ is the sample size for the estimation of parameters of an independent stationary stochastic component. If a hydrologic series is 50 years long, then the independent stationary stochastic component of daily flows, ξ_i, has $N = 50 \times 365 = 18{,}250$ values. The estimates of parameters of distribution of the ξ_i variable become sufficiently reliable, even taking into account the loss of degrees of freedom in estimating the parameters of periodic functions and of the dependence model. For monthly flows in this case $N = 50 \times 12 = 600$ values. Since ξ_i represents one of the main stochastic variations of a series, estimates of its properties have sufficient accuracy. However, the estimates of various parameters of periodic components and of dependence models, whose parameters are also subject to sampling variations, may be more in error than the estimates of parameters of the ξ_i variable. If it is true that estimates of parameters in basic deterministic periodic components involve only the second statistical moments (or no higher-order moment is periodic), while ξ_i requires the estimates of the third or even the fourth moment, the above approach of providing a large effective sample for ξ_i seems attractive, especially when the number of previously estimated parameters (as degrees of freedom lost) is small.

The significance of this investigation is the derivation of additional information about the structure of hydrologic time series in order to obtain a more realistic and accurate mathematical description of hydrologic stochastic processes.

1.8 Generation of New Samples

The results of this study should enhance the application of the experimental statistical method (the Monte Carlo or data generation method) to hydrology. Since the important stochastic part in hydrologic time series is the independent stationary stochastic component, the above discussed fourth equation in the form of a probability density function of this component is the basic random part of a series. The sampling fluctuation of parameters of deterministic components, and of the dependence model of the stochastic part, are further sources of randomness. If a hydrologic stochastic process can be decomposed in such a way that the stochastic component ξ_i results in an independent stationary or identically distributed variable over all values τ inside the year, it could substantially contribute to a better application of the data generation method to hydrology. If this component is lognormally distributed it is sufficient to generate as large a number of independent standard normal random numbers as is required by the problem to be solved, and then simply transform them to independent lognormal random numbers. By performing a sequence of transformations of deterministic mathematical models, the new samples of the x variable of Equation 1.1 may be produced.

In generating new samples of the variable of monthly flows, daily flows, monthly precipitation, or similar variables by application of the experimental method, various water resources problems may be solved. The first prerequisite for reproducing properties of time series in new samples is the proper generation of samples of independent stochastic components. Generating new samples of time series can be regarded as a reversible process of decomposition of a time series into their various components. Analysis in this study may contribute to a more realistic and more accurate method for generating hydrologic series for purposes of solving those problems that cannot be solved with sufficient accuracy either by classical empirical methods, extensively used in hydrology and water resources, or by analytical methods.

1.9 Two Techniques in the Primary Analysis of Time Series

Primary analysis of a hydrologic time series is defined here as steps that lead to the identification of its structure. The existence of deterministic and stochastic components should be ascertained and the hypotheses about their connecting mathematical model should result from this primary analysis. This existence may be ascertained by several methods but primarily by two techniques, the autocorrelation analysis, represented by a correlogram, and the spectral analysis, represented by the variance density spectrum.

The correlogram is a function between the serial correlation coefficients ρ_k, as ordinates, against the lag k as abscissae, with ρ_k given by

$$\rho_k = \frac{\operatorname{cov}(x_i, x_{i+k})}{(\operatorname{var} x_i \operatorname{var} x_{i+k})^{1/2}}, \qquad (1.7)$$

as the ratio of the population covariance to the population variance. The correlogram shows the general character of a time series. When the values ρ_k are estimated by the sample values r_k, the statistical inference is then performed to ascertain whether these serial correlation coefficients are significantly different from the correlogram of a mathematical model. An advantage of this procedure is that a direct relation can be established between the shape of the autocorrelation function and the type of the time series. The expected correlograms are known for basic time series. Visual inspection of the correlogram shows, in general, what type of series may be dealt with, and a hypothesis about its structure may be advanced. It can be determined by statistical tests whether a given hypothesis about the structure of a series should be accepted or rejected. If rejected, new hypotheses may be advanced and tested.

If the correlogram gives the general type of dependence, the variance density spectrum may be used to better discern some aspects of the structure of a time series, particularly the identification of periodic components or some dependence models. The spectral function is defined in relation to the autocorrelation coefficients ρ_k as

$$\gamma(f) = \sum_{k=1}^{m} D(k)\, \rho_k \cos 2\pi f k \quad , \qquad (1.8)$$

in which $\gamma(f)$ is the variance density, f is the ordinary frequency, m is the number of autocorrelation coefficients used in this transformation,

and D(k) is the smoothing function for the correlogram. If the series is N long, often m = N/10 to N/5, though this is an arbitrary cutoff point. Instead of using Equation 1.8 the fast Fourier transforms (or the basic and original concepts of Fourier transforms of a series) are used at present to estimate the spectral densities with much less computer time.

Spectral densities are Fourier transforms of the autocorrelation function. Referring to the correlogram, Equation 1.8 shows that both techniques-autocorrelation and spectral analysis--are basically identical, and that limitations in the correlogram accuracy affect the accuracy of variance spectrum densities. They are different pictures of the same properties of basic data, with the correlogram showing better some aspects of time dependence while the variance spectrum, smoothed by a filtering process, discerns other aspects better--particularly the periodic components and some dependence models of a time series.

Since considerable experience shows that the periodicity of the year exists in nearly all continuous hydrologic time processes, it is usually unnecessary to identify it every time a new series of the same type of variables is investigated. Thus, the Fourier analysis in the form of discrete or line-spectrum, known as the periodogram, should be used for estimating amplitudes of harmonics instead of the continuous variance spectrum. The Fourier discrete-spectrum harmonic analysis of the periodicity of the year is the technique used in this study for mathematically describing periodic components in hydrologic time series of known basic periods.

1.10 Stationarity of Stochastic Components

A stochastic process is strictly stationary if the distribution of the set $x_1, x_2,...,x_n$ is the same as the set $x_{1+k}, x_{2+k},...,x_{n+k}$ for every n and k. In practice, such stationarity is difficult to prove for a hydrologic time series. The problem is not to obtain this stationarity by removing trends, jumps and periodic components, but to approch it closely through valid statistical tests. These statistical tests either relate to the first two statistical moments in order to detect periodicities, and by removing them to obtain an approximation to the second-order stationarity of the remaining part, or relate also to higher-order moments to approximate the higher-order stationarity.

The process is second-order stationary if its expected value and covariance are independent of the position in a time series, or

$$E[x_i] = \text{constant} \qquad (1.9)$$

for the first-order stationarity, or the stationarity in the mean, and

$$E[(x_i - Ex_i)(x_{i+k} - (Ex_{i+k})] = 0 \qquad (1.10)$$

for a given k, for the second-order stationarity, or the stationarity in the covariance. The investigation of this paper tends in all its aspects to reduce, by structural analysis, the stochastic component to a final result of a second-order or higher-order stationary and independent random variable, identically distributed over the time series positions.

1.11 Organization of Material in this Paper

Theoretical aspects of structural analysis concerning periodicity in the time series parameters, dependence models, distribution functions and parameters of stochastic components, and the effects of various types of bias in the original series on its structure, are presented in subsequent chapters, after the hypotheses underlying this investigation are presented in Chapter 2. Examples are given in support of derivations and discussions. Conclusions are presented in the last chapter.

The three topics are crucial in the structural analysis of hydrologic time series: the inference about the significance of harmonics in the periodic parameters, the inference about the best dependence models for the stochastic components of a given order of stationarity, and the inference for the probability distribution of the best fit to the frequency distribution of the independent stationary stochastic component. These three topics are discussed in details. Besides, a special attention is given to the problem of biases of the sampling type in order not to be perpetuated by the structural analysis and mathematical description of a time series.

Chapter 2

HYPOTHESES UNDERLYING THE STRUCTURAL ANALYSIS

No structural analysis, nor the mathematical models that have been advanced and tested, can be better than the hypotheses underlying them. Therefore, this chapter presents hypotheses used in this study.

2.1 Selection of Working Hypotheses

A distinction should be made between the following four concepts: basic data, information contained in the data, hypotheses that underly the extraction of information from data, and methods used to extract this information. The development of mathematical models for the description of hydrologic time series and the estimation of model parameters represent an advanced form of extraction, condensation and description of information contained in the data. To accomplish this, methods are necessary. However, no method can be developed without postulated hypotheses. These hypotheses are most often developed from experience with a large number of hydrologic time series, from the physical properties of the underlying processes, and from the general understanding of phenomena. Methods can be no better than the hypotheses upon which they are based. The description and justification of these hypotheses is a first step for better understanding the investigative line followed in this study. The following hypotheses are first briefly outlined and then discussed in detail.

(1) A hydrologic continuous time series is composed of deterministic components, in the form of periodic parameters, and of a stochastic component. The basic hypothesis is that a series can be separated into these components without an adverse effect on the final understanding and description of the time series structure and extraction of information. The periodic part of the series is encompassed by a general term of cyclicity or periodicity (the signal in communication engineering language), while the random part is called the stochasticity or the rendomness (the noise in communication engineering language).

(2) From the total variation of a variable nearly all of random variation (the stochasticity in the series) is allocated to the stochastic component, while only the unavoidable sampling errors--within the limits of the best estimation techniques used--is left inside the estimated parameters of periodic parameters (the cyclicity in the series).

(3) By removing the inferred periodic components in various parameters from the series, or by removing the cyclicity, the hypothesis is that the stochastic component of the series is approximately a stationary random variable of a given order of stationarity, provided improvements in the order of stationarity and data processing costs justify attaining that given order of stationarity. In other words, the stochasticity of a series is reduced to a stationary stochastic process of a given, required or justified order of stationarity.

(4) A hydrologic series may contain biases produced by man-made processes or other sudden or slow casual changes in nature. These changes are defined as nonhomogeneity. Also, inconsistency (systematic errors) is often present in data. The hypothesis is that nonhomogeneity and/or inconsistency in series are detected, described, and removed prior to the structural analysis of time series as conceived in this study. Misunderstandings are often produced when some models are tested on nonhomogeneous and inconsistent series.

(5) A hydrologic time series may have various sampling biases in the form of a long-range trend (pseudo-trend), a long-range periodicity (pseudo-cyclicity), unrepresentatively short or long drought or wet periods for the sample size available, exceptionally high or low flood events, which are not representative of the sample size, and similar sampling biases. Regional investigations and some experimental statistical analyses can show that these particular occurrences in time series have small probabilities to be or not to be exceeded in the future for the same sample sizes. The hypothesis is that the structural analysis of these samples of time series, their mathematical description, and the generation of new samples by the experimental statistical method do not perpetuate these sampling biases.

(6) The inferred structural mathematical models, the estimated coefficients of periodic components, and the estimates of population parameters

of the stochastic component of a time series are subject to sampling errors. The regional information from a set of time series can improve the information about these models, coefficients, and/or parameters. The hypothesis is that the structural analysis of a set of series in a region can improve significantly the models and estimated coefficients and/or parameters, if the proper regional information replaces the information at a point of its observed individual time series.

(7) If a structural analysis and simulation method are developed for small time series units (say one-day values or even smaller) the methodology should be applicable to any other, and particularly to larger units of time series, which still preserves the basic properties of cyclicity and stochasticity of that series. This hypothesis infers that the selection of the time unit (1-day, 2-day, 3-day, 7-day, 13-day, 14-day, 15-day, 1-month, etc.) does not affect the applicability of the developed methods.

(8) The structural analysis is pursued to such an extent that all pertinent information about a hydrologic periodic-stochastic process may be extracted in the form of a set of mathematical models and estimated statistics (coefficients, parameters, descriptors) that describe these models. The generation of new samples of time series from these mathematical models should reproduce all basic inferred population properties, and do it so well that the original sample cannot yield any more substantial information about the process than the models and, consequently, the generated new samples. In other words, the reliability of generated samples reproducing well the properties of the process depends on the correct methodology and the extent to which the structural analysis, the corresponding mathematical description, and the estimation of parameters from the data of original time series are applied to attain a given order of stationarity of the stochastic component.

(9) Mathematical models are necessary in the analysis and description of time series. For these models the coefficients and/or parameters must be estimated from the data. The hypothesis of this structural analysis is that minimum of coefficients and/or parameters should be estimated, because the more statistics estimated the lesser their overall reliability, and the smaller is the remaining degrees of freedom for other estimates. An optimization is made, by statistical inference, between the number and the reliability of estimates of these parameters.

(10) The dependent stationary stochastic component is fitted by a mathematical model of dependence and from it a stationary independent stochastic component may be determined. The hypothesis is that this independent series contains the major random variation in a series, as the independent stochasticity or noise. Because of a very large independent sample thus produced, with relatively small loss of degrees of freedom due to already estimated other parameters, the parameters of distribution of this independent random variable can be estimated accurately. The hypothesis is that the mathematical models of dependence have a sound physical background and/or justification, and are usually applicable to a large number of series of the same variable, under the conditions of acceptable methods of testing this type of statistical hypotheses.

2.2 Periodic-Stochastic Structure of Hydrologic Time Series

Assume a hydrologic continuous periodic-stochastic time process, $\{\zeta_t\}$, to be composed of periodic functions in some of its parameters and a stochastic component. A realization of this ζ_t process is given in the form of a finite discrete series, $x_{p,\tau}$, with given values at intervals Δt apart. The symbol $\tau = 1, 2, \ldots, \omega$ are discrete values in the basic cycle, ω. This period ω is either a day or a year (or both) in the majority of hydrologic time series. The symbol p, with $p = 1, 2, \ldots, n$, represents the successive values of the period ω, with n their total number. For ω being a year, n is the number of years in the sample time seris. The total sample size is then $N = n\omega$. Estimating the parameters that are significant in the periodic functions inside the $x_{p,\tau}$ series is the first problem to study.

The population mean at a position τ is designated by $\mu_\tau(\zeta_t)$ and the corresponding population standard deviation by $\sigma_\tau(\zeta_t)$ for the ζ_t process. As the consequence of the above basic assumption, $\mu_\tau(\zeta_t)$ and $\sigma_\tau(\zeta_t)$ are two periodic functions of τ. Similarly, the population autocorrelation coefficients $\rho_{k,\tau}(\zeta_t)$ may be periodic as functions of τ. By removing these periodic functions from the ζ_t series, the remaining part of the series should be the second-order stationary stochastic component. However, the third-order parameters at positions τ may also be periodic. By removing periodicities in these parameters, the remaining part of the ζ_t process should be the third-order

stationary stochastic component. This approach may be continued to the fourth- and higher-order parameters, and the stationarity of the fourth- or higher-orders of the stochastic component may be obtained by removing periodicities in all these parameters. This leads to the basic application of the first hypothesis, namely that any hydrologic periodic-stochastic ζ_t process can be decomposed into the periodic parameters of a given order, and of all other smaller-order parameters than this order, and a stationary stochastic component of this given order of stationarity. A realization of the ζ_t process as a finite $x_{p,\tau}$ series is used for the estimation of periodic functions in the parameters of a given order and of the properties of stochastic component of the same order of stationarity.

The basic principle in applying this hypothesis is that the order of the moments used in defining parameters that may be periodic functions of τ, and the corresponding order of stationarity of the stochastic component, should be selected by some criteria that determine how well the mathematical description of a time stochastic process ought to approximate the real structure of this process.

The estimation of a given $\mu_\tau(\zeta_t)$ at the position τ of the period ω from an $x_{p,\tau}$ series, with p = 1, 2, . . ., n, is

$$m_\tau = \frac{1}{n} \sum_{p=1}^{n} x_{p,\tau} \quad . \qquad (2.1)$$

The sampling difference between an estimate m_τ from a given sample of size $n\omega$ and the corresponding population value $\mu_\tau(\zeta_t)$ is then

$$e_\tau(m_\tau) = m_\tau - \mu_\tau(\zeta_t) \quad . \qquad (2.2)$$

Because n is usually small for most hydrologic time series, if n represents the number of years, the sampling errors $e_\tau(m_\tau)$ of Equation 2.2 are often large. Besides, if ω is large, say 365 for daily time series, all 365 values of m_τ cannot be estimated accurately. The question may be posed whether the appropriate fit of a periodic function μ_τ to ω values of m_τ, with μ_τ as the joint estimates of $\mu_\tau(\zeta_t)$ or as the estimate of the periodic function $\mu_\tau(\zeta_t)$, and with the new sampling errors

$$e_\tau(\mu_\tau) = \mu_\tau - \mu_\tau(\zeta_t) \quad , \qquad (2.3)$$

reduces the overall variance of sampling errors in the means? If this is the case the variance of errors of Eq. 2.3 should be

$$\text{var } [e_\tau(\mu_\tau)] = \alpha \text{ var } e_\tau(m_\tau) \quad , \qquad (2.4)$$

with α much smaller than unity. If m_τ is used as the estimate of $\mu_\tau(\zeta_t)$, this is equivalent to stating that a larger part of sampling variation of the stochastic component in the $x_{p,\tau}$ series is retained in the estimate values m_τ of the ζ_t process than in the case where the estimates μ_τ of the periodic function are used.

As a consequence of the first hypothesis, based on experience and physical analyses of the responses of various hydrologic environments, the periodicity in the mean should be a smooth function. It is sufficient to estimate m_τ of a series for the samples of different sizes, and to find that the smoothness of the m_τ series increases with an increase of the sample size n.

Figure 2.1 demonstrates the basic current experience of hydrology, namely that the variation of m_τ along τ becomes smoother with an increase of the number of years n. The means m_τ of daily flow series of the Tioga River are given as the example, in four graphs and for: (1) the 10-year period (1921-1930); (2) the 20-year period (1921-1940); (3) the 30-year period (1921-1950), and (4) the 40-year period (1921-1960). This smoothness should correspond to the smooth astronomic periodic functions of heat supply over regions of the earth, which are only modified by the responses and interactions of hydrologic environments for the resulting processes of various hydrologic variables. This is the major reason why a smooth μ_τ periodic function should be estimated instead of using the sequence of ω values of m_τ. The second reason for using μ_τ instead of m_τ comes from the second hypothesis, namely the requirement of attaching as much as possible of the sampling fluctuations in the $x_{p,\tau}$ series to its stationary stochastic component rather than to the coefficients of periodic functions of various parameters. The third reason for using μ_τ instead of m_τ is the consequence of the ninth hypothesis, namely developing mathematical models of periodic-stochastic processes with a minimum of estimated coefficients and/or parameters. For these three reasons, based on three hypotheses, it is considered that the proper statistical fit of the μ_τ periodic function to ω values of m_τ is superior to using ω estimates of m_τ.

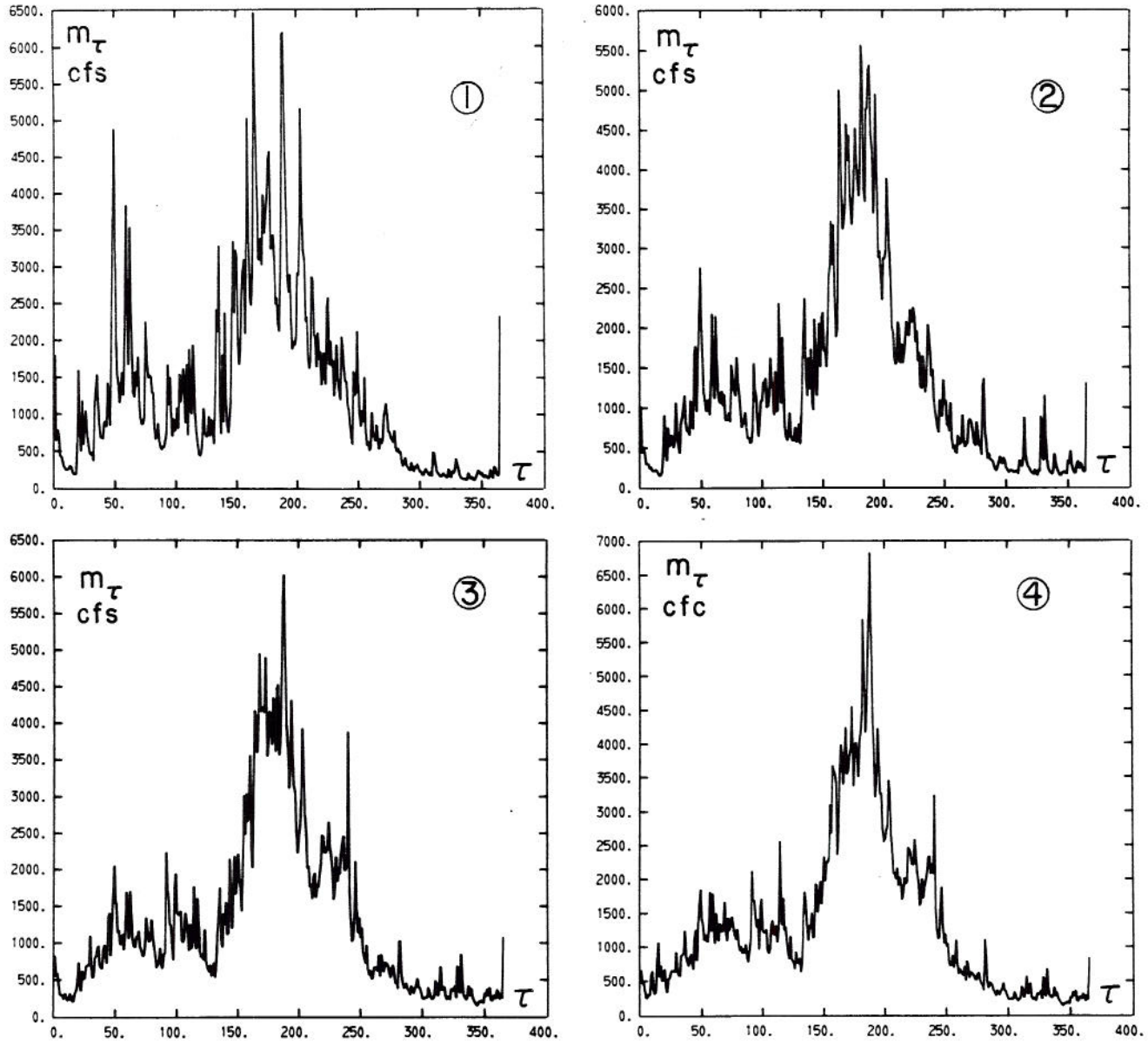

Fig. 2.1 The increase of smoothness of the sequence of mean daily flows of the Tioga River near Erwins, New York, with an increase of the sample size: (1) 10-year period (1921-1930); (2) 20-year period (1921- 1940); (3) 30-year period (1921-1950); and (4) 40-year period (1921-1960).

The estimate of the population standard deviations, $\sigma_\tau(\zeta_t)$, at any given position τ of the period ω from the $x_{p,\tau}$ series is by

$$s_\tau = \left[\frac{1}{n}\sum_{p=1}^{n}(x_{p,\tau} - m_\tau)^2\right]^{1/2} , \quad (2.5)$$

if $n \geqslant 30$, or by an unbiased estimate $\hat{s}_\tau^2 = n\, s_\tau^2/(n-1)$ if $n < 30$.

Another approach for estimating s_τ is the use of μ_τ instead of m_τ in Equation 2.5 by

$$s_\tau^* = \left[\frac{1}{n}\sum_{p=1}^{n}(x_{p,\tau} - \mu_\tau)^2\right]^{1/2} , \quad (2.6)$$

if the μ_τ periodic function has been already estimated. However, this approach will produce larger values of s_τ than Equation 2.5 gives because s_τ is a minimum when the first moment m_τ of the sample is used. The estimates by Equation 2.5 are used in this study for the fitting of the periodic function σ_τ.

Just as for m_τ, a smooth function σ_τ may be fitted to the estimated ω values of s_τ, so that the variance of difference $e_\tau(\sigma_\tau) = \sigma_\tau - \sigma_\tau(\zeta_t)$ becomes much smaller, on the average, than the variance of difference $e_\tau(s_\tau) = s_\tau - \sigma_\tau(\zeta_t)$. The differences $e_\tau(s_\tau) - e_\tau(\sigma_\tau) = s_\tau - \sigma_\tau$ are then retained in the stochastic component of the $x_{p,\tau}$ series instead of being left in the periodic function s_τ of the standard deviation, just as the sampling differences $e_\tau(m_\tau) - e_\tau(\mu_\tau) = m_\tau - \mu_\tau$ are retained in the stochastic component.

This procedure is followed for any other periodic parameter, generally designated by $\nu_\tau(\zeta_t)$, of a given periodic-stochastic process $\{\zeta_t\}$, and its available sample series, $x_{p,\tau}$, by fitting the periodic functions ν_τ to the estimated values v_τ at discrete positions τ of the period ω.

2.3 Two Fundamental and Opposite approaches in the Analysis of Periodic Components of Hydrologic Time Series

The general equation of the periodic deterministic function of any parameter ν_τ in using the Fourier series approach, is

$$\nu_\tau = \mu_\nu + \sum_{j=1}^{m} C_j \cos\left(\frac{2\pi j\tau}{\omega} + \theta_j\right) \quad , \quad (2.7)$$

in which ν_τ is the symbol for any periodic parameter related to the ζ_t process and estimated from the $x_{p,\tau}$ series, μ_ν is the mean of ν_τ or v_τ over the ω positions of τ, j is the sequential number of any harmonic out of the $\omega/2$ possible harmonics, m is the number of significant harmonics (or of the harmonics that have the amplitudes statistically significantly greater than for the nonperiodic series), C_j is the amplitude and θ_j the phase of the j-th harmonic.

The first basic hypothesis of this structural analysis is that several parameters, ν_τ, of the ζ_t process are deterministic-periodic functions of the type of Eq. 2.7, with μ_ν, C_j's, θ_j's, and m being constant coefficients for any periodic parameter of the ζ_t process, with these coefficients estimated by sample statistics of the $x_{p,\tau}$ series for a given ω. Once μ_ν, C_j's, θ_j's, and m are estimated for a periodic parameter, values of ν_τ at any position τ are uniquely defined. This property for all periodic parameters is understood in this study to represent the deterministic-periodic component of the process.

An opposite hypothesis of composition of a series is with

$$\nu_\tau = \mu_\nu + \sum_{i=1}^{m} C^i_{p,\tau} \cos\left(\frac{2\pi\tau i}{\omega} + \theta^i_{p,\tau}\right) \quad , \quad (2.8)$$

in which μ_ν is the mean of the parameter ν_τ, while $C^i_{p,\tau}$ and $\theta^i_{p,\tau}$ are random amplitudes and random phases, respectively, and i = 1, 2, . . ., m is the number of harmonics. In other words, any parameter ν_τ has only the mean μ_ν as a constant, and as many pairs of random variables, $C^i_{p,\tau}$ and $\theta^i_{p,\tau}$, as there are harmonics. For only one harmonic, ω, there is a particular amplitude value, $C^1_{p,\tau}$, and a particular phase value, $\theta^1_{p,\tau}$, for each time position, say each day of the sequence of years if the yearly cyclicity is present, or each hour of the sequence of days, if the daily cyclicity is studied. The random variables $C^i_{p,\tau}$ and $\theta^i_{p,\tau}$ should be serially correlated and mutually dependent. The simplest case of application of Equation 2.8 is when only the means m_τ along the τ positions are assumed to follow "random periodicity."

Equation 2.7 requires the estimation of 2m + 1 coefficients for each periodic parameter, while Equation 2.8 requires the estimation of $1 + k_c + k_\theta + \lambda_c + \lambda_\theta + \lambda_{c,\theta}$ coefficients or parameters, with k_c and k_θ the number of parameters for the probability distributions of $C^1_{p,\tau}$ and $\theta^1_{p,\tau}$, respectively, λ_c and λ_θ the corresponding number of statistics for the autocorrelation models of $C^1_{p,\tau}$ and $\theta^1_{p,\tau}$, respectively, and $\lambda_{c,\theta}$ the number of statistics for measuring the mutual dependence between $C^1_{p,\tau}$ and $\theta^1_{p,\tau}$, and/or between $C^1_{p,\tau}$ and $\theta^1_{p,\tau}$ and other parameters. The total of $n\omega$ values of the $x_{p,\tau}$ series may not be sufficient to reliably estimate all the above parameters. The minimum expected number of parameters is when $k_c = k_\theta = 2$ (say normal distributions for $C^1_{p,\tau}$ and $\theta^1_{p,\tau}$), $\lambda_c = \lambda_\theta = 1$ (say the first-order autoregressive linear model), and $\lambda_{c,\theta} = 1$ (the linear correlation coefficient between, $C^1_{p,\tau}$ and $\theta^1_{p,\tau}$), or a minimum of eight parameters. For m = 3, this is about equivalent to 2m + 1 = 7 coefficients in harmonics of ν_τ of Equation 2.7.

These two opposite approaches should be compared from several standpoints, namely from the physical justification, the estimation theory, the future use of these models, the generation of new samples by the experimental statistical method, and similar. The difference is in the basic concepts of how

to divide the total information of a series. The first approach divides the series contained in the deterministic-periodic functions of parameters and all the rest of the information contined in the stochasticity of the process. The second approach contains nearly all the information in various stochastic parts, such as the random variables of periodic parameters and the random variable of stochasticity.

Environments within the earth in which the main hydrologic processes occur can be considered as closed systems. The outputs of one environment represents input into another. The only exception is the open system of solar energy supply and irradiation from the earth into space. The solar energy input is a deterministic-periodic process for any unit area at the limits of atmosphere. However, the energy output of irradiation at the same unit area at the limits of atmosphere is a periodic-stochastic process. The various responses of the earth's environments produce the earth's energy output into space which has a high stochastic component and a modified periodic process. Turbulence, cloudiness, transparency, and other phenomena of the atmosphere add considerable randomness to the deterministic-periodic solar energy influx, so that the solar energy supplies to most of the atmosphere, to oceans, seas, and to continental areas are periodic-stochastic processes. The effects of winds, ocean currents, randomness in the mass and energy transfers between oceans and the atmosphere, between the atmosphere and continental areas, and other phenomena of continental areas further produce randomness and modify periodic components. Any environment (the atmosphere, oceans and seas, continental surfaces, underground spaces, etc.) of importance to hydrologic processes has responses that add or smooth randomness of input processes, by modifying the periodic process, by attenuating or amplifying the amplitudes of harmonics, and by shifting the phases of these harmonics.

The hypothesis that all of the earth's hydrologic processes are composed of deterministic-periodic and stochastic components seems supported by the basic periodic influx of solar energy. The attractive explanation of environmental responses to various inputs of hydrologic relevance is that they modify the properties of solar periodicity, but not the period itself, while adding substantial randomness. These assumptions about the environmental responses give support to the first hypothesis, namely that the hydrologic time series are composed of a deterministic-periodic process and a stochastic process.

The second approach of considering amplitudes and phases as random variables in the periodic function of any parameter would require complex responses of earth environments to the deterministic-periodic inputs of solar energy to various areas of the earth's surface. It is easier to conceive of the responses of hydrologic environments to given inputs as being composed of deterministic (linear and non-linear) parts and superimposed random parts, than to conceive of all responses as being stochastic. For example, the transparency (opacity) of the atmosphere for solar energy, because of cloudiness and other factors, should be periodic-stochastic in character, one season having larger cloudiness on the average than the other, with the average transparency over the seasons representing the component of deterministic-periodic process, and the chance variations about the averages representing the component of stochastic process.

The greater the average of a random input into a hydrologic environment is, the larger the expected variations in its responses. Since most inputs are positively valued variables (only positive values or zeros occur), the boundary of zero (or any other boundary) greatly limits the possible variations on the lower side of small average inputs, while for large average inputs the range of variations on both sides is substantial, infinite on the higher side and relatively large on the lower side. This range of values requires differences in the standard deviations along the positions of the year, with greater values for large average inputs and smaller values for low average inputs. However, this does not imply a proportionality between the standard deviations and the means, though that case may be true for several hydrologic periodic-stochastic processes. In summary, the physical considerations, and particularly the clear deterministic-periodic character of the solar energy input, give more support--at least at the present level of experience--for the first approach than for the second approach with regard to treating the cyclicity of hydrologic periodic-stochastic processes.

The future use of mathematical models of strucutral analysis of time series, say as the form of condensation of information, is much more attractive with all models either deterministic (cyclicity in

parameters), or functions of random variables (autoregressive models), or only as the independent stationary stochastic component, than the models with many random variables. In the generation of large samples by the experimental method the first approach requires the fitting, the testing of the goodness of fit of the probability distribution functions for only one variable, and the generating of only one random variable, while the second approach requires the same work for several random variables with some being mutually dependent.

In summary, there is justification for using the first approach to consider each periodic-stochastic hydrologic process as composed of deterministic-periodic functions in various basic parameters of the process, and an independent or dependent stationary stochastic component of a given order of stationarity. The only uncertainties in the deterministic-periodic components result from sampling errors in the Fourier coefficients or their equivalent coefficients of the amplitude and the phase, and the sampling errors in the estimated parameters of autoregressive models of dependent stochastic components.

2.4 Reduction of Stochasticity of Hydrologic Series to an Independent Stationary Component

The procedure for separating deterministic–periodic components in the various parameters and the stochastic component can be pursued to different levels of stationarity. If only the cyclicity in the mean over the positions $\tau = 1, 2, \ldots, \omega$ is inferred and removed, the remaining process

$$x^{*}_{p,\tau} = x_{p,\tau} - \mu_{\tau} \tag{2.9}$$

could be considered the first-order stationary process, with the cyclicity only in the mean to be removed. If the cyclicity in the standard deviation is inferred and removed only, the remaining process,

$$x^{*}_{p,\tau} = \frac{x_{p,\tau}}{\sigma_{\tau}} \quad , \tag{2.10}$$

could be considered as stationary in the variance, with σ_{τ} the cyclicity in the standard deviation. If both μ_{τ} and σ_{τ} are removed the remaining process,

$$\epsilon_{p,\tau} = \frac{x_{p,\tau} - \mu_{\tau}}{\sigma_{\tau}} \tag{2.11}$$

is both stationary in the mean and the variance, but still is not the second-order stationary process.

The general dependence of the $\epsilon_{p,\tau}$ process is

$$\epsilon_{p,\tau} = f(\epsilon_{p,\tau-1}, \epsilon_{p,\tau-2}, \ldots ; \alpha_{1,\tau}, \alpha_{2,\tau}, \ldots) + \sigma \xi_{p,\tau} \quad , \tag{2.12}$$

in which $\epsilon_{p,\tau}$ is a function of some or all previous $\epsilon_{p,\tau}$ values; $\alpha_{1,\tau}$, $\alpha_{2,\tau}$, . . . are the various periodic and/or nonperiodic parameters in this dependence model, σ is the standard deviation as a function of $\alpha_{j,\tau}$ - coefficients in order that $\xi_{p,\tau}$ is a second-order standardized stationary (and independent) stochastic component.

If the skewness coefficient and any other third-order parameter, such as the product of any three values of the second-order stationary process, show cyclicities, the periodicity can be tested and removed to obtain the third-order stationary stochastic component. Similarly, the fourth-moment properties may be tested for cyclicity along the $\tau = 1,2,\ldots,$ ω values, and when proven significant they can be removed to obtain the fourth-order stationary stochastic component.

The basic hypothesis is that successive investigations can raise the order of stationarity of an independent stochastic component by proper statistical analysis, inference, and mathematical description of the periodic parameters and the dependent stochastic component. This hypothesis, then, is a problem of optimization, optimization between the accuracy of structural analysis and the required economy in data processing. The greater the requirement for a reliable description and a full reproducibility of properties of a hydrologic process in the generation of new samples, the more justified become efforts for attaining the third- and/or higher-order stationarity of the independent stochastic component.

It is expected that parameters that are the functions of third-or higher-order moments will have, in general, a sufficient "signal-to-noise" ratio, in order to detect the periodic components. In other words, the ratio of explained variance of fluctuations of a parameter along the τ positions by the deterministic-periodic component and the variance explained by the sampling noise should not be small. This ratio usually decreases with an increase of the highest moment necessary to define a parameter. This is the same as saying the power of detecting the periodicity in parameters by statistical inference decreases with an increase of the order of moments which define these parameters.

2.5 Effects of Nonhomogeneity and Inconsistency

Because water resources are subject to man-made changes both in the natural processes and in environmental responses, nonhomogeneity in data is extremely common in hydrologic time series. Hydrologic data also have often systematic errors, thereby adding this inconsistency to the non-homogeneity which is either man-made or produced by some significant natural disruptive factors.

The detection, description, and removal of non-homogeneity and inconsistency should have both statistical significance and physical or historical support and justification. Nonhomogeneity and inconsistency may be in any or in all of the basic parameters of a hydrologic time series. Some investigations show [6,7] that all parameters are usually affected whenever a trend and/or a positive or negative jump are produced in a hydrologic series by nonhomogeneity and inconsistency. The discussion of methods of detecting, describing, and removing non-homogeneity and inconsistency, before a series is analyzed for its structural composition, is outside the scope of this paper.

2.6 Sampling Biases in Historical Time Series

As discussed in the introduction, the "trends" and "periodicities" in the parameters of time series may be produced only by sampling variations because of small historical samples. The smaller a sample the greater is the probability for it to exhibit some biased property, such as an upward or downward trend, sampling or pseudo-cyclicity, unrepresentative high or low extreme values for the size of the sample, and similar. Each sampling statistic has a distribution, and the sample estimates may be at the tails instead of being around the mean or median of this distribution.

Studies of annual precipitation, annual effective precipitation, and annual runoff series show them to be stationary time processes [1,2]. Therefore, trends and pseudo-cyclicity in small samples of annual time series can be mainly the result of sampling variations, provided nonhomogeneity and inconsistency in the series are removed.

The hypothesis is advanced here that a good structural analysis of time series should be such as to not perpetuate sampling trends and pseudo-cyclicities, either by inferred mathematical models or by estimated parameters, and, consequently, perpetuated in the new generated samples by Monte Carlo method. This hypothesis can be fulfilled by avoiding those procedures of analyses of time series that perpetuate both pseudo-trends and pseudo-cyclicities. The sampling properties of short time series may explain some past unproductive research efforts in searching for significant hidden periodicities and trend-type persistencies in basically the stationary stochastic processes by using improper analyses of short and/or nonhomogeneous time series.

Biases in extremes also may affect the time series analysis. If a prolonged drought or a wet period or both occur in a small sample, they will affect all parameters describing the time series properties. Similarly, if the largest drought or the largest wet period in a sample are relatively short, or their total deviations from the mean are small, these unrepresentative extremes also affect all parameters which describe time series properties. The concepts of representative extremes for given sample sizes, and for such properties as peak discharge, lowest discharge, drought, wet period, etc., must be introduced. Therefore, the hypothesis is advanced here that only those methods of structural analysis should be used that detect the unrepresentative extremes in the sample available, and eventually enable a correction for some parameters, particularly by using the available regional information on a hydrologic random variable.

2.7 Regional Information on Models, Coefficients, and Parameters

If hydrologic variables are observed as time series at a number of points in a region the regional information in the form of mathematical models, jointly estimated coefficients of periodic parameters, and jointly estimated parameters of stochastic components, is usually much more reliable than for estimates made separately for any individual time series. The hypothesis underlying investigations in this paper is that mathematical models, coefficients and parameters estimated for an individual series may be improved by regional analysis of all available time series. The methods of jointly estimating parameters and coefficients on a regional basis, to improve the corresponding values for the series at a given station, are not discussed in this paper.

The basic hypothesis in this regional joint estimation of coefficients and parameters of mathematical models, and in testing the goodness of

fit of these models to data, is that the parameters or coefficients of these models change smoothly over the region from one point to the next. For a given hydrologic random variable with observed time series at a number of points in the (x,y) - plane, the basic parameters of these series change with x,y - coordinates by a relatively smooth trend surface; for short time series available the parameter values must exhibit the sampling variations about these trend surfaces. By inferring for each basic parameter what is the most reliable regional equation for this surface, and by estimating its coefficients, the joint estimates of basic parameters are obtained; these estimates should have, on the average, much smaller sampling errors than the individual estimates obtained independently.

By removing periodicities from all periodic parameters of a set of regional time series of a hydrologic random variable, and by using the parameters or coefficients from the regional surface models, the stochastic components of observed series at this set of points may be reduced to stationary time series, and by removing the regional trend surfaces in basic parameters, the ensemble of series of this set may be considered to be drawn from a stationary and ergodic stochastic process.

2.8 Concluding Remarks on Hypotheses

Although hypotheses (7) through (10) of section 2.1 are self-explanatory, some additional remarks are warranted. The present approach in generating the new samples of monthly time series by nonparametric methods (say, by using 12 monthly means, 12 monthly standard deviations, and 12 times the monthly autoregressive coefficients) is not feasible for generating new samples of daily time series, with the number 12 replaced by 365, or hourly time series with 12 replaced by 8,780. Therefore, a parametric method is needed that does not depend on the time unit used. The parametric method of analysis permits an optimization between the number of statistics to be used and the accuracy in their estimation.

The dependence models for the stochastic components should be either based on a physical background or justified by experience on a large number of analyzed dependent stochastic components of a given hydrologic variable and its periodic-stochastic components of a given hydrologic variable and its periodic-stochastic process, or both. The time dependence models are conveniently divided into linear and nonlinear. There is an infinite possible number of linear and nonlinear dependence models. Without a physical background or the experience of successful fitting particular models to a large number of stochastic components, or both, the unnecessary proliferation of models is unavoidable.

Chapter 3

PERIODICITY IN PARAMETERS OF HYDROLOGIC TIME SERIES

3.1 Periodicity

It is assumed that any eventual inconsistency and nonhomogeneity in time series, particularly man-made nonhomogeneity, are identified by the proper techniques, supported by physical or historical investigations, and removed from a hydrologic series prior to its structural analysis. The process of separating the deterministic-periodic components from the independent or dependent stationary stochastic component may then be undertaken.

The analysis in this chapter is mainly concerned with periodic parameters of hydrologic time series. Continuous variables and discrete time series are only considered. Statistical inference in detecting significant harmonics in these periodic components is a major part of this chapter.

Periodicity of a hydrologic time series may be present in one, two, or several of its parameters, such as the mean, the standard deviation, the auto-covariances or the coefficients of the autocorrelation function, the higher-order moments or the parameters which are functions of these moments, and similar. An independent or dependent stationary stochastic component is assumed always present in any hydrologic time series, while the periodic parameters may or may not be present. The stationarity of the stochastic component is assumed either of the second order, which is weak stationarity, or of the higher order which is strong stationarity. This means that the expected values of the corresponding moments or parameters are independent of the absolute position i for discrete time series, but depend only on the position differences k when these differences are relevant for definitions of moments or parameters.

3.2 Determination of Periodic Parameters

Experience shows that each month, day, or hour, or any multiple of these units of the year has a different expected value and different standard deviation in a hydrologic time series.

A value of variable x in the year p and at the position τ inside the year is $x_{p,\tau}$, with p = 1,2,..., n, and τ = 1,2,...,ω. This $x_{p,\tau}$ value is for month τ, day τ, or hour τ, of year p following the beginning of records, with n the number of year of record, and ω the total number of discrete values in a year. The individual monthly, daily, or hourly mean values m_τ, or m_τ of any multiple of these time intervals, for a given τ are estimated by Equation 2.1, while the individual monthly, daily, or hourly standard deviations s_τ are estimated by Equation 2.5.

Similarly, ω intervals of the year are used to estimate the other parameters as they vary throughout the year. The series of ω values of any parameter may be periodic, stochastic, a constant, and a combination of periodic and stochastic parts, with sampling random variations superposed. Because of these sampling variations, particularly in small samples, the random fluctuations in a parameter are always superposed to a periodic movement or to a constant. The estimation of Fourier coefficients or of amplitudes and phases of the population periodic parameters is always affected by sampling fluctuations. In other words, because a periodic parameter is superposed by a stochastic fluctuation for a finite sample, the sampling variation associated with the stochastic part of a series does not permit the computation of population coefficients of equations of deterministic-periodic parameters. These coefficients are then statistics subject to sampling distributions.

3.3 Nonparametric Method of Separating Periodic and Stochastic Components

The simple transformation

$$\epsilon_{p,\tau} = \frac{x_{p,\tau} - m_\tau}{s_\tau} \quad , \tag{3.1}$$

in which m_τ and s_τ are the sample means and sample standard deviations at the positions τ, computed by Equations 2.1 and 2.5 respectively, is the nonparametric method of standardization of the $x_{p,\tau}$ variable. This is also a way to remove the periodic components in m_τ and s_τ. It requires the use of 2ω statistics, ω of m_τ and ω of s_τ. For monthly values 2ω = 24, for daily values 2ω = 730, and for hourly values 2ω = 17,520. These two latter

cases require the estimation of 730 or 17,520 statistics of m_τ and s_τ, respectively. If v parameters of the $x_{p,\tau}$ process contain periodicity to be identified, described and separated from the stochastic component of a series, then this nonparametric method requires the use of $v\omega$ statistics for all periodic parameters of the $x_{p,\tau}$ series. For example, with periodicity in five parameters of daily series the number of statistics to be estimated is $v\omega = 1825$. This is an unnecessary large number. The nonparametric method removes from a series the periodicity in parameters but also removes all sampling variations associated with the coefficients of the periodic functions of parameters.

Looking on this nonparametric method from the standpoint of sampling theory, these $v\omega$ values cannot be accurately determined. They must have large sampling errors. Besides, they decrease significantly the number of degrees of freedom. If the objectives of time series analysis are either the condensation of information or the generation of new samples, there is no point in perpetuating these detailed sampling variations in parameters. One of the objectives of statistical analysis and development of mathematical stochastic models is to economize on the number of parameters used to describe any random variable or its stochastic process. As long as the control variable ω and the number of periodic parameters v are small, say $\omega = 12$ for monthly values and $v = 2$ for the mean and the standard deviation, the nonparametric description and removal of periodicity in time series does not present serious difficulties because of the limited number of statistics involved.

This nonparametric method may be very useful in any preliminary analysis or in detecting the character of the stationary stochastic component. For various tests of hypotheses the effective sample size of a dependent stationary stochastic component may be needed. The approximate computation of this size can be obtained by using this nonparametric method, before testing for significance of harmonics.

3.4 Parametric Method of Separation of Periodic and Stochastic Components

To economize on the number of statistics needed for the mathematical description of a series, the periodic series m_τ and s_τ may be approximated for large ω by a relatively small number of harmonics of ω. For example, if the periodic components of daily means and daily standard deviations are well approximated each by six harmonics, and all other fluctuations in m_τ and s_τ are assumed or inferred to be sampling variations, then the Fourier series approximation of a periodic parameter requires only the mean plus 12 values of A_j and B_j Fourier coefficients for each parameter, with a total of 26 statistics. This is a significant savings in the number of statistics used, 26 instead of 730 for the case of daily flows. The savings is even greater for hourly means and standard deviations.

The classical approach in estimating the significant harmonics in composed series is of the type

$$x_{p,\tau} = \mu_\tau + \sigma_x \epsilon_{p,\tau} \quad , \tag{3.2}$$

in which μ_τ is the periodicity in the mean and σ_x is the standard deviation assumed to be a constant. The periodic component is then given in the form

$$\mu_\tau = \mu_x + \sum_{j=1}^{m} (A_j \cos \lambda_j \tau + B_j \sin \lambda_j \tau) \quad , \tag{3.3}$$

for m harmonics with amplitudes significantly different from those of stochastic series, and μ_x is the general mean of $x_{p,\tau}$.

The coefficients, A_j and B_j, $j = 1, 2, \ldots, m$, in Equation 3.3 are estimated from $N\omega$ values of $x_{p,\tau}$ by

$$A_j = \frac{2}{n\omega} \sum_{p=1}^{n} \sum_{\tau=1}^{\omega} (x_{p,\tau} - \mu_x) \cos \frac{2\pi j \tau}{\omega} \quad , \tag{3.4}$$

and

$$B_j = \frac{2}{n\omega} \sum_{p=1}^{n} \sum_{\tau=1}^{\omega} (x_{p,\tau} - \mu_x) \sin \frac{2\pi j \tau}{\omega} \quad , \tag{3.5}$$

in which $\lambda_j = 2\pi j/\omega$ with $\min \lambda_j = 2\pi/\omega$; m is the number of significant harmonics in the range of variation of j with $j = 1, 2, \ldots, \omega/2$, and τ is the time series sequence inside each period, $\tau = 1, 2, \ldots, \omega$, with $N = n\omega$ being the size of sample series. For the last harmonic, $j = \omega/2$ or $j = (\omega-1)/2$ for ω an odd number, the Fourier coefficients are $A_{\omega/2} = A_j/2$ and $B_{\omega/2} = B_j = 0$, for j in Eqs. 3.4 and 3.5 being $\omega/2$.

Because σ_x of Eq. 3.2 is rarely a constant in hydrology if μ_τ is periodic, two cases arise: (1) $\mu_\tau(x)$ and $\sigma_\tau(x)$, as the population periodic components, estimated by μ_τ and σ_τ, are proportional, and (2) $\mu_\tau(x)$ and $\sigma_\tau(x)$ are not

proportional. In the first case, $\sigma_\tau \approx \eta_o \mu_\tau$, with η_o the proportionality constant, so that

$$x_{p,\tau} = \mu_\tau + \sigma_\tau \epsilon_{p,\tau} = \mu_\tau (1 + \eta_o \epsilon_{p,\tau}) = \mu_\tau \epsilon^*_{p,\tau} \quad . \quad (3.6)$$

This is the case of a multiplication of a periodic parameter and a stochastic component, with $\epsilon^*_{p,\tau} = 1 + \eta_o \epsilon_{p,\tau}$, which is a linear transformation of $\epsilon_{p,\tau}$. Equations 3.3 through 3.5 are not applicable in this case. However, by using

$$\ln x_{p,\tau} = \ln \mu_\tau + \ln \epsilon^*_{p,\tau} \quad , \quad (3.7)$$

for $x_{p,\tau} > 0$, $\mu_\tau > 0$, and $\epsilon^*_{p,\tau} > 0$, the case of Equation 3.6 is reduced to the case of applying Equations 3.3 through 3.5 to logarithms of Equation 3.7. The case of applying Equations 3.6 and 3.7 explains why studying logarithms of a hydrologic $x_{p,\tau}$ variable may give, in some cases, more meaningful results than studying the $x_{p,\tau}$ values.

If σ_τ is not proportional to μ_τ, the simple composition model of the periodic and stochastic components is

$$x_{p,\tau} = \mu_\tau + \sigma_\tau \epsilon_{p,\tau} \quad , \quad (3.8)$$

in which case Equations 3.3 through 3.7 are not directly applicable, because μ_τ and σ_τ may have different significant harmonics, nonproportional amplitudes of the same harmonics, and/or different phases in the case of the same significant harmonics.

To avoid these difficulties in the application of the classical approach of Equation 3.2, or its application in using logarithms in the form of Equation 3.7, when the model of Equation 3.8 is required, various parameters that may be periodic along the sequence points $\tau = 1, 2, \ldots, \omega$ should be first computed and the significant harmonics fitted to them.

That the basic periodicity $\omega = 2\pi/\lambda$ is always known in advance in hydrologic time series, say a year or a day, facilitates the fit of significant harmonics of μ_τ and σ_τ to the $x_{p,\tau}$ series. These harmonics may be fitted directly to m_τ and s_τ computed by Eqs. 2.1 and 2.5 respectively. This second approach of direct fits of periodic functions μ_τ and σ_τ to m_τ and s_τ is discussed in detail in the ensuing text. The basic procedure is in computing the ω values of a periodic parameter, in estimating the amplitudes and phases of various harmonics, and in selecting by an appropriate procedure those harmonics that have amplitudes considered significantly greater than those of a random process by any criterion.

Instead of fitting a periodic function σ_τ to the periodic standard deviation s_τ, the periodic function may be fitted to the estimated periodic variance, s^2_τ or $\hat{s}^2_\tau$, as σ^2_τ, and then σ_τ of Eq. 3.8 is determined from this periodic function as $\sqrt{\sigma^2_\tau}$. This approach has an advantage that the mean σ^2_x of σ^2_τ or s^2_τ is easier related to the general variance of $x_{p,\tau}$, while σ_x as the mean of σ_τ or s_τ is not simply related to the general standard deviation, s_x, of $x_{p,\tau}$.

The periodic component in any parameter ν may be approximated by m harmonics of its basic period ω in the form

$$\nu_\tau = \nu_x + \sum_{j=1}^{m} (A_j \cos \lambda_j \tau + B_j \sin \lambda_j \tau) \quad , \quad (3.9)$$

in which $\lambda_j = 2\pi j/\omega$ is the angular (circular) frequency, ω is the basic period in ν, m is the number of harmonics inferred as significant in the Fourier series mathematical description of the periodic parameter ν, and ν_x is the mean of ν_τ fitted to the ω values of the estimated v_τ-values from the sample series, or it also is the mean of v_τ.

The standardization by Equation 3.1 but using the mathematical models of μ_τ and σ_τ, with a limited number of harmonics of Equation 3.9 as the fitted periodic components to m_τ and s_τ, is defined here as the parametric method of standardization:

$$y_{p,\tau} = \frac{x_{p,\tau} - \mu_\tau}{\sigma_\tau} \quad . \quad (3.10)$$

Because of difficulties in estimating the coefficients A_j and B_j of Equation 3.9 directly from the $x_{p,\tau}$ series, they can be estimated from the ω values v_τ by

$$A_j = \frac{2}{\omega} \sum_{\tau=1}^{\omega} v_\tau \cos \frac{2\pi j \tau}{\omega} \quad , \quad (3.11)$$

and

$$B_j = \frac{2}{\omega} \sum_{\tau=1}^{\omega} v_\tau \sin \frac{2\pi j \tau}{\omega} \quad . \quad (3.12)$$

For the last harmonic, $j = \omega/2$ for ω an even number and $j = (\omega-1)/2$ for ω an odd number,

$A_{\omega/2} = A_j/2$, and $B_{\omega/2} = B_j = 0$. This is important in cases when all possible harmonics are computed (say all six harmonics of the monthly series).

The maximum number of harmonics in this discrete series of ω values of v_τ for monthly series is m = 6, and for daily series is m = 182. However, the daily series rarely show significant harmonics beyond the first 6 to 12 harmonics. The Fourier series are rapidly convergent, with the amplitudes of high overtones, particularly over the fourth harmonic for monthly series and over the sixth harmonic for daily series, being small enough to be neglected. For daily data, this circumstance is significant by looking at the explained variances of m_τ and s_τ by harmonics over the fourth or the sixth. It is also implemented in monthly series by looking at the mean and the variance of residuals after the significant harmonics are removed. To illustrate this point, several cases of harmonic analysis were performed for monthly series by removing between two and six harmonics. The results of the differences in explained variances were often small. The fitted periodic functions to m_τ and s_τ, as given by Equation 3.9, are designated by μ_τ and σ_τ, because random sampling fluctuations in m_τ and s_τ are supposed to be greatly reduced, and left remaining inside the stochastic component.

For the parametric method, Equation 3.10 is only approximately a standardized variable, because $E(y_{p,\tau})$ and var $y_{p,\tau}$ are somewhat different from the expected value of zero and the variance of unity, respectively. To obtain a standardized variable in case the parametric method is used, a further transformation produces

$$\epsilon_{p,\tau} = \frac{y_{p,\tau} - \mu_y}{\sigma_y} \quad , \qquad (3.13)$$

in which μ_y is the mean of $y_{p,\tau}$ (estimated by $\bar{y}_{p,\tau}$) and σ_y is its standard deviation (estimated by s_y). The autocorrelation coefficients, the skewness and excess coefficients of distributions for each month or day are not affected by the transformation of Equation 3.13.

The refinement from using the variable $y_{p,\tau}$ of Equation 3.10 and standardizing it to obtain $\epsilon_{p,\tau}$ by Equation 3.13 requires the estimates of the two new parameters, μ_y and σ_y. In the case of six harmonics used for each μ_τ and σ_τ, the total number of parameters to be estimated and used in the standardization procedure of Equations 3.9 through 3.13 is now 28 instead of 26. Similarly as for μ_τ and σ_τ the significant harmonics of the periodicity in other parameters may be determined, and equations of the type of Equation 3.9 may be derived.

If one would like to preserve in $\epsilon_{p,\tau}$ the general mean μ_x and the general standard deviation s_x of the $x_{p,\tau}$ series, one can do so by

$$\epsilon^o_{p,\tau} = \frac{x_{p,\tau} - \mu_\tau + \sigma_\tau \bar{y}}{\sigma_\tau s_y} s_x + \mu_x \quad , \qquad (3.14)$$

in which s_x is different from σ_x with σ_x the mean of ω values of s_τ.

3.5 General Information on Testing the Significance of Harmonics of Periodic Parameters

Assume that ω values have been obtained for a parameter as a new discrete series. To fit any harmonic of the angular frequency $\lambda_j = 2\pi j/\omega$, Eqs. 3.11 and 3.12 can be used to get the estimates of coefficients A_j and B_j. It is not necessary to compute all harmonics for large values of ω. Experience shows that j should be not greater than about m $\approx$ 6 - 12 for ω = 365 of a series of daily values.

The square of amplitude C_j of any harmonic is given by

$$C_j^2 = A_j^2 + B_j^2 \quad , \qquad (3.15)$$

and the mean square of deviations from the mean μ_x, as the variance of that harmonic designated by h_j, is

$$\text{var } h_j = \frac{C_j^2}{2} = \frac{A_j^2 + B_j^2}{2} \quad . \qquad (3.16)$$

Three approaches for determining the significant harmonics in the periodicity of parameters are discussed in this paper: (1) the classical Fisher's approach of a process composed of the sum of a harmonic and a normal independent process; (2) approximate approaches by using either the first m harmonics until P percent of the variation of ω values of v_τ about the mean ν_x of the parameter ν_τ is explained by these first m harmonics, and (3) use of a special property of cumulative periodograms. The search for new theoretical and/or experimentally determined distributions of amplitudes of harmonics of a complex periodic-stochastic process is necessary in the future.

3.6 Fisher's Approach to Testing the Significance of Harmonics

The parameter that can be used in testing the significance of various harmonics of Eqs. 3.2 and 3.7 is the variance of individual harmonics, $C_j^2/2$, provided the Fourier coefficients A_j and B_j are estimated by Eqs. 3.4 and 3.5. If a test shows that a given $C_j^2/2$ value is not greater than a critical $C_c^2/2$ value of an independent stochastic process, this j-th harmonic is considered insignificant. Sampling distribution of the testing parameter, $C_j^2/2$, is needed. Once a given $C_j^2/2$ is found significant, the phase of the harmonic is estimated from the computed A_j and B_j values.

In the case where the variance σ_x^2 of the $x_{p,\tau}$ series must be estimated from the sample data, which is the usual case, Fisher's test of significance should be applied [8] when Equations 3.2 and 3.7 are applicable. Fisher's test uses the statistic in the form of the ratio

$$g = \frac{C_{max}^2}{2\, s_x^2} = \frac{C_{max}^2}{\sum_{j=1}^{m} C_j^2} \quad , \qquad (3.17)$$

for testing the significance of the harmonic with the largest value C_{max}^2 of a sequence of C_j^2 values, with m the total number of harmonics and s_x^2 the estimate of the variance σ_x^2 of the $x_{p,\tau}$ series. For $m = N/2$ in case N is an even number or $m = (N-1)/2$ in case N is an odd number with $N = n\omega$ the total sample size, the probability P that the g value of Equation 3.17 would exceed a critical value g_c is given by

$$P = m(1-g_c)^{m-1} - \frac{m(m-1)}{2}(1-2g_c)^{m-1} + \ldots + (-1)^{k-1}\frac{m!}{k!\,(m-k)!}(1-kg_c)^{m-1} \quad , \qquad (3.18)$$

in which k is the greatest integer less than $1/g_c$. In most cases, the first term on the right side of Equation 3.18 gives a sufficient approximation for g_c. Fisher's test has dominated the detection of significant harmonics in the cases where Equations 3.2 and 3.7 are applicable, and the series is composed of a sum of the periodic and stochastic components. The problem in practice is reversed, with P given and g_c computed either exactly by Equation 3.18 or approximately by the first term on the right side of the equation. Figure 3.1 and Table 3.1 give values of g_c of Equation 3.18 as functions of m for two values of P, P = 0.05 and P = 0.01. If g of Equation 3.17 is greater than g_c of Equation 3.18 for given P, m, and k, the h_1 harmonic with C_{max} is significant; otherwise it is not.

An example of applying Equation 3.7 is when monthly precipitation series, for which the mean m_τ, and the standard deviation s_τ, may be considered proportional, with all $x_{p,\tau}$ values greater than zero (for $x_{p,\tau} = 0$, is should be replaced, say by $x_{p,\tau} = 0.001$, so that the logarithm of $x_{p,\tau}$ is a finite negative value). For simple river regimes (say when only rain produces runoff), the monthly runoff series often have m_τ and s_τ approximately proportional, with Equation 3.7 applicable.

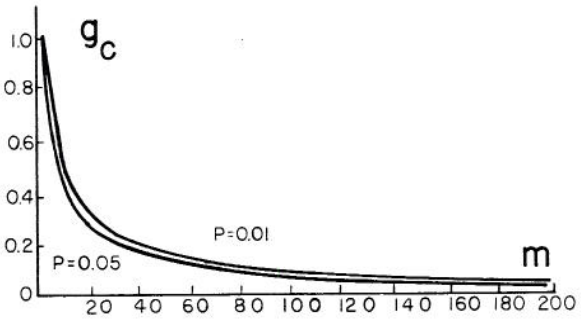

Fig. 3.1 The relations of Fisher's g-critical to the number of possible harmonics for two given probability levels, P = 0.01 and P = 0.05.

If two or more harmonics are significant, two practical approaches may be used. When $C_{max}^2 = C_1^2$ is found to be significant with $g_1 > g_c$ for a selected probability level P (say P = 0.05 or P = 0.01), this harmonic is computed and subtracted from the series. Then the next highest value, C_2^2, is tested but this time against $2\, s_x^2 - C_1^2$, so that the new value g_2 is

$$g_2 = \frac{C_2^2}{2\, s_x^2 - C_1^2} \quad . \qquad (3.19)$$

Similarly for any i-th harmonic from all harmonics sorted in the descending order of C_j^2, the g_i value is

$$g_i = \frac{C_i^2}{2\, s_x^2 - \sum_{j=1}^{i-1} C_j^2} \quad , \qquad (3.20)$$

in which C_j^2 values in the sum are all greater than C_i^2. This approach has biases. First, when a significant harmonic is subtracted from $x_{p,\tau}$ of Equation 3.2 or from $\ln x_{p,\tau}$ of Equation 3.7, the part of variance at its frequency corresponding to the random variable $\epsilon_{p,\tau}$ is also deducted so that the

TABLE 3.1

FISHER'S g-CRITICAL VALUES, FOR m THE TOTAL NUMBER OF HARMONICS

m	P = 0.05	P = 0.01	m	P = 0.05	P = 0.01	m	P = 0.05	P = 0.01
5	.68377	.78874	20	.27040	.32971	35	.17513	.21338
6	.61615	.72179	21	.26060	.31783	36	.17124	.20860
7	.56115	.66440	22	.25155	.30683	37	.16754	.20405
8	.51569	.61517	23	.24315	.29661	38	.16400	.19970
9	.47749	.57271	24	.23534	.28709	39	.16062	.19554
10	.44495	.53584	25	.22805	.27819	40	.15738	.19156
11	.41688	.50357	26	.22123	.26986	41	.15429	.18776
12	.39240	.47510	27	.21483	.26205	42	.15132	.18411
13	.37085	.44982	28	.20883	.25470	43	.14847	.18060
14	.35172	.42722	29	.20317	.24778	44	.14573	.17724
15	.33461	.40689	30	.19784	.24124	45	.14310	.17401
16	.31922	.38851	31	.19280	.23506	46	.14057	.17089
17	.30529	.37180	32	.18803	.22921	47	.13814	.16789
18	.29262	.35655	33	.18351	.22366	48	.13579	.16501
19	.28104	.34257	34	.17921	.21839	49	.13353	.16222
20	.27040	.32971	35	.17513	.21388	50	.13135	.15954

TABLE 3.2

FISHER'S g-CRITICAL FOR CURRENT HYDROLOGIC TIME INTERVAL DISCRETE SERIES

Time Interval	ω	m	g_c P = 0.05	g_c P = 0.01
1-day	365	182	0.04429	0.05275
2-day	182	91	0.08002	0.09632
7-day	52	26	0.22131	0.26986
14-day	26	13	0.37085	0.44982
1-month	12	6	0.61615	0.72179
2-month	6	3	0.87090	0.94226
3-month	4	2	0.97500	0.99500

denominators of Equations 3.19 and 3.20 are somewhat smaller than they should be, or the expected value of g_i is somewhat greater than its true mean. This bias may produce some significant marginal harmonics. Second, all harmonics that by the sampling variation are shown to have squares of amplitudes at the tails of their distributions are automatically accepted as significant, though they belong to a stationary stochastic process. On the other side, the significant harmonics beyond h_1 result in a greater sample value s_x^2 of Equation 3.17 in comparison with a variable without periodicity, or in a smaller g.

The second approach is to investigate the probability that two or more harmonics simultaneously have significant amplitudes. Such an approach is discussed by Fisher [9] for i harmonics being significant at the same time, with the case of two harmonics (g_1 and g_2) shown as an example. In this approach, the i-th harmonic has the critical value g given by

$$P = \sum_{j=i}^{k} (-1)^{j-i} \frac{m!\,(1-jg_c)^{m-1}}{(m-j)!(j-i)!(i-1)!\,j} \quad . \qquad (3.21)$$

For i = 1, Equation 3.21 is identical to Equation 3.18, where k is the largest integer less than $1/g_c$. For i = 2, Equation 3.21 becomes

$$P = \frac{m(m-1)}{1!}\left[\frac{(1-2g_c)^{m-1}}{2} - (m-2)\frac{(1-3g_c)^{m-1}}{3} + \cdots \pm \frac{(m-2)!}{(k-2)!\,(m-k)!}\,\frac{(1-kg_c)^{m-1}}{k}\right], \qquad (3.22)$$

and similar equations for i = 3, 4, . . .,. In this case, the g_c value for i = 1, 2, . . ., may be obtained as soon as P is selected, with k the largest integer less than $1/g_c$.

This method of testing significance of several harmonics simultaneously has difficulties. The critical values of Equation 3.21 for i = 1 and i = 2, and for P = 0.05 and m = 20, are $g_{c,1} = 0.27046$ and $g_{c,2} = 0.17599$, respectively. Assume that $C^2_{max} = C^2_1$ gives $g_1 = 0.253$ and C^2_2 gives $g_2 = 0.183$. In this case, $g_1 < g_{c,1}$, while $g_1 > g_{c,2}$ and $g_2 > g_{c,2}$. Should both C^2_1 and C^2_2 be considered as significant? Fisher realized this difficulty and briefly discusses it at the end of reference [9]. An approach to solving this difficulty is to select all C_j values that produce for a given i, with $j \leq i$, the g_i values that are greater than the $g_{c,i}$ values of Eq. 3.21. However, some other C^2_j values with $j > i$ may also produce $g_i > g_{c,i}$.

The first approach of using Equations 3.17 and 3.20 is an approximate procedure, but regardless of the built-in biases it is simple to apply, and may satisfy certain needs in the analysis of hydrologic time series with known periodicities. The case of Equation 3.2 is not current in hydrologic practice, though it is often assumed and treated as such. For instance, the spectral analysis of periodic-stochastic processes in hydrology is often performed under the hypothesis that Equation 3.2 is applicable. However, Equation 3.7 is a much more current case in hydrology, because the assumption of m_τ and s_τ being proportional may be close to

physical reality. This circumstance is then a justification for the analysis of $\ln x_{p,\tau}$ instead of $x_{p,\tau}$, with all three series, $x_{p,\tau}$, m_τ, and $\epsilon_{p,\tau}$ being the positively valued quantities. In case any one of their values is zero, the zero is replaced by a very small positive value, such as 0.001 or 0.0001 or some other similar small value. Because the transformation $\ln x_{p,\tau}$ and $\ln \epsilon_{p,\tau}$ make their distributions less skewed, or they are close to symmetrical distributions, Fisher's approach is then applicable provided that $\ln \epsilon^*_{p,\tau}$ is an independent variable. This is often satisfied for monthly precipitation series, but rarely fulfilled for the monthly runoff series for which $\ln \epsilon^*_{p,\tau}$ is a time dependent random variable. By applying a proper dependence model to $\ln \epsilon^*_{p,\tau}$, the effective sample length N_e can be determined, and m of Eqs. 3.18, 3.21, and 3.22 becomes now $m = N_e/2$, or $m = (N_e-1)/2$ depending on whether N_e is an even or odd number. Because N_e rarely comes out to be an integer when computed from the dependence model of $\ln \epsilon^*_{p,\tau}$, it should be approximated by the nearest integer.

3.7 Determining Significant Harmonics by Fisher's Test in Cases of Normal Dependent Stochastic Components

Fisher's test, as described, is based on the distribution of the parameter g for the normal independent process. The hypothesis is that μ_τ, the means along τ positions, are a constant, or $\mu_\tau = \mu_x$ When a significant harmonic is found, then the opposite hypothesis, $\mu_\tau \neq \mu_x$, is accepted. Two approaches can be used in the case of normal dependent processes: (1) determining the effective sample size N_e of the dependent stochastic component, and the use of the same procedure as in the case of normal independent process, and (2) producing from the dependence model a new variable, $z_{p,\tau}$, which is an independent stochastic component, approximately normal, and testing the significant harmonics in $z_{p,\tau}$.

If $\epsilon_{p,\tau}$ in Equation 3.2 or $\ln \epsilon^*_{p,\tau}$ in Equation 3.7 are close to a normal dependent process, they may be assumed to follow approximately the first-order linear autoregressive model. The above procedure of choosing an effective series length is then applicable. Assume in this case that $\epsilon_{p,\tau} = \rho_\cdot \epsilon_{p,\tau-1} + \sqrt{1-\rho^2}\ \xi_{p,\tau}$, in which ρ is the first autocorrelation coefficient of $\epsilon_{p,\tau}$ values, so that Equation 3.2 becomes

$$x_{p,\tau} = \mu_\tau + \sigma_x\ (\rho\ \epsilon_{p,\tau-1} + \sqrt{1-\rho^2}\ \xi_{p,\tau}) \quad , \tag{3.23}$$

and Equation 3.7 rewritten in the form

$$\ln x_{p,\tau} = \ln \mu_\tau + (\rho^* \ln \epsilon^*_{p,\tau-1} + \sqrt{1-\rho^{*2}}\ \xi^*_{p,\tau}) \quad , \tag{3.24}$$

in which ρ^* is the first autocorrelation coefficient of $\ln \epsilon^*_{p,\tau}$. The coefficients ρ or ρ^* are estimated by the sample first serial correlation coefficient, r_1 or r_1^* as biased or unbiased estimates.

Assuming as an approximation that $\mu_\tau = \mu_x$ in Equation 3.23 and $\ln \mu_\tau = \ln \mu_x$ in Equation 3.24, the effective sample size for the study of the mean, in case the first-order autoregressive model is applicable, is approximately,

$$N_e = \frac{N(1-\rho)}{1+\rho} \quad , \tag{3.25}$$

with N_e rounded off to the nearest integer. Then $m = N_e/2$ or $m = (N_e-1)/2$, depending on whether the rounded N_e is an even or an odd number, is used in Equations 3.18, 3.21, or 3.22, whichever is appropriate, to compute g_c for a given P. The g values of Equations 3.17, 3.19, and 3.20 are computed either for $x_{p,\tau}$ of Equation 3.2 or for $\ln x_{p,\tau}$ of Equation 3.7, as the case may be. If ν_τ represents the standard deviation, the variance, or autocorrelation coefficients, the effective sample size in case of the variance may be used as an approximation for all these parameters as

$$N_e = \frac{N(1-\rho^2)}{1+\rho^2} \quad . \tag{3.26}$$

Because $N_e < N$, then $m = N_e/2$ in Equation 3.18 means an increase of g - critical in comparison with $m = N/2$. Because $s_x^2 >> \text{var } v_\tau$, g_1 of Equation 3.17 is much smaller in the case of using s_x^2 than in the case of using var v_τ in Equations 3.17 or 3.20. However, because $N_e/2 >> \omega/2$, the critical g values of Equations 3.18 or 3.21 are much smaller also if $N_e/2$ is used instead of $\omega/2$.

It is recommended for the periodic, hydrologic time series to use s_x^2 and $m = N_e/2$ if $\epsilon_{p,\tau}$ is dependent, and ω is small (say $\omega < 52$), and to use var v_τ and $m = \omega/2$ in case $\epsilon_{p,\tau}$ is independent, and ω is large (say $\omega \geq 52$). As

the $\epsilon_{p,\tau}$ components of precipitation series are close to being independent the use of var v_τ and m = $\omega/2$ may be applied for this case.

The other approach of using $z_{p,\tau}$ permits Equation 3.23 to be rewritten as

$$z_{p,\tau} = x_{p,\tau} - \rho\sigma_x \epsilon_{p,\tau-1}$$

$$= \mu_\tau + \sigma_x \sqrt{1-\rho^2}\, \xi_{p,\tau} \quad , \qquad (3.27)$$

with $z_{p,\tau}$ the new variable reducing the problem to the case of Equation 3.2. Similarly, Equation 3.24 gives

$$z_{p,\tau} = \ln x_{p,\tau} - \rho^* \ln \epsilon^*_{p,\tau}$$

$$= \ln \mu_\tau + \sqrt{1-\rho^{*2}}\, \xi_{p,\tau} \quad , \qquad (3.28)$$

with $z_{p,\tau}$ as the new variable reducing the problem to the case of Equation 3.7. The variance of $z_{p,\tau}$ is then

$$\text{var } z = (1-\rho^2) \text{ var } x + \rho^2 \text{ var } \mu_\tau \qquad (3.29)$$

for Equation 3.27, and a similar equation is obtained in the case of Equation 3.28 with x replaced by ln x, μ_τ by ln μ_τ and ρ by ρ^*. Starting from the hypothesis that $\mu_\tau = \mu_x$, then var μ_τ can be assumed to be s_x^2/n, with n the number of ω periods. In this case,

$$\text{var } z_{p,\tau} = (1-r_1^2)\, s_x^2 + \frac{s_x^2}{n} \quad , \qquad (3.30)$$

in which r_1 is the estimate of ρ or ρ^*, and s_x^2 is the estimate of var x or var (ln x) respectively for Equation 3.27 and Equation 3.28. The term s_x^2/n can be neglected if r_1^2 is not close to one. The procedure is as follows. The new variable $z_{p,\tau}$ is computed by Eq. 3.27 or Eq. 3.28, with ρ or ρ^* estimated by r_1, σ_x by s_x, and $\epsilon_{p,\tau}$ is obtained by the standardization $\epsilon_{p,\tau} = (x_{p,\tau} - m_\tau)/s_x$. Then C_j^2 values are computed for the $z_{p,\tau}$ variable, and, consequently, the g values of Eqs. 3.18, 3.21, and 3.22 are obtained, the corresponding g_c values are then computed as it is done for the normal independent process.

In summary, the Fisher's test is carried out by first selecting P value in Equation 3.18, with g_c computed for given P and m. If g computed by Equation 3.17 is smaller than this g_c, the harmonic with C_{max} is considered as insignificant. The original Fisher's table gives g_c values up to m = 50. It is extended in Table 3.1 to m = 182, because of the use of ω = 365 for daily values. Table 3.2 gives the most important values of ω in the analysis of periodic components of hydrologic time series and their m and g_c values for both P(g) = 0.05 and P(g) = 0.01 probability levels, computed by using only the first term on the right side of Equation 3.18. If g of the harmonic with the largest value C^2_{max} is shown to be insignificant, or g of Equation 3.17 has been shown to be smaller than g_c of Equation 3.18 for given P and m, then the values μ_τ (of $x_{p,\tau}$, or $\ln x_{p,\tau}$, or $z_{p,\tau}$) are inferred to be nonperiodic. If C^2_{max} is shown to be significant, the test is repeated for the harmonic with the second largest C_j^2, and so on. For each successive harmonic or the ranked values C_j^2, the values g are computed by Equation 3.20.

3.8 An Approximate Empirical Approach for Testing the Significance of Harmonics

Because of difficulties in applying Fisher's test for inferring the significant harmonics in various periodic parameters under the conditions of complex compositions of hydrologic time series, an approximate testing method is developed in this study as an empirical procedure. This procedure is as follows.

Any parameter of a hydrologic series is assumed to be periodic until proven that it has no significant harmonic, with the periodicity defined by the ω discrete values of τ, with $\tau = 1,2,\ldots,\omega$. A periodic parameter is designated by ν_τ and described mathematically by Eq. 3.9, with m being the number of inferred significant harmonics. The ω values of v_τ, as the estimates of ν_τ, have the variance $s^2(v_\tau)$. A total of $\omega/2$ or $(\omega-1)/2$ harmonics, for ω an even or an odd number respectively, can be estimated by Eqs. 3.11 and 3.12 from the v_τ estimates.

Because a limited number of harmonics of the lowest frequencies is sufficient to explain the major part of variance $s^2(v_\tau)$ of a periodic parameter ν_τ, it is not necessary to always estimate all $\omega/2$ or $(\omega-1)/2$ harmonics. The maximum number of potential significant harmonics in a series of monthly values is six. It is assumed in this empirical procedure that only the first six harmonics of a periodic parameter for time series of any interval $\Delta t \leq 30$ days should be tested for significance. In other words, if all six harmonics of monthly time series may be found significant, only 6 harmonics also may be found to be significant for 15-day, 7-day, 3-day, or 1-day time interval series. However, it

should be expected that the population periodic functions may need a larger number of harmonics for describing these periodicities as the time interval Δt decreases. Regardless of this general expected pattern, the present experience in studying the periodicities in parameters of daily flow or daily precipitation series shows that several harmonics beyond the sixth harmonic add relatively small additional explanation of the variance of estimated v_τ values. Besides, when the eventual significant harmonics beyond the sixth harmonic are not included in the mathematical model of a periodic parameter of Equation 3.9, they are retained in the stochastic component in case the periodicity in ν_τ up to six significant harmonics is removed from the original time series. This is equivalent to stating that a small part of the periodic function in the v_τ values is not removed from the stochastic component.

For the Fourier coefficients of the first six harmonics, estimated by Equations 3.11 and 3.12 from ω values of v_τ, the variances of harmonics are computed by Equation 3.16 as var h_j, j = 1,2,...,6. The mean square deviations of the values of discrete harmonic functions from the general mean ν_x of a parameter ν_τ, with h_j the symbol of a harmonic, are called here the variances of harmonics. The ratio

$$\Delta p_j = \frac{\text{var } h_j}{s^2 (v_\tau)} \qquad (3.31)$$

represents the part of the variation of v_τ which is explained by the j-th harmonic. The sum of Δp_j, j = 1 2, . . .,6, gives p, the part of variation of v_τ which is explained by the first six harmonics.

This empirical procedure is based on the selection of two critical p-values, p_{min}, and $p_{max} = 1 - p_{min}$. If $p \leq p_{min}$, no significant harmonic exists in the sequence of v_τ values, or $\nu_\tau = \nu_x$ is a nonperiodic parameter. If $p_{min} < p \leq p_{max}$, all six harmonics are inferred to be significant. However, if $p > p_{max}$, only some of the six harmonics are considered significant. The values of var h_j are then sorted by magnitude from the highest to the lowest. Only those harmonics with the highest var h_j are selected, which when summed up first exceed p_{max}. As an example, if the three harmonics with highest var h_j have $\Sigma \Delta p_j < p_{max}$, but the four harmonics with highest var h_j have $\Sigma \Delta p_j > p_{max}$, these four harmonics are inferred to be significant.

The general expected pattern is that p_{min} is dependent on the length ω of the basic period (or on the time interval, Δt), the sample size or the number of periods in the series available (say, n years), and the order c of the highest moment used in the definition of a parameter ν. This pattern serves as the basis for deriving an empirical expression for p_{min}, and $p_{max} = 1 - p_{min}$. The greater ω the larger should be p_{min} in order to include all six harmonics in the periodic component, because the larger ω the more pronounced should be the periodicity in a parameter. The larger the sample size or the number n of years of data the smoother should be v_τ values along τ and the smaller will be p_{min} as the critical value for the rejection of significance of harmonics. Similarly, the higher the order of the highest moment used in the definition of a parameter ν, the larger is the sampling variation of the computed v_τ about the periodic function ν_τ, and the smaller should be p_{min}. The empirical expressions of p_{min} and p_{max} are then

$$p_{min} = a \sqrt{\frac{\omega}{cn}} \quad , \qquad (3.32)$$

and consequently

$$p_{max} = 1 - a \sqrt{\frac{\omega}{cn}} \quad . \qquad (3.33)$$

The suggested empirical constant to use is a = 0.033. The practical ranges of using Equations 3.32 and 3.33 are $12 \leq \omega \leq 365$, or, Δt = 1 day to Δt = 30 days, and $10 \leq n \leq 160$. This constant may be taken somewhat smaller than 0.033 if the periodicity of six harmonics should be retained; because the smaller p_{min} is, and the larger p_{max} is the chances are greater for all six harmonics to be significant. If the constant is taken greater than 0.033, the rejection region $p \leq p_{min}$ will be increased, and the region for less than six harmonics being significant, $p > p_{max}$, will also be increased.

As an example, $p_{min} = 0.10$ for a = 0.033, n = 40, ω = 365 and c = 1; $p_{min} = 0.07$ for the same values of a, n, and ω, and c = 2. For n = 160, ω = 12, c = 1, and a = 0.033, then $p_{min} = 0.009$ (1 percent). For a time series of 160 years the periodicity in monthly means would be rejected only if all six monthly means are nearly equal. For n = 10, ω = 365, c = 2, and a = 0.033, then $p_{min} = 0.145$ (14.5 percent). The chances are much greater that the periodicity in a parameter based on the second order moment will be rejected, because the first six harmonics out of 182 possible harmonics would explain less than 14.5 percent of

the total variation of v_τ. Though this empirical approach is based on several arbitrary decisions, it may be useful until good theoretical approaches, or experimental statistical (Monte Carlo) methods of testing significance of harmonics in the periodic parameters of complex hydrologic series are developed.

3.9 Use of the Cumulative Periodogram and the Breaking Point in a Graphical Estimation Procedure

Because periodicities in hydrologic time series are known, with no need to estimate frequencies that may or may not be significant, the line-spectrum (periodogram) is an appropriate technique in that case. The ratio of the cumulative variance of the first m harmonics in relation to the variance of estimates v_τ of a parameter ν_τ , gives the line-spectrum cumulative information

$$p_m = \frac{\sum_{j=1}^{m} \text{var } h_j}{\text{var } v_\tau} \qquad (3.34)$$

The symbol j may refer to a sequence of harmonics from the smallest to the highest frequencies, say j = 1, 2, . . .,$\omega/2$, or $(\omega\text{-}1)/2$, so that a harmonic with a large amplitude may be added to p after a harmonic with a smaller amplitude. However, the harmonics may be sorted according to the magnitude of their amplitudes, from the largest to the smallest. In that case the symbol j, j = 1, 2, . . .,ω, or to $(\omega\text{-}1)/2$, refers to this ordered sequence of amplitudes. In the latter case, the cumulative sum has a convex upward shape. However, in the examples of the use of graphical estimates of significant harmonics the first rather than the second approach is used.

The graphical method is based on the concept that the variation of p_m as a function of m, $p_m = f(m)$, is composed of the two distinct parts: (1) the periodic part of a fast rising of p_m with m, and (2) the sampling part of a slow rising of p_m with m. Two approaches are feasible. First, the two parts are approximated by smooth curves that intersect at a point. The critical frequency of that point then gives the number of significant harmonics, which are all harmonics with lower frequency than this critical frequency. The second approach is to assume the approximate mathematical models of the two parts, estimate the parameters of these models, and find the intersection of the two curves. The frequency nearest to the intersection point is then the critical frequency.

In this second approach the independent stochastic components would produce a straight line part of sampling variation, while the autoregressive models would produce the equations of the corresponding cumulative curve of sampling errors. The change of squares of amplitude from the first to the sixth harmonic, for example, by fitting a one-parameter gamma function, would give an approximate law of the change of sequence of squares of amplitudes. The integral of this equation would produce the mathematical description of the rising limb due to the periodicity in a parameter ν.

Figures 3.2 and 3.3 present the above concepts of fitted curves graphically, or the fitted functions in determining the intersection point A for the critical frequency (f_c) for a periodic-stochastic process in the case of an independent and a dependent stochastic component, respectively. The vertical position (p_m) of the point A is determined by the sample size, while its horizontal position (f_c) should be little affected by the sample size and the sampling variations. Difficulties arise when the point A of Figure 3.3 for a dependent stochastic component is in such a position that both fitted curves, (3) and (4), come out to be nearly one continuous curve,

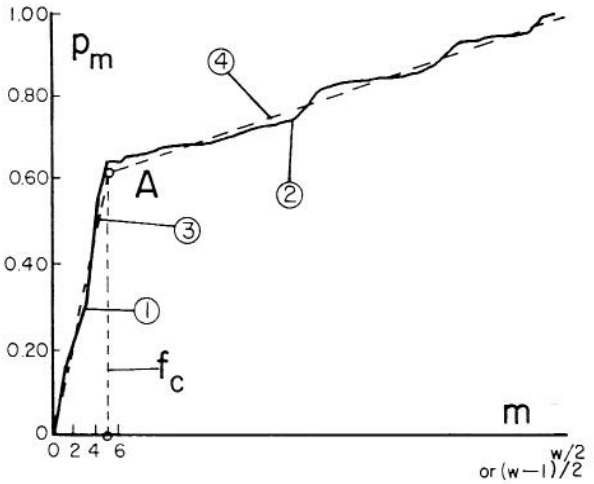

Fig. 3.2 Separation of the cumulative relative periodogram into the periodic part, both the observed (1) and the fitted (3), and the sampling variation part, also both the observed (2) and the fitted (4), in case of an independent stochastic component in a periodic-stochastic process.

implying that the separation of two basic parts of the cumulative relative periodogram becomes uncertain. Examples show that this case is less common in practice.

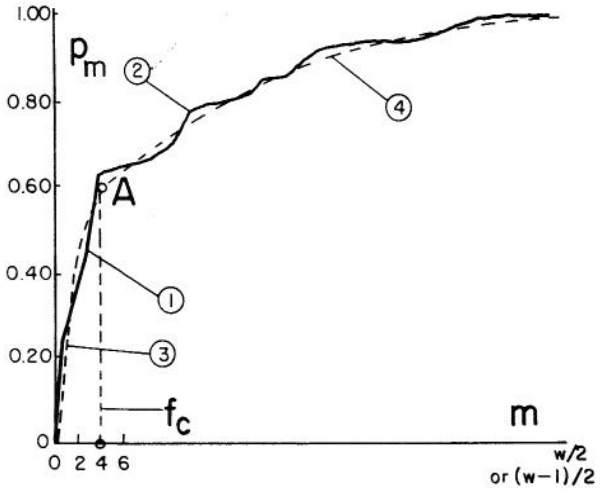

Fig. 3.3 Separation of the cumulative relative periodogram into the periodic part, observed (1) and fitted (3), and the sampling variation part, observed (2) and fitted (4), in case of dependent autoregressive linear stochastic component of a periodic-stochastic part.

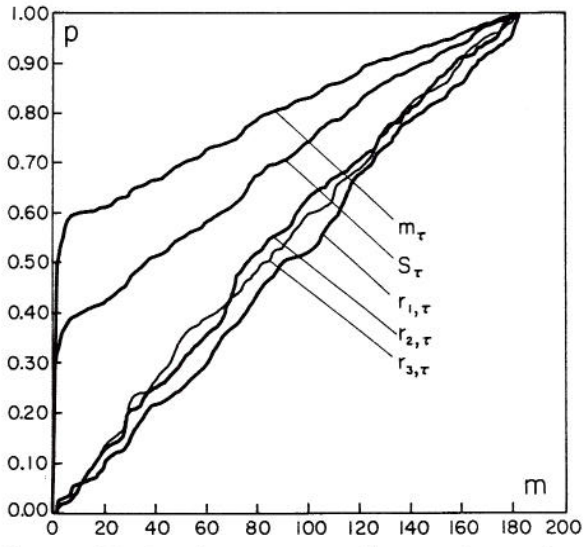

Fig. 3.4 Explained variance p by m harmonics in five parameters of daily precipitation series, Fort Collins, Colorado.

Figures 3.4 through 3.8 as examples give the relation of p_m to m for five parameters: the mean m_τ, the standard deviation s_τ, and the first, second, and third serial correlation coefficients, $r_{1,\tau}$, $r_{2,\tau}$, and $r_{3,\tau}$, for five discrete series: (1) the daily precipitation at Fort Collins, Colorado, from 1898 to 1966, or for 69 years, Figure 3.4; (2) the 3-day precipitation at Austin, Texas, from 1898 to 1967, or for 70 years, Figure 3.5; (3) the 7-day precipitation at Ames, Iowa, from 1949 to 1966, or for 18 years, Figure 3.6; (4) the daily discharge of the Tioga River near Erwins, New York, from 1921 to 1960, or for 40 years, Figure 3.7; and (5) the 3-day

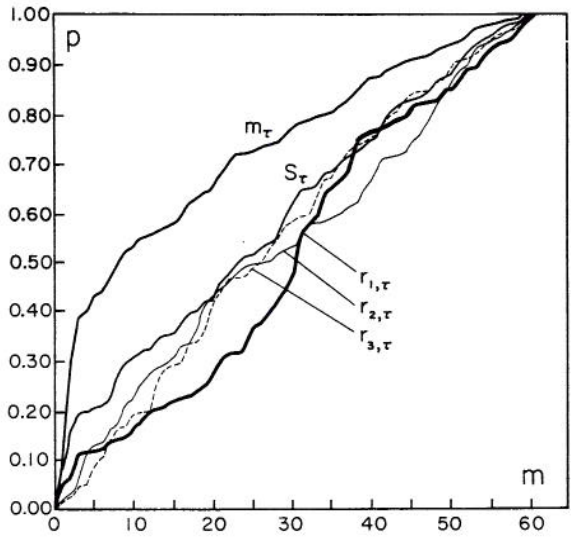

Fig. 3.5 Explained variance p by m harmonics in five parameters of 3-day precipitation series, Austin, Texas.

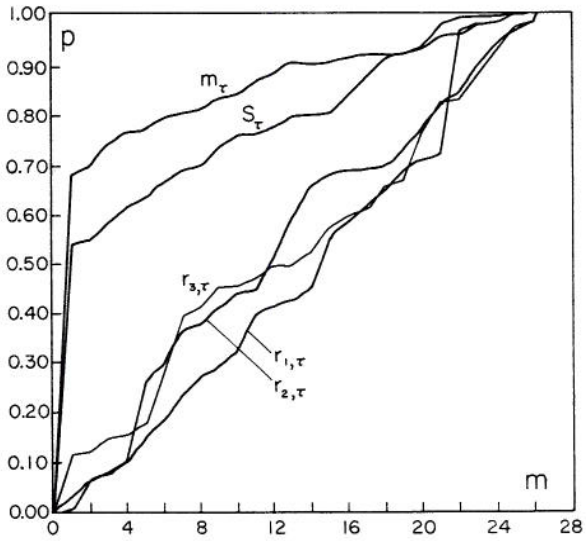

Fig. 3.6 Explained variance p by m harmonics in five parameters of 7-day precipitation series, Ames, Iowa.

discharge of the McKenzie River at McKenzie Bridge, Oregon, from 1924 to 1960, or for 37 years, Figure 3.8. The parameter m has the range m = 1 - 182 for daily series, m = 1 - 60 for three-day series, and m = 1 - 26 for seven-day series. Because other precipitation and river gauging stations for one-day, three-day, and seven-day discrete series show results which are similar to those of Figs. 3.4 through 3.8, the following conclusions drawn from these figures are generally valid.

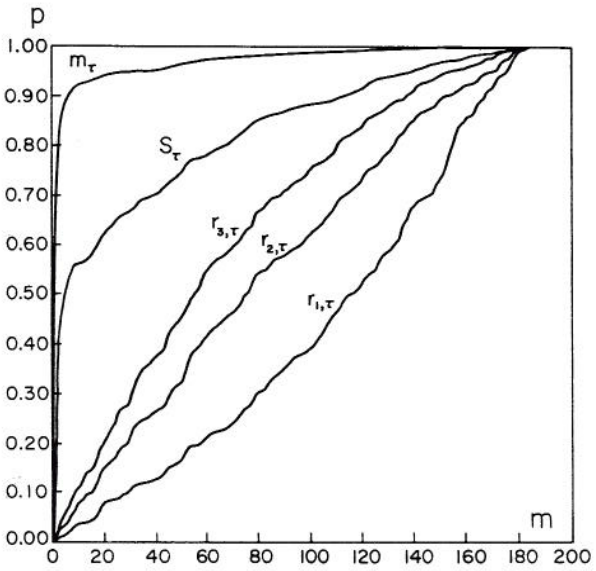

Fig. 3.7 Explained variance p by m harmonics in five parameters of daily flow series of the Tioga River.

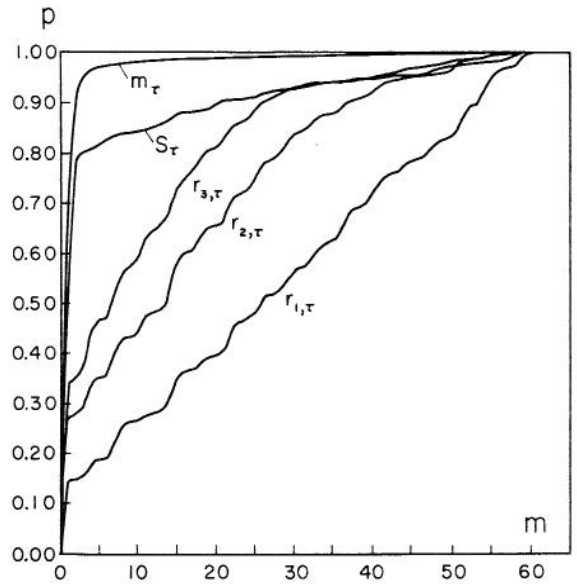

Fig. 3.8 Explained variance p by m harmonics in five parameters of 3-day flow series of the McKenzie River.

(1) The mean m_τ and standard deviation s_τ for the precipitation series are periodic, with the $p_m = f(m)$ curve composed of two basic parts: a steep rise from m = 1 to an m value of up to m = 6 as the periodic part, and a slow rise beyond that m following approximately a straight line as the sampling variation part with an independent stochastic component. The rapid rise can be attributed to periodicities in m_τ or s_τ, and the slowly rising straight line may be considered only the sampling variation. The curve $p_m = f(m)$ for m_τ is always above the corresponding curve for s_τ; this difference is attributed to a larger sampling variation of the second moment, s_τ^2, than of the first moment, m_τ.

(2) The three serial correlation coefficients, $r_{1,\tau}$, $r_{2,\tau}$, and $r_{3,\tau}$, after the periodicities in the mean μ_τ and in the standard deviation σ_τ are removed, show approximately the straight line $p_m = f(m)$ relations that are expected for nonperiodic and sequentially independent estimates v_τ of a parameter ν.

(3) The estimated means m_τ and the estimated standard deviations s_τ of one-day and three-day series (and also of seven-day series) of the examples of runoff series show a sharp rise of the curve $p_m = f(m)$ up to m = 3-6, and then a slow rise to the maximum value of m, with this upper part of the curve mostly convex upward. The first part of the $p_m = f(m)$ curve indicates the significant periodicity, whereas the second part indicates sampling effects for a dependent stochastic component of approximately the autoregressive type.

(4) Rivers with runoff predominately produced by rainfall demonstrate no periodicity in the serial correlation coefficients, as shown by Figure 3.7. However, the $p_m = f(m)$ curves do not follow closely the straight line from m = 1 on, showing that the sampling variation of serial correlation coefficients about their mean values ρ_1, ρ_2, and ρ_3 are affected by the sequential dependence in $r_{1,\tau}$, $r_{2,\tau}$, and $r_{3,\tau}$ as the result of sequential dependence of the underlying stochastic component after the periodicities μ_τ and σ_τ are removed.

(5) Rivers greatly affected by snow accumulation and melt, or river regimes with combined runoff from rainfall and snowmelt, usually show periodicity in serial correlation coefficients, as shown by Figure 3.8. All three parameters, $r_{1,\tau}$,

$r_{2,\tau}$, and $r_{3,\tau}$, exhibit the same sudden rise for small m as do m_τ and s_τ. However, their p_m critical intersection values are lower.

(6) As expected, the sample size affects the smoothness and reliability of the p_m = f(m) curves, as shown in a comparison between Figure 3.6 with 18 years of data and Figures 3.4 and 3.5 with about 70 years of data, though Figure 3.6 refers to seven-day series, whereas Figure 3.4 and 3.5 refer to one-day and three-day series.

(7) Precipitation discrete series with time intervals as fractions of the year show clearly that their nonstationarity basically results from the periodicity in the mean and the standard deviation, while the nonstationarity of the corresponding discrete series of runoff often results from the periodicity in the serial correlation coefficients, and likely in periodicities of third and fourth order parameters.

(8) The p_m = f(m) graphs enable an investigator to advance hypotheses about the significant harmonics present by finding the breaking points between the fast and slow rising parts of the curve.

This graphical method, or the use of fitted functions of cumulative relative periodogram for its periodic and sampling variation parts to find the critical dividing frequency f_c is a promising approach for inferring the significance of harmonics in periodic hydrologic parameters. The detailed treatment of this method is not the subject of this paper.

3.10 Explained Variance of a Periodic-Stochastic Process by its Components

The simplest periodic-stochastic structure of a hydrologic time series is a summation of the periodic mean and a stochastic component given by

$$x_{p,\tau} = \mu_\tau + \epsilon_{p,\tau} \quad , \qquad (3.35)$$

in which μ_τ is the periodic mean at any position τ, $\tau = 1,2,...,\omega$, and $\epsilon_{p,\tau}$ is a stationary stochastic component. It is assumed here that $E(\epsilon_{p,\tau}) = 0$ and var $\epsilon_{p,\tau} = \sigma_\epsilon^2$. For the periodic mean μ_τ Eq. 3.9 is applicable with ν_τ replaced by μ_τ, and ν_x by μ_x. This case is applicable when Eqs. 3.6 and 3.7 are applicable, so that it is warranted to start the analysis of explained variance by this simple case.

As a consequence of Equation 3.9 and 3.35, the property of $x_{p,\tau}$ is

$$E(x_{p+i,\tau}) = E(x_{p,\tau}) \qquad (3.36)$$

with i an integer, for any p and τ. The variance of $\epsilon_{p,\tau}$ is

$$E(x_{p,\tau} - \mu_\tau)^2 = E(\epsilon_{p,\tau}^2) = \sigma_\epsilon^2 \qquad (3.37)$$

for any τ. The variance of $x_{p,\tau}$ about the general mean μ_x in case μ_τ and $\epsilon_{p,\tau}$ are independent is

$$\text{var } x_{p,\tau} = \text{var } \mu_\tau + \text{var } \epsilon_{p,\tau}$$

$$= \frac{1}{2} \sum_{i=1}^{m} (A_i^2 + B_i^2) + \sigma_\epsilon^2 \cdot \qquad (3.38)$$

The covariance of $\epsilon_{p,\tau}$ at any position τ is

$$\text{cov } (\epsilon_{p,\tau}, \epsilon_{p,\tau+k}) = \rho_k(\epsilon)\, \sigma_\epsilon^2 \quad , \qquad (3.39)$$

which is independent of τ and is zero for a given $k \neq 0$ and for $\epsilon_{p,\tau}$ an independent random variable; $\rho_k(\epsilon)$ is a constant for a given $k \neq 0$, independent of τ in case $\epsilon_{p,\tau}$ is a dependent stationary random variable. The covariance of $x_{p,\tau}$ has a superposition of a periodic function to the correlogram of Equation 3.39.

The portion of variance of the variable $x_{p,\tau}$ explained by its periodic mean μ_τ is

$$\frac{\text{var } \mu_\tau}{\text{var } x_{p,\tau}} = \frac{1}{1 + 2\sigma_\epsilon^2 / \sum_{i=1}^{m} C_i^2} \quad , \qquad (3.40)$$

with C_i^2 given by Equation 3.25. The complement to unity of the value of Equation 3.40 represents the portion of variance of $x_{p,\tau}$ explained by the $\epsilon_{p,\tau}$ variable.

The application of Equations 3.35 and 3.40 to the sum of Equation 3.7 carries to bias because of the logarithmic transformation in passing from Equation 3.6 to Equation 3.7. However, the concept of variances of a dependent variable, explained by the other variable, is based on a linear relation between all these variables. When the relation is a power function or a product of a set of variables, the logarithmic transformation reduces the relation to a linear function.

It may be postulated that the periodicity is only in the standard deviation σ_τ (and consequently in the variance σ_τ^2 or in the covariances C_τ) but not in the mean, in autocorrelation coefficients or in

any other parameter. This case is presented here for the sake of completeness. The connection between the periodic and stochastic components is given by

$$x_{p,\tau} = \sigma_\tau \epsilon_{p,\tau} \quad , \tag{3.41}$$

in which $\epsilon_{p,\tau}$ and σ_τ are independent, $\epsilon_{p,\tau}$ has the same stationarity properties as in Equation 3.35, and σ_τ is the periodic component in the standard deviation. In this case ν_τ of Equation 3.9 is replaced by σ_τ, and ν_x by σ_x.

As a consequence of Equations 3.9 and 3.41

$$E(x_{p,\tau}) = E(x_\tau) = \sigma_\tau E(\epsilon_{p,\tau}) \quad , \tag{3.42}$$

which is zero if $E(\epsilon_{p,\tau}) = 0$ for any τ. The general variance of $x_{p,\tau}$, with σ_τ and $\epsilon_{p,\tau}$ independent, is

$$\begin{aligned} \text{var } x_{p,\tau} &= \text{var } \sigma_\tau \epsilon_{p,\tau} = \text{var } \sigma_\tau \text{ var } \epsilon_{p,\tau} \\ &= \sigma_\epsilon^2 \text{ var } \sigma_\tau \quad , \end{aligned} \tag{3.43}$$

with var $\sigma_\tau = \Sigma C_i^2/2$, i = 1, 2, . . .,m, and m the number of significant harmonics in σ_τ. For a given τ, var $x_{p,\tau} = \sigma_\tau^2$ var $\epsilon_{p,\tau}$, or a periodic function of τ. The covariance function at a given τ is

$$\begin{aligned} \text{cov } (x_{p,\tau}, x_{p,\tau+k}) &= \sigma_\tau \sigma_{\tau+k} \text{ cov } (\epsilon_{p,\tau}, \epsilon_{p,\tau+k}) \\ &= \sigma_\tau \sigma_{\tau+k} \rho_k(\epsilon) \sigma_\epsilon^2 \quad , \end{aligned} \tag{3.44}$$

also a periodic function of τ. The autocorrelation coefficients for a given τ are

$$\rho_k(x) = \frac{\sigma_\tau \sigma_{\tau+k} \text{ cov } (\epsilon_{p,\tau}, \epsilon_{p,\tau+k})}{\sigma_\tau \sigma_{\tau+k} \sigma_\epsilon^2} = \rho_k(\epsilon) \quad , \tag{3.45}$$

or they are independent of τ.

Because the logarithmic transformation gives

$$\ln x_{p,\tau} = \ln \sigma_\tau + \ln \epsilon_{p,\tau} \quad , \tag{3.46}$$

provided $\epsilon_{p,\tau} > 0$, the portion of variance of $x_{p,\tau}$ explained by the periodic component is approximately

$$e_\sigma = \frac{\text{var } (\ln \sigma_\tau)}{\text{var } (\ln x_{p,\tau})} \, . \tag{3.47}$$

The most current hydrologic case is the occurrence of periodicity in both the mean and the standard deviation. There are several reasons for this joint occurrence, which are not discussed in this paper.

The structure of a time series of the type

$$x_{p,\tau} = \mu_\tau + \sigma_\tau \epsilon_{p,\tau} \quad , \tag{3.48}$$

in which μ_τ and σ_τ are independent of $\epsilon_{p,\tau}$, with μ_τ and σ_τ given by Equation 3.9 with the corresponding changes for each case, represents the general case of periodic mean and standard deviation.

The variance of $x_{p,\tau}$ is

$$\begin{aligned} \text{var } x_{p,\tau} &= \text{var } \mu_\tau + \text{var } \sigma_\tau \epsilon_{p,\tau} \\ &\quad + 2 \text{ Cov } (\mu_\tau, \sigma_\tau \epsilon_{p,\tau}) \quad . \end{aligned} \tag{3.49}$$

For σ_τ and $\epsilon_{p,\tau}$ independent, and $\epsilon_{p,\tau}$ an independent stationary random variable, then

$$\text{var } x_{p,\tau} = \text{var } \mu_\tau + \sigma_\epsilon^2 \text{ var } \sigma_\tau \quad , \tag{3.50}$$

so that the portion of variance of $x_{p,\tau}$ explained by μ_τ is

$$e_\mu = \frac{\text{var } \mu_\tau}{\text{var } x_{p,\tau}} = \frac{1}{1 + \sigma_\epsilon^2 \sum_{i=1}^{m_2} C_i^2 / \sum_{i=1}^{m_1} C_i^2} \, , \tag{3.51}$$

with m_1 the significant harmonics in μ_τ and m_2 in σ_τ.

To find the portion of variance explained by σ_τ, $1 - e_\mu$ should be divided in proportion to var $(\ln \sigma_\tau)$ and var $(\ln \epsilon_{p,\tau})$, though this logarithmic transformation is biased and is only an approximation, so that

$$e_\sigma = \frac{(1 - e_\mu) \text{ var } (\ln \sigma_\tau)}{\text{var } (\ln \sigma_\tau) + \text{var } (\ln \epsilon_{p,\tau})} \quad , \tag{3.52}$$

and

$$e_\epsilon = 1 - e_\mu - e_\sigma \quad , \tag{3.53}$$

with e_μ, e_σ and e_ϵ being the portions of variance of $x_{p,\tau}$ explained by μ_τ, σ_τ and $\epsilon_{p,\tau}$, respectively.

The next combination of a stochastic stationary component and the periodic parameters is the case of periodicity in autocorrelation coefficients $\rho_k(\epsilon)$. Assume that the $\rho_k(\epsilon)$ correlogram is given for any linear type dependence of a discrete time series such as the first, second or higher order autoregressive linear models. For the m-th order linear model the series is

$$\epsilon_{p,\tau} = \sum_{j=1}^{m} \alpha_{j,\tau-j} \epsilon_{p,\tau-j} + \sigma \xi_{p,\tau} \quad , \tag{3.54}$$

in which, $\alpha_{1,\tau}, \ldots, \alpha_{m,\tau}$ are parameters (in general either periodic functions or constants) to be estimated from the $\epsilon_{p,\tau}$ series, and $\xi_{p,\tau}$ is the independent second-order stationary random variable. The periodicity may be in any or all of m coefficients,, $\alpha_{1,\tau}, \ldots, \alpha_{m,\tau}$. For the simple case of the

first-order autoregressive linear model, with

$$\epsilon_{p,\tau} = \rho_{1,\tau}\,\epsilon_{p,\tau-1} + \sigma\,\xi_{p,\tau} \quad , \quad (3.55)$$

in which $\alpha_1 = \rho_{1,\tau}$, the parameter $\rho_{1,\tau}$ is assumed to be periodic, and $\sigma = (1-\rho_{1,\tau}^2)^{1/2}$. In that case, ν_τ of Eq. 3.9 is replaced by $\rho_{1,\tau}$, and ν_x by ρ_ϵ, where ρ_ϵ is the mean first autocorrelation coefficients of ω values of $\rho_{1,\tau}$, and m is the number of harmonics in the description of $\rho_{1,\tau}$, which is not necessarily the same as for the μ_τ and σ_τ series. The structure of the time series is expressed by

$$x_{p,\tau} = \mu_\tau + \sigma_\tau(\rho_{1,\tau}\,\epsilon_{p,\tau-1} + \sigma\,\xi_{p,\tau}) \quad (3.56)$$

in which μ_τ, σ_τ and $\rho_{1,\tau}$ are periodic and independent of $\xi_{p,\tau}$, with $E(\xi_{p,\tau}) = 0$, var $\xi_{p,\tau} = 1$, and cov $(\xi_{p,\tau}, \xi_{p,\tau+k}) = 0$ for any $k \neq 0$. The periodicity in $\rho_{1,\tau}$ implies the periodicity in σ, in order that $\xi_{p,\tau}$ has no second-order parameter periodic. The composition of a series given by Equation 3.56 has the periodic mean, standard deviation and autocorrelation coefficients, with $\rho_{k,\tau} = \rho_{1,\tau}^k = f(k,\tau)$, because the periodic $\rho_{1,\tau}$ makes all $\rho_{k,\tau}$ periodic. The case of Equation 3.56 is a combination of an independent second-order standardized stationary stochastic variable and the periodic second-order parameters. The case of $\rho_{1,\tau}$ a constant, and $\rho_{k,\tau}$ coefficients a sequence of constants, but μ_τ and σ_τ periodic is called here the quasi second-order periodicity.

Similarly as for Equation 3.56, the second-order, third-order or higher-order autoregressive linear models may have periodic autoregressive coefficients.

The portions of variance of $x_{p,\tau}$ explained by μ_τ, σ_τ, $\alpha_{j,\tau}$'s, and $\sigma\,\xi_{p,\tau}$ may be determined as above by using the logarithmic transformation whenever a product of two terms is involved.

3.11 Testing The Significance of Harmonics in $\rho_{k,\tau}$ Coefficients by the Split-Sample Technique

The ω autocorrelation coefficients $\rho_{k,\tau}$, of $\epsilon_{p,\tau}$, are most currently estimated by the sample serial correlation coefficients

$$r_{k,\tau} = \frac{\sum_{p=1}^{n} (\epsilon_{p,\tau} - \bar{\epsilon}_{p,\tau})(\epsilon_{p,\tau+k} - \bar{\epsilon}_{p,\tau+k})}{\left[\sum_{p=1}^{n} (\epsilon_{p,\tau} - \bar{\epsilon}_{p,\tau})^2 \sum_{p=1}^{n} (\epsilon_{p,\tau+k} - \bar{\epsilon}_{p,\tau+k})^2\right]^{1/2}} \quad (3.57)$$

for $\tau = 1, 2, \ldots, \omega$, and k = 1, 2, 3, . . ., though these estimates may be biased. To test whether $\rho_{1,\tau}$, or its estimate $r_{1,\tau}$ with k = 1 in Equation 3.57, is periodic or not the following procedure may be used. The series of ω values $r_{1,\tau}$ is split into two sub-samples: (1) all $r_{1,\tau}$ values with τ odd numbers and (2) all $r_{1,\tau}$ with τ even numbers.

The reason for this split-sample approach is the dependence between the successive values of $\rho_{1,\tau}$ introduced by the manner of computing them. As an example, when $r_{1,\tau}$ is computed, say for τ = 25th and $\tau + 1$ = 26th April for daily values, the two computed $r_{1,\tau}$ values are dependent even for independent time series. For $\rho_{1,\tau}$ of April 25, values of $\epsilon_{p,\tau}$ for the 25th and 26th April are used. For $\rho_{1,\tau}$ of April 26, values of $\epsilon_{p,\tau}$ for the 26th and 27th April are used. Because $\epsilon_{p,\tau}$ values of April 26 figure in both $\rho_{1,\tau}$ values, they must be dependent. By putting every second value of $\rho_{1,\tau}$ in the first subsample, and the remaining values in the second subsample, this split-sample approach avoids the problem of spurious correlation between the successive $\rho_{1,\tau}$ values. However, due to the time dependence in $\epsilon_{p,\tau}$, there must also be the corresponding dependence in $\rho_{1,\tau}$, which is a separate dependence from the spurious correlation introduced. Similarly as for daily values, the $r_{1,\tau}$ for the month of April and $r_{1,\tau+1}$ for the month of May for monthly values are dependent because of spurious correlation, apart from the eventual dependence in $\rho_{k,\tau}$ resulting from the dependence in $\epsilon_{p,\tau}$.

For each of the two subsamples of $\omega/2$ values if ω is even, or of $\omega/2 + 1$ and $\omega/2$ if ω is odd, the test of significance of harmonics is carried out as it was done for other parameters, and as shown in the previous text. If both subsamples show the same harmonics to be significant, those harmonics are considered to be significant.

A similar test is carried out whether $\rho_{2,\tau}$, for k = 2 in Equation 3.57, is periodic or not. The sample of ω values of $r_{2,\tau}$ is split into two independent subsamples. The first subsample contains all values $r_{2,\tau}$ starting with $r_{2,1}$ and includes the $r_{2,\tau}$ values two at a time with the next two deleted so that the first subsample is $r_{2,1}$, $r_{2,2}$, $r_{2,5}$, $r_{2,6}$, $r_{2,9}$, $r_{2,10}$,..., or approximately $\omega/2 + 1$ or $\omega/2$ values, whatever comes out of this selection. The second subsample starts

with $r_{2,3}$ and includes the $r_{2,\tau}$ values two at a time with the next two deleted, so that the second subsample is $r_{2,3}, r_{2,4}, r_{2,7}, r_{2,8}, r_{2,11}, r_{2,12}, \ldots$, or approximately $\omega/2 + 1$ or $\omega/2$ values. The tests of significance are made as described in the previous text, but on each of the two subsamples. If both subsamples show the same significant harmonics, they are accepted as such.

Similarly, the test whether $r_{3,\tau}$, for $k = 3$ in Eq. 3.57, is periodic or not can be performed. The ω values of $r_{3,\tau}$ are split into two subsamples. The first subsample starts with $r_{3,1}$, and takes three consecutive values of $r_{k,\tau}$ at a time with the next three deleted, or it is $r_{3,1}, r_{3,2}, r_{3,3}, r_{3,7}, r_{3,8}, r_{3,9}, r_{3,13}, \ldots$. The second subsample starts with $r_{3,4}$ and uses the remaining part of the series, as $r_{3,4}, r_{3,5}, r_{3,6}, r_{3,10}, r_{3,11}, r_{3,12}, r_{3,16}, \ldots$. The test is made on each of the two subsamples, whether or not both have the same significant harmonics.

The above procedure of testing the significance of harmonics in $\rho_{1,\tau}$, $\rho_{2,\tau}$ and $\rho_{3,\tau}$ by the split-sample technique can be generalized to any parameter $\rho_{k,\tau}$. For a given k there are two split samples, the first subsample consisting of k values in sequence, while the next k values are a part of the second subsample, alternating until all ω values are used.

In the case of tests of harmonics in $\rho_{1,\tau}$, $\rho_{2,\tau}, \ldots, \rho_{k,\tau}$ with the two subsamples, the criterion used is that both subsamples should have the same harmonic significant to be accepted as such. This is a somewhat stronger criterion for accepting a harmonic as significant than if it is shown as such only in one subsample. Regardless of this, the approach of both subsamples showing a harmonic to be significant should be used as a stronger test.

3.12 Periodicity In Parameters Which are Functions of Higher Order Moments

The independent components $\xi_{p,\tau}$ may not be the third-or higher-order stationary, though they are the second-order stationary. If $\xi_{p,\tau}$ is normally distributed the second-order stationarity implies also the higher-order stationarity. The $\xi_{p,\tau}$ distribution must be either skewed and/or non-normal symmetrical to have periodicities in the third, fourth or other higher order moments, or in the parameters derived from these moments. Assuming that $\xi_{p,\tau}$ is an independent but non-normal random variable, the skewness and excess coefficients may be periodic.

To have the skewness coefficient of $\xi_{p,\tau}$ vary with τ independently of $E(\xi_{p,\tau})$ and var $\xi_{p,\tau}$, the probability distribution functions with three parameters must be used for $\xi_{p,\tau}$. Assume that $\xi_{p,\tau}$ is an independent random variable, with cov $(\xi_{p,\tau}, \xi_{p,\tau+k}) = 0$ for all $k \neq 0$. If it is the second-order stationary variable, the mean and variance of $\xi_{p,\tau}$ are constants independent of τ.

The two-parameter lognormal distribution with the mean μ_n and the variance σ_n^2 of logarithms of ξ has the skewness coefficient

$$\beta_\xi = \left(\frac{\sigma_\xi}{\mu_\xi}\right)^3 + \frac{3\sigma_\xi}{\mu_\xi} \quad , \tag{3.58}$$

with μ_ξ the mean, and σ_ξ the standard deviation of $\xi_{p,\tau}$, with β_ξ also a constant. Only the three-parameter lognormal distribution enables the mean and the standard deviation to be constants and independent of τ but the skewness coefficient β_ξ to be either a constant or to vary with τ as required by the definition of the third-order non-stationarity in the skewness coefficient. The lower boundary γ of the three-parameter lognormal distribution must change with τ in order that β_ξ changes with τ while μ_ξ and σ_ξ are kept constants for any τ. If μ_ξ in Equation 3.58 is replaced by $(\mu_\xi - \gamma)$, and the periodic parameters are β_ξ and γ designated by β_τ and γ_τ, then Equation 3.58 gives

$$\beta_\tau (\mu_\xi - \gamma_\tau)^3 - 3\sigma_\xi (\mu_\xi - \gamma_\tau)^2 = \sigma_\xi^3 \quad . \tag{3.59}$$

If β_τ is periodic, while the mean μ_ξ and the standard deviation σ_ξ are constants independent of τ, then γ_τ is also periodic. The three-parameter lognormal distribution of $\xi_{p,\tau}$ is then

$$f(\xi) = \frac{1}{(\xi-\gamma_\tau)\sigma_n\sqrt{2\pi}} \; e^{[\ln(\xi-\gamma_\tau)-\mu_n]^2/2\sigma_n^2}, \tag{3.60}$$

in which

$$\mu_n = \frac{1}{2}\frac{(\mu_\xi - \gamma_\tau)^4}{(\mu_\xi - \gamma_\tau)^2 + \sigma_\xi^2} \tag{3.61}$$

and

$$\sigma_n^2 = \ln\left[1 + \frac{\sigma_\xi^2}{(\mu_\xi - \gamma_\tau)^2}\right] \quad . \tag{3.62}$$

Therefore, Equation 3.59 gives γ_τ for any β_τ, and Equations 3.61 and 3.62 enable the computations of μ_n and σ_n of Equation 3.59 for a given value of γ_τ. In other words, if the three-parameter lognormal distribution is used for the distribution of $\xi_{p,\tau}$, it is an independent and second-order stationary random variable, while the third-order non-stationarity in the skewness coefficient β_τ may be accomplished only for the lower boundary γ_τ being a periodic parameter. Note that in this case, μ_n the mean of logarithms of $\xi_{p,\tau}$, and σ_n^2 the variance of logarithms of $\xi_{p,\tau}$, are also periodic, while μ_ξ the mean and σ_ξ^2 the variance of $\xi_{p,\tau}$ are constants independent of τ. If γ_τ is significantly different from a constant γ, then $\xi_{p,\tau}$ has the three-parameter lognormal distribution of Equation 3.60, with periodic parameters μ_n, σ_n, and γ_τ.

If the gamma distribution is used for $\xi_{p,\tau}$, with $E(\xi_{p,\tau}) = \mu_\xi$ and var $\xi_{p,\tau} = \sigma_\xi^2$ as constants independent of τ, then

$$\mu_\xi = \alpha\beta + \gamma \qquad (3.63)$$

and

$$\sigma_\xi^2 = \alpha\beta^2 \quad , \qquad (3.64)$$

in which $\alpha > 0$ is the shape parameter, $\beta > 0$ is the scale parameter, and γ is the lower boundary. From Equations 3.63 and 3.64 then

$$\alpha = \left(\frac{\mu_\xi - \gamma_\tau}{\sigma_\xi}\right)^2 \quad , \qquad (3.65)$$

and the skewness coefficient is

$$g_\tau = \frac{2}{\sqrt{\alpha}} = \frac{2\sigma_\xi}{\mu_\xi - \gamma_\tau} \qquad (3.66)$$

so that

$$\gamma_\tau = \mu_\xi - \frac{2\sigma_\xi}{g_\tau} \quad , \qquad (3.67)$$

in which g_τ is the periodic skewness coefficient. In order to have a gamma distribution of $\xi_{p,\tau}$, with μ_ξ and σ_ξ^2 constants independent of τ, the three-parameter gamma distribution must be used so that the skewness coefficient g_τ may be periodic. This is equivalent of having the lower boundary γ_τ of this distribution periodic. The three-parameter gamma distribution of $\xi_{p,\tau}$ is

$$f(\xi) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{\xi - \gamma_\tau}{\beta}\right)^{\alpha-1} e^{-\left(\frac{\xi-\gamma_\tau}{\beta}\right)} \quad , \qquad (3.68)$$

in which α is given by Equation 3.65, β by Equations 3.64 and 3.65, and γ_τ by Equation 3.67 as a function of g_τ. For g_τ a function of τ, γ_τ is also a function of τ.

As $\xi_{p,\tau}$ in the above two cases of three-parameter distributions is not assumed to be a positively-valued random variable, then the lower boundaries are either positive or negative constants or they are periodic. If β_τ or consequently γ_τ for the lognormal distribution, or g_τ or consequently γ_τ for the gamma distribution, are periodic, then Eq. 3.9 is valid with ν_τ replaced by γ_τ, and ν_x by γ_ξ the means of lower boundaries respectively for the lognormal and gamma three-parameter distributions. Once β_τ and the corresponding γ_τ, or g_τ and the corresponding γ_τ, are inferred to be periodic, and the Fourier coefficients of Eq. 3.9 for γ_τ estimated the distribution of $\xi_{p,\tau}$ may be simply reduced to the stationarity in the skewness coefficient by the transformation $\theta_{p,\tau} = \xi_{p,\tau} - \gamma_\tau$. The algebraic equation which connects the periodic parameters and stochastic component, in the case of the first order autoregressive linear model for $\epsilon_{p,\tau}$ with $\rho_{1,\tau} = \rho_\tau$, becomes

$$x_{p,\tau} = \mu_\tau + \sigma_\tau\left[\rho_\tau \epsilon_{p,\tau-1} + \sqrt{1-\rho_\tau^2}\,(\theta_{p,\tau} - \gamma_\tau)\right] \qquad (3.69)$$

with $\theta_{p,\tau}$ a random variable, with either the two-parameter lognormal or the two-parameter gamma distribution, which is second-order stationary but also stationary in the skewness coefficient.

A further analysis of the third-order stationarity is using the cross-product $C(\xi_{p,\tau}, \xi_{p,\tau+1}, \xi_{p,\tau+2})$, or the similar third-order moments, in the form of

$$C_{\tau,\tau+1,\tau+2} = \frac{1}{n\omega}\sum_{p=1}^{n}\sum_{\tau=1}^{\omega}(\xi_{p,\tau} - \bar{\xi}_{p,\tau})(\xi_{p,\tau+1} - \bar{\xi}_{p,\tau+1})(\xi_{p,\tau+2} - \bar{\xi}_{p,\tau+2}) \quad , \qquad (3.70)$$

and testing whether the ω values of this cross-product are periodic or not. In Equation 3.70 the lags are $k_1 = 1$ and $k_2 = 2$ in the simple form of the third-order cross-product $C_{\tau,\tau+k_1,\tau+k_2}$ $(\xi_{p,\tau}, \xi_{p,\tau+k_1}, \xi_{p,\tau+k_2})$. If the cross-product of Equation 3.70 shows no significant harmonic, it is expected that the covariances of other values of k_1 and k_2 will not be periodic either.

Chapter 4

TESTING PARAMETERS FOR NOT BEING SIGNIFICANTLY DIFFERENT FROM CONSTANTS

The analysis given in the previous chapter was based on the hypothesis that each parameter over the τ positions of the basic period ω of a series is periodic until proven that it is not. Therefore, periodicity is assumed to be nearly always present in the basic parameters. This hypothesis results mainly from the complexity of runoff time series. It may be rightfully claimed that the river flow time series belong to the most complex time series of geophysics, and that fact is the reason for various techniques available at present for the analysis of streamflow time series. The more complex a geophysical process, the more varied are the approaches used and the techniques available for its analysis.

The economy in the number of parameters and coefficients necessary to be estimated in the mathematical description of a hydrologic time process requires another approach, namely, the hypothesis that the variation of some parameters along the τ positions is not significantly different from a constant. This assumes a priori the hypothesis of both the nonperiodicity and the constancy of a parameter. The tests of this type of hypotheses are different than in the previous chapter. They are outlined for some parameters in this chapter.

The proportionality of s_τ and m_τ, or the constant value of the coefficient of variation, is one of these tests. The constant values of $r_{k,\tau}$, the autocorrelation coefficients of the $\epsilon_{p,\tau}$ series, is another. The identically distributed $\xi_{p,\tau}$ variables at all τ positions, with the skewness and excess coefficients being constants independent of τ, is still another type of test, and so on.

4.1 Properties of the Coefficient of Variation Along the Positions of the Basic Period

The general coefficient of variation of the $x_{p,\tau}$ variable is defined as the ratio of its general standard deviation $\sigma(x_{p,\tau})$ and its general mean $\mu(x_{p,\tau})$, estimated by $s(x_{p,\tau})$ and $\bar{x}(x_{p,\tau})$ of the available $x_{p,\tau}$ series. This value of the coefficient of variation is only approximately the ratio σ_x/μ_x with σ_x and μ_x given as the averages of s_τ and m_τ, the periodic standard deviation and mean, respectively. Because of these periodicities in s_τ and m_τ the coefficient of variation at each position τ is estimated by

$$V_\tau = \frac{s_\tau}{m_\tau} \quad , \tag{4.1}$$

with V_τ a function of τ and $\tau = 1, 2, \ldots, \omega$. This new series may not be periodic in many hyrologic time series.

The first approach in testing this hypothesis may be to use the methods described in Chapter 3, or by estimating C_j coefficients of a number of harmonics, say for the six harmonics of monthly values and for up to about first 12 harmonics of daily values. If the tests show that V_τ is not a periodic parameter, an economy in estimated parameters is accomplished. This means that the significant harmonics in σ_τ and μ_τ have the same frequencies and phases, and proportional amplitudes. If V_τ does not show any significant harmonic it does not imply that the V_τ series is independent.

Instead of using the test with the hypothesis of periodicity, two other tests may be used to ascertain whether ω values of V_τ are or are not independent in sequence, and are or are not significantly different--on a prescribed probability level--from a constant $\overline{V}$, as the average of ω values of V_τ. The other alternative is to use V of the entire $x_{p,\tau}$ series given by

$$V = \frac{s(x_{p,\tau})}{\bar{x}(x_{p,\tau})} \quad . \tag{4.2}$$

Usually V and $\overline{V}$ are not equal though they are close values. The value of Equation 4.2 is highly affected by various biases, sampling or otherwise.

To perform the test whether ω values of V_τ are or are not significantly different from a constant, $\overline{V}$, the distribution of ω values of V_τ, each computed for n years of observations, is assumed to be approximately normal with the mean $\overline{V}$ and the standard deviation of V_τ given as s_v, estimated by $\sqrt{s_v^2}$, and s_v^2 as an approximation given by [10, p. 358]

$$s_v^2 = \frac{\mu^2(\mu_4 - \mu_2^2) - 4\mu\mu_2\mu_3 + 4\mu_2^3}{4\mu^4\mu_2 n} \quad , \tag{4.3}$$

in which μ is the mean, and μ_2, μ_3 and μ_4 are the central moments of the variable for the sample size of n years. To estimate μ_2, μ_3 and μ_4, the $x_{p,\tau}$ series is first transformed to a new variable by

$$z_{p,\tau} = \frac{[x_{p,\tau} - (\mu_\tau - \mu_x)]\, s_x}{\sigma_\tau} \qquad (4.4)$$

thus removing the harmonics in m_τ and s_τ, where μ_τ is the fitted equation for the periodic mean m_τ, μ_x is the mean of m_τ and $x_{p,\tau}$, σ_τ is the fitted equation for the periodic standard deviation s_τ, and s_x is the standard deviation of $x_{p,\tau}$. The difference $(\mu_\tau - \mu_x)$ and the ratio s_x/σ_τ are used here to obtain the $z_{p,\tau}$ series without the periodicities in the mean and the standard deviation, with the mean of μ_x and the standard deviation of about s_x. The mean μ_z and the standard deviation s_z may not be exactly equal to μ_x and s_x, because σ_τ is the fitted function to the periodic standard deviation by a selected number of harmonics, which function does not pass exactly through all values of s_τ. However, differences between μ_z and μ_x, and s_z and s_x are expected to be small for the majority of time series analyzed. For the hypothesis of s_τ and m_τ proportional, this should be reflected in the estimated significant harmonics in both μ_τ and σ_τ.

Estimates of $\mu(z_{p,\tau})$ and of the central moments $\mu_2(z_{p,\tau})$, $\mu_3(z_{p,\tau})$, and $\mu_4(z_{p,\tau})$ are obtained from the entire series, $N = n\omega$ values, of the $z_{p,\tau}$ variable. This should give some reliability to the above estimates of second, third and fourth central moments, provided there are no significant sampling biases. The underlying hypothesis is that the $z_{p,\tau}$ series has the same population value of V_τ independent of τ. Then estimates of moments in Equation 4.3 are

$$\mu_i(z_{p,\tau}) = \frac{1}{n\omega} \sum_{p=1}^{n} \sum_{\tau=1}^{\omega} (z_{p,\tau} - \mu_z)^i \qquad (4.5)$$

with i = 2, 3, and 4. The estimate of s_v from Equation 4.3 by using μ_i's of Equation 4.5, together with $V_z = s_z / m_z$, gives the two parameters of the normal distribution of V_τ, $N[V_z, s_v]$.

The χ^2 test may be performed by comparing this theoretical normal distribution with ω values of V_τ computed for the $z_{p,\tau}$ series for τ = 1,2,...,ω over n year. If χ^2 statistic comes out to be smaller than a prescribed value χ_c^2 for a selected probability level, the V_τ values are considered as not being significantly different from a constant. The reason that V_τ of $z_{p,\tau}$ should be tested for not being significantly different from the mean V_z, and not V_τ of $x_{p,\tau}$ from V_x, is the fact that $x_{p,\tau}$ is periodic, while $z_{p,\tau}$ has at least the major periodicities in m_τ and s_τ removed. The random variations $(\mu_\tau - m_\tau)$ and $(\sigma_\tau - s_\tau)$ are preserved in $z_{p,\tau}$ and carried over to the new values $m_\tau(z_{p,\tau})$ and $s_\tau(z_{p,\tau})$. Because $z_{p,\tau}$ may be dependent in sequence, the V_τ values are then also dependent in sequence. The effect of this dependence may be taken into account by computing the effective size ω_e and using it in tests instead of ω, provided the model of dependence in V_τ is simple.

For the $z_{p,\tau}$ distribution close to normal, Equation 4.3 may be approximated by

$$n\, s^2 = \frac{n^2}{2} \left(1 + 2\, \frac{\mu_2}{\mu}\right) , \qquad (4.6)$$

with $\mu = \mu_z$, and both μ_2 and μ_z estimated from $n\omega$ values of $z_{p,\tau}$.

This result of V_τ being not significantly different from a constant implies that m_τ and s_τ are proportional, with the correlation coefficient $\rho(m_\tau, s_\tau)$ estimated from the ω concurrent values of m_τ and s_τ by $r(m_\tau, s_\tau)$, measuring this proportionality. It is then expected that

$$s_\tau = V_z\, m_\tau \quad , \qquad (4.7)$$

or with $E(s_\tau/m_\tau) = V_z$. However, the relation between s_τ and m_τ may be more complex, and if linear, then it is

$$s_\tau = A + V_z\, m_\tau + \phi_\tau \quad , \qquad (4.8)$$

with ϕ_τ the residuals. Equations 4.7 and 4.8 mean that the corresponding Fourier coefficients A_j and B_j of μ_τ and σ_τ of the $x_{p,\tau}$ variable are either approximately proportional, or are simply related. If A is close to zero and ϕ_τ has a small variance, or both are not significantly different from zero, then all coefficients A_j and B_j are proportional. The study of relations between s_τ and m_τ in hydrologic series represents a topic of high interest in the future efforts for a better structural analysis of hydrologic time series.

It can often occur in hydrology that the sequence of ω values of V_τ is not statistically distinguishable from a constant value V . However,

the variance of V_τ along the τ positions may change significantly from season to season. In this case a test, say by the split-sample technique, may be performed to determine whether the variance of V_τ is independent of the τ position. The practical approach would be to divide the ω values of V_τ into 2-4 sections, say be seasons, and test their means or variances for equality. The division by seasons should follow approximately the physical seasonal variations, say the changes in the type and origin of precipitation, snow accumulation and melting, or rain-producing runoff, with transitions between these typical season, or by similar criteria.

4.2 Properties of Autocorrelation Coefficients of the $\epsilon_{p,\tau}$ Series

The tests of the hypothesis that the $r_{k,\tau}$ autocorrelation coefficients along the τ positions are not significantly different from constants, require the distributions of sample serial correlation coefficients, both for the dependent and independent $\epsilon_{p,\tau}$ series. These distributions are available for independent series, as well as for series with simple dependence models [2,4]. The split-sample techniques, as described in the previous chapter, should be used to avoid the spurious correlation. In the case of dependent $\epsilon_{p,\tau}$ series, two approaches are feasible: first, by whitening $\epsilon_{p,\tau}$ and investigating $r_{k,\tau}$ of the inferred independent component $\xi_{p,\tau}$, with the use of sampling distributions of r_k for independent series; and second, by using the available sampling distributions of r_k of dependent $\epsilon_{p,\tau}$ processes in cases these distributions are available. The use of the effective period length ω_e instead of ω in a case of dependent series may be also a simplified approach convenient for the use.

4.3 Properties of Skewness Coefficient of Independent Stochastic and Second-order Stationary Components

The skewness coefficient computed along the τ positions, $\tau = 1, 2, \ldots, \omega$, as a dimensionless parameter, is defined by

$$\beta_\tau = \frac{{}_\tau\mu_3}{({}_\tau\mu_2)^{3/2}} = \frac{{}_\tau\mu_3}{\sigma_\tau^3} \quad , \qquad (4.9)$$

in which ${}_\tau\mu_2$ and ${}_\tau\mu_3$ are the second and the third central moment at each position τ of the independent second-order stationary series $\xi_{p,\tau}$. For small n (the number of years) the unbiased estimates of β_τ are [10,357]

$$\hat{\beta}_\tau = \frac{n^2\ {}_\tau\mu_3}{(n-1)(n-2)\ {}_\tau\hat{\mu}_2^{\,3/2}} \quad , \qquad (4.10)$$

in which ${}_\tau\hat{\mu}_2$ is the unbiased estimate of the second central moment and ${}_\tau\mu_3$ is the biased third moment. Along the positions τ there is a series of ω values of $\hat{\beta}_\tau$, and $\hat{\beta}_\xi$ is the mean of $\hat{\beta}_\tau$.

By using Equation 3.9 with ν_τ and ν_x replaced by $\hat{\beta}_\tau$ and $\hat{\beta}_\xi$, and Equation 3.11 and 3.12 with $(\nu_\tau - \nu_x)$ replaced by $(\hat{\beta}_\tau - \hat{\beta}_\xi)$, the estimates of m pairs of Fourier coefficients (A_j, B_j) are obtained, and the tests for significant harmonics in the hypothesized periodicity may then be performed as described in Chapter 2. For the hypothesis that the $\hat{\beta}_\tau$ is not periodic, for no significant periodicity found in $\hat{\beta}_\tau$, or for the inferred periodic component removed with the remaining ω values being $\theta_\tau = \hat{\beta}_\tau - \beta_\tau$, where β_τ is the fitted periodic function then $\hat{\beta}_\tau$ or the θ_τ residuals may be tested for not being significantly different from a constant. A new variable $u_{p,\tau}$ may be obtained similarly as it was done with $z_{p,\tau}$ of Equation 4.4, so that the differences between the series $\hat{\beta}_\tau$ and the fitted periodic function β_τ carried along into the $u_{p,\tau}$ variable.

For the hypothesis of $\hat{\beta}_\tau$ not being different from a constant and if the variable $\xi_{p,\tau}$ is approximately normal, then

$$\text{var}\ \hat{\beta} = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \approx \frac{6}{n} \quad . \qquad (4.11)$$

If $\xi_{p,\tau}$ is sufficiently skewed and also a dependent variable, the approximation of var $\hat{\beta}$ by Equation 4.11 may contain a substantial error. The average skewness coefficient $\hat{\beta}_\xi$ of $\xi_{p,\tau}$ may be very large for daily river flows, sometimes of the order of 3.00 - 4.00, and $\hat{\beta}_\tau$ may fluctuate in large limits about $\hat{\beta}_\xi$. A good approximation of var $\hat{\beta}$ is given by [10]:

$$4\mu_2^5 n\ \text{var}\ \beta = 4\mu_2^2\mu_6 - 12\mu_2\mu_3\mu_5 - 24\mu_2^3\mu_4 + 9\mu_3^2\mu_4 + 35\mu_2^2\mu_3^2 + 36\mu_2^5 \quad . \qquad (4.12)$$

Equation 4.12 requires the estimation of five central moments, μ_i, with i = 2, 3, 4, 5, and 6. They are estimated from the sample of size $N = n\omega$ of $\xi_{p,\tau}$. In a case of monthly flows $N = 12n$, and for daily flow $N = 365n$, where n is the number of years. For daily series

of n = 50, N = 18, 250, so that even μ_5 and μ_6 may be considered reasonably accurate, provided no biases, sampling or otherwise, are present in extreme values because these biases influence disproportionately the values of high central sample moments.

To test whether the $\hat{\beta}_\tau$ series is significantly different from β_τ, or whether the remaining $\hat{\theta}_\tau$ series, after the periodic β_τ is removed, is significantly different from zero, $\hat{\beta}_\tau$ or $\hat{\theta}_\tau$ are assumed to be normally distributed, $N[\beta_\xi,(\text{var}\hat{\beta})^{1/2}]$, or $N[0,(\text{var}\hat{\theta})^{1/2}]$. The χ^2 test may be used in this case with ω values of $\hat{\beta}_\tau$ or $\hat{\theta}_\tau$.

If $\hat{\beta}_\tau$ or $\hat{\theta}_\tau$ are autocorrelated along ω values, with ρ their first autocorrelation coefficient, then for the test the effective length (in case the first-order autoregressive linear model for this dependence is a good approximation) is

$$\omega_e = \frac{(1-\rho)\omega}{1+\rho} \quad , \tag{4.13}$$

and for the variance this effective length is

$$\omega_e = \frac{(1-\rho^2)\omega}{1+\rho^2} \quad , \tag{4.14}$$

so that in the χ^2 tests ω_e replaces ω.

For the two-parameter lognormal distribution fitted to $\xi_{p,\tau}$, the following analysis may help to make inference about the properties of $\hat{\beta}_\tau$ or $\hat{\theta}_\tau$. For a two-parameter lognormal distribution fitted to $\xi_{p,\tau}$, $\Lambda(\mu_n, \sigma_n)$, where μ_n and σ_n are the mean and the standard deviation of logarithms of $\xi_{p,\tau}$, the coefficient of variation η is given for this distribution by $\eta^2 = e^{\sigma_n^2} - 1$.

The skewness coefficient β is then a function only of η, $\beta = \eta^3 + 3\eta$. (4.15)

The main bias in β may come from the unrepresentative extremes which highly affect the estimates of σ_n.

Moments of the two-parameter lognormal probability function about the origin are

$$m_j = e^{j\mu_n + \frac{1}{2} j^2 \sigma_n^2} \quad . \tag{4.16}$$

Equation 4.12 requires estimates of μ_2, μ_3, μ_4, μ_5, and μ_6. The second central moment is $\mu_2 = \mu^2\eta^2$, with μ the mean of $\xi_{p,\tau}$. By denoting $1 + \eta^2$ as v, then $\mu_2 = \mu^2(v\text{-}1)$. Terms μ_4/μ_2, $\mu_5/\mu_2^{5/2}$, and μ_6/μ_2^3 are obtained by using moments about the origin of Eq. 4.16.

$$\frac{\mu_4}{\mu_2^2} = \frac{1}{(v\text{-}1)^2}(v^6 - 4v^3 + 6v - 3) \quad , \tag{4.17}$$

$$\frac{\mu_5}{\mu_2^{5/2}} = \frac{1}{(v\text{-}1)^5}(v^{10} - 5v^6 + 10v^3 - 10v + 4) \quad , \tag{4.18}$$

and

$$\frac{\mu_6}{\mu_2^3} = \frac{1}{(v\text{-}1)^6}(v^{15} - 6v^{10} + 15v^6 - 20v^3 + 15v - 5) \quad . \tag{4.19}$$

The condition is $|\beta| \leqslant \sqrt{n}$.

Daily flows have β_ξ which are not much greater than 4.00 (in a case of 15 daily flow series only one station had β_ξ = 4.082). The case of β_ξ = 4.00 and n = 40 is shown here as an example for the application of Equations 4.12 and 4.17 through 4.19. In this case η = 1, η^2 = 1, and v = 2, so that $\mu_4/\mu_2^2 = 41$, $\mu_5/\mu_2^{5/2} = 768$, and $\mu_6/\mu_2^3 = 27{,}449$. This gives var β = 493, or the standard deviation s_β = 22.1. It is evident from this computation that either the approximation of Equation 4.12 is very poor in the case β = 4.00, or the two-parameter lognormal function is not applicable for high skewness values, or the variation of β is really large in this case of high skewness of lognormal distributions. If Equation 4.11 is used for var $\hat{\beta}$ instead of Equation 4.12, it gives for n = 40 the value var β = 0.140, or s_β = 0.374. This is a very small value valid only for β_ξ = 0 and the normal variable $\xi_{p,\tau}$.

4.4 Relationships Between the Skewness Coefficient and the Coefficient of Variation

A simple linear correlation analysis may be applied between β_τ (the skewness coefficient) and η_τ (the coefficient of variation) if both come out to be close to constants by various tests. In this case, the ratio

$$\alpha_\tau = \frac{\beta_\tau}{\eta_\tau} \tag{4.20}$$

may be not statistically distinguishable from a constant even if η_τ and β_τ are different from constants. Then $r(\beta_\tau,\eta_\tau)$, the correlation coefficient between β_τ and η_τ, and the simple regression equation $\beta_\tau = \alpha_1 + \alpha_2 \eta_\tau + e_i$ may be used, and the corresponding tests performed that α_1 may be

close to zero, and that e_i has a very small variance, so that Equation 4.20 may be applicable, with α_2 being the mean of α_τ. This approach may also make economy in the number of parameters necessary to estimate in the structural analysis of time series.

4.5 Properties of the Excess Coefficient of Independent Stochastic Component

The excess coefficient along the τ positions, $\tau = 1, 2, \ldots, \omega$, as a dimensionless parameter, is defined by

$$\gamma_\tau = \frac{{}_\tau\mu_4}{{}_\tau\mu_2^2} - 3 \qquad (4.21)$$

in which ${}_\tau\mu_2$ and ${}_\tau\mu_4$ are the second and fourth central moments of $\xi_{p,\tau}$, to be estimated by the corresponding sample central moments.

The unbiased estimate of γ_τ for n years is [10,p. 357]

$$\hat{\gamma}_\tau = \frac{n^2(n+1)\,{}_\tau\mu_4 - 3n^2(n-1)\,{}_\tau\hat{\mu}_2^2}{(n-1)(n-2)(n-3)\,{}_\tau\hat{\mu}_2^2} \quad , \qquad (4.22)$$

where ${}_\tau\mu_4$ is estimated by the biased sample fourth central moment, and ${}_\tau\hat{\mu}_2$ by the unbiased sample central second moment.

For $\xi_{p,\tau}$ normally distributed, the variance of $\hat{\gamma}$ is[10, p. 357]

$$\text{var}\,\hat{\gamma} = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)} \qquad (4.23)$$

However, as $\xi_{p,\tau}$ is often highly skewed (and may be also autocorrelated), then [10, p. 387]

$$n\mu_2^6\,\text{var}\,\hat{\gamma} = \mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 - 8\mu_2^2\mu_3\mu_5 + 4\mu_4^3 - \mu_2^2\mu_4^2 + 16\mu_2\mu_3^2\mu_4 + 16\mu_2^3\mu_3^2 \qquad (4.24)$$

Central moments μ_i, i = 2, 3, 4, 5, 6, and 8 must be estimated from the sample. If daily flows are used, N = nω = 365n, so that there still may be feasible to estimate even the eighth moment from the data, provided the original data is not very biased at its extremes. Then $\hat{\gamma}_\tau$ may be tested by the χ^2 statistic, with $\hat{\gamma}_\tau$ normally distributed, $N[\gamma_\xi, (\text{var}\,\hat{\gamma})^{1/2}]$, whether they are or are not significantly different from the mean excess coefficient, γ_ξ.

The analysis of whether the skewness and kurtosis or excess coefficients are or are not different from a constant, if the periodicity ω is present in the mean and standard deviation, may be based on a regional approach for some variables. By properly using all $\xi_{p,\tau}$ series of a region, a large amount of data may permit the reliable estimates of most of moments in equations for the variances of estimates of these coefficients.

Chapter 5

DEPENDENCE MODELS OF STOCHASTIC COMPONENTS

Many hydrologic variables have stochastic components which are dependent time series. The analysis of this dependence, with the proper inference about dependence models is the subject of this chapter. Though the autoregressive coefficients may be periodic, or even the higher-order parameters may be periodic, this chapter is related to the dependence models of stochastic part after the periodicities in the mean and standard deviation are removed.

5.1 Investigation of Dependence Models

Previous studies [3,4,5] have shown that the variable $\epsilon_{p,\tau}$, obtained by removing the periodicity in the mean and standard deviation, is only approximately a second-order stationary dependent or independent time series. The dependence can be often approximated by the first-, second-, third-, or higher-order autoregressive linear models. Higher-order models beyond the third show a significant advantage in comparison with the first three models only when the series are sufficiently long. Physical explanations exist for the use of autoregressive models in hydrology [2], though other ideas exist on this topic among hydrologists. Short hydrologic series rarely justify an investigation of higher-order autoregressive linear models though they may be indicated by physical processes. Linear models seem sufficiently accurate for practical purposes, though the real physical stochastic models may be nonlinear.

The general m-th order autoregressive linear model is

$$\epsilon_{p,\tau} = \sum_{j=1}^{m} \alpha_{j,\tau}\, \epsilon_{p,\tau-j} + \sigma\, \xi_{p,\tau} \quad , \qquad (5.1)$$

with $\alpha_{j,\tau}$ the autoregressive coefficients, either periodic as $\alpha_{j,\tau}$ or nonperiodic as constants α_j, and σ is a standard deviation, periodic or nonperiodic, which enables $\xi_{p,\tau}$ to be a second-order stationary and standard (0,1) random independent variable if $\epsilon_{p,\tau}$ is a standard random but dependent variable.

The value of σ is

$$\sigma = \left[1 - \sum_{j=1}^{m} \alpha_{j,\tau}^2 - 2 \sum_{i>k} \alpha_{i,\tau}\, \alpha_{k,\tau}\, \rho_{i-k} \right]^{-1/2} . \qquad (5.2)$$

If $\epsilon_{p,\tau}$ is a standard variable (0,1), then σ is periodic as soon as any value $\alpha_{j,\tau}$, i = 1,2,...,m, is periodic.

Parameters of the linear dependence models are either periodic or nonperiodic. These two alternatives are discussed as two separate cases.

5.2 Case of Nonperiodic Autocorrelation Coefficients

If tests show that $r_{k,\tau}$ is not significantly different from a constant independent of τ for a given k, equivalent to tests rejecting the hypothesis of periodicity in the autocorrelation coefficients, two alternatives in estimating the autocorrelation coefficients in the selected models and in computing the $\xi_{p,\tau}$ series may come into consideration: (a) the use of ρ_1, ρ_2, and ρ_3 (or ρ_k in general) as the means of ω values of $r_{1,\tau}$, $r_{2,\tau}$ and $r_{3,\tau}$ (or of $r_{k,\tau}$ in general); and (b) the use of the values $\rho_{1,\epsilon}$, $\rho_{2,\epsilon}$, and $\rho_{3,\epsilon}$ (or $\rho_{k,\epsilon}$ in general) of the overall correlogram of $\epsilon_{p,\tau}$, estimated by $r_{1,\epsilon}$, $r_{2,\epsilon}$, and $r_{3,\epsilon}$ (or $r_{k,\epsilon}$ in general) from the data of the available sample of the $\epsilon_{p,\tau}$ variable. This latter case neglects the positions $\tau = 1,2,...,\omega$ inside the year, and $\epsilon_{p,\tau}$ is treated as a stationary series, ϵ_i, i = 1,2,...,N, with $N = n\omega$.

Computations show that the means of ω values of $r_{k,\tau}$, designated here as $\rho_1, \rho_2, \rho_3, \ldots, \rho_k$, are often significantly greater or smaller than the corresponding values $r_{1,\epsilon}, r_{2,\epsilon}, r_{3,\epsilon}, \ldots, r_{k,\epsilon}$ of the overall $\epsilon_{p,\tau}$ correlogram. This difference may be due to one or both of the following two factors, namely that $r_{1,\epsilon}$ underestimates ρ, more or less than ρ_1 does, or that the nonstationarity and sampling biases in $\epsilon_{p,\tau}$ make ρ_k, as the mean of ω values of $r_{k,\tau}$, greater or smaller than $r_{k,\epsilon}$. Whether one or the other alternative is used for the dependence model and for the computation of $\xi_{p,\tau}$ from $\epsilon_{p,\tau}$ depends on the character of the $\epsilon_{p,\tau}$ series. If $\rho_1, \rho_2, \rho_3, \ldots, \rho_k$ sequence of the mean autocorrelation coefficients is used, then the ρ_k values of the computed autocorrelation coefficient of the independent $\xi_{p,\tau}$ variable should oscillate around the expected values $E(\rho_{k,\tau}) = 0$ for $k > 0$. In that case, the values $r_{k,\xi}$ of the

general correlogram of $\xi_{p,\tau}$ or the values v_f of the variance density spectrum of the computed $\xi_{p,\tau}$ may show some deviations from $E(r_k) = 0$ for $k > 0$, or from $E(v_f) = 2$, for $0 \leqslant f \leqslant 0.50$, for the correlogram and the spectrum of an independent and standardized time series, respectively.

It can be shown that a stationary series $\epsilon_{p,\tau}$ would produce a stationary $\xi_{p,\tau}$ if the proper dependence model for $\epsilon_{p,\tau}$ is used, and that $E(\rho_k) = E(r_{k,\xi})$ = the population value of that correlation coefficient, with ρ_k the mean of ω values of $r_{k,\tau}$ of the $\xi_{p,\tau}$ series. Therefore, biases in estimates and biases in the series (sampling or otherwise) are factors which make differences between ρ_k and $r_{k,\xi}$, besides the effects of the basic remaining nonstationarity in the derived $\xi_{p,\tau}$ series. Besides, the skewness of $\xi_{p,\tau}$ as well as nonstationarity of the higher-order moments or parameters may account for part of the differences between ρ_k and $r_{k,\epsilon}$.

5.3 Selection of Mathematical Dependence Model of Stochastic Components for Constant Autocorrelation Coefficients

The technique of statistical tests for fitting the autoregressive linear models is given for large samples by Quenouille [11]. This technique is a laborious method of computing two sets of constants and a test parameter, which require more computer time than the simplified method proposed in this text. As the structural analysis and the sample size limit the order of linear models, or presume that data available do not justify the use of the higher-order models, a simplified, practical method is considered here as a feasible approach.

Another approach is by whitening the series or by assuming a model of the autoregressive linear type, by estimating its parameters and by computing the presumed independent $\xi_{p,\tau}$ component. Then $\xi_{p,\tau}$ is tested for independence. If this hypothesis is accepted, the hypothesis of the model fitting well the time dependence is also accepted. This approach does not compare the various models and it requires large computations. The following simplified method removes these two shortcomings of the "whitening series" approach.

The measure of the goodness of fit of the autoregressive linear models by this simplified method is the determination coefficient, R_i^2, $i = 1,2,3,...$ It tells what portion of the total variation of $\epsilon_{p,\tau}$ is explained by each term of the autoregressive equations, the remaining portion of the variance of $\epsilon_{p,\tau}$ being explained by the term $\sigma\xi_{p,\tau}$. Because $R_m^2 >...> R_3^2 > R_2^2 > R_1^2$, a criterion can be developed when a model of a given order should be selected in comparison with the other models. For the purpose of this study and if only the first three models are studied, with R_1^2, R_2^2 and R_3^2, the third model is accepted if $R_3^2 - R_2^2 > 0.01$, or 1 percent of additional value in the portion of the explained variation, when the third model is tested against the second model. The second model is accepted if $R_2^2 - R_1^2 \geqslant 0.01$, for the second model tested against the first model. Whichever small value ΔR_i^2 is used, for $i = 1,2$, and 3, say 0.01 or 0.02 or a similar small difference, it does not make a significant impact on the final results. This approach simplifies the selection of the model without too much loss of information. In this text $\Delta R_i^2 = 0.01$ is selected. Similar criterion can be used when models of a greater order than the third are used.

Whether ρ_k, as the mean of $r_{k,\tau}$, or $r_{k,\epsilon}$, as the r_k-th value of the entire $\epsilon_{p,\tau}$ series, is used, the determination coefficients of the first three order autoregressive linear models are computed by

$$D_1 = R_1^2 = \rho_1^2 \quad , \tag{5.3}$$

$$D_2 = R_2^2 = \frac{\rho_1^2 + \rho_2^2 - 2\rho_1^2\rho_2}{1 - \rho_1^2} \quad , \tag{5.4}$$

and

$$D_3 = R_3^2 = \frac{\rho_1^2+\rho_2^2+\rho_3^2+2\rho_1^3\rho_3+2\rho_1^2\rho_2^2+2\rho_1\rho_2^2\rho_3 - 2\rho_1^2\rho_2-4\rho_1\rho_2\rho_3 -\rho_1^4 - \rho_2^4 - \rho_1^2\rho_3^2}{1 - 2\rho_1^2 - \rho_2^2 + 2\rho_1^2\rho_2} \tag{5.5}$$

in which ρ_1, ρ_2 and ρ_3 are either the means $\bar{r}_{k,\tau}$ of ω values of $r_{k,\tau}$, for $k = 1,2$, and 3, and $\tau = 1,2,...,\omega$, or are estimated by the first three values $r_{1,\epsilon}$, $r_{2,\epsilon}$ and $r_{3,\epsilon}$ of the general correlogram.

If $R_2^2 - R_1^2 \leqslant 0.01$ and $R_3^2 - R_1^2 \leqslant 0.02$, the first-order model is selected. If $R_2^2 - R_1^2 > 0.01$ but $R_3^2 - R_2^2 \leqslant 0.01$, the second-order model is selected. If $R_2^2 - R_1^2 > 0.01$ and $R_3^2 - R_2^2 > 0.01$, the third-order model is selected.

For the linear autoregressive models with $m > 3$ in Equation 5.1, the α_j constant coefficients may be estimated by using the multiple linear regression estimation techniques with the remaining "error" term $\sigma\xi_{p,\tau}$. Because var $\epsilon_{p,\tau} = 1$, the difference of $1 - \sigma^2$ gives the explained variances, D_i, $i = 4, 5, \ldots, m$, by the autoregressive terms. Then a criterion of $D_i - D_{i-1} \geq e$ with $e = 0.01$ or a similar small number, may be used to determine the m-th order model of the best fit, if the F - test is unfeasible for finding the order of the autoregressive linear model.

5.4 Estimates of Nonperiodic Autoregressive Coefficients and Computation of Independent Stochastic Series, $\xi_{p,\tau}$

If the first-order model is selected, the estimate of the autoregressive coefficient is either $r_1 = \rho_1 = \bar{r}_{1,\tau} = a_1$, or $r_1 = r_{1,\epsilon} = a_1$, whichever of the two approaches is selected. The new series of $n\omega$-values of the standardized $\xi_{p,\tau}$ series is computed from Equation 5.1 for $m = 1$ by

$$\xi_{p,\tau} = \frac{\epsilon_{p,\tau} - a_1 \epsilon_{p,\tau-1}}{\sqrt{1-a_1^2}} \quad , \qquad (5.6)$$

with $a_1 = \rho_1$ or $a_1 = r_{1,\epsilon}$.

If the second-order model is selected, the two autoregressive coefficients, α_1 and α_2 of Equation 5.1, are estimated by

$$a_1 = \frac{r_1 - r_1 r_2}{1 - r_1^2} \quad \text{and} \quad a_2 = \frac{r_2 - r_1^2}{1 - r_1^2} \quad , \qquad (5.7)$$

with r_1 and r_2 replaced either by ρ_1 and ρ_2, or by $r_{1,\epsilon}$ and $r_{2,\epsilon}$. In that case, the new series of $n\omega$-values of $\xi_{p,\tau}$ is computed from Equation 5.1 for $m = 2$ by

$$\xi_{p,\tau} = \frac{\epsilon_{p,\tau} - a_1 \epsilon_{p,\tau-1} - a_2 \epsilon_{p,\tau-2}}{\sqrt{1 - (a_1^2 + a_2^2 + 2a_1 a_2 r_1)}} \quad , \qquad (5.8)$$

with a_1 and a_2 obtained by Equation 5.7, and r_1 being replaced either by ρ_1 or by $r_{1,\epsilon}$ of the general correlogram of $\epsilon_{p,\tau}$.

If the third-order model is selected, the three autoregressive coefficients, α_1, α_2, and α_3 of Equation 5.1, are estimated by

$$a_1 = \frac{(1-r_1^2)(r_1 - r_3) - (1-r_2)(r_1 r_2 - r_3)}{(1-r_2)(1-2r_1^2 + r_2)} \quad ,$$

$$a_2 = \frac{(1-r_2)(r_2 + r_2^2 - r_1^2 - r_1 r_3)}{(1-r_2)(1-2r_1^2 + r_2)} \quad ,$$

and

$$a_3 = \frac{(r_1 - r_3)(r_1^2 - r_2) - (1-r_2)(r_1 r_2 - r_3)}{(1-r_2)(1-2r_1^2 + r_2)} \quad , \qquad (5.9)$$

with r_1, r_2, and r_3 replaced either by ρ_1, ρ_2, and, ρ_3, or by $r_{1,\epsilon}$, $r_{1,\epsilon}$, and $r_{3,\epsilon}$, whichever approach is used. The new series of $n\omega$ values of $\xi_{p,\tau}$ is computed from Equation 5.1 for $m = 3$ by

$$\xi_{p,\tau} = \frac{\epsilon_{p,\tau} - a_1 \epsilon_{p,\tau-1} - a_2 \epsilon_{p,\tau-2} - a_3 \epsilon_{p,\tau-3}}{\sqrt{1 - (a_1^2 + a_2^2 + a_3^2 + 2a_1 a_2 r_1 + 2a_1 a_3 r_2 + 2a_2 a_3 r_1)}} \qquad (5.10)$$

with a_1, a_2, and a_3, computed by Equation 5.9, and r_1 and r_2 being replaced either by ρ_1 and ρ_2, or by $r_{1,\epsilon}$ and $r_{2,\epsilon}$.

In the case of fitting an autoregressive model with $m > 3$ and by using the multiple linear regression analysis in estimating α_j coefficients, the residual

$$\epsilon_{p,\tau} - \sum_{j=1}^{m} \alpha_j \epsilon_{p,\tau-j}$$

give the values $\sigma\xi_{p,\tau}$, from which for var $\xi_{p,\tau} = 1$ the parameter σ is estimated. Equation 5.1 permits the computation of the $\xi_{p,\tau}$ series.

Once the $\xi_{p,\tau}$ series is obtained, it is advisable to compute either its general correlogram, $r_k(\xi_{p,\tau})$, for $k = 1,2,\ldots,n\omega/10$, and its variance density spectrum, $v_f(\xi_{p,\tau})$, for $0 \leq f \leq 0.50$. The ω values of $r_{k,\tau}(\xi_{p,\tau})$ for $k = 1,2,\ldots,m$ should also be computed, with their means $\rho_1, \rho_2, \ldots, \rho_m$.

For this latter case of ω values of $r_{1,\tau}, r_{2,\tau}, \ldots r_{m,\tau}$, the tests should be performed whether or not they significantly depart from the corresponding ω values of an independent time series. Similarly, the tests should be performed for $r_k(\xi_{p,\tau})$ and $v_f(\xi_{p,\tau})$ whether or not they significantly depart from the correlogram or the spectrum of an independent time series, respectively. Because of various biases in the original series, it is not expected that both approaches, by using $\rho_1, \rho_2, \ldots, \rho_m$, as the means of, $r_{1,\tau}, r_{2,\tau}, \ldots, r_{m,\tau}$, in the first case, or by using, $r_{1,\xi}, r_{2,\xi}, \ldots$, in the second case, would always comply with the results of $\xi_{p,\tau}$ being an independent time series.

5.5 Case of Periodic Autocorrelation Coefficients

The first autocorrelation coefficient of $\epsilon_{p,\tau}$ usually has the greatest influence in autoregressive linear dependence models of hydrology. It mostely affects the variance of $\epsilon_{p,\tau}$ which is explained by all terms of the dependence model, and the determination coefficient R_1^2 of the first-order model or of the first term of other models of Equation 5.1 is usually large. The criterion, whither or not the dependence parameters are periodic, should be basically decided whether the estimated ω values of $r_{1,\tau}$ are periodic or not. This represents a simple criterion whether the periodicity is or is not present in the autoregressive coefficients of the $\epsilon_{p,\tau}$ series.

It may come out that $r_{1,\tau}$ is periodic while neither $r_{2,\tau}$, $r_{3,\tau}$,..., are periodic. The decision on periodicity in the autocorrelation coefficients should be made by finding the significant harmonics in $r_{1,\tau}$, and not in $r_{2,\tau}$, $r_{3,\tau}$..., if they are periodic. If any of $r_{2,\tau}$, $r_{3,\tau}$, is nonperiodic, constants ρ_2, ρ_3..., as the means of $r_{2,\tau}$, $r_{3,\tau}$,..., respectively may be used in these models, though their population values may be periodic. If $r_{1,\tau}$ is nonperiodic, the mean ρ_1 of $r_{1,\tau}$ is used together with constants ρ_2, ρ_3,...,regardless of whether $r_{2,\tau}$, $r_{3,\tau}$,..., are periodic or not. This approach is a simplification of statistical tests in selecting the first-, second-, third-or higher-order linear autoregressive model to fit the time dependence patterns in $\epsilon_{p,\tau}$.

It is difficult to visualize some physical hydrologic conditions which would make $r_{2,\tau}$, $r_{3,\tau}$,..., periodic but $r_{1,\tau}$ equal to a constant ρ_1. Usually, if $r_{1,\tau}$ is periodic all other serial correlation coefficients should be periodic, and if $r_{1,\tau}$ is not periodic the other coefficients also should not be. When the opposite results are produced by the testing method for $r_{1,\tau}$ and other coefficients, the likelihood is high that this is a product either of the sampling errors or because of the approximate estimation and testing methods applied.

The question arises how to make practical tests to determine which model should be selected for a particular $\epsilon_{p,\tau}$ variable. It is sufficiently accurate to compute the means ρ_1, ρ_2, ρ_3,..., of ω values of $r_{1,\tau}$, $r_{2,\tau}$, $r_{3,\tau}$,..., and select the model by the simple procedure given in the previous text (computing the determination coefficient of each regression model, $D_i = R_i^2$, i = 1,2,3,..., and selecting i by the criterion established, ΔR_i^2). Periodic functions of Equation 3.9 are applied to $r_{1,\tau}$, $r_{2,\tau}$, $r_{3,\tau}$...,if they are found also to be periodic, or the constants, ρ_2, ρ_3,..., are used if $r_{2,\tau}$, $r_{3,\tau}$..., are not found periodic.

5.6 Estimates of Periodic Autoregressive Coefficients and Computation of Independent Stochastic Series $\xi_{p,\tau}$

The estimates of periodic autocorrelation coefficients are made by using Equation 3.9, and replacing v_τ by $r_{k,\tau}$, with k = 1,2,3,... . Then the periodic autoregressive coefficients $\alpha_{i,\tau}$ are estimated by the sample values, $a_{i,\tau}$, in the following way.

For m = 1 in Equation 5.1, the first-order model,

$$\epsilon_{p,\tau} = \alpha_{1,\tau}\,\epsilon_{p,\tau-1} + \sigma\xi_{p,\tau} \quad ,$$

if multiplied by $\epsilon_{p,\tau-1}$ gives for var ϵ = 1 and cov $\xi_{p,\tau}\epsilon_{p,\tau-1}$ = 0,

$$\rho_{1,\tau-1} = \alpha_{1,\tau} \quad , \qquad (5.11)$$

with $\rho_{p,\tau-1}$ the autocorrelation coefficient between $\epsilon_{p,\tau-1}$ and $\epsilon_{p,\tau}$, and $\alpha_{1,\tau}$ estimated by $a_{1,\tau} = r_{1,\tau-1}$.

For m = 2 in Equation 5.1, the second-order model,

$$\epsilon_{p,\tau} = \alpha_{1,\tau}\,\epsilon_{p,\tau-1} + \alpha_{2,\tau}\,\epsilon_{p,\tau-2} + \sigma\,\xi_{p,\tau}$$

if multiplied first by $\epsilon_{p,\tau-1}$ and secondly by $\epsilon_{p,\tau-2}$ produces the following relations of $\alpha_{1,\tau}$ and $\alpha_{2,\tau}$ to the $\rho_{k,\tau-j}$ autocorrelation coefficients

$$\alpha_{1,\tau} = \frac{\rho_{1,\tau-1} - \rho_{1,\tau-2}\,\rho_{2,\tau-2}}{1 - \rho_{1,\tau-2}^2} \quad , \qquad (5.12)$$

and

$$\alpha_{2,\tau} = \frac{\rho_{2,\tau-2} - \rho_{1,\tau-1}\,\rho_{1,\tau-2}}{1 - \rho_{1,\tau-2}^2} \quad , \qquad (5.13)$$

with $\rho_{1,\tau-1}$ the symbol for ρ_1 between $\epsilon_{p,\tau}$ and $\epsilon_{p,\tau-1}$, $\rho_{1,\tau-2}$ the symbol for ρ_1 between $\epsilon_{p,\tau-1}$ and $\epsilon_{p,\tau-2}$, and $\rho_{2,\tau-2}$ the symbol for ρ_2 between $\epsilon_{p,\tau}$, and $\epsilon_{p,\tau-2}$.

Similarly, for m = 3 in Equation 5.1, the third-order linear model, the relations between, $\alpha_{i,\tau}$ and $\rho_{k,\tau-j}$ are obtained in multiplying

$$\epsilon_{p,\tau} = \alpha_{1,\tau}\,\epsilon_{p,\tau-1} + \alpha_{2,\tau}\,\epsilon_{p,\tau-2} + \alpha_{3,\tau}\,\epsilon_{p,\tau-3} + \sigma\xi_{p,\tau}$$

by $\epsilon_{p,\tau-1}$, $\epsilon_{p,\tau-2}$, and $\epsilon_{p,\tau-3}$, one after another, and obtaining the following functions:

$$A\,\alpha_{1,\tau} = \rho_{1,\tau-2}(1-\rho^2_{1,\tau-3}) + \rho_{1,\tau-3}\,\rho_{1,\tau-2}\,\rho_{3,\tau-3} - \rho_{1,\tau-2}\,\rho_{2,\tau-2} - \rho_{2,\tau-3}\,\rho_{3,\tau-3} + \rho_{1,\tau-3}\,\rho_{2,\tau-2}\,\rho_{2,\tau-3}, \quad (5.14)$$

$$A\,\alpha_{2,\tau} = \rho_{2,\tau-2}(1-\rho^2_{2,\tau-3}) + \rho_{1,\tau-2}\,\rho_{2,\tau-3}\,\rho_{3,\tau-3} - \rho_{1,\tau-2}\,\rho_{1,\tau-1} - \rho_{1\,\tau-3}\,\rho_{3,\tau-3} + \rho_{1,\tau-3}\,\rho_{2,\tau-3}\,\rho_{1,\tau-1}, \quad (5.15)$$

and

$$A\,\alpha_{3,\tau} = \rho_{3,\tau-3}(1-\rho^2_{1,\tau-2}) + \rho_{1,\tau-3}\rho_{1,\tau-2}\,\rho_{1,\tau-1} - \rho_{1,\tau-3}\cdot\rho_{2,\tau-2} - \rho_{2,\tau-3}\,\rho_{1,\tau-1} + \rho_{1,\tau-2}\,\rho_{2,\tau-2}\,\rho_{2,\tau-3}, \quad (5.16)$$

in which

$$A = 1 + 2\rho_{1,\tau-2}\,\rho_{2,\tau-3}\,\rho_{1,\tau-3} - \rho^2_{1,\tau-3} - \rho^2_{1,\tau-2} - \rho^2_{2,\tau-3}, \quad (5.17)$$

with $\alpha_{j,\tau}$ estimated by $a_{j,\tau}$, and $\rho_{k,\tau-j}$ by $r_{k,\tau-j}$. For nonperiodic autocorrelation coefficients, Eqs. 5.11 through 5.16 reduce to Eqs. 5.7 and 5.9.

To compute $\xi_{p,\tau}$ from Equation 5.1 and the periodic values of $\alpha_{j,\tau}$ of Equations 5.11 through 5.16, σ of Equation 5.1 must be obtained as a function of $\alpha_{j,\tau}$ and $\rho_{k,\tau-j}$ coefficients. The values of σ are:

for m = 1

$$\sigma = (1-\alpha^2_{1,\tau})^{1/2}, \quad (5.18)$$

for m = 2

$$\sigma = (1 - \alpha^2_{1,\tau} - \alpha^2_{2,\tau} - 2\alpha_{1,\tau}\,\alpha_{2,\tau}\,\rho_{1,\tau-2})^{1/2}, \quad (5.19)$$

and for m = 3

$$\sigma = (1-\alpha^2_{1,\tau} - \alpha^2_{2,\tau} - \alpha^2_{3,\tau} - 2\alpha_{1,\tau}\alpha_{2,\tau}\rho_{1,\tau-2} - 2\alpha_{1,\tau}\alpha_{3,\tau}\rho_{2,\tau-3} - 2\alpha_{2,\tau}\alpha_{3,\tau}\rho_{1,\tau-3})^{1/2}. \quad (5.20)$$

In the case ω is large, say ω = 365 for daily series, the differences between $\rho_{1,\tau-1}$, $\rho_{1,\tau-2}$, $\rho_{1,\tau-3}$, or $\rho_{2,\tau-2}$ and $\rho_{2,\tau-3}$, and $\rho_{3,\tau-3}$ from $\rho_{1,\tau}$, $\rho_{2,\tau}$, and $\rho_{3,\tau}$ are small, so that for large ω and small k and j values (say k = 1, 2, and 3 and j = 1, 2, and 3), Equations 5.7 and 5.9 may replace sufficiently accurately Equations 5.12 through 5.16. Similarly, Equations 5.18, 5.19, and 5.20 may be replaced by the denominators of Equations 5.6, 5.8, and 5.10, respectively, and using only the values of $\rho_{1,\tau}$, $\rho_{2,\tau}$, and $\rho_{3,\tau}$ as fitted periodic functions.

5.7 Some Current Misinterpretations of Autoregressive Models in Hydrology

Several misinterpretations are related to autoregressive models in their applications to hydrology. First, when the $\epsilon_{p,\tau}$ values are expressed by a linear relation to the m previous values, it is often stated that the "memory of the process is only for m terms". This is not correct, because this relation to m previous values is the method by which a process with an infinite memory is expressed by a finite number of terms. The fact that the dependence of the m-th order in autoregressive chains or models may be expressed either by a m-th order matrix of transitional probabilities or by a mathematical dependence model of the variable, being a function of the m previous terms, respectively, does not mean that the memory is only for m terms.

The next misinterpretation, and sometimes a surprise to those who oversimplify the concept of autoregressive models, is the fact that the autoregressive processes are equivalent to the moving average schemes with an infinite number of terms of ξ_i and its β_j coefficients, with the definite relations of this infinite number of β_j coefficients to the finite number of α_j coefficients of an m-th order autoregressive scheme. The recurrence equations, replacing the ϵ_{i-j} term by its expression, j = 1, 2,..., leads to a new dependence

$$\epsilon_i = \beta_o\,\xi_i + \beta_1\,\xi_{i-1} + \ldots + \beta_j\,\xi_{i-j} + \ldots, \quad (5.21)$$

which is a moving average type model with an infinite number of terms, and, therefore, memory.

A third current misinterpretation is that autoregressive processes must necessarily be linear models. It is quite likely that the hydrologic physical reality imposes nonlinear models. Therefore, the general autoregressive discrete models may be $\epsilon_i = f(\epsilon_{i-1},\ldots,\epsilon_{i-m}) + \xi_i$ with m being a sufficient length of previous values that no information is necessary for the values previous to ϵ_{i-m} in order to use the autoregressive models, and ξ_i is an independent random variable with a given variance. Though it may be difficult mathematically to design the nonlinear models which preserve the stationarity

of the process, it does not mean that there are not in nature some nonlinear models which preserve the general character of stationarity of $\epsilon_{p,\tau}$ series. Before one rejects the autoregressive models, they should try the nonlinear functions if the physical conditions are such that they require the nonlinear models, and the sample sizes are such to justify or enable the estimates of parameters in these models.

A fourth misinterpretation may be found in the method by which the autoregressive coefficients are estimated. The use of m values of r_k, k = 1, 2,...,m, to estimate the m values of α_j coefficients may produce sufficiently accurate estimates, through r_k underestimates ρ_k. This bias in estimating α_j coefficients may be significant for small samples and/or large ρ_k. In this case, the corrections for the bias in this estimation may be justified and should be applied.

A fifth misinterpretation may be the current attitude that it is sufficient and efficient to estimate the m autoregressive coefficients by the first m autocorrelation coefficients. Though r_1 may be sufficient estimate for ρ in the first-order model, the fit of the correlogram $\rho_k = \rho^k$ to a large number of r_k values, k = 1, 2,..., n, may show significant differences $\rho^k - r_k$, though the population model is of the first-order linear model. A correction of r_1, which is a kind of correction for the bias, may be made by fitting the ρ^k function to the r_k function by some weighting procedure for r_k's, because their accuracy decreases with an increase of k.

A sixth misinterpretation is the application of autoregressive models to hydrologic time series which are evidently either subject to systematic errors or have the man-made nonhomogeneity in a significant manner. It can be shown that the linear autoregressive models are not applicable to series with added trends or jumps, if they are applicable to stationary hydrologic stochastic processes.

5.8 Bias in Estimated Serial Correlation Coefficients

The bias in the estimates of serial correlation coefficients occupied the interests of statisticians during the decade of 1950's. Kendal and Steward [12, p. 431-435] summarized the results of various investigations. The biases in underestimating the serial correlation coefficients of stationary processes are of interest in this study, particularly for independent and autoregressive processes, these latter processes represented by the first-order linear model. The underestimate of $E(r_k)$ for $\rho_k = 0$, $k > 0$ for the independent processes is of the order of 1/(N-k), with N the sample size. For the first-order autoregressive linear model, Kendal [12, p. 435] gives

$$E(r_1) = \rho - \frac{1+3\rho}{N-1} \quad , \tag{5.22}$$

with the bias of $(1 + 3\rho)/(N-1)$, and

$$E(r_k) = \rho^k - \frac{1}{N-k}\left[\frac{1+\rho}{1-\rho}(1-\rho^k) + 2k\rho^k\right] , \tag{5.23}$$

for $k > 1$. For $\rho = 0.50$ in this model with N = 25, the mean value of r_1 would be about 0.40 instead 0.50. This is a serious bias not often recognized by those who reject the application of autoregressive models without first trying to apply them properly.

Quenouille [12, p. 435] proposed the estimate of ρ by

$$R = 2r - \frac{1}{2}\left[r_{(1)} + r_{(2)}\right] \quad , \tag{5.24}$$

in which r is the first serial correlation coefficient of the entire series (N), and $r_{(1)}$ is the value of the first half (N/2) of the series and (r_2) of the second half (N/2). Because r with N has a smaller negative bias than $r_{(1)}$ and $r_{(2)}$ with N/2, Equation 5.24 adds a correction for the bias, so that R is now in error of about N^{-1}. Equation 5.24 may be used as an approximation for any r_k in autoregressive linear models.

Several other methods are available for the corrections of bias in estimated serial correlation coefficients, with some of them decreasing the bias to the order of N^{-3}.

5.9 Estimate of ρ of the First-Order Linear Model as an Indirect Correction for the Bias

The population correlogram of the first-order linear model,

$$\epsilon_i = \rho\,\epsilon_{i-1} + \sqrt{1-\rho^2}\,\xi_i \quad , \tag{5.25}$$

is

$$\rho_k = \rho^k \quad , \tag{5.26}$$

with $\rho = \rho_1$. Usually in practice, the parameter ρ in Equations 5.25 and 5.26 is estimated by the sample first serial correlation coefficient, r_1. Using this estimate often leads to the conclusion that the model of Equation 5.23 is not applicable to a series,

though the correlogram of an observed series, $r_k = f(k)$, may be well fitted by a power function of Equation 5.26, provided a different estimate of ρ is used, considering the difference as a bias in the estimation.

Lets assume that δ is the bias in r_1 so that $E(r_1 + \delta) = \rho$

$$E(r_1 + \delta)^k = \rho_k \qquad (5.27)$$

By considering δ as a small quantity in comparison with r_1, and by neglecting all terms in Equation 5.27 of δ^j with $j > 1$, then Equation 5.27 as an approximation, becomes

$$E[r_1^{k-1}(r_1 + k\delta)] \approx \rho_k \quad . \qquad (5.28)$$

With an increase of k the term r^{k-1} decreases rapidly if r_1 is not very close to unity, while $k\delta$ increase linearly. Therefore, the ρk of the model of Equation 5.28 has to depart more and more from r_k as k increases, if the bias δ in r_1 is not negligible. The relative bias

$$\frac{k\delta}{r_1} = f(k) \qquad (5.29)$$

increases rapidly with k, while the sampling errors in r_k may be much smaller.

When the Wiener-Khintchine equation is used for the estimates of spectral densities, all coefficients, the major part, or the first part of the correlogram is used, and not only one value, if the autoregressive models are applicable or investigated.

If k is not too large, the direct estimate of ρ_k by r_k may be less in error than an estimate by r_1 with the error $k\delta$. This may be true regardless that the sample size in estimating ρ_k by r_k is N-k instead of N or N-1, when an open-series approach is used in computing r_k. There is a point k = q at which the error $k\delta$ is of the same order of magnitude as the sampling error due to the decreasing sample size, N-k. If either these first q values of r_k, or all r_k values, k = 1, 2,..., N-1 are used in estimating ρ, this estimate may be less biased than the estimate obtained by using only r_1.

Let assume that a correlogram, $r_k = f(k)$, is estimated either up to $q = N/\alpha$, $\alpha = 5, 6,...,10$, or up to q = N-1. The square of differences $\rho_k - r_k$ gives

$$S = \sum_{k=1}^{q} (\rho^k - r_k)^2 \quad . \qquad (5.30)$$

Because of different accuracy of r_k, the square of differences are weighted by N-k, the sample size of each r_k, so that the larger weight is given to the first r_k values, and Equation 5.30 becomes

$$S = \sum_{k=1}^{q} (N-k)(\rho^k - r_k)^2 \quad . \qquad (5.31)$$

The smallest value of S, by equating $dS/d\rho$ with zero, produces another estimate of ρ. This $dS/d\rho = 0$ gives

$$\sum_{k=1}^{q} k(N-k)\rho^{2k-1} = \sum_{k=1}^{q} k(N-k)r_k\rho^{k-1} \quad . \qquad (5.32)$$

The solution of this equation gives the estimate ρ. The left term may be expressed as a function of ρ, q, and N, while the right term uses the r_k values. However, in Equation 5.32 eliminating the sum on the left side is not simple.

The problem is in the selection of q as it is with the selection of a truncation point of the correlogram in estimating spectral densities. Two methods are feasible. Either an objective selection of q, or the value α in $q = N/\alpha$ is selected such as $\alpha = 6$ or similar, as it is often done in the approximate estimates of variance densities by using the correlogram and the Wiener-Khintchine equation.

The solution of Equation 5.32 is most feasible by an iterative, computer oriented procedure. First $\hat{\rho}$ is assumed (say as r_1) and both terms are computed. If the left side terms results are greater than the right side term, $\hat{\rho}$ is reduced; otherwise it is increased. A tolerance level between the two sums of Equation 5.32 in the successive approximations of $\hat{\rho}$ should be established in this iterative procedure.

Figure 5.1 presents the example of annual run-off series of the St. Lawrence River at Ogdensburg, N. Y., with a long time series (N = 97). The use of $r_1 = 0.705$ in the first-order linear autoregressive model produces the fitted correlogram below the observed one. For $\hat{\rho} = r_1 + b$, with b the bias, $b = (1 + 3\hat{\rho})/(N - 1)$, which when solved for $\hat{\rho}$ gives

$$\hat{\rho} = \frac{r_1(N-1) + 1}{N - 4} \quad , \qquad (5.33)$$

so that for $r_1 = 0.705$ and N = 97 this gives $\hat{\rho} = 0.748$. By using Eq. 5.32, the estimate of ρ gives 0.780. The computed correlogram, r_k, and the correlograms of Eq. 5.26, with three estimates of ρ, 0.705, 0.747, and 0.780, are given in Fig. 5.1. Though $r_1 = 0.780$ fits well Eq. 5.16 to the observed correlogram, it may be questioned for its lack

of theoretical background, and some preference may be then given to $r_1 = 0.747$, as corrected for the bias of Eqs. 5.22 and 5.33.

5.10 Estimates of α_1 and α_2 of the second-order Linear Model with a Decrease of Bias

The general equation relating the estimates a_1 and a_2 of α_1 and α_2 with the correlogram estimates, r_k, is

$$r_k = a_1 r_{k-1} + a_2 r_{k-2} \quad . \tag{5.34}$$

The classical estimates of α_1 and α_2 are calculated by using k = - 2 and k = - 1, producing the two equations

$$\left.\begin{aligned} r_2 &= a_1 r_1 + a_2 \quad , \\ \text{and} \qquad & \\ r_1 &= a_1 + a_2 r_1 \quad , \end{aligned}\right\} \tag{5.35}$$

from which a_1 and a_2 are expressed in function of r_1 and r_2. Only these two values determine the fit, and all information contained in r_3, r_4,..., r_{11} is neglected though there is a bias in r_1 and r_2 as the estimates of ρ_1 and ρ_2.

By using Equation 5.34 for the population, k = 1,...,n, then

$$\left.\begin{aligned} \rho_1 &= \alpha_1 + \alpha_2 \rho_1 \\ \rho_2 &= \alpha_1 \rho_1 + \alpha_2 \\ &\cdots\cdots\cdots \\ \rho_q &= \alpha_1 \rho_{q-1} + \alpha_2 \rho_{q-2} \quad , \end{aligned}\right\} \tag{5.36}$$

and by using the weights N-k for the differences $\rho_k - r_k$, then their sum of squares is

$$S = \sum_{k=1}^{q} (N-k)(\rho_k - r_k)^2 \quad . \tag{5.37}$$

The two partial derivatives $\partial S/\partial\alpha_1$ and $\partial S/\partial\alpha_2$ equated to zero should produce the new estimates α_1 and α_2, which are

$$2 \sum_{k=1}^{q} (N-k)(\rho_k - r_k) \frac{\partial \rho_k}{\partial \hat{\alpha}_1} = 0 \tag{5.38}$$

and

$$2 \sum_{k=1}^{q} (N-k)(\rho_k - r_k) \frac{\partial \rho_k}{\partial \hat{\alpha}_2} = 0 \quad . \tag{5.39}$$

By determining $\partial\rho_k/\partial\hat{\alpha}_1$ and $\partial\rho_k/\partial\hat{\alpha}_2$ from Equations 5.36, Equations 5.38 and 3.39 may be rewritten as

$$\sum_{k=1}^{q} (N-k)(\hat{\rho}_k - r_k)\, \hat{\rho}_{k-1} = 0 \tag{5.40}$$

and

$$\sum_{k=1}^{q} (N-k)(\hat{\rho}_k - r_k)\, \hat{\rho}_{k-2} = 0 \quad . \tag{5.41}$$

Then

$$\sum_{k=1}^{q} (N-k)\, \hat{\rho}_k\, \hat{\rho}_{k-1} = \sum_{k=1}^{q} (N-k)\, r_k\, \hat{\rho}_{k-1} \quad , \tag{5.42}$$

and

$$\sum_{k=1}^{q} (N-k)\, \hat{\rho}_k\, \hat{\rho}_{k-2} = \sum_{k=1}^{q} (N-k)\, r_k\, \hat{\rho}_{k-2} \quad . \tag{5.43}$$

As $\hat{\rho}_k$, $\hat{\rho}_{k-1}$ and $\hat{\rho}_{k-2}$ are functions of $\hat{\alpha}_1$ and $\hat{\alpha}_2$, Equations 5.42 and 5.43 are two equations with two parameters, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, which must be estimated.

Equations 5.42 and 5.43 can be solved by an iterative procedure by first assuming $\hat{\alpha}_1$ and $\hat{\alpha}_2$, and repeating the computations with an incremental change of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ until the equations are satisfied with a prescribed tolerence level. The initial pair of values can be estimated by Equations 5.35 by using only r_1 and r_2. Several approaches may be used to find these estimates assumed here to be less biased than when only r_1 and r_2 are used in estimating α_1 and α_2.

5.11 Estimates of Autoregressive Coefficients of the m-th Order Model With a Decrease of Bias

Starting from the basic equation of relationship between ρ_k and α_j parameters of the m-th order linear model, then

$$\rho_k = \sum_{j=1}^{m} \alpha_j \rho_{k-j} \quad . \tag{5.44}$$

Then q equations of the type of Equations 5.36 may be obtained with $q >> m$.

The fitting of ρ_k function of Equation 5.44, with k = 1, 2,...,q, to the first q values of R_k- correlogram by the least-square method gives first a sum of deviations weighted by N-k as

$$S = \sum_{k=1}^{q} (N-k)(\rho_k - R_k)^2 \quad . \tag{5.45}$$

with q selected in a proper manner to cover the major part of the R_k correlogram before it practically converges to zero.

The general method of finding the least value of S of Equation 5.45 and the corresponding set of $\hat{\alpha}_j$ coefficients is in obtaining the m partial

derivatives, $\partial S/\partial \hat{\alpha}_j$, j = 1,2,...,m, equate them to zero and solve for $\hat{\alpha}_j$'s. This gives a set of m equations

$$\sum_{k=1}^{q} (N-k)\, \hat{\rho}_k\, \rho_{k-j} = \sum_{k=1}^{q} (N-k)\, R_k\, \rho_{k-j} \qquad (5.46)$$

with j = 1, 2,...,m. Coupled with Equation 5.44 for ρ_k, k = 1, 2,...,q, there are m equations in $\hat{\alpha}_j$ for m values of $\hat{\alpha}_j$, with k-j replaced by j-k whenever k-j is negative.

The solution of m equations of Equation 5.46 are often difficult and combersome even for m as low as 2. There exist various methods in the literature for solving them.

The use of Equation 5.46 coupled with Equation 5.44 is unnecessary, because the methods exist in finding a minimum of Equation 5.45 and estimating $\hat{\alpha}_j$'s directly. In order to minimize S of Equation 5.45 in an efficient way, a good optimization routing in selecting $\hat{\alpha}_j$'s is required. A routine due to Rosenbrock may be used, as described in details in reference [13].

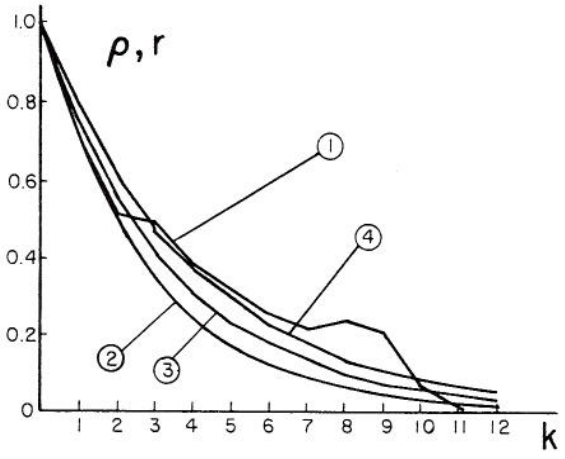

Fig. 5.1 Fitting the first-order linear autoregressive model to the correlogram of the annual runoff series of the St. Lawrence River at Ogdensburg, New York: (1) The estimated correlogram from data, $r_k = f(k)$; (2) By using Eq. 5.26 and the estimate $r_1 = 0.705$; (3) By using Eq. 5.26 and the value $r_1 = 0.747$, corrected for the bias of Eq. 5.22; and (4) By using Eq. 5.26 and the estimate $r_1 = 0.780$ in applying Eq. 5.32.

Chapter 6

PROBABILITY DISTRIBUTIONS OF INDEPENDENT STOCHASTIC COMPONENTS

6.1 Fitting Probability Functions to Empirical Frequency Distributions of Independent Stochastic Components

Once $\xi_{p,\tau}$ is accepted as a stationary independent variable of either of the second-order or third-order stationarity, the $n\omega$ values serve to determine a probability function of the best fit to the empirical frequency distribution. If the $n\omega$ values are to be stored in any usual way for future use, such as tables, graphs, magnetic tapes, punched cards, or similar, the basic objective of reducing the information in the $x_{p,\tau}$ process to the mathematical models and the estimation of their parameters would be defeated. Also, the eventual generation of large samples of $x_{p,\tau}$, using the experimental statistical (Monte Carlo) method starting from the generation of new samples, can still be easily performed by transforming the uniformly distributed random numbers by using the sample frequency distribution curve directly to the random numbers of the $\xi_{p,\tau}$ variable with the same probabilities. This second route in the use of the Monte Carlo method means that all sampling zig-zag deviations of the empirical distribution around a smooth curve and all eventual biases in $\xi_{p,\tau}$ would be perpetuated in the generated samples. Besides, the larger and the smaller values of $\xi_{p,\tau}$ than those observed will not be generated if some adjustments on the extremes are not made. Therefore, the fitting of a probability function to the frequency distribution curve of $\xi_{p,\tau}$ is the approach followed in this study to structurally analyze and mathematically describe a hydrologic time series.

The transformation of $x_{p,\tau}$ to produce the standardized variable $\epsilon_{p,\tau}$ and the treatment of $\epsilon_{p,\tau}$ to produce the independent stochastic variable $\xi_{p,\tau}$ make the positively-valued variable $x_{p,\tau}$ as a $\xi_{p,\tau}$ variable with both negative and positive falues. However, the minimal values of $\xi_{p,\tau}$ may not have a lower bound. If the lower limit of $x_{p,\tau}$ is zero, then the expected range of negative values of $\xi_{p,\tau}$ can be computed from various transformations. This leads to two alternatives in selecting the lower bound of the $\xi_{p,\tau}$ probability distribution function: (a) to compute or estimate what is the approximate lower bound of $\xi_{p,\tau}$ and use it as a fixed boundary of a selected probability function; and (b) to estimate the lower boundary of $\xi_{p,\tau}$ with the estimate of its other parameters.

To find the properties of the negative values of $\xi_{p,\tau}$, a lower boundary should be assigned to $x_{p,\tau}$. In this latter case three solutions are feasible: (a) min $(x_{p,\tau}) = 0$; (b) min $(x_{p,\tau}) = x_o$, a positive value; and (c) $x_{p,\tau}$ has a truncated distribution, so that $x_{p,\tau} \geqslant 0$, but min $(x_{p,\tau})$ of untruncated distribution may have the lower boundary $x_o \leqslant 0$.

For min $(x_{p,\tau}) = 0$, Equation 3.10 for min $(x_{p,\tau}) = 0$ gives the minimum value, min $(y_{p,\tau}) = -\mu_\tau/\sigma_\tau$, so that the two periodic components μ_τ and σ_τ must be divided to find the largest value of μ_τ/σ_τ or of $1/\eta_\tau$:

$$\min (y_{p,\tau}) = - \max (\mu_\tau/\sigma_\tau) = - 1/\min \eta_\tau \quad . \qquad (6.1)$$

Similarly, Equation 3.13 shows that

$$\min \epsilon_{p,\tau} = - \max \left[\frac{\mu_\tau + \mu_y \sigma_\tau}{\sigma_\tau \sigma_y} \right] \quad . \qquad (6.2)$$

Finally, the m-th order linear autoregressive model gives the properties of the lowest value of $\xi_{p,\tau}$ as

$$\min \xi_{p,\tau} = \min \left[\frac{1}{\sigma} \epsilon_{p,\tau} - \frac{\mu_\tau + \mu_y \sigma_\tau}{\sigma_y \sigma_\tau \sigma} - \frac{1}{\sigma} \sum_{j=1}^{m} \alpha_{j,\tau} \epsilon_{p,\tau-j} \right], \qquad (6.3)$$

in which $\alpha_{1,\tau}, \alpha_{2,\tau},\ldots,\alpha_{m,\tau}$ the autoregressive coefficients of the linear models, either periodic or not, and σ is a function of autoregression coefficients. By using m = 3 as the eventual maximum value of j in Equation 6.3 in practical cases, the smallest value of $\xi_{p,\tau}$ may be approximately determined. Assuming several values of $x_{p,\tau}$ zero in a sequence, and assuming that $\mu_y \approx 0.00$, $\sigma_y \approx 1.00$, and $\mu_\tau/\sigma_\tau \approx 1/\eta_o$ with η_o a constant the Eq. 6.3 becomes approximately

$$\min \xi_{p,\tau} = \min \left[\frac{1}{\sigma} \epsilon_{p,\tau} - \frac{1}{\eta_o \sigma} \left(1 + \sum_{j=1}^{3} \alpha_{j,\tau}\right) \right]. \qquad (6.4)$$

In conclusion, the periodic components transform the lower boundary of $x_{p,\tau}$ to a lower boundary of $\epsilon_{p,\tau}$. However, the autoregressive linear models

transform the variable $\epsilon_{p,\tau}$ which is bounded on the left side to a new variable $\xi_{p,\tau}$ which theoretically may be unbounded.

Though $\xi_{p,\tau}$ theoretically can have large negative values, the practical considerations still limit them. It is unlikely that the minimum value $\epsilon_{p,\tau}$ of Equation 6.2 would be preceded by a very large value $\epsilon_{p,\tau-1}$ under the condition of $\alpha_{1,\tau}$ being also very large or close to unity. For daily flows $\alpha_{1,\tau} = \rho_{1,\tau}$ is very large, often of the order 0.80 - 0.98, but the river basin response rarely permits $\epsilon_{p,\tau}$ to go from a very large value suddenly to zero. If this is the case, however, $\rho_{1,\tau}$ is then small, so that the product $\rho_{1,\tau}\ \epsilon_{p,\tau-1}$, for a large $\epsilon_{p,\tau-1}$, also becomes small and contributes little to the negative values of Equations 6.3 and 6.4. This fact permits still an application of bounded distribution functions to $\xi_{p,\tau}$. An approximation in estimating the practical lower boundary of $\xi_{p,\tau}$ may be used, so that distribution functions with lower boundaries can be also tested for the goodness of fit to the empirical frequency distributions of $\xi_{p,\tau}$.

Five probability density functions may be used for fitting the empirical frequency density curves of $\xi_{p,\tau}$: (1) the general transformation of the normal function by using the polynomial of a given order, as the transform; (2) the two-parameter normal distribution; (3) the three-parameter lognormal distribution with the lower boundary different from zero; (4) the three-parameter gamma distribution (with the lower boundary different from zero); and (5) the double-branch gamma distribution with a total of six parameters.

The extreme values of $\xi_{p,\tau}$ may have high sampling errors, or they may not be representative of the sample size, being either too large or too small for that sample size. They greatly affect the estimates of various parameters. To avoid this bias, the values at both extremes of the empirical $\xi_{p,\tau}$ distributions may be deleted from the estimation of parameters. For example, 0.50 percent of the largest and 0.50 percent of the smallest values of the total sample may be discounted in the estimation. The new sample N is then shorter, 0.99N or 0.99nω. However, the entire series N = nω should be used in testing the goodness of fit of various distribution functions by χ^2 tests or any other test.

6.2 General Fit of a Normal Distribution Transformed by an m-th Order Polynomial

The most general fit of a probability density function to a frequency distribution curve of $\xi_{p,\tau}$ is obtained by using the normal probability density function transformed by an m-th order polynomial. The normal function has two parameters to be estimated, μ and σ, and the m-th order polynomial has m+1 parameters, β_k, with k = 0, 1,...,m, so that the total number is m + 3. The larger m, is the better fit, but the more degrees of freedom are lost. Then the total number of degrees of freedom is

$$DF = n\omega - m - 3 - \nu \qquad (6.5)$$

with ν the number of all parameters previously estimated. Because nω may be very large, say for daily or monthly values, and ν, the number of parameters or coefficients which are estimated from data in order to obtain $\xi_{p,\tau}$, may also be small, the m-th order polynomial may be selected with m as large as feasible without a significant loss of degrees of freedom.

The polynomial is given by

$$\xi_i = \beta_o + \sum_{k=1}^{m} \beta_k\ z_i^k \quad , \qquad (6.6)$$

where ξ_i replaces $\xi_{p,\tau}$ (no important periodicity is left in the series), z_i is the random variable normally distributed, $N(\mu,\sigma)$, and $\beta_o, \beta_1,...,\beta_m$ are the polynomial parameters to be estimated.

Fitting the polynomial of Equation 6.6 is the same as fitting the following probability function, $f(\xi)$, to the frequency distribution of ξ,

$$f(\xi) = f(z) \left\{ \left| \sum_{k=1}^{m} kb_k\ z^{k-1} \right| \right\}^{-1} \quad , \qquad (6.7)$$

in which f(z) is the normal probability function of z, (μ_z, σ_z), β_k are the polynomial coefficients of Eq. 6.6, with k = 0, 1, 2,...,m, and m is the selected order of the polynomial.

One of the practical methods of estimating β_k coefficients of Equation 6.6 is as follows. A parameter u is selected, which is the number of class intervals of equal probability 1/u of the empirical frequency distribution of $\xi_{p,\tau}$. In practice, the following values may be used: u = 200 for daily series, u = 160 for 3-day series, u = 100 for 7-day series, u = 80 for 13-day series, u = 60 for monthly series, and u = 24 for 3-month series. In order to apply this procedure, the $\xi_{p,\tau}$ series is ranked in

ascending order and denoted as the ξ_i series, i = 1, 2,...,N, with $N = n\omega$ being the sample size. The class limits ξ_j, with j = 1,2,..., u-1, of the selected u class intervals of equal absolute frequency $N/u = n\omega/u$ of the $\xi_{p,\tau}$ variable are determined from the ascended sequence ξ_i. The ξ_j-th class limit is computed as the midpoint of i-th and (i+1)-th values of the ranked ξ_i series, with $i = Nj/u$ and $i + 1 = Nj/u + 1$, if i is an integer. If Nj/u is not an integer, i is designated as the integer segment of Nj/u, and d as the decimal segment of Nj/u; the j-th class limit is then computed by

$$\xi_j = \frac{\xi_i + \xi_{i+1}}{2} + (\xi_{i+1} - \xi_i)\, d \quad , \; j = 1, 2, \ldots, u\text{-}1 . \tag{6.8}$$

The computed (u-1)-values of ξ_j are then used to relate them to the normal variable, and particularly by using their cumulative frequency distribution, with $f(\xi \leqslant \xi_j) = j/u$. The (u-1) class limits, t_j, of the standard normal distribution function are computed, with u the same selected number of class intervals as for the $\xi_{p,\tau}$ variable, or each with the same probability 1/u. To compute t_j values, with j = 1, 2,...,u-1, the following approximation is used

$$t_j = s - \frac{c_o + c_1 s + c_2 s^2}{1 + d_1 s + d_2 s^2 + d_3 s^3} \quad , \tag{6.9}$$

in which

$$s = \sqrt{\ln \frac{u^2}{(u-j)^2}} \quad , \tag{6.10}$$

and $c_o = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$, and $d_3 = 0.001308$.

The t_j-values of Equation 6.9 may be transformed to z_j-values, as the (u-1)-class limits with equal class probabilities 1/u of the nonstandard normal function, if the mean and standard deviation of $\xi_{p,\tau}$ are different from zero and unity, respectively, by

$$z_j = \bar{\xi}_{p,\tau} + s_\xi\, t_j \quad . \tag{6.11}$$

The second, third, fourth, or higher order polynomial is used to find the best fit of the (u-1)-values of the ξ_j and the t_j or z_j class limits, representing the same probability P_j, with $P_j = j/u$, by estimating the first three, four, five, or more parameters, $\beta_o, \beta_1, \beta_2, \beta_3, \beta_4$,..., respectively for the selected polynomial, by Eq. 6.6 in the form

$$\xi_j^* = b_o + \sum_{k=1}^{m} b_k z_j^k \quad , \tag{6.12}$$

in which b_k are coefficients,with m = 2, 3, 4,..., the order of the best fit polynomial and ξ_j the prediction values of ξ_j for given t_j or z_j.

The (u-1)-values of ξ_j^* are then computed for the second, third, fourth, or higher order polynomials by Equation 6.12 by using the (u-1)-values of t_j of Equation 6.9 or z_j of Equation 6.11. For these polynomials the variance of estimates is computed by

$$e_\xi^2 = \frac{1}{u\text{-}1} \sum_{j=1}^{u-1} (\xi_j - \xi_j^*)^2 \quad . \tag{6.13}$$

These variances are designated by V_2, V_3, V_4,..., respectively for m = 2,3,4,..., in Equation 6.12.

The practical method of selecting the order of the fitted polynomial is as follows. For only V_2, V_3, and V_4 computed, then if conditions

$$\frac{V_2 - V_3}{V_2} \leqslant 0.01, \text{ and } \frac{V_2 - V_4}{V_2} \leqslant 0.02 \tag{6.14}$$

are satisfied, the second order polynomial is selected. If the conditions of Equation 6.14 are not satisfied, but the condition

$$\frac{V_3 - V_4}{V_3} \leqslant 0.01 \tag{6.15}$$

is satisfied, the third order polynomial is selected; otherwise, the fourth order polynomial is selected. If V_5, V_6, V_7 ..., are computed, then the differences $(V_{p+1} - V_p)/V_p$ may be the criteria whether any higher order polynomial may be a feasible fit. If all these values, with p = 5, 6, 7,..., are smaller than 0.01, then the lower order polynomial may be selected. However, the chi-square test should be used as the final criterion whether Equation 6.7, with the selected polynomial of the m-th order, is a good fit or not. A summary of this chi-square test is given here for the sake of completeness.

The critical chi-square value, χ_c^2 should be computed for a number, ν, of class intervals of equal probability, with ν a parameter selected so that it satisfies $u = q.\ \nu$, with u the selected number of class intervals, and q and ν integers. The suggested values of q and ν are: q = 4 for all series while ν = 50 for daily series, $\nu = 40$ for 3-day series, $\nu = 25$ for 7-day series, $\nu = 20$ for 13-day series, $\nu = 20$ for 13-day series, $\nu = 15$ for monthly series, and ν = 6 for 3-month series.

If $\nu \geqslant 30$, χ^2 is approximately normally distributed with the degrees of freedom,

$$D_f = \nu - (m+3) \; , \tag{6.16}$$

in which m is the order of the selected polynomial of Equation 6.12. The critical value χ_c^2 is computed by the approximation

$$\chi_c^2 = D_f \left[1 - \frac{2}{9D_f} + t_p \sqrt{\frac{2}{9D_f}} \right]^3 , \quad (6.17)$$

in which $t_p = 0.84162$ as the value for the one-tail probability rejection level, for P = 0.20, of the standard normal variable.

If $\nu < 30$, the $\chi_c^2 = x$ is determined by the inverse of the integral

$$P(\chi_c^2 > x) = \frac{1}{2\Gamma(D_f/2)} \int_x^\infty \left(\frac{x}{2}\right)^{\frac{D_f}{2} - 1} e^{-x/2} \, dx , \quad (6.18)$$

in which $P(\chi_c^2 > x) = 0.20$, or it is the same probability rejection level as selected for Equation 6.17, and D_f is given by Equation 6.16.

Every q-th value is selected from the ξ_j^* array of (u-1) values of Equation 6.12, denoted here as the ξ_s^* values, with s = 1, 2,...,ν-1. All $\xi_{p,\tau}$-values, N = nω, are sorted in the ν class intervals limited by class limits ξ_s^*, with s = 1, 2,...ν-1 ; the first interval is $\xi_{p,\tau} \leq \xi_1^*$ and the last interval is $\xi_{p,\tau} > \xi_{\nu-1}^*$. Then, the absolute frequencies f_k with k = 1, 2,...,ν, are obtained.

The χ^2-statistic is then obtained by

$$\chi^2 = \frac{\nu}{N} \sum_{k=1}^{\nu} \left(f_k - \frac{N}{\nu}\right)^2 . \quad (6.19)$$

If χ^2 of Equation 6.19 is smaller than χ_c^2 of Equations 6.17 or 6.18, whichever is relevant for a selected ν, the fit obtained by the selected polynomial of Equation 6.12, or the fit obtained for $f(\xi)$ by Equation 6.7 is accepted.

If χ^2 of Equation 6.19 is larger than χ_c^2 of Equations 6.17 or 6.18, as applicable for the selected ν, the fit of the selected order m of the polynomial is rejected. If this occurs, the order of the polynomial is increased, the number u of class intervals may also be increased, say by 50 percent, and the parameter q may be changed while ν is increased, say q to be 3-5, while ν may be 75 for daily series, 60 for 3-day series, 30 for 7-day series, 30 for 13-day series, 30 for monthly series, 9 for 3-month series. Then the test is repeated. The obtained values u, m, b_o, b_1,...,b_m, represent the final estimates for the model of Eq. 6.7, with either the t_i normal-variable (0,1), or the z_i normal variable $(\bar{\xi}, s_\xi)$ used in Eq. 6.7.

6.3 Fitting The Two-Parameter Normal Probability Function to $\xi_{p,\tau}$ Variable

The estimates of the mean of the $\xi_{p,\tau}$ variable is computed by

$$E\xi = \frac{1}{n\omega} \sum_{p=1}^{n} \sum_{\tau=1}^{\omega} \xi_{p,\tau} \quad (6.20)$$

and the standard deviation by

$$s_\xi = \left[\frac{1}{n\omega} \sum_{p=1}^{n} \sum_{\tau=1}^{\omega} (\xi_{p,\tau} - \bar{\xi})^2 \right]^{1/2} . \quad (6.21)$$

For testing the goodness of fit of the normal distribution to the $\xi_{p,\tau}$ variable, with N = nω, the ν class intervals with equal probabilities $1/\nu$ are selected. The ν-1 class limits, t_j of the standard normal variable are determined by the inverse of the integral

$$\frac{j}{\nu} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t_j} e^{-t^2/2} \, dt , \quad (6.22)$$

with j = 1, 2,...,ν-1, as t_1, t_2,...,$t_{\nu-1}$.

When the $\xi_{p,\tau}$ series is not standardized, the ν-1 values of t_j are easily transformed to ν-1 values of ξ_j by

$$\xi_j = s_\xi t_j + \bar{\xi} . \quad (6.23)$$

Sorting the total $\xi_{p,\tau}$ series into ν class intervals produces the f_k frequencies, so that Equation 8.19 enables the computation of the χ^2 statistic. By comparing this χ^2 with χ_c^2, obtained either by Equation 6.17 or Equation 6.18, which depends on the selected ν value, the test of fitting the two-parameter normal function is performed. When $\chi^2 \leq \chi_c^2$, the normal function is satisfactory and there is no practical need for testing the fit of other probability functions to the $\xi_{p,\tau}$ frequency distribution, which have more parameters, though their χ^2-values may be smaller.

6.4 Fitting The Three-Parameter Lognormal Probability Function to $\xi_{p,\tau}$ Variable

Assuming the lower boundary is ξ_o, and $\xi_{p,\tau}$ is replaced by the symbol ξ_i, then the lower boundary is estimated by

$$\left(\sum_{i=1}^{N} \frac{1}{\xi_i - \xi_o} \right) \left\{ \frac{1}{N} \sum_{i=1}^{N} \ln^2 (\xi_i - \xi_o) - \right.$$

$$\left[\frac{1}{N}\sum_{i=1}^{N}\ln(\xi_i-\xi_o)\right]^2 - \frac{1}{N}\sum_{i=1}^{N}\ln(\xi_i-\xi_o)\Bigg\} + \sum_{i=1}^{N}\frac{\ln(\xi_i-\xi_o)}{\xi_i-\xi_o} = 0 \quad , \qquad (6.24)$$

with $N = n\omega$.

The lower boundary ξ_o must be estimated from Equation 6.24 by an iterative procedure and by prescribing how much the left side of Equation 6.24 may deviate from zero of the right side. When ξ_o is obtained, the other two parameters, the mean of logarithms of the deviations $(\xi_i-\xi_o)$ are computed by

$$\mu_n = \frac{1}{N}\sum_{i=1}^{N}\ln(\xi_i-\xi_o) \quad , \qquad (6.25)$$

and the standard deviation of logarithms of $(\xi_i-\xi_o)$ by

$$\sigma_n = \left\{\frac{1}{N}\sum_{i=1}^{N}[\ln(\xi_i-\xi_o)-\mu_n]^2\right\}^{1/2} . \qquad (6.26)$$

Then the probability density function of distribution of $\xi_{p,\tau}$ is

$$f(\xi) = \frac{1}{\sigma_n(\xi-\xi_o)\sqrt{2\pi}}\exp\left\{-[\ln(\xi-\xi_o)-\mu_n]^2/2\sigma_n^2\right\} \qquad (6.27)$$

In order to test how good the fit is by Eq. 6.27 to the $\xi_{p,\tau}$ empirical distribution, the chi-square test is also used. Starting with Eq. 6.22, the ν-1 class limits of t_i are determined. Then the class limits of Eq. 6.27 are obtained by the transformation

$$\xi_j = \xi_o + e^{\sigma_n t_j + \mu_n} \qquad (6.28)$$

The χ^2-test can be performed either on $\xi_{p,\tau}$ or on $\xi_{p,\tau}-\xi_o$.

6.5 Fitting The Three-Parameter Gamma Probability Function To $\xi_{p,\tau}$ Variable

For the estimation of the three parameters, α (shape), β (scale), and ξ_o (lower boundary) of the three-parameter gamma function, the maximum likelihood estimation method should be used. The boundary is estimated by an iterative procedure from

$$\frac{1+(1+\frac{4}{3}A)^{1/2}}{1+(1+\frac{4}{3}A)^{1/2}-4A} - (\bar{\xi}-\xi_o)\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\xi_i-\xi_o} = 0 \qquad (6.29)$$

in which

$$A = \ln(\bar{\xi}-\xi_o) - \frac{1}{N}\sum_{i=1}^{N}\ln(\xi_i-\xi_o) \quad , \qquad (6.30)$$

and $\bar{\xi}$ = the mean of N values of ξ_i.

Once ξ_o is estimated, the parameter α is estimated by

$$\alpha = \frac{1+(1+\frac{4}{3}A)^{1/2}}{4A} - \Delta\alpha \quad , \qquad (6.31)$$

with A given by Equation 6.30 and $\Delta\alpha$ approximated by

$$\Delta\alpha = 0.04475(0.26)^{\alpha} \quad . \qquad (6.32)$$

The parameter β is estimated by

$$\beta = \frac{1}{\alpha}\sum_{i=1}^{N}(\xi_i-\xi_o) = \frac{1}{\alpha}(\bar{\xi}-\xi_o) \quad . \qquad (6.33)$$

With all three parameters estimated, the probability density function of $\xi_{p,\tau}$ is

$$f(\xi) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{\xi-\xi_o}{\beta}\right)^{\alpha-1} e^{(\xi-\xi_o)/\beta} \quad . \qquad (6.34)$$

To determine how good the fit of the $\xi_{p,\tau}$ empirical distribution is by Equation 6.34, the chi-square test is performed. Consider ν class intervals each with the equal probability $1/\nu$. The ν class limits are determined from the one-parameter gamma distribution by the inverse of the integral

$$P(X > x_j) = \frac{1}{\Gamma(\alpha)}\int_{x_j}^{\infty} x_j^{\alpha-1} e^{-x_j}\, dx_j \qquad (6.35)$$

for $x_j \geq 0$, for $j = 0,1,\ldots,\nu-1$. The computed values $x_o, x_1,\ldots,x_{\nu-1}$, are then transformed to ξ_j- class limits by

$$\xi_j = \xi_o + \beta x_j \quad , \qquad (6.36)$$

if $\xi_{p,\tau}$ are tested; or $\xi_j - \xi_o = \beta x$ are the limits if $(\xi_{p,\tau} - \xi_o)$-variable is tested for the goodness of fit; β is the parameter of Eq. 6.34.

Instead of Equations 6.35 and 6.36, the inverse of the integral may be written as

$$\frac{j}{\nu} = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_{\xi_o}^{\xi_j}(\xi-\xi_o)^{\alpha-1} e^{-(\xi-\xi_o)/\beta}\, d\xi \quad , \qquad (6.37)$$

in which ξ_o, α and β are estimated by Equations 6.29 through 6.33 and the variable $(\xi_{p,\tau}-\xi_o)$ is tested for the goodness of fit.

6.6 Fitting The Double-Branch Gamma Probability Function to the $\xi_{p,\tau}$ Variable

Some $\xi_{p,\tau}$ variables of daily flow series may show that the three previous probability functions (normal, lognormal and gamma) do not fit well their empirical frequency density curves. These curves are highly peaked at a given value ξ_o, and rapidly decrease in both branches the positive and the negative values of $(\xi_{p,\tau} - \xi_o)$.

Three alternatives have been studied for these cases of high peakedness:

(1) the use of Laguerre polynomials applied to the gamma distribution;

(2) the use of other distributions, particularly various Pearson curves, except the Type III (three-parameter gamma); and

(3) the use of the two two-parameter gamma distributions with either $\alpha < 1$ or $\alpha = 1$, for the two branches of the highly peaked empirical $\xi_{p,\tau}$ distribution.

This latter case has been shown to be practical, provided that the six parameters to be estimated, if $\alpha < 1$, may be replaced by only four parameters, if $\alpha = 1$ is selected for both branches.

First, the position of the peak, ξ_o, and the probabilities $P = P(\xi_{p,\tau} \leq \xi_o)$ and $P(\xi_{p,\tau} > \xi_o) = 1 - P$ should be determined in order to obtain the point where the two branches intersect for $\alpha = 1$, or have the vertical asymptote, for $0 < \alpha < 1$. Then the total areas, P and 1-P, of the left and the right branch are obtained, respectively. The parameters α_1 and β_1 for the left branch, and α_2 and β_2 for the right branch, are estimated by the following equations,

$$f(\xi_o - \xi) = \frac{P}{\beta_1^{\alpha_1}\,\Gamma(\alpha_1)}\,(\xi_o - \xi)^{\alpha_1 - 1}\; e^{-(\xi_o - \xi)/\beta_1}, \quad (6.38)$$

for $\xi < \xi_o$, for the left branch, and

$$f(\xi - \xi_o) = \frac{1 - P}{\beta_2^{\alpha_2}\,\Gamma(\alpha_2)}\,(\xi - \xi_o)^{\alpha_2 - 1}\; e^{-(\xi - \xi_o)/\beta_2}, \quad (6.39)$$

for $\xi > \xi_o$, for the right branch. For $\alpha_1 = \alpha_2 = 1$, Equations 6.38 and 6.39 are simpler to use.

For the ξ_i-series, the six parameters of this double-branch gamma functions to be estimated are ξ_o, P, α_1, β_1, α_2 and β_2, if $0 < \alpha < 1$, but only ξ_o, P, β_1, and β_2, if $\alpha_1 = \alpha_2 = 1$.

The mode ξ_o, or the value of ξ_i with the largest frequency density, can be estimated by selecting the class interval $\Delta\xi$, say 0.001 or smaller, and by finding the lower boundary ξ_ℓ of the interval $\Delta\xi$ with the largest frequency of the ξ_o series. The estimate of ξ_o is then approximately

$$\xi_o = \xi_\ell + \Delta\xi/2 \quad . \quad (6.40)$$

The parameter P, or the probability of all values $\xi_{p,\tau} \leq \xi_o$, is determined by

$$P = \frac{N_1}{n\omega} \quad , \quad (6.41)$$

with N_1 the number of all values $\xi_{p,\tau} \leq \xi_o$.

Parameters α_1 and β_1 of the left branch are estimated by using only N_1 values of the $\xi_{p,\tau}$ series; and α_2 and β_2 are estimated by using the $N - N_1 = N_2$ values of $\xi_{p,\tau}$, with $\xi_{p,\tau} > \xi_o$.

Then α_1 and β_1 are estimated by

$$\alpha_1 = \frac{1 + \sqrt{1 + \frac{4}{3}A_1}}{4A_1} - 0.04475(0.26)^{\alpha_1} \quad (6.42)$$

with

$$A_1 = \ln\left[\xi_o - \frac{1}{N_1}\sum_{i=1}^{N_1}\xi_i\right] - \frac{1}{N_1}\cdot\sum_{i=1}^{N_1}\ln(\xi_o - \xi_1) \quad , \quad (6.43)$$

and

$$\beta_1 = \frac{1}{\alpha_1}\left[\xi_o - \frac{1}{N_1}\sum_{i=1}^{N_1}\xi_i\right] \quad . \quad (6.44)$$

Parameters α_2 and β_2 of the right branch are estimated by using only N_2 values of the $\xi_{p,\tau}$ series and equations similar to Eqs. 6.42 through 6.44, with

$$A_2 = \ln\left[\frac{1}{N_2}\sum_{i=1}^{N_2}(\xi_i - \xi_o)\right] - \frac{1}{N_2}\sum_{i=1}^{N_2}\ln(\xi_i - \xi_o) \quad . \quad (6.45)$$

For testing the goodness of fit of the double-branch gamma function to the frequency distribution of the $\xi_{p,\tau}$ variable, with $N = n\omega$, the number of class intervals selected is ν, with equal probabilities $1/\nu$. The class limits of ξ_j values are determined by the inverse of the integrals, for the left branch in the form

$$\frac{j}{\nu} = \frac{P}{\beta_1^{\alpha_1}\,\Gamma(\alpha_1)} \int_{-\infty}^{\xi_j} (\xi_o - \xi)^{\alpha_1 - 1}\, e^{-(\xi_o - \xi)/\beta_1}\, d\xi\,, \tag{6.46}$$

and for the right branch in the form

$$\frac{\nu - j}{\nu} = \frac{-(1-P)}{\beta_2^{\alpha_2}\,\Gamma(\alpha_2)} \int_{\infty}^{\xi_{\nu-j}} (\xi - \xi_o)^{\alpha_2 - 1}\, e^{-(\xi - \xi_o)/\beta_2}\, d\xi\,. \tag{6.47}$$

In this case, the class interval of the two closest class limits to ξ_o, say ξ_j and ξ_{j+1}, should have the probability $1/\nu$, if the left branch is integrated from ξ_j to ξ_o and the right branch is integrated from ξ_{j+1} to ξ_o, and the two areas summed. The χ^2-test is performed as for the other probability functions.

Chapter 7

BIAS RETAINED BY INAPPROPRIATE STRUCTURAL ANALYSIS OF TIME SERIES

Results obtained in structural anslysis and mathematical description of hydrologic time series, and in the preservation of basic properties of a stochastic process in generating new samples, with these properties inferred from an available sample, may be assessed by answering the question of whether or not any bias in the existing sample is carried into the new generated samples. The bias in this chapter is defined either as a sample characteristic which has a very small probability to be repeated in the new observed samples (or to be the population property), or as some other property of the observed sample which makes the analysis and description distorted.

7.1 Leap Year Effect

In using small time intervals (day, 3-day, 7-day), the leap year effect represents a shift in the period of the year. However, one-day shift every four years does not significantly affect the results either of the periodicity in several parameters, or of the time dependence for its stationary stochastic component. The simplest approach seems to be the deletion of one day in leap years, making it 365 instead of 366 days. This means the period of every year is shifted 1/4 of day for the first three years in comparison to the previous year, and the fourth year, shifted back for a full day, has the period 3/4 of a day in advance of the previous year.

7.2 Sampling Trends and Cycles

The most current bias in structural analysis and mathematical description of hydrologic time series is the preservation in generated samples of the sampling trends and sampling cyclicities of an available sample. By removing the within-the-year periodicity in the mean and standard deviation, the upward or downward chance trends and the pseudo-cyclical chance fluctuations over periods greater than the year are not removed from the $\epsilon_{p,\tau}$ series. These chance patterns are reflected in the general correlogram r_k or the general spectrum v_f of an $\epsilon_{p,\tau}$ series. Because a combination of an autoregressive linear model and a trend, or a combination of an autoregressive linear model and a pseudo-cyclicity result in a complex correlogram or spectrum, the fitting of a simple mathematical autocorrelation function does not usually produce an independent and second-order stationary $\xi_{p,\tau}$ series. If a non-stationary approach of structural analysis is used, by which $r_{1,\tau}$ is investigated for $\epsilon_{p,\tau}$ at each position value τ, say for the first-order autoregressive model, the reproduction of $\epsilon_{p,\tau}$ by these ω values of $r_{1,\tau}$, with autoregressive linear equations, may not be significantly affected by the sampling trend or pseudo-cyclicity, so that the $r_{1,\tau}$ values may be realistic estimates of population values $\rho_{1,\tau}$.

To remove this type of bias, the use of $r_{k,\tau}$ values, $k = 1, 2, \ldots, m$ and $\tau = 1, 2, \ldots, \omega$, as column values for the year as the period, instead of r_k of the general $\epsilon_{p,\tau}$ correlogram, decreases or eliminates the effects of chance trends or chance pseudo-cyclicities, though it may not remove their effects completely. The correlogram of the means $r_{k,\tau}$ values, $k = 1, 2, \ldots$, represents then the basic dependence structure, though the trends and pseudo-cyclicities may still affect some of these mean values. This is a basic reason which has influenced the position in this study, namely to use $r_{k,\tau}$ and its means rather than the general correlogram r_k of the $\epsilon_{p,\tau}$ series.

7.3 Biases at Extremes

The case is frequent when an extreme high or extreme low value represent a bias because their probabilities to be exceeded or not exceeded in a sample of limited size are not sufficiently high, or these extremes may not be representative of the sample of a given size. For example, a flood peak which has occurred in a sample of 30 years might have a return period of only 10-15 years, or the opposite, of a recurrence interval of 80-100 years or even more. If the maximum observed annual peak of annual flood series has a return period somewhere between 20 and 40 years for a sample series of 30 years, this maximum value would be considered representative of the sample. Similarly, the analysis of runs of a series, with observed dry or wet spells for a given crossing level which defines the runs (measured either by duration, by total deficit or by maximum or minimum intensity), may not be representative of a sample size. The average return period of such extreme runs may be either much

greater or much smaller than the sample size in which they are observed.

To avoid these unrepresentative extremes becoming perpetuated either in the mathematical description of a time series, or in the new generated samples, an appropriate structural analysis should be selected. The basic approach in the structural analysis, as described in this paper, is the fitting of probability distribution functions to empirical frequency distribution curves of the independent stochastic component, so that the tails of these functions are approximated by a decreasing function, basically of the exponential or simple exponential type. By replacing the sample frequency distribution curve by an inferred probability density function, the problem of extreme unrepresentative values may partly be alleviated. By using the maximum likelihood method of estimation instead of the method of moments in fitting the probability distribution functions, the effects of unrepresentative extremes are further reduced.

It is presently current among some investigators, who apply stochastic processes in hydrology and/or develop the methods of generating new samples of hydrologic processes, to measure these methods by how well they reproduce the extreme values of observed samples. Some of the dependence models used currently in hydrology, like the autoregressive or Markov linear models, have been criticized in that the reproduction of extremes in new generated samples is not good. It is claimed that proposed new models and methods would better reproduce the extremes; reproducing extremes as closely as possible may represent the perpetuating of unrepresentative extremes, or retaining various sampling biases. The issue is reduced to the question of what should be best reproduced in the generation of new samples from the basic properties of historic samples. The reproduction of every sampling variation of historic series in the new generated samples defeats the basic objective of the application of Monte Carlo method, namely to investigate how water resources systems would perform under new realizations of a process, realizations which have a sufficiently high probability to occur in the future.

These criticisms of autoregressive linear models for representing stochastic dependence in hydrologic time processes lead to the impression that these arguments have been derived mainly to justify proposed new dependence models rather than being based on an objective analysis of natural physical processes and the applicability of autoregressive linear models to them.

A stochastic mathematical model, to be applicable in hydrology, should satisfy the three basic tests:

a. That it has a sound theoretical background based on general properties of hydrologic natural processes;

b. That the type of responses of hydrologic environments to various inputs, which responses are the main factors of time dependence, predetermine the type of models; and

c. That the investigation of a large number of observed homogeneous series of a random hydrologic variable over a sufficiently large region supports, by the use of best statistical inference techniques, the good fit of a given model for the dependence characteristics of these series. The support should occur at least the percent equivalent to the selected probability level of the statistical inference technique.

By neglecting any of the three tests, it is necessary for some models advanced in the literature to be justified by superficial or dubious arguments, when they are compared with models of dependence of already proven applications.

Chapter 8

CONCLUSIONS

The topics analyzed in this paper are basically approaches with a fundamental hypothesis: that any continuous hydrologic time process, or its discrete series approximations, can be separated into deterministic component parameters, and a stationary stochastic component. The two rationales of this approach are:

(a) Physical hydrologic processes in nature support this hypothesis, both from the point of view of solar energy input, and hydrologic environment, particularly river basin responses. They justify periodicities in parameters and mathematical dependence models in stochastic components.

(b) Mathematical descriptions of deterministic components and of stationary stochastic components are most feasible when treating complex processes with the presently available methods for stationary and ergodic processes.

Three parts of the analysis presented are crucial. The methods, approximate or exact, are outlined in treating these three basic parts.

1. The available statistical inference techniques are used to infer the presence of periodicities in basic parameters, while the mathematical description of these periodicities is made by using Fourier series analysis, with a limited number of low frequency harmonics and their estimated coefficients.

2. When periodicities in periodic parameters are removed from the original series, and the resulting stationary stochastic component of a given order of stationarity is analyzed for dependence, then autoregressive linear models are used. They are used under the assumption that they have been inferred as being the closest approximation of reality either by investigating the physical processes in hydrologic environments, or by statistical analysis of hundreds of available time series of a hydrologic random variable.

3. When a stochastic stationary independent component is computed from the dependence model of the stochastic component, the probability distribution functions which best fit the frequency distribution curves are analyzed, sorting them from the simplest to the more complex probability distribution functions. The larger an interval of a discrete time series, the simpler is this function.

Historic samples of hydrologic time processes are subject to various biases. The basic approach in this study was to shape the techniques of structural analysis, mathematical description and data generation such that these biases are not perpetuated. It is expected that the real difference between the use of already proven practical mathematical dependence models, and abstract untested models is that the preservation of various biases in the latter approach is not given a proper critical assessment.

Methods available at present for the structural analysis and mathematical description of hydrologic time series may be divided into two broad groups: the analysis of series as nonstationary processes, and the analysis of series as composed of a stationary process and inferred deterministic components. The first dilemma in selecting an approach is always of this type. The position taken in this paper is that any technique of structural analysis, mathematical description and data generation in hydrology cannot be better than the basic hypothesis which underlies these three practical aspects of hydrologic time series.

REFERENCES

1 Yevjevich, Vujica, Fluctuations of Wet and Dry Years, Part I, Research Data Assembly and Mathematical Models, Colorado State University Hydrology Papers, No. 1, July 1963, Fort Collins, Colorado.

2 Yevjevich, Vujica, Fluctuations of Wet and Dry Years Part II, Analysis by Serial Correlation, Colorado State University Hydrology Papers, No. 4, June 1964, Fort Collins, Colorado.

3 Roesner, L. A., and V. Yevjevich, Mathematical Models for Time Series of Monthly Precipitation and Monthly Runoff, Colorado State University Hydrology Papers, No. 15, October 1966, Fort Collins, Colorado.

4 Quimpo, G. Rafael, Stochastic Model of Daily River Flow Sequences, Colorado State University Hydrology Paper, No. 18, February 1967, Fort Collins, Colorado.

5 Quimpo, R. G., and V. Yevjevich, Stochastic Description of Daily River Flows, Proceedings International Hydrology Symposium, Fort Collins, Colorado, September 6-8, 1967, Vol. I, p. 290-297.

6 Yevjevich, V., and R. I. Jeng, Effects of Inconsistency and Non-Homogeneity of Hydrologic Time Series, Proceedings International Hydrology Symposium, Fort Collins, Colorado, September 6-8, 1967, Vol. I, p. 451-458.

7 Yevjevich, V., and R. I. Jeng, Properties of Non-Homogeneous Hydrologic Series, Colorado State University Hydrology Papers, No. 32, April 1968, Fort Collins, Colorado.

8 Fisher, R. A., Tests of Significance in Harmonic Analysis, Proc. of Royal Soc. of London, Ser. A, v. 125, p. 54-59, 1929. (Also, Contributions to Mathematical Statistics, John Wiley and Sons, Inc., New York, 1950 Paper 16.53-16.59).

9 Fisher, R. A., On the Similarity of the Distributions Found for the Test of Significance in Harmonic Analysis, and in Stevens' Problem in Geometrical Probability, Annals of Engenics, v. X, p. E.1, p. 14-17, 1940; Also, Contributions to Mathematical Statistics, John Wiley and Sons, Inc., New York, 1950, paper 37.13 a-37.17.

10 Cramer, H., Mathematical Statistics, 8th printing, Princeton University Press, 1958.

11 Quenouille, M. H., A Large Sample Test for the Goodness of Fit of Auto-regressive Schemes, Journal Royal Statistical Society, v. 110, p. 123-129, 1949.

12 Kendall, M. G., and H. Stuart, The Advanced Theory of Statistics, Volume 3, Design and Analysis and Time-Series, 2nd Edn., Hafner, New York, 1968.

13 Rosenbrock, H. H., An Automatic Method for Finding the Greatest or Least Value of a Function, The Computer Journal, v. 3, 1960, pp. 175-184.

KEY WORDS: Time series, hydrologic series, analysis of series, structural composition of series, hydrologic time processes.

ABSTRACT: Structural analysis and mathematical description of hydrologic time series are based on a set of well defined hypotheses, with the assumption that development cannot be better than hypotheses underlying it. Techniques are presented on how to infer the existence of periodic deterministic component parameters in time series. The unavailability of exact inference techniques is replaced by approximations, whenever the complexity of hydrologic time series does not justify the use of existing statistical inference techniques. Once periodicities are inferred, Fourier analysis is used to mathematically describe periodicities in parameters by a minimum number of low frequency harmonics and their estimated coefficients.

KEY WORDS: Time series, hydrologic series, analysis of series, structural composition of series, hydrologic time processes.

ABSTRACT: Structural analysis and mathematical description of hydrologic time series are based on a set of well defined hypotheses, with the assumption that development cannot be better than hypotheses underlying it. Techniques are presented on how to infer the existence of periodic deterministic component parameters in time series. The unavailability of exact inference techniques is replaced by approximations, whenever the complexity of hydrologic time series does not justify the use of existing statistical inference techniques. Once periodicities are inferred, Fourier analysis is used to mathematically describe periodicities in parameters by a minimum number of low frequency harmonics and their estimated coefficients.

KEY WORDS: Time series, hydrologic series, analysis of series, structural composition of series, hydrologic time processes.

ABSTRACT: Structural analysis and mathematical description of hydrologic time series are based on a set of well defined hypotheses, with the assumption that development cannot be better than hypotheses underlying it. Techniques are presented on how to infer the existence of periodic deterministic component parameters in time series. The unavailability of exact inference techniques is replaced by approximations, whenever the complexity of hydrologic time series does not justify the use of existing statistical inference techniques. Once periodicities are inferred, Fourier analysis is used to mathematically describe periodicities in parameters by a minimum number of low frequency harmonics and their estimated coefficients.

KEY WORDS: Time series, hydrologic series, analysis of series, structural composition of series, hydrologic time processes.

ABSTRACT: Structural analysis and mathematical description of hydrologic time series are based on a set of well defined hypotheses, with the assumption that development cannot be better than hypotheses underlying it. Techniques are presented on how to infer the existence of periodic deterministic component parameters in time series. The unavailability of exact inference techniques is replaced by approximations, whenever the complexity of hydrologic time series does not justify the use of existing statistical inference techniques. Once periodicities are inferred, Fourier analysis is used to mathematically describe periodicities in parameters by a minimum number of low frequency harmonics and their estimated coefficients.

Inferred dependence models for the stationary stochastic components of a given order of stationarity, after all periodicities in parameters are removed, are basically of the autoregressive linear type. The assumption is that the autoregressive coefficients may be both periodic and nonperiodic. Several misinterpretations of autoregressive linear models are discussed.

The frequency distribution curve of the independent stochastic stationary component, derived from the inferred dependence model is approximated by best fit as one among various probability distribution functions studied.

Biases in time series which should not be reproduced or perpetuated by structural analysis, mathematical description and generation of new samples, are outlined and discussed.

REFERENCE: Yevjevich, Vujica, Colorado State University, Hydrology Paper No. 56 (October, 1972) "Structural Analysis of Hydrologic Time Series."

Inferred dependence models for the stationary stochastic components of a given order of stationarity, after all periodicities in parameters are removed, are basically of the autoregressive linear type. The assumption is that the autoregressive coefficients may be both periodic and nonperiodic. Several misinterpretations of autoregressive linear models are discussed.

The frequency distribution curve of the independent stochastic stationary component, derived from the inferred dependence model is approximated by best fit as one among various probability distribution functions studied.

Biases in time series which should not be reproduced or perpetuated by structural analysis, mathematical description and generation of new samples, are outlined and discussed.

REFERENCE: Yevjevich, Vujica, Colorado State University, Hydrology Paper No. 56 (October, 1972) "Structural Analysis of Hydrologic Time Series."

Inferred dependence models for the stationary stochastic components of a given order of stationarity, after all periodicities in parameters are removed, are basically of the autoregressive linear type. The assumption is that the autoregressive coefficients may be both periodic and nonperiodic. Several misinterpretations of autoregressive linear models are discussed.

The frequency distribution curve of the independent stochastic stationary component, derived from the inferred dependence model is approximated by best fit as one among various probability distribution functions studied.

Biases in time series which should not be reproduced or perpetuated by structural analysis, mathematical description and generation of new samples, are outlined and discussed.

REFERENCE: Yevjevich, Vujica, Colorado State University, Hydrology Paper No. 56 (October, 1972) "Structural Analysis of Hydrologic Time Series."

Inferred dependence models for the stationary stochastic components of a given order of stationarity, after all periodicities in parameters are removed, are basically of the autoregressive linear type. The assumption is that the autoregressive coefficients may be both periodic and nonperiodic. Several misinterpretations of autoregressive linear models are discussed.

The frequency distribution curve of the independent stochastic stationary component, derived from the inferred dependence model is approximated by best fit as one among various probability distribution functions studied.

Biases in time series which should not be reproduced or perpetuated by structural analysis, mathematical description and generation of new samples, are outlined and discussed.

REFERENCE: Yevjevich, Vujica, Colorado State University, Hydrology Paper No. 56 (October, 1972) "Structural Analysis of Hydrologic Time Series."