DISTRIBUTIONS OF HYDROLOGIC INDEPENDENT STOCHASTIC COMPONENTS

by
Pen-chih Tao, V. Yevjevich
and N. Kottegoda

January 1976

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

82

DISTRIBUTIONS OF HYDROLOGIC INDEPENDENT STOCHASTIC COMPONENTS

By
Pen-chih Tao, V. Yevjevich and N. Kottegoda

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80523

January 1976 No. 82

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The authors express their thanks to Dr. M. M. Siddiqui, Dr. M. C. Bryson, Professors, and Dr. J. Whittaker, Visiting Professor, of the Department of Statistics at Colorado State University for their valuable suggestions given during the studies leading to this paper. The financial support of the U.S. National Science Foundation, through the research grants GK-31512X and ENG74-17396, with V. Yevjevich as the principal investigator, is gratefully acknowledged.

ABSTRACT

The structural analysis and mathematical models used for evaluation and removal of the periodicity and dependence from the hydrologic time series are reviewed, summarized and discussed. Records of daily runoff for seventeen river basins in the United States are used as the basic research data and mathematical models were applied to analyze their periodicity and dependence. Periodicities in serial correlation coefficients, $r_{k,\tau}$, of the stochastic components are also analyzed and found to be not negligibly small. Independent stochastic components (or the residuals) are obtained by removing the periodicity and dependence from the daily runoff series. Methods of testing the distributions of the tails of empirical frequency distributions are developed. Tails do not seem to belong to the class of heavy tails. Seven groups of probability distribution functions: classical, Pearson's family, probability density functions modified by polynomials, Weibull, double-branch gamma, mixture of probability functions, and family of stable distribution functions, are applied to fit the frequency distributions of the independent stochastic variables. The same techniques of removing periodicity and dependence were applied to series with larger time intervals, such as the 3-day, 7-day, 13-day and monthly series, which are derived from the daily runoff series, and the distributions of these variables are compared. It was found that the 3-parameter lognormal function fits well the frequency distributions of monthly independent stochastic variables. Since the frequency distributions of variables with small time intervals are more skewed than for the large time interval series and since they have sharper peaks and longer tails, the probability distribution functions with more parameters should be used to fit these empirical distributions. For 13-day variables, the 3-parameter lognormal and the 3-parameter gamma functions are found to fit the frequency distributions quite well, while for 7-day and 3-day variables the double-branch gamma function with six parameters is found most applicable. However, no distribution is found to fit consistently the frequency distributions of daily variables, because of the sharp peak and high skewness of these empirical distributions; hence, an empirical method of fitting is suggested.

PREFACE

The hydrologic time processes, either continuous or discrete for time intervals of a fraction of the year, are composed processes (with periodic parameters, non-homogeneous and/or inconsistent parameters, and a stochastic component). When a sample of such a process is mixed by using known deterministic and stochastic components, its decomposition (disegregation, separation, structural analysis) never leads to exact characteristics of the individual components. Therefore, even if one starts with a normal independent process for the stochastic component, and mixes it with the periodic and/or trend parameters, the analysis of the sample rarely produces a conclusive evidence that the resulting stochastic residuals of the sample decomposition are normally distributed. Therefore, determination of probability distributions of obtained residuals in the form of independent stochastic component of complex hydrologic series is subject to bias and/or incorrect conclusions due to difficulties inherent in the decomposition.

The following paper had as an objective the analysis of stochastic residuals in their two aspects: (1) the complete distribution of residuals, and (2) the character of distribution tails. It was shown how difficult it is to fit simple probability distribution functions to residuals, because the inference on periodicity of parameters, on dependence model of remaining series after periodicity is removed, with an eventually unremoved nonhomogeneity and inconsistency in data, all lead to a mixed distribution of complex analytical expressions for residuals and not to simple functions as expected in practice. Stochastic residuals of daily flow series showed to be the most difficult to fit by simple probability distribution functions.

The analysis of tails showed, regardless of above difficulties in fitting residuals by simple probability distribution functions, that they are exponential. The over-removed and/or under-removed harmonics of periodic parameters tend to make tails heavier and peaks sharper, than the true distribution would show. Regardless of this, one may conclude from the research results given in this paper, that hydrologic independent stochastic components, represented in the paper by the independent residuals of daily runoff series, are exponential, and both, for the sharp rising left tail and the slow decreasing right tail of frequency density functions of residuals. This result does not support the theory of so-called heavy tails of stable distributions as the true characteristics of stochastic component of hydrologic time processes. Therefore, mathematical models based on the heavy-tail concept may be nothing else than a fit to biased estimations of independent stochastic components.

January 1976
Fort Collins, Colorado

Vujica Yevjevich
Professor-in-Charge
Hydrology and Water Resources Program

Chapter 1
INTRODUCTION

The prediction of future water supply is one of the basic goals of a program of water resources management. In the absence or inadequacy of physical theory concerning the atmosphere-earth water cycle relation, hydrologists are predisposed to use statistics and simulation.

The analysis of observed chronological sequences of physical phenomena lies within the time series analysis. The process could be either probabilistic in nature, or deterministic with stochastic components superimposed [35]. The deterministic and stochastic components of the process could be statistically analyzed using historical records.

In this chapter, general properties of a hydrologic process are briefly reviewed. The objectives and the procedures of this study are outlined, and the research data used are briefly described.

1-1 General Composition of Hydrologic Time Series.

The characteristics of hydrologic time series [34, 35] can be divided into long-range trends and other long-range persistencies, periodicities of the year, and randomness with time dependence in the stochastic variation. These characteristics are considered as basic components of hydrologic time series.

Long-range trends and other long-range persistencies. Trend is defined as a systematic and continuous change over an entire sample in any parameter of a series. Inconsistency (systematic error) and nonhomogeneity (changes in nature by either man-made or natural processes) are the main causative factors for the long-range trends or eventual long-range persistencies. They must be identified and removed before hydrologic structural analysis is initiated. Trends and cyclicities may often result from sampling fluctuations in short time series. When cyclicity is only a result of sampling variations, it is called the sampling cyclicity. For a regional study it is necessary to determine whether there is any significant trend or cyclicity to be assigned to a particular series in the area. In addition, different hydrological, meteorological and geophysical time series may be compared. Without such a broadly based confirmation the apparent long-range trends and cyclicity should not be considered as permanent properties of any series of annual values of a hydrologic variable, even though the factors of the known nonhomogeneity and inconsistency are removed. Consequently, those factors which are the result of sampling variation, should not be perpetuated in structural analysis and mathematical description of time series.

Within-the-year periodicity. Astronomic cycles produce the periodicities in various hydrologic time series. In a given river basin, for example, high precipitation in summer seasons and low precipitation in winter seasons, or vice versa, may be expected. River flows are high or low in different seasons. Means and variances of stream flows are large in wet seasons and small in dry seasons. This phenomenon indicates the within-the-year periodicity in the means and variances. Usually, periodicities in a hydrologic time series would appear in one, two or several of its parameters, especially in the means, standard deviations and autocorrelation coefficients. Periodic components are deterministic properties of time series and their parameters.

Randomness and time dependence. Randomness of hydrologic time series is caused by such factors as turbulence, large-scale vorticities, heat transfer, air opacity for radiation waves and other sources of randomness of atmospheric, oceanic and continental air and water movements. Time dependence in stochastic variation is created or increased by water storage of various types in hydrologic environments. A stationary stochastic process is assumed superimposed on a periodic or deterministic process in a given manner which can be described by an algebraic equation of time series composition. Therefore, hydrologic time series are basically nonstationary and could be decomposed into deterministic components and a stationary stochastic process.

Definition of independent stochastic components. The above assumption of periodic components in a hydrologic time series being deterministic parameter processes implies that they can be removed by means of appropriate mathematical models. The remaining stationary stochastic component is in general sequentially dependent. This dependence is often found to be approximately linear. Linear models such as the autoregressive type are commonly used in hydrology. Residuals in mathematical models of sequential dependence of stationary stochastic series are called the independent stochastic component, the ξ variable. For a discrete time series, the independent stochastic components are designated by $\xi_{p,\tau}$, for which p and τ indicate the τ-th discrete time interval position from a total of ω positions in one cycle for the p-th period. One calendar year is, for example, one period; ω is the number of interval subdivisions of the year, i.e. $\omega = 365$ for daily series, $\omega = 12$ for monthly series, etc., with $\tau = 1,2,\ldots,\omega$, as discrete values of the basic period; $p = 1,2,\ldots,n$, is the successive year number and n is the number of years in the sample. The $\xi_{p,\tau}$ variables are assumed initially to be independent and identically distributed at all positions τ of the period ω. These $\xi_{p,\tau}$ variables should be nearly stationary and sequentially independent for subsequent analysis and statistical inference.

1-2 Study Objectives.

The objectives of this study are:

1. To select mathematical models with the related error analysis to be employed in the structural analysis of hydrologic time series.

2. To condense the information in the data of the independent stochastic component by fitting appropriate probability distribution functions to their frequency distributions.

3. To study the extreme or unusual events of hydrologic time series, which affect the selection of probability distribution functions for the independent stochastic components. An exceptionally high flood, for example, may produce a correspondingly very high value in the sample data of independent stochastic components. One of the objectives of this study is to find the best way of treating these extreme values of the independent stochastic components, by a proper selection or a modification of the probability distribution functions selected to fit their frequency distribution curves.

4. To apply the central limit theorem so that the sums of random variables become asymptotically normally distributed, and in such a way that the probability distribution functions selected for the small-interval independent stochastic variables converge to the normal function as the Δt-interval increases. For the random variables following a stable distribution with the characteristic exponent, α, the distribution of sums of these random variables is also a stable distribution with the same characteristic exponent, α. For the variables to follow a stable distribution requires that an independent stochastic variable has the same characteristic exponent, α, for all interval values of Δt. Another objective of this study is to investigate jointly the applicability of the central limit theorem and the stable distributions to hydrology.

1-3 Procedures of Investigation.

The approach for this investigation is to study both the structure of deterministic components and the sequential dependence of resulting stochastic series. Statistical tests are used to infer the independence of resulting stochastic variables, to allow testing the adequacy of mathematical models used. In studying the frequency distribution of the independent stochastic variables, seven groups of probability distribution functions are studied and tested: classical probability distribution functions, Pearson's family of probability distribution functions, Weibull probability distribution functions, classical probability distribution functions modified by the polynomials, double-branch gamma functions, and a mixture of some probability distribution functions and stable distributions. The procedures used are shown schematically in Fig. 1-1.

1-4 Research Data.

The data used in this study [29] contain seventeen sets of daily runoff series, with the 3-day, 7-day, 13-day, and monthly runoff series derived from the daily series by taking averages of the daily values over the intervals of 3 days, 7 days, 13 days, and one month.

These 17 daily runoff series are from the records published by the U.S. Geological Survey under the condition that the flows are virgin or have not been altered by significant man-made diversions or flow regulations. Minor diversions, up to a maximum of one percent of the average annual flow, are tolerated. The names of stations selected are given in Table 1-1, with the approximate geographic location of stations shown in Fig. 1-2.

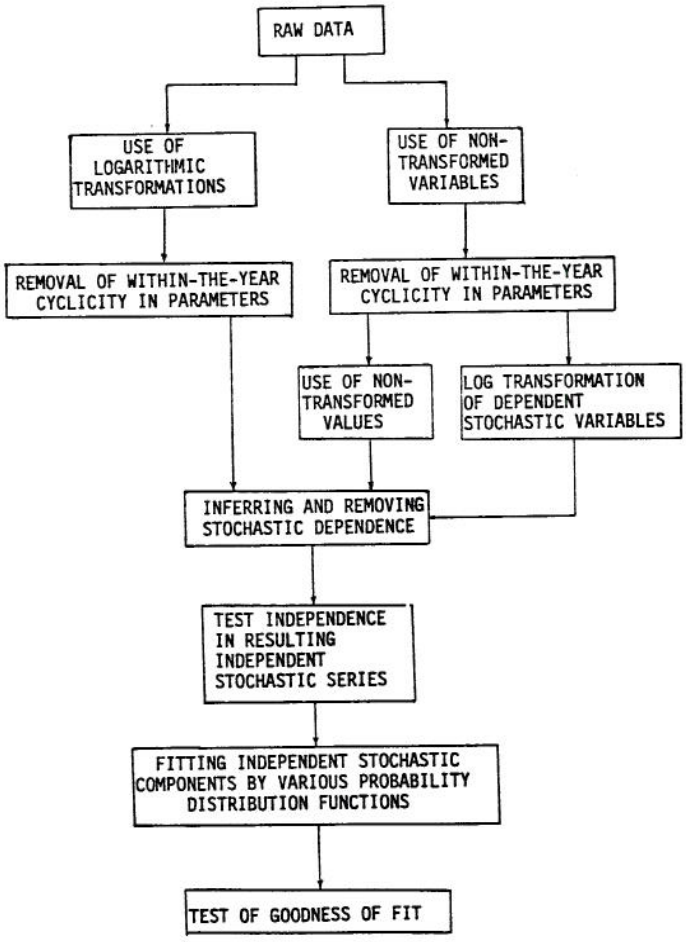

Fig. 1-1 Schematic flow chart of investigations.

Table 1-1 STATIONS SELECTED FOR INVESTIGATION.

Station Number	River	Location		Area (Sq. Mi.)	Records Available	Mean Daily Flow	Standard Deviation	Remarks on Accuracy of Record *
		Latitude	Longitude					
1B.5265	Tioga near Erwins, N. Y.	42°07'	77°08'	1370.0	1921-1960	1378.6	2777.8	Excellent. Fair during periods of ice effect.
4.0710	Oconto near Gillett, Wisconsin	44°52'	88°18'	678.0	1921-1960	543.5	441.0	Good. Fair during periods of ice effect.
7.0670	Current at Van Buren, Mo.	37°00'	91°01'	1667.0	1922-1960	1921.0	2694.3	Good. Poor during periods of ice effect.
14.1590	Mckenzie at Mckenzie Br., Ore.	44°11'	122°08'	345.0	1924-1960	1638.2	744.4	Excellent
8.0335	Neches near Rockland, Tex.	31°02'	94°24'	3539.0	1924-1960	2385.2	3813.0	Good
13.1850	Boise near Twin Springs, Idaho	43°40'	115°44'	830.0	1921-1960	1172.7	1458.6	Excellent. Good during periods of ice effect.
11.2750	Falls Creek near Hetch-hetchy, Cal.	37°58'	119°46'	45.2	1923-1960	141.2	234.2	Good. Fair during periods of ice effect.
3A.1835	Greenbrier near Alderson, W. Va.	37°44'	80°38'	1357.0	1921-1960	1885.5	3053.4	Good. Poor during periods of ice effect.
6B.8905	Delaware at Valley Falls, Kansas	39°21'	95°27'	922.0	1923-1960	375.9	1617.7	Good. Fair during periods of ice effect.
6A.0375	Madison near W. Yellowstone, Mont.	44°39'	111°04'	419.0	1924-1960	458.6	190.7	Excellent. Good during periods of ice effect.
3B.5320	Powell near Arthur, Tenn.	36°32'	83°38'	683.0	1921-1960	1116.1	1739.0	Good
12.4150	St. Maries near Lotus, Idaho	47°15'	116°38'	437.0	1923-1960	515.0	762.3	Good. Poor during periods of ice effect.
2A.0160	Cowpasture near Clifton Forge, Va.	37°48'	79°46'	456.0	1926-1960	515.6	762.3	Good
3A.2695	Mad near Springfield, Ohio	39°55'	83°52'	1474.0	1921-1960	487.2	686.7	Good
11.2665	Merced at Pohono Br., Yosemite, Cal.	37°43'	119°40'	321.0	1921-1960	595.7	979.4	Good
1B.3295	Batten kill at Battenville, N. Y.	43°06'	73°25'	394.0	1923-1960	722.9	722.9	Good. Fair during periods of ice effect.
5.3620	Jump near Sheldon, Wisconsin	45°18'	90°57'	574.0	1921-1960	505.0	1162.0	Good. Fair during periods of ice effect.

*According to U.S.G.S., the classification of the records are excellent, good, fair, or poor depending on whether errors in them are less than 5, 10, or 15 percent or greater than 15 percent, respectively.

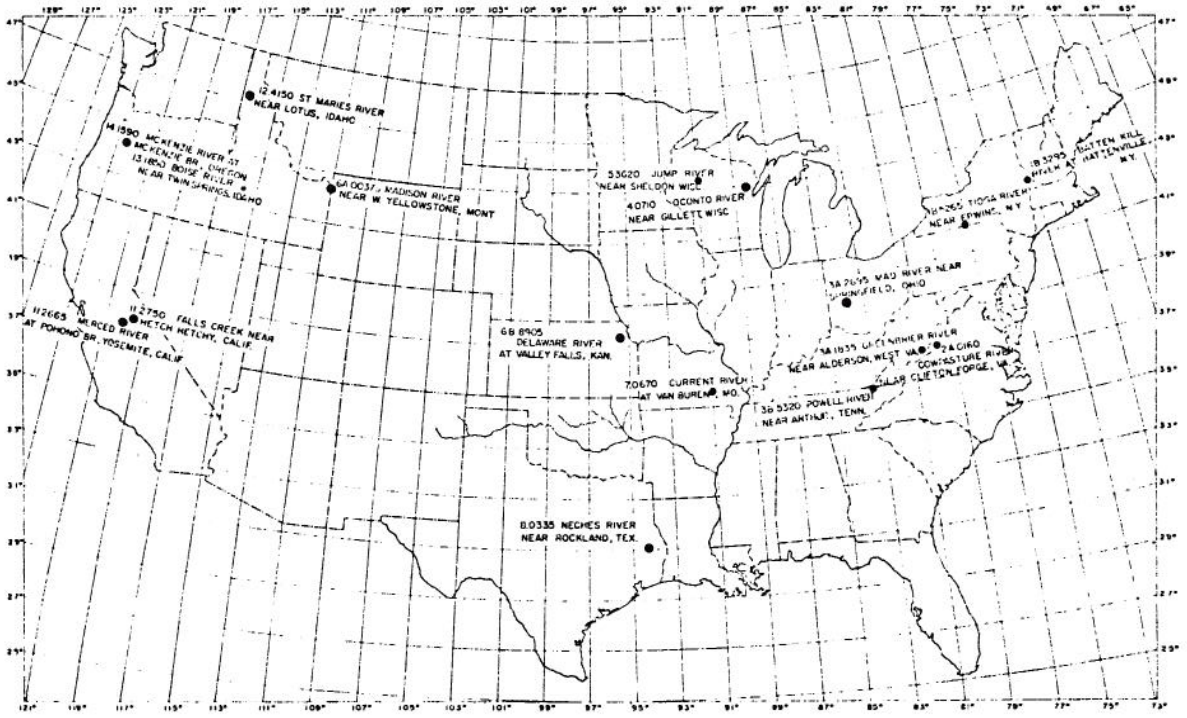

Fig. 1-2 Geographic distribution runoff stations used in the study.

Chapter 2
BRIEF REVIEW OF PERTINENT LITERATURE

The first part of this chapter includes the mathematical models used for identifying and describing the deterministic characteristics of the hydrologic time series. The second part involves the probability distribution functions applied to fitting the independent stochastic hydrologic components of time series.

2-1 Mathematical Models of Periodicity and Dependence in Hydrologic Time Series.

Early attempts of modelling the hydrologic events were made by Sudler [32] and Barnes [2]. Sudler's model was simply to collect all the historical annual series followed by a random rearrangement of the order of these events in order to obtain a sequence of new combinations of the original series. Barnes assumed the observed annual flows were normally distributed. He used a table of random numbers to synthesize the annual flows with the same mean and standard deviation as the original record. Although Barnes' work made an improvement on Sudler's procedure, it still neglected the serial correlation that usually exists in annual flows.

Thomas and Fiering [19] used the first-order autoregressive linear model to approximate the persistence in the series of monthly runoff. In providing a justification for the application of autoregressive linear models to hydrologic series, Yevjevich [34] showed that, if a simple exponential function will fit the recession curves of the runoff, the persistence among the annual runoff values follows the first-order linear autoregressive model. However, recession curves following the form; $Q_t = Q_o\, e^{-ct^n}$ (with c and n constants, Q_o the initial runoff of the recession curve, and Q_t the recession discharge at the time t) will be fit better by the second-order or third-order autoregressive linear models. Chow and Ramaseshen [6] found that autoregressive linear models were applicable to storm rainfall.

All these applications of the autoregressive linear models are valid under the assumption that the processes studied are stationary stochastic processes. Since the solar radiation input to hydrologic environments is periodic, the periodicity in the environmental hydrologic inputs and outputs cannot be avoided. For example, the correlogram of monthly stream flows is often periodic. Yevjevich [17] showed that for the monthly stream flow transformed by $\varepsilon_{p,\tau} = (Q_{p,\tau} - Q_\tau)/s_\tau$, (with $Q_{p,\tau}$ the stream flow of month τ, and s_τ the standard deviation of the month τ stream flow), the correlogram of $\varepsilon_{p,\tau}$ may not be periodic. Quimpo [28] used the model of $Q_t = P_t + \varepsilon_t$ (with P_t the periodic component; and ε_t the stochastic component) to approximate a periodic hydrologic process Q_t. He found that an approximate second-order stationary stochastic series is usually obtained by identifying and removing the periodicities in the mean and standard deviation. The dependence of stochastic components of daily flows could be approximated well by the second-order autoregressive linear model.

In order to identify and remove the periodicities in parameters, the harmonic analysis is commonly used. In spite of availability of some statistical methods for testing the significance of harmonics, these theoretical test methods show difficulties when applied to complex hydrologic time series. Quimpo [28] assumed that only the first six harmonics are potentially significant, and used the variance explained by each of these six harmonics as a criterion to determine how many harmonics are necessary to approximate the periodic components. Yevjevich [36] improved this procedure of testing the significance of harmonics by using the sum of the explained variance of all harmonics and two predefined critical values, as an empirical method for testing the significance of harmonics.

2-2 Probability Distribution Functions for Fitting Frequency Distributions of Hydrologic Stochastic Components.

The normal density function has been used for centuries as the theoretical probability distribution of error residuals. The gamma and lognormal probability density functions are widely used in hydrology and related fields. Markovic [20] investigated the distribution of annual precipitation and runoff series and pointed out that their frequency distributions may be fitted by normal, lognormal and gamma probability density functions. Matalas [22] assumed the historical series to follow either gamma or lognormal distributions and developed the simulation schemes with the condition that the residuals of the dependence models are normally distributed. Attempts were made by Bonne [3] to use the lognormal, Pearson Type III (gamma probability density function), and logarithmic Pearson Type III distribution functions to fit the frequency distributions of monthly runoff. Normal variates are often obtained by various transformations which depend on the original distributions of stream flows.

Distribution functions of more complex forms have been used in fitting the frequency distributions of hydrologic variables, when a simple probability density function fails. The normal function, modified by using the Hermite polynomials given by Edgeworth [23], and the gamma function modified by using the Laguerre polynomials investigated by Llamas and Siddiqui [18] are examples of more complex probability functions.

The mixture of two or more probability density functions, under the assumption that the variables result from two or more populations, is also used to fit the frequency distributions. The mixture of two normal density functions was first studied by Karl Pearson in 1894, while the parameter estimation for this mixture was improved and summarized by Cohen [7].

Bryson [5] suggested that a probability density function becomes "heavy-tailed" when it converges to zero less rapidly than an exponential probability density function. He also found that some hydrologic variables may follow the probability distributions which fall within the heavy-tailed category of functions.

Mandelbrot [21] and Fama [9] applied the stable probability distributions to economic variables such as the stock prices. Since the independent stochastic components, derived for the daily flow series, may have the similar properties as the stable distributions with heavy tails, the stable distributions are investigated also in this study.

Chapter 3
MATHEMATICAL MODELS USED FOR OBTAINING INDEPENDENT STOCHASTIC COMPONENTS, AND RELATED ERROR ANALYSIS

Mathematical models used for the periodic components of hydrologic time series and the time dependence of the stochastic components are summarized in this chapter, which served to obtain the independent stochastic variables. Even though the parameters of mathematical models are estimated from the sample data series by the best available estimation methods, they are inevitably subject to various sources of errors. Therefore, the analysis of these errors and their propagation through each step in obtaining the independent stochastic components is also presented in this chapter.

3-1 Removal of Periodicity in Parameters.

Two practical methods (nonparametric and parametric) may be used to remove the periodic components from a time series. The errors involved in the harmonic analysis of these periodic components are discussed.

The symbol $Q_{p,\tau}$ stands for discrete values of an observed hydrologic time series with τ and p previously defined. The symbol ν_τ denotes any periodic parameter to be estimated from the $Q_{p,\tau}$ series.

The nonparametric method. The removing of periodicities in the mean and variance of $Q_{p,\tau}$ is by

$$\varepsilon_{p,\tau} = \frac{Q_{p,\tau} - Q_\tau}{s_\tau}, \qquad (3\text{-}1)$$

in which $\varepsilon_{p,\tau}$ is the stochastic component of $Q_{p,\tau}$, Q_τ and s_τ are the sample mean and sample standard deviation of $Q_{p,\tau}$ at the position τ. Q_τ and s_τ are computed by

$$Q_\tau = \frac{1}{n}\sum_{p=1}^{n} Q_{p,\tau}, \qquad (3\text{-}2)$$

$$s_\tau = \left[\frac{1}{n-1}\sum_{p=1}^{n}(Q_{p,\tau} - Q_\tau)^2\right]^{\frac{1}{2}} \qquad (3\text{-}3)$$

The nonparametric method of removing the periodicities in Q_τ and s_τ by Eq. 3-1 is equivalent to the standardization of $Q_{p,\tau}$ variable. It requires 2ω statistics: ω of Q_τ and ω of s_τ. The nonparametric method is useful in any preliminary analysis when ω is small. For monthly series, Eq. 3-1 is often used because $2\omega = 24$ is not a very large number of statistics. However, there are two major drawbacks in using the nonparametric method:

1. In case of daily series, the number of statistics of Eq. 3-1 becomes $2\omega = 730$. When some other parameters of daily series, such as the autocorrelation coefficients, or the skewness coefficients are also periodic, and the periodicities must be identified, described and removed to obtain a second-order or third-order stationary stochastic component of daily values, the total number of statistics in nonparametric approach is drastically increased. Since it is impossible to estimate so many parameters accurately from a limited size of sample series, these parameters must be subject to large sampling errors.

2. Sampling errors are propagated by the nonparametric method.

The general objectives of mathematical modelling of the deterministic-stochastic process are: (a) To effectively separate the deterministic and stochastic components; and (b) To condense the information by employing such models which use the number of parameters parsimoniously. Since the nonparametric method does not satisfy these objectives, at least for large values of ω, this method, therefore, is not used in the further investigations.

The parametric method. Two types of periodic functions can be used in the description of periodic parameters, ν_τ, with the Fourier coefficients of significant harmonics either constants or also random variables. The first is

$$\nu_\tau = \nu_x + \sum_{j=1}^{m} C_j \cos\left(\frac{2\pi j\tau}{\omega} + \theta_j\right), \qquad (3\text{-}4)$$

and the second is

$$\nu_\tau = \nu_x + \sum_{j=1}^{m_o} C^j_{p,\tau} \cos\left(\frac{2\pi j\tau}{\omega} + \theta^j_{p,\tau}\right), \qquad (3\text{-}5)$$

in which ν_x is the mean of ν_τ, C_j and θ_j the constant amplitude and phase of the j-th harmonics, $C^j_{p,\tau}$ and $\theta^j_{p,\tau}$ the random amplitude and random phase of the j-th harmonic, and $j = 1,2,\ldots,m$ and $j = 1,2,\ldots,m_o$ the number of harmonics in the two cases. Once ν_x, C_j, and θ_j, with $j = 1,2,\ldots,m$, are estimated from the sample series, Eq. 3-4 uniquely defines the ν_τ values at any τ. However, ν_x is the only constant of the whole process in Eq. 3-5, with the m_o pairs of random amplitudes and phases for m_o harmonics. The random variables $C^j_{p,\tau}$ and $\theta^j_{p,\tau}$ may not only be serially but also mutually correlated.

The first periodic function, Eq. 3-4, is based on the assumption that the deterministic periodic parameters of the process can be separated, with the residuals considered as the stochasticity of the process. The second periodic function, Eq. 3-5, means that nearly all the information is contained in various stochastic parts, such as the random Fourier coefficients and the random variable of the stochastic part.

The hypothesis that all of the earth's hydrologic processes are composed of deterministic-periodic parameters and stochastic component seems supported by the basic periodic influx of solar energy. The environmental responses to solar energy input modify certain properties of input cyclicity without changing the cyclicity itself, adding randomness to the output.

It seems logical on the basis of geophysical evidence to conceive the responses of hydrologic environments as producing the outputs which are composed of periodic parameters and a stochastic component than to conceive that the responses produce only a set of stochastic variables in output. Therefore, Eq. 3-4 is used to describe the periodic parameters of hydrologic time series.

An alternative form to Eq. 3-4 is

$$v_\tau = v_x + \sum_{j=1}^{m} \left(A_j \cos \frac{2\pi j\tau}{\omega} + B_j \sin \frac{2\pi j\tau}{\omega} \right), \qquad (3\text{-}6)$$

with m the number of significant harmonics, and A_j and B_j the Fourier coefficients estimated from the ω values of V_τ (with v_τ sample values), by

$$A_j = \frac{2}{\omega} \sum_{\tau=1}^{\omega} V_\tau \cos \frac{2\pi j\tau}{\omega} \qquad (3\text{-}7)$$

$$B_j = \frac{2}{\omega} \sum_{\tau=1}^{\omega} V_\tau \sin \frac{2\pi j\tau}{\omega} \quad , \qquad (3\text{-}8)$$

with the amplitude and phase

$$C_j = \sqrt{A_j^2 + B_j^2} \quad ; \text{ and } \theta_j = \tan^{-1}\left(-\frac{B_j}{A_j} \right) . \qquad (3\text{-}9)$$

The empirical method is used in selecting the significant harmonics. Let $s^2(v_\tau)$ be the variance of computed v_τ. For a harmonic j, var $h_j = (A_j^2 + B_j^2)/2$, with A_j and B_j computed by Eqs. 3-7 and 3-8. The ratio

$$\Delta P_j = \frac{\text{var } h_j}{s^2(v_\tau)} \qquad (3\text{-}10)$$

represents the part of the variance of v_τ explained by the j-th harmonic. These ratios ΔP_j are then ordered in a descending sequence, and summed to

$$P_j = \sum_{i=1}^{j} \Delta P_i \quad , \text{ for } j = 1,2,3,\ldots,m \quad , \qquad (3\text{-}11)$$

with $\max(m) = \omega/2$. The harmonics of any parameter of hydrologic daily series, such as the daily mean, daily standard deviation, or daily autocorrelation coefficient, are usually not significant beyond the first six harmonics. The critical values of the sequence P_j are two empirical constants, P_{min} and P_{max}. With the critical values P_{min} and P_{max} selected, the criteria used to determine the significant harmonics are: (a) if $P_6 \le P_{min}$, no harmonic is significant, or $v_\tau = v_x$ as a constant; (b) if $P_{min} < P_6 \le P_{max}$, all six harmonics are significant; and (c) if $P_6 > P_{max}$, the first j harmonics, whose P_j values first exceed P_{max}, are considered as significant. The empirical expression of P_{min} and P_{max} are:

$$P_{min} = a\sqrt{\frac{\omega}{nc}} \quad , \qquad (3\text{-}12)$$

and

$$P_{max} = 1 - P_{min} \, , \qquad (3\text{-}13)$$

in which c is the order of the highest moment used in estimating the parameter v_τ, ω and n are the length of basic period and the number of years in record, and a is a constant (with the suggested value $a = 0.033$).

With the periodic mean μ_τ and the periodic standard deviation σ_τ, the model for $Q_{p,\tau}$ is

$$Q_{p,\tau} = \mu_\tau + \sigma_\tau \varepsilon_{p,\tau} \, , \qquad (3\text{-}14)$$

with $\varepsilon_{p,\tau}$ the stochastic component, which is stationary at least in the mean and standard deviation.

The procedure for inferring the periodic mean and standard deviation and removing them from the sample $Q_{p,\tau}$ series, is as follows:

a. Estimate the sample means Q_τ and the sample standard deviations s_τ, $\tau = 1,2,\ldots,\omega$, by using $Q_{p,\tau}$ and Eqs. 3-2 and 3-3, respectively.

b. Replace v_τ in Eqs. 3-7 and 3-8 by m_τ and s_τ, and compute the Fourier Coefficients A_j and B_j, for $j = 1,2,\ldots,6$, respectively.

c. Use the empirical test for the significant harmonics in m_τ and s_τ.

d. Denote by μ_τ and σ_τ the periodic parts in m_τ and s_τ, respectively, the equations to estimate μ_τ and σ_τ are

$$\mu_\tau = m_x + \sum_{i}^{m_1} \left(A_i \cos \frac{2\pi\tau i}{\omega} + B_i \sin \frac{2\pi\tau i}{\omega} \right) \, , \qquad (3\text{-}15)$$

and

$$\sigma_\tau = s_x + \sum_{i}^{m_2} \left(A_i \cos \frac{2\pi\tau i}{\omega} + B_i \sin \frac{2\pi\tau i}{\omega} \right) \, , \qquad (3\text{-}16)$$

with m_x and s_x the averages of m_τ and s_τ, and m_1 and m_2 the number of significant harmonics in μ_τ and σ_τ, respectively.

e. The periodic components are removed from $Q_{p,\tau}$ by

$$y_{p,\tau} = \frac{Q_{p,\tau} - \mu_\tau}{\sigma_\tau} \, . \qquad (3\text{-}17)$$

Equation 3-17 is similar to Eq. 3-1 with the periodic models having a limited number of harmonics, as the parametric method. The $y_{p,\tau}$ variable of Eq. 3-17 is approximately standardized. With a further transformation,

$$\varepsilon_{p,\tau} = \frac{y_{p,\tau} - \mu_y}{\sigma_y} \, , \qquad (3\text{-}18)$$

the standardized variable $\varepsilon_{p,\tau}$ is obtained, with μ_y and σ_y the mean and standard deviation of $y_{p,\tau}$ variable.

Error analysis. Since the true number of harmonics, necessary to describe the periodicity in v_τ, is unknown, and since the Fourier coefficients estimated by Eqs. 3-7 and 3-8 are subject to the sampling errors in v_τ, the harmonic analysis in inferring the periodic components produces the following errors:

a. The number of significant harmonics inferred may be either underestimated or overestimated in comparison with the number of harmonics necessary for describing the true periodicity in v_τ. This kind of error results from the deficiency in the testing method for significant harmonics, and the sampling errors in v_τ, which distort the test decisions.

b. Due to the sampling errors in v_τ, any inferred significant harmonic may either underestimate or overestimate the explained variance of v_τ by that harmonic.

Considering the two sources of errors, Eq. 3-6 can be modified as

$$v_\tau \pm \Delta v_\tau = v_x + \sum_{j=1}^{m} \left(A_j \cos \frac{2\pi j\tau}{\omega} + B_j \sin \frac{2\pi j\tau}{\omega}\right) \pm \sum_{i}^{k} \left(A_i \cos \frac{2\pi j\tau}{\omega} + B_i \sin \frac{2\pi j\tau}{\omega}\right) + \sum_{j=1}^{m} \left(\pm\Delta A_j \cos \frac{2\pi j\tau}{\omega} \pm \Delta B_j \sin \frac{2\pi j\tau}{\omega}\right), \quad (3\text{-}19)$$

with m the inferred number of significant harmonics, k the number of harmonics which should or should not be included into m selected harmonics, ΔA_j and ΔB_j the errors in the Fourier coefficients of selected significant harmonics, and $\pm\Delta v_\tau$ the total error in v_τ. The errors made in the harmonic analysis greatly affect the accuracy of v_τ; for example, if a harmonic explains only one percent of the total variance of v_τ, and were incorrectly treated by the test of significance, such as

$$\Delta v_\tau = \pm \left(A_k \cos \frac{2\pi k\tau}{\omega} + B_k \sin \frac{2\pi k\tau}{\omega}\right)$$

and

$$\frac{1}{2}\left(A_k^2 + B_k^2\right) = 0.01\ \sigma_v^2 ,$$

with σ_v^2 the variance of v_τ. If A_k and B_k are equal, then $A_k = B_k = 0.1\ \sigma_v$, when $2\pi k\tau/\omega = \pi/4$, and Δv_τ will be equal to $0.14\ \sigma_\tau$. The errors made in the estimated harmonics are propagated into the stochastic component.

3-2 Dependence Models of Stationary Stochastic Components.

The $\varepsilon_{p,\tau}$ variable obtained by removing the periodicities in the mean and standard deviation from $Q_{p,\tau}$, is approximately a second-order stationary time series, provided the autoregression coefficients are not periodic. The dependence models for $\varepsilon_{p,\tau}$ may be autoregressive of the moving average type, a combination of the two, and of other schemes, even of a nonlinear regression. The autoregressive linear models are found to be most useful in hydrology.

Moving average and linear autoregressive schemes. Two types of linear equations, or their combinations, are used to describe the dependence in hydrologic stochastic series such as $\varepsilon_{p,\tau}$ variable. It is assumed that each value of $\varepsilon_{p,\tau}$ is only a combined effect of previous values of the independent stochastic component, $\xi_{p,\tau}$. Symbolically,

$$\varepsilon_{p,\tau} = \xi_{p,\tau} - \sum_{i=1}^{\infty} b_i\ \xi_{p,\tau-i} \quad , \quad (3\text{-}20)$$

is the scheme of moving averages, or the linear autoregression

$$\varepsilon_{p,\tau} = \sum_{i=1}^{\infty} a_i\ \varepsilon_{p,\tau-i} + \xi_{p,\tau} \quad . \quad (3\text{-}21)$$

Theoretically, an infinite sum of Eqs. 3-20 and 3-21 may be needed. However, as the effect of previous values on the current value decreases with time in all processes, a finite sum is often sufficient to approximate the dependence process. In other words, if the degree of required precision is fixed, only a finite number of terms in Eqs. 3-20 and 3-21 are necessary, with the alternative representation

$$\varepsilon_{p,\tau} = \xi_{p,\tau} - \sum_{i=1}^{q} b_i\ \xi_{p,\tau-i} \quad , \quad (3\text{-}22)$$

and

$$\varepsilon_{p,\tau} = \sum_{i=1}^{p} a_i\ \varepsilon_{p,\tau-i} + \xi_{p,\tau} \quad , \quad (3\text{-}23)$$

with p and q finite. The b_i coefficients of Eq. 3-22 are related to coefficients a_i of Eq. 3-23 with one set defining the other [28]. It is easy to show that a finite scheme of the moving average process can be converted into an infinite autoregressive process and vice versa.

Sometimes, both the autoregressive and moving average terms are combined in a model

$$\xi_{p,\tau} = \varepsilon_{p,\tau} - \sum_{i=1}^{p} a_i\ \varepsilon_{p,\tau-i} + \sum_{j=1}^{q} b_j\ \xi_{p,\tau-j} \quad , \quad (3\text{-}24)$$

as a mixed autoregressive-moving average scheme of the order (p,q), abbreviated by ARMA (p,q) [4].

The application of autoregressive linear models. The dependence of the stochastic hydrologic series can be approximated by various orders of linear autoregressive models. The first-, second- and third-order autoregressive linear models are most commonly applied rather than the higher-order models. These higher-order models may show advantage only in case of very long sample series. Short samples hardly justify the application of higher-order autoregressive linear models [36]. Linear models seem sufficiently accurate for all practical purposes even though the true physical stochastic dependence may be nonlinear.

The general m-th order autoregressive linear model is

$$\varepsilon_{p,\tau} = \sum_{k=1}^{m} \alpha_{k,\tau}\, \varepsilon_{p,\tau-k} + \sigma_{\xi,\tau}\, \xi_{p,\tau} , \qquad (3\text{-}25)$$

with $\alpha_{k,\tau}$ the autoregression coefficients as functions of serial correlation coefficients $\rho_{k,\tau}$, either periodic or nonperiodic, $\sigma_{\xi,\tau}$ the standard deviation of $\xi_{p,\tau}$ which is periodic if $\alpha_{k,\tau}$ are periodic, with $\xi_{p,\tau}$ a standardized variable. The serial correlation coefficient $\rho_{k,\tau}$ of lag k, is

$$\rho_{k,\tau} = \frac{\text{cov}\,(\varepsilon_{p,\tau}, \varepsilon_{p,\tau+k})}{\text{var}\,(\varepsilon_{p,\tau})} \qquad (3\text{-}26)$$

The coefficients $\sigma_{\xi,\tau}$ and $\alpha_{k,\tau}$, with k = 1,2, and 3, are: For the first order model,

$$\alpha_{1,\tau} = \rho_{1,\tau-1} , \qquad (3\text{-}27)$$

with

$$\sigma_{\xi,\tau} = (1 + \alpha_{1,\tau}^2 - 2\alpha_{1,\tau}\, \rho_{1,\tau-1})^{\frac{1}{2}} . \qquad (3\text{-}28)$$

For the second order model,

$$\alpha_{1,\tau} = \frac{\rho_{1,\tau-1} - \rho_{1,\tau-2}\, \rho_{2,\tau-2}}{1 - \rho_{1,\tau-2}^2} , \qquad (3\text{-}29)$$

and

$$\alpha_{2,\tau} = \frac{\rho_{2,\tau-2} - \rho_{1,\tau-1}\, \rho_{1,\tau-2}}{1 - \rho_{1,\tau-2}^2} , \qquad (3\text{-}30)$$

with

$$\sigma_{\xi,\tau} = (1 + \alpha_{1,\tau}^2 + \alpha_{2,\tau}^2 - 2\,\alpha_{1,\tau}\, \rho_{1,\tau-1} - 2\,\alpha_{2,\tau}\, \rho_{2,\tau-2} + 2\,\alpha_{1,\tau}\, \alpha_{2,\tau}\, \rho_{1,\tau-2})^{\frac{1}{2}} . \qquad (3\text{-}31)$$

For the third order model,

$$\alpha_{1,\tau} = \frac{\rho_{1,\tau-2}(1-\rho_{1,\tau-2}^2) + \rho_{1,\tau-3}\rho_{1,\tau-2}\rho_{3,\tau-3} - \rho_{1,\tau-2}\rho_{2,\tau-2} - \rho_{2,\tau-3}\rho_{3,\tau-3}}{1 + 2\rho_{1,\tau-2}\rho_{2,\tau-3}\rho_{1,\tau-3} - \rho_{1,\tau-3}^2 - \rho_{1,\tau-2}^2 - \rho_{2,\tau-3}^2} + \frac{\rho_{1,\tau-3}\rho_{2,\tau-2}\rho_{2,\tau-3}}{1 + 2\rho_{1,\tau-2}\rho_{2,\tau-3}\rho_{1,\tau-3} - \rho_{1,\tau-3}^2 - \rho_{1,\tau-2}^2 - \rho_{2,\tau-3}^2} , \qquad (3\text{-}32)$$

$$\alpha_{2,\tau} = \frac{\rho_{2,\tau-2}(1-\rho_{2,\tau-3}^2) + \rho_{1,\tau-2}\rho_{2,\tau-3}\rho_{3,\tau-3} - \rho_{1,\tau-2}\rho_{1,\tau-1} - \rho_{1,\tau-3}\rho_{3,\tau-3}}{1 + 2\rho_{1,\tau-2}\rho_{2,\tau-3}\rho_{1,\tau-3} - \rho_{1,\tau-3}^2 - \rho_{1,\tau-2}^2 - \rho_{2,\tau-3}^2} + \frac{\rho_{1,\tau-3}\rho_{2,\tau-3}\rho_{1,\tau-1}}{1 + 2\rho_{1,\tau-2}\rho_{2,\tau-3}\rho_{1,\tau-3} - \rho_{1,\tau-3}^2 - \rho_{1,\tau-2}^2 - \rho_{2,\tau-3}^2} , \qquad (3\text{-}33)$$

and

$$\alpha_{3,\tau} = \frac{\rho_{3,\tau-3}(1-\rho_{1,\tau-2}^2) + \rho_{1,\tau-3}\rho_{1,\tau-2}\rho_{1,\tau-1} - \rho_{1,\tau-3}\rho_{2,\tau-2} - \rho_{2,\tau-3}\rho_{1,\tau-1}}{1 + 2\rho_{1,\tau-2}\rho_{2,\tau-3}\rho_{1,\tau-3} - \rho_{1,\tau-3}^2 - \rho_{1,\tau-2}^2 - \rho_{2,\tau-3}^2} + \frac{\rho_{1,\tau-2}\rho_{2,\tau-2}\rho_{2,\tau-3}}{1 + 2\rho_{1,\tau-2}\rho_{2,\tau-3}\rho_{1,\tau-3} - \rho_{1,\tau-3}^2 - \rho_{1,\tau-2}^2 - \rho_{2,\tau-3}^2} , \qquad (3\text{-}34)$$

with

$$\sigma_{\xi,\tau} = (1 + \alpha_{1,\tau}^2 + \alpha_{2,\tau}^2 + \alpha_{3,\tau}^2 - 2\alpha_{1,\tau}\rho_{1,\tau-1} - 2\alpha_{2,\tau}\rho_{2,\tau-2} - 2\alpha_{3,\tau}\rho_{3,\tau-3} + 2\alpha_{1,\tau}\alpha_{2,\tau}\rho_{1,\tau-2} + 2\alpha_{1,\tau}\alpha_{3,\tau}\rho_{2,\tau-3} + 2\alpha_{2,\tau}\alpha_{3,\tau}\rho_{1,\tau-3})^{\frac{1}{2}} . \qquad (3\text{-}35)$$

The serial correlation coefficients $\rho_{k,\tau}$ are estimated from the sample series by

$$r_{k,\tau} = \frac{\sum_{p=1}^{n} \left[\varepsilon_{p,\tau} - \frac{1}{n}\sum_{p=1}^{n} \varepsilon_{p,\tau}\right] \left[\varepsilon_{p,\tau+k} - \frac{1}{n}\sum_{p=1}^{n} \varepsilon_{p,\tau+k}\right]}{\left[\sum_{p=1}^{n} (\varepsilon_{p,\tau} - \frac{1}{n}\sum_{p=1}^{n} \varepsilon_{p,\tau})^2\right]^{\frac{1}{2}} \left[\sum_{p=1}^{n} (\varepsilon_{p,\tau+k} - \frac{1}{n}\sum_{p=1}^{n} \varepsilon_{p,\tau+k})^2\right]^{\frac{1}{2}}} , \qquad (3\text{-}36)$$

when n the number of years of record, and k the lag. When the index τ is equal to $(\omega-k+1)$, the constant n in Eq. 3-36 is replaced by (n-1) and $\varepsilon_{p,\tau+k}$ is replaced by $\varepsilon_{p+1,k}$.

If estimated $r_{k,\tau}$ values are analyzed for periodicity by using the harmonic analysis, and are found to be periodic, the periodic function $\rho_{k,\tau}$ should be used to replace the $r_{k,\tau}$ values. If no periodicity is found in $r_{k,\tau}$, the mean of $r_{k,\tau}$, denoted by r_k, is used instead of $r_{k,\tau}$ or $\rho_{k,\tau}$. Consequently, $\alpha_{k,\tau}$ is replaced by α_k and $\sigma_{\xi,\tau}$ by σ_ξ in all previous equations, Eqs. 3-27 through 3-26.

The statistical test of the adequacy of autoregressive linear models for large samples is given by Quenouille [27]. This technique is a laborious method which includes computations of two sets of constants and a test parameter. Another approach is on the assumption that the model is of a given order. It is then performed by estimating the parameters and the computation of the presumed independent stochastic component. This component is then tested for independence. If the hypothesis of independence is accepted, the hypothesis that the model is of a given order is also accepted. This approach does not compare the various models, and may require large computations.

A simplified, practical method for determining the order of the autoregressive linear model to be used is suggested by Yevjevich [36]. The measure of the goodness of fit of the model used in this method is expressed by the determination coefficients, D_i, $i = 1,2,3,\ldots$, which give the percentage of the total variance of $\varepsilon_{p,\tau}$ explained by the i-th order term of an autoregressive equation, while the remaining portion of the variance of $\varepsilon_{p,\tau}$ is explained by $(\sigma_{\xi,\tau}\ \xi_{p,\tau})$-term.

Since $D_m \geq \ldots \geq D_3 \geq D_2 \geq D_1$, a criterion can be developed so that a model of the given order could be selected in comparison with the higher order models. For the purpose of this study, and with the first three autoregressive models used, the criteria are:

a. If $D_2 - D_1 \leq \Delta D$, and $D_3 - D_1 \leq 2\,\Delta D$, the first-order model is selected;

b. If $D_2 - D_1 > \Delta D$, but $D_3 - D_2 \leq \Delta D$, the second-order model is selected; and

c. If $D_2 - D_1 > \Delta D$, and $D_3 - D_2 > \Delta D$, the third-order model is selected, in which $(D_j - D_i)$, with $j > i$, is the difference between the percentages of the explained variance by the j-th and i-th order terms of the model, ΔD is a constant (suggested value is $\Delta D = 0.01$, or one percent of the total variance). The determination of coefficients for the first three order models are:

$$D_1 = r_1^2 , \tag{3-37}$$

$$D_2 = \frac{r_1^2 + r_2^2 - 2r_1^2 r_2}{1 - r_1^2} , \tag{3-38}$$

and

$$D_3 = (r_1^2 + r_2^2 + r_3^2 + 2r_1^3 r_3 + 2r_1^2 r_2^2 + 2r_1 r_2^2 r_3 - 2r_1^2 r_2 - 4r_1 r_2 r_3 - r_1^4 - r_2^4 - r_1^2 r_3^2)/(1 - 2r_1^2 - r_2^2 + 2r_1^2 r_2) \tag{3-39}$$

in which r_1, r_2, and r_3 are the mean values of $r_{1,\tau}$, $r_{2,\tau}$, and $r_{3,\tau}$, respectively.

A test parameter of independence of the independent stochastic component, obtained by subtracting the terms of the autoregressive scheme from the stochastic components $\varepsilon_{p,\tau}$, provides a measure of the adequacy of autoregressive schemes. The two-tail test for the autocorrelation coefficients for significant differences from the expected values at a given significant level is given by Anderson [1]. This is derived for a circular time series. However, it may also be applied to an open series if the sample is sufficiently long which is the case in a daily series, bearing in mind other possible limitations [29]. Usually, the first autocorrelation coefficient is of the greatest importance. The tolerance limits (u,l) at 95 percent level of significance are

$$S_{u,\ell} = \frac{-1 \pm 1.96\sqrt{N-3}}{N-2} \tag{3-40}$$

with N the number of observations. By applying an open series approach, Siddiqui (unpublished study in 1957) gave the distribution of r_1 for normal independent variables

$$f(r_1) = \frac{(1-r_1)^{(N-1)/2}\,(1+r_1)^{(N-3)/2}}{2^{N-1}\,B\left(\frac{N+1}{2}\,,\,\frac{N-1}{2}\right)} + O\left(\frac{1}{N^2}\right) , \tag{3-41}$$

with N the sample size, $B(\frac{N+1}{2}\,,\,\frac{N-1}{2})$ the beta function and $O(\frac{1}{N^2})$ a small quantity decreasing rapidly as N increases. The mean and variance of r_1 are

$$E(r_1) = -\frac{1}{N} ; \tag{3-42}$$

and

$$\mathrm{var}(r_1) = \frac{N^3 - 3N^2 + 4}{N^2\,(N^2-1)} . \tag{3-43}$$

With sufficiently large $N(N \geq 30)$, Eq. 3-41 is well approximated by a normal function. When N is very large and k is very small, Eqs. 3-41 through 3-43 may be applied to the distributions of r_k [37]. The tolerance limits for r_k at 95 percent significant level are

$$S_{u,\ell} = -\frac{1}{N} \pm 1.96\sqrt{\frac{N^3-3N^2+4}{N^2(N^2-1)}} . \tag{3-44}$$

If Fisher's z-transformation of r_k is used with N large $(N > 30)$ and k relatively small, the transformed r_k is normally distributed with mean equal to zero and variance equal to $1/n$. The z-transform is

$$z_k = \frac{1}{2}\ln\frac{1+r_k}{1-r_k} . \tag{3-45}$$

The tolerance limits for z_k at 95 percent significant level are

$$S_{u,\ell} = \pm 1.96\,\frac{1}{\sqrt{N}} . \tag{3-46}$$

Error propagations. Let $\pm e_x$ and $\pm e_s$ denote the error terms in Eq. 3-19 corresponding to the periodic mean and periodic standard deviation, respectively. If these error terms enter the stochastic component of $Q_{p,\tau}$, then

$$\varepsilon_{p,\tau} \pm \Delta\varepsilon = \frac{Q_{p,\tau} - (u_\tau \pm e_x)}{\sigma_\tau \pm e_s} , \tag{3-47}$$

with $\Delta\varepsilon$ the error in the stochastic component. For a Taylor's series expansion applied to the denominator of the right hand side of Eq. 3-47 and if all the second or higher order error terms are negligible, the total error $\Delta\varepsilon$ of the stochastic component becomes

$$\Delta\varepsilon = \frac{\pm\,\varepsilon_{p,\tau}\,e_s \pm e_x}{\sigma_\tau} . \tag{3-48}$$

Equation 3-48 shows the error in the stochastic component to be proportional to the stochastic component itself but inversely proportional to the standard

deviation σ_τ. If e_s and e_x are of the same order of magnitude and both proportional to σ_τ, in the order of 5 percent of σ_τ, while $\varepsilon_{p,\tau}$ a standard normal variable, then $\Delta\varepsilon = \pm 0.148$ for the 95 percent confidence interval.

Since the serial correlation coefficients, $r_{k,\tau}$, computed by Eq. 3-36 are scale and location invariant, the $r_{k,\tau}$-values are the same as those estimated directly from $Q_{p,\tau}$. Consequently, the errors in the periodic mean and standard deviation do not affect the estimates of $r_{k,\tau}$. However, the general autocorrelation coefficients (computed for the total series and not for each τ separately) of the stochastic variable will be smaller by $1/[1 + \text{var}(\Delta\varepsilon)]$.

Chapter 4
THEORETICAL DISTRIBUTION FUNCTIONS, TESTS OF GOODNESS OF FIT, AND HEAVY TAILS

4-1 Selection of Theoretical Distribution Functions with Estimation of Their Parameters.

In this study, the probability distribution functions (abbreviated in the further text as PDF) selected to fit the frequency distributions obtained for the independent stochastic components are classified into seven groups. There are some other common PDF's such as Cauchy PDF, Pareto PDF, beta PDF,..., etc., which are not used in this study either because they are not suitable for fitting the frequency distributions of the independent stochastic components or they are special cases of the seven groups. The classification of these seven groups depends mainly upon the characteristics of the PDF's selected for the analysis. Based on the properties of an independent stochastic component, the theoretical probability distribution function of best fit to an observed frequency distribution should have the following properties: (1) it is continuous and defined for all positive and negative values of the variable; (2) the upper tail is unbounded; (3) the density curve is asymptotic to the axis on the positive side and is also asymptotic on the negative side in case the lower tail is unbounded; and (4) it should be representative of a large range of skewness and kurtosis coefficients.

Classical PDF's. The commonly used PDF's in hydrology include the normal, lognormal and gamma density functions. A voluminous literature related to these three probability density functions is available. Therefore, only the functions and their parameters are summarized here.

a. Normal PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \quad -\infty \leq x \leq \infty \tag{4-1}$$

with x the variable, μ the population mean, and σ the population standard deviation. The maximum likelihood estimates of μ and σ, with N the sample size, are

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad , \tag{4-2}$$

and

$$\hat{\sigma} = \left[\frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \right]^{\frac{1}{2}} \quad , \tag{4-3}$$

b. Lognormal PDF

$$f(x) = \frac{1}{(x-x_0)\sigma\sqrt{2\pi}} e^{-\frac{[\ln(x-x_0) - \ln\mu\ ]^2}{2\sigma^2}} \quad , \quad x_0 \leq x \leq \infty \tag{4-4}$$

with μ the population geometric mean of $(x-x_0)$, σ the population standard deviation of $\ln(x-x_0)$, and x_0 the lower boundary. Equation 4-4 is a three-parameter lognormal function. It becomes a two-parameter function for $x_0 = 0$. The lower boundary $\hat{x}_0$ is the maximum likelihood estimate of x_0, and is obtained by solving the following equation by an iteration procedure:

$$\sum_{i=1}^{N} \frac{1}{x_i - \hat{x}_0} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ln^2 (x_i - \hat{x}_0) - \left[\frac{1}{N} \sum_{i=1}^{N} \ln (x_i - \hat{x}_0) \right]^2 - \frac{1}{N} \sum_{i=1}^{N} \ln(x_i - \hat{x}_0) \right\} + \sum_{i=1}^{n} \frac{\ln(x_i - \hat{x}_0)}{x_i - x_0} = 0 \ . \tag{4-5}$$

The maximum likelihood estimates of $\ln \mu$ and σ are

$$\widehat{\ln\mu} = \frac{1}{N} \sum_{i=1}^{N} \ln (x_i - \hat{x}_0) \quad , \tag{4-6}$$

and

$$\hat{\sigma} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[\ln(x_i - \hat{x}_0) - \widehat{\ln\mu} \right] 2 \right\}^{\frac{1}{2}} \tag{4-7}$$

c. Gamma or Pearson's Type III PDF

$$f(x) = \frac{1}{\beta\Gamma(\alpha)} \left(\frac{x-x_0}{\beta} \right)^{\alpha-1} e^{-\frac{x-x_0}{\beta}} \quad , \quad x_0 \leq x \leq \infty \tag{4-8}$$

with α the shape parameter, β the scale parameter, x_0 the location parameter of the lower boundary, and $\Gamma(\alpha)$ the gamma function of α. The maximum likelihood estimate of the lower boundary is obtained by solving the following equation by an iterative procedure

$$\frac{1+(1+\frac{4}{3}A)^{\frac{1}{2}}}{1+(1+\frac{4}{3}A)^{\frac{1}{2}}-4A} - (\bar{x} - \hat{x}_0) \frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i - \hat{x}_0} = 0 \quad , \tag{4-9}$$

in which

$$A = \ln (\bar{x} - \hat{x}_0) - \frac{1}{N} \sum_{i=1}^{N} \ln (x_i - \hat{x}_0) \quad , \tag{4-10}$$

with $\bar{x}$ the mean of N values of x. Once $\hat{x}_0$ is determined, the parameter α is estimated by

$$\hat{\alpha} = \frac{(1 + \frac{4}{3}A)^{\frac{1}{2}}}{4A} - \Delta\alpha \quad , \tag{4-11}$$

with A given by Eq. 4-10 and $\Delta\alpha$ approximated by

$$\Delta\alpha = 0.04475\ (0.26)^{\alpha} \quad . \tag{4-12}$$

The parameter β is then estimated by

$$\beta = \frac{1}{\hat{\alpha}} (\bar{x} - \hat{x}_0) \quad . \tag{4-13}$$

Pearson's Family of PDF's. From the general class of functions originated by Karl Pearson [8], Types IV, VI and VII were selected for application. The parameters of these three selected functions are estimated by using the method of moments, with the highest order of moment being four. It was not practical in this study to use the maximum likelihood method for estimation of parameters because it

requires the solution of simultaneous nonlinear equations.

a. Pearson's Type IV PDF

$$f(x) = y_o \left[(1 + \frac{x}{a} - \frac{v}{r})^2\right]^{-m} e^{[v \tan^{-1}(\frac{x}{a} - \frac{v}{r})]} \quad (4\text{-}14)$$

with

$$r = \frac{6(\beta_2-\beta_1-1)}{2\beta_2-3\beta_1-6} \quad ; \; m = \frac{1}{2}(r+2);$$

$$v = \frac{r(r-2)\sqrt{\beta_1}}{\sqrt{16(r-1)-\beta_1(r-2)^2}} \quad ;$$

$$a = \frac{1}{4}\sqrt{\mu_2[16(r-1)-\beta_1(r-2)^2]} \quad , \text{ and}$$

$$y_o = \left[a \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \cos^r\theta \, e^{-v\theta} \, d\theta\right]^{-1} \quad . \quad (4\text{-}15)$$

Five parameters in Eqs. 4-14 and 4-15 are functions of the first four central moments of x. The curve of Eq. 4-13 is skewed with unbounded range on both tails, and v and μ_3 have opposite signs. The origin of the curve is at the mean, with the mode at $x = va/r$.

b. Pearson's Type VI PDF

$$f(x) = y_o(1 + \frac{x}{A_1})^{-q_1}(1 + \frac{x}{A_2})^{q_2} \quad , \quad (4\text{-}16)$$

with

$$q_1 = -\frac{r-2}{2} + \frac{r(r+2)}{2}\sqrt{\frac{\beta_1}{\beta_2(r+2)^2+16(r+1)}} \quad ,$$

$$q_2 = q_1 + (r-2) \; ; \; A_1 = \frac{a(q_1-1)}{(q_1-1)-(q_2+1)} \; ;$$

$$A_2 = \frac{a(q_2-1)}{(q_1-1)-(q_2+1)} \; ; \text{ and}$$

$$y_o = \frac{(q_2+1)^{q_2}(q_1-q_2-2)^{q_1-q_2}\Gamma(q_1)}{a(q_1-1)^{q_1}\Gamma(q_1-q_2-1)\Gamma(q_2+1)} \quad , \quad (4\text{-}17)$$

in which

$$r = \frac{6(\beta_2-\beta_1-1)}{6+3\beta_1-2\beta_2} \quad , \text{ and}$$

$$a = \frac{1}{2}\sqrt{\mu_2[\beta_1(r+2)^2+16(r+1)]} \quad (4\text{-}18)$$

These five parameters are functions of the first four central moments of x. The curve of Eq. 4-16 is skewed within the range $-A$ to $+\infty$. If μ_3 is negative, the range is from A_1 to $-\infty$. The origin of the curve is at the mean with the mode at $x = -a(q_1 - 1)/(q_1 - q_2 - 2)$. For $q_1 = 2$ and $q_2 < 1$, Eq. 4-16 becomes the Pareto distribution.

c. Pearson's Type VII PDF

$$f(x) = y_o(1 + \frac{x^2}{a^2})^{-m} \quad , \quad (4\text{-}19)$$

with $m = \frac{5\beta_2-9}{2(\beta_2-3)}$; $a^2 = \frac{2\mu_2\beta_2}{\beta_2-3}$, and

$$y_o = \frac{1}{a\sqrt{\pi}} \frac{\Gamma(m)}{\Gamma(m-\frac{1}{2})} \quad . \quad (4\text{-}20)$$

Weibull probability distribution function. The distribution of a random variable is a Weibull distribution [16] when $y = [(x-x_o)/\alpha]^c$ is exponentially distributed with $c > 0$, $\alpha > 0$, and x_o the lower boundary. The probability density function is

$$f(x) = \frac{c}{\alpha}\left(\frac{x-x_o}{\alpha}\right)^{c-1} e^{-\left(\frac{x-x_o}{\alpha}\right)^c} \quad . \quad (4\text{-}21)$$

The maximum likelihood estimates $\hat{c}$, $\hat{\alpha}$, and $\hat{x}_o$ of c, α, and x_o should satisfy the three equations

$$\hat{\alpha} = \left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{x}_o)^{\hat{c}}\right]^{1/\hat{c}} \quad , \quad (4\text{-}22)$$

$$\hat{c} = \left\{\sum_{i=1}^{N}\left[(x_i-\hat{x}_o)^{\hat{c}} \ln(x_i-\hat{x}_o)\right]\frac{1}{\sum_{i=1}^{N}(x_i-\hat{x}_o)^{\hat{c}}} - \frac{1}{N}\sum_{i=1}^{N}\ln(x_i-\hat{x}_o)\right\}^{-1} ,$$

and (4-23)

$$(\hat{c}-1)\sum_{i=1}^{N}(x_i-\hat{x}_o)^{-1} - \hat{c}\hat{\alpha}^{-\hat{c}}\sum_{i=1}^{N}(x_i-\hat{x}_o)^{\hat{c}-1} = 0 \quad , \quad (4\text{-}24)$$

Since $\min(x_1,x_2,\ldots,x_n)$ is a maximum likelihood estimate of x_o, the practical way of solving for $\hat{x}_o$, $\hat{\alpha}$ and $\hat{c}$ in the above three equations is by the following four steps: (a) Let $x_o = \min(x_1,x_2,\ldots,x_n) - \Delta x$, with Δx a small positive quantity, like 0.0001; (b) Solve $\hat{c}$ from Eq. 4-23 by an iterative procedure; (c) Substitute $\hat{x}_o$, $\hat{c}$ into Eq. 4-22 and compute $\hat{\alpha}$; and (d) Substitute $\hat{x}_o$, $\hat{c}$, and $\hat{\alpha}$ into Eq. 4-24. If the absolute value of the left side of Eq. 4-24 is not sufficiently close to zero, another value of Δx in the step (a) should be selected and the above procedure repeated.

It is difficult to find physical reasons for using the Weibull probability distribution function to fit the frequency distributions of independent stochastic components. However, this function gives a power transformation of the original variable, and becomes a practical and convenient way of introducing the flexibility in the fitting model which leads to an exponential probability distribution.

Probability distribution functions modified by polynomials. With a given density function, $f(x)$, another density function, $g(x)$, might be expressed as a product of $f(x)$ and a series expansion of x in the form

$$g(x) = f(x) \sum_{j=0}^{k} q_j(x) \quad , \tag{4-25}$$

with $q_j(x)$ a function of x, and k finite or infinite number of terms. Two methods are used in solving Eq. 4-25. The first method restricts the polynomial $q_j(x)$, $j = 1,2,\ldots,k$, to the orthogonal relations, namely

$$\int q_i(x)\, q_j(x)\, f(x)\, dx \begin{cases} = 0 \ , & \text{if } i \neq j \\ = 1 \ , & \text{if } i = j \end{cases} . \tag{4-26}$$

The second method uses the Edgeworth series [23], starting with the characteristic function $\phi_g(t)$ of $g(x)$,

$$\phi_g(t) = \int_{-\infty}^{\infty} e^{\theta x}\, g(x)\, dx = e^{\left[\sum_{j=1}^{\infty} (k_j\, \theta^j/j!)\right]} , \tag{4-27}$$

with $\theta = it$, $i = \sqrt{-1}$, k_j the j-th cumulant of $g(x)$, assuming that all k_j ($j = 1,2,\ldots$) exist. If all cumulants η_j of $f(x)$ exist, $f(x)$ has the characteristic function

$$\phi_f(t) = e^{\left[\sum_{j=1}^{\infty} (\eta_j\, \theta^j/j!)\right]} , \tag{4-28}$$

which can be written as

$$\phi_g(t) = e^{\left[\sum_{j=1}^{N} (k_j - \eta_j)\, \theta^j/j!\right]} \phi_f(t) \quad . \tag{4-29}$$

Using the inverse Fourier transform of $\phi_f(t)$,

$$f(x) = \int_{-\infty}^{\infty} e^{-\theta x}\, \phi_f(t)\, dt \quad , \tag{4-30}$$

and assuming it differentiable under the integral sign, then

$$(-D)^j f(x) = \int_{-\infty}^{\infty} \theta^j\, e^{-\theta x}\, \phi_f(t)\, dt \quad , \tag{4-31}$$

in which $D = d./dx$, $j = 1,2,\ldots$. Since all the cumulants exist, one may combine Eqs. 4-29 through 4-31 and take the inverse Fourier transformation of $\phi_g(t)$ giving

$$g(x) = \left[1 + \sum_{j=1}^{\infty} a_j\, (-D)^j/j!\right] f(x) \quad , \tag{4-32}$$

with a_j depending on k_j and η_j.

The normal density function modified by the Hermite polynomials [23] is derived from this second method. The gamma density function modified by the Laguerre polynomials [30] is derived by the first method. They are

a. Normal function modified by Hermite polynomials.

$$f(x) = N(\mu,\sigma) \left\{1 + \frac{\gamma_3}{3!} H_3(x) + \frac{\gamma_4}{4!} H_4(x) + \ldots\right\} , \tag{4-33}$$

with x the standardized random variable, of the normal probability density function, $N(\mu,\sigma)$ defined by Eq. 4-1, γ_m the m-th cumulant of x, and H_m the Hermite polynomial terms of order m, given by

$$H_m = (-1)^m\, e^{-\frac{1}{2}x^2}\, \frac{d^m}{dx^m} \left(e^{-\frac{1}{2}x^2}\right) . \tag{4-34}$$

For $m = 3$ and 4, Eq. 4-34 gives $H_3 = x^3 - 3x$, and $H_4 = x^4 - 6x^2 + 3$.

b. Gamma function modified by Laguerre polynomials.

$$f(x) = G(\alpha,\beta,x_o) \sum_{m=0}^{\infty} \left[\frac{m!\,\Gamma(\alpha)}{\Gamma(m+\alpha)}\, \frac{d_m}{\beta^m}\, L_m^{(\alpha-1)}\left(\frac{x-x_o}{\beta}\right)\right] , \tag{4-35}$$

with $G(\alpha,\beta,x_o)$ the gamma probability density function defined by Eq. 4-8, $L_m^c(y)$ the Laguerre polynomial terms of order m, given by

$$L_m^c(y) = \sum_{j=0}^{m} \binom{m+c}{m-j} \frac{(-y)^j}{j!} \ , \text{ and}$$

$$d_m = \sum_{j=0}^{m} \binom{m-1+x}{m-j} (-1)^j\, (\beta)^{m-j}\, \frac{\gamma_j}{j!} \ , \text{ with}$$

$$\gamma_j = E(x - x_o)^j \ . \tag{4-36}$$

With the above polynomial modifications, only the first few polynomials are important. As a general rule, the order of the last polynomial term considered must be such that: (a) no significant oscillations occur in the probability density function; (b) the coefficient with the x^m term should be very small in comparison with the coefficients of the lower order terms. Considering the above two conditions, the probability density function of x is usually truncated with $m = 3$ or $m = 4$.

Double-branch gamma probability density function.

$$f(x) = \frac{P}{\beta_1^{\alpha_1}\, \Gamma(\alpha_1)}\, (x_o - x)^{\alpha_1 - 1}\, e^{-\frac{x_o - x}{\beta_1}}\, I_{(-\infty,x_o)} + \frac{1-P}{\beta_2^{\alpha_2}\, \Gamma(\alpha_2)}\, (x - x_o)^{\alpha_2 - 1}\, e^{-\frac{x - x_o}{\beta_2}}\, I_{(x_o,\infty)} \ , \tag{4-37}$$

with P defined as $P = P_r(x \le x_o)$ and $1-P = P_r(x > x_o)$, x_o the mode, α_1 and β_1 the parameters of the left branch, and α_2 and β_2 the parameters of the right branch [36]. The mode x_o is best estimated by selecting the class interval Δx, like 0.001 or smaller, which has the largest frequency, and is approximately computed by $\hat{x}_o = x_L + \Delta x/2$, in which x_L is the lower bound of the interval Δx with the largest frequency. The parameter P is determined by $P = n_1/N$, with n_1 the number of x values satisfying $x_i \le x_o$. The parameters α_1 and β_1 of the left branch are estimated by using only the n_1 values with $x_i \le x_o$, while α_2 and β_2 of the right branch are estimated by using the remaining n_2 sample values with $n_2 = N-n_1$, for $x_i > x_o$. The parameters α_1 and β_1 are estimated by

$$\hat{\alpha}_1 = \frac{1+\sqrt{1+\frac{4}{3}A_1}}{4A_1} - 0.04475\ (0.26)^{\alpha_1} \qquad (4\text{-}38)$$

with

$$A_1 = \ln\left[\hat{x}_o - \frac{1}{n_1}\sum_{i=1}^{n_1} x_i\right] - \frac{1}{n_1}\sum_{i=1}^{n_1} \ln(\hat{x}_o - x_i)\ , \qquad (4\text{-}39)$$

and

$$\hat{\beta}_1 = \frac{1}{\hat{\alpha}_1}\left(\hat{x}_o - \frac{1}{n_1}\sum_{i=1}^{n_1} x_i\right)\ . \qquad (4\text{-}40)$$

Similarly, parameters α_2 and β_2 are estimated by using Eqs. 4-38 through 4-40 with the term $x_o - x_i$ substituted by $x_i - x_o$ and n_1 by n_2. The shape parameters α_1 and α_2 of the two branches should be equal to or less than unity. If Eq. 4-38 gives a value greater than unity, the one-peak gamma function should be replaced by a j-shaped exponential function.

Mixture of probability distribution functions. Distributions resulting from mixing of two or more component distribution functions are denoted as the mixed distributions. Intuitively, a mixture may be conceived as two or more populations of random variables, physically mixed but neither the population distribution nor the proportion of the component population distributions of the mixture are known. Difficulties arise in parameter estimations of these distributions. Karl Pearson attempted the estimation of parameters in a mixture of two normal populations, considering two means, two variances, and a proportion factor. He equated the first five moments with their sample values. In order to solve these equations for five unknown parameters, a ninth-order polynomial equation is required. The computations are not difficult with a digital computer, but more than one real root may exist in the ninth-order polynomial, which demands a correct selection between two or more sets of estimates [7]. The estimation procedures for parameters of the mixture of k normal functions by the maximum likelihood method results in $(3k - 1)$ nonlinear equations involving the $(3k - 1)$ unknown parameters [15]. This is an obstacle to both the general and practical applications of the mixture of distribution functions. However, the problem may be simplified by assuming that some properties of these mixed functions are known. Three examples are given here.

a. Mixtures of two normal density functions with the same means.

$$f(x) = \frac{P}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma_1^2}} + \frac{1-P}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma_2^2}}\ , \qquad (4\text{-}41)$$

with P the parameter of the proportion of the two normal functions, μ the population mean, σ_1 and σ_2 the populations standard deviations of these two normal functions. The parameter μ is estimated by $\hat{\mu}$ given by Eq. 4-2, σ_1, σ_2, and P by $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{P}$, respectively, by the following equations [7]:

$$\hat{\sigma}_1 = t_1 + \hat{\sigma}^2\ ,\quad \hat{\sigma}_2 = t_2 + \hat{\sigma}^2\ ,\ \text{and}$$

$$\hat{P} = \frac{-t_2}{t_1 - t_2}\ , \qquad (4\text{-}42)$$

with $\hat{\sigma}$ the sample standard deviation given by Eq. 4-3, t_1 and t_2, with $t_2 > t_1$, the roots of quadratic equation

$$Y^2 - \frac{k_6}{5k_4}\, Y - \frac{k_4}{3} = 0, \qquad (4\text{-}43)$$

and k_4 and k_6 the 4-th and 6-th sample cumulants, respectively.

Equation 4-41 is applicable for fitting the frequency distributions of x under the conditions that the distributions are symmetrical in respect to their means, and the kurtosis coefficients are greater than three.

b. Mixtures of normal and gamma density functions.

$$f(x) = \frac{P}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x-x_o)^2}{2\sigma_1^2}}\, I_{(-\infty,+\infty)} + \frac{1-P}{\beta\Gamma(\alpha)}\left(\frac{x-x_o}{\beta}\right)^{x-1} e^{-\frac{x-x_o}{\beta}}\, I_{(x_o,\infty)} \qquad (4\text{-}44)$$

with x_o the mode defined by Eq. 4-37, P the weighting factor denoting the proportion of the two functions, estimated by $P = 2n_1/N$, with n_1 the number of all values of x_i and $x_i \le x_o$, σ_1 the standard deviation of the normal function computed by

$$\hat{\sigma}_1 = \left[\frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \hat{x}_o)^2\right]^{\frac{1}{2}} \qquad (4\text{-}45)$$

α and β the parameters of the gamma function estimated by

$$\hat{\alpha} = \frac{A_g^2}{V_g}\ ,\ \text{and}\quad \hat{\beta} = \frac{V_g}{A_g}\ , \qquad (4\text{-}46)$$

with

$$A_g = \frac{\hat{\mu}-P\,\hat{x}_o}{1-P}, \text{ and } V_g = \frac{\hat{\sigma}^2+\hat{\mu}^2-P(\sigma_1^2+\hat{x}_o^{\,2})}{1-P} - A_g^{\,2}, \quad (4\text{-}47)$$

with $\hat{\mu}$ and $\hat{\sigma}$ the mean and standard deviation of x.

Equation 4-48 is valid to fit the frequency distribution of x under the conditions that $N>2n_1$, V_g computed by Eq. 4-47 is positive, and the statistic

$$\beta_2 = \left[\frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\hat{x}_o)^4\right] \Big/ \left[\frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\hat{x}_o)^2\right]^2 \quad (4\text{-}48)$$

is close to three.

c. Mixtures of Pearson's Type VII and gamma density functions.

$$f(x) = P\,y_o\left(1+\frac{x^2}{a^2}\right)^{-m} I_{(-\infty,+\infty)} + \frac{1-P}{\beta\Gamma(x)}\left(\frac{x-x_o}{\beta}\right)^{\alpha-1} e^{-\frac{x-x_o}{\beta}} I_{(x_o,\infty)}, \quad (4\text{-}49)$$

with y_o, a, and m the parameters of Pearson's Type VII function, to be estimated by using Eqs. 4-20 and 4-48, for m, a^2, y_o, and β_2, respectively. The other parameters in Eq. 4-49 can be estimated by using Eqs. 4-46 and 4-47. Equation 4-49 is applicable under the conditions of $N>2n_1$ and $\beta_2>3$.

Family of stable distributions. The family of stable distributions [9,12] is defined as the logarithm of their characteristic functions which have the general form

$$\ln[\phi_x(t)] = \ln[E(e^{ixt})] = i\delta t-\gamma|t|^\alpha\left[1+i\beta\left(\frac{t}{|t|}\right)w(t,\alpha)\right], \quad (4\text{-}50)$$

with x the random variable, t any real number, $i=\sqrt{-1}$, and

$$w(t,\alpha) = \begin{cases} \tan\frac{\pi\alpha}{2} & \text{for } \alpha\neq 1, \\ \frac{2}{\pi}\ln|t| & \text{for } \alpha=1. \end{cases} \quad (4\text{-}51)$$

The stable distributions have four parameters: α, β, δ, and γ.

The parameter α is called the characteristic exponent of the distribution determining the rate of convergence of the extreme tails of distributions. It can take any value in the interval $0<\alpha\leq 2$. For $\alpha=2$ and $\alpha=1$, the stable distributions are the normal and Cauchy distributions, respectively. For $0<\alpha<2$, the extreme tails of stable distributions are higher than those of the normal distribution. The larger the total probability in the extreme tails for given x, the smaller the value of α. The variance only exists in a limit case for $\alpha=2$. The mean exists for $\alpha>1$.

The parameter β is an index of skewness taking any value in the interval $-1\leq\beta\leq 1$. The distribution is symmetric for $\beta=0$ and positively skewed for $\beta>0$, with the positive skewness increasing as β increases. Similarly, for $\beta<0$ the distribution is negatively skewed, and the absolute skewness increases as the absolute value of β increases.

The parameter δ is the distribution location parameter. For $\alpha>1$, δ is the expected value, or the mean of distribution; however, for $\alpha\leq 1$, the mean is not defined. In this case δ is some other parameter that describes the location.

The parameter γ defines the scale of a stable distribution. For example, if $\alpha=2$, γ is one half of the variance. For $\alpha<2$, the variance of the distribution is finite. In this case a finite γ parameter still defines the scale of the distribution, however, γ is not one half of the variance.

The three most important properties of stable distributions are: the extreme tail areas follow the asymptotic form of the Pareto law; they are stable or invariant under addition; and these distributions are the only limiting distributions for sums of independent identically distributed random variables.

a. Asymptotic form of the Pareto law. Because the tails of stable distributions follow a weak or asymptotic form of the Pareto law, then

$$1-F(x) \to c_1 x^{-\alpha}, \text{ for } x\to\infty \quad (4\text{-}52)$$

and

$$F(x) \to c_2\,|x|^{-\alpha}, \text{ for } x\to-\infty, \quad (4\text{-}53)$$

with x the random variable, and c_1 and c_2 constants. This implies that if $\ln[1-F(x)]$ is plotted against $\ln x$ for the right tail, or $\ln[F(x)]$ is plotted against $\ln(-x)$ for the left tail, the resulting curve should be asymptotic to a straight line with the slope equal to $-\alpha$, as x approaches infinity.

b. Stability or invariance under addition. The distributions of sums of independent, identically distributed stable variables are themselves stable with the same distribution as the individual variables. The logarithm of the characteristic function of the sum of independent, identically distributed stable variables is

$$n\ln[\phi_x(t)] = i(n\delta)t - (n\gamma)|t|^\alpha\left[1+i\beta\frac{t}{|t|}w(t,x)\right], \quad (4\text{-}54)$$

with n the number of variables in the sum and $\ln[\phi_x(t)]$ the logarithm of the characteristic function of individual variables. Equation 4-54 is the same as Eq. 4-50, except that the parameters δ and γ are multiplied by n. Except for the origin and the scale, the distribution of the sum is exactly the same as that of individual variables. Briefly, the stability means that the parameters α and β remain constants after addition.

c. Limit distributions. Stability or invariance under addition is related to an important corollary property of stable distributions, namely, they are the only possible limiting distributions for sums of

independent, identically distributed random variables. If these variables have the finite variance, the limiting distribution for their sum is a normal distribution. If their sum follows a limiting distribution, the limiting distribution must be stable, with $0 < \alpha \leq 2$.

In summary, the sum of independent identically distributed stable variables is also a stable variable, with the same characteristic exponent as the distribution of the individual variables. The process of taking the sum changes only the scale of the distribution. To find a constant weight each variable in the sum so that the scale parameter of the distribution of the sum is the same as that of the individual variables, the constant b must satisfy

$$n\gamma\,|bt|^{\alpha} = \gamma\,|t|^{\alpha}\,, \tag{4-55}$$

giving

$$b = n^{-\frac{1}{\alpha}}\,. \tag{4-56}$$

This implies that each of the component variables must be divided by $n^{1/\alpha}$. The converse proposition is that the scale of the distribution of an unweighted sum is $n^{1/\alpha}$ times that of the individual variables.

For example, the interquartile range (0.75 fractile to 0.25 fractile) of the distribution of the sum of n independent, identically distributed stable variables will be $n^{1/\alpha}$ times that of the individual variables. This property provides the basis of the spacing of the order statistics approach to the estimation of α. It is not possible to express the stable density function in a closed form except for the cases of $\alpha = 1/2$, $\alpha = 1$, (Cauchy), and $\alpha = 2$, (normal). However, Bergstrom, Fama and Roll [10], present the expansions series capable of approximating the stable functions. Since the parameter estimations of stable distributions for both the symmetric and asymmetric cases are of importance, they are discussed below.

a. Symmetric stable distributions. The logarithmic characteristic function of symmetric stable distributions is given by setting $\beta = 0$ in Eq. 4-50 so that

$$\ln \phi_x(t) = i\,\delta\,t - \gamma\,|t|^{\alpha} \tag{4-57}$$

Considering the transformation by

$$u = \frac{x-\delta}{c}\,, \tag{4-58}$$

with $c = \gamma^{1/\alpha}$, then for Eq. 4-57 (with the application of properties of the characteristic function, u is stable with the parameters α, unaffected by the transformation, $\delta = 0$ and $\gamma = c = 1$), the logarithmic characteristic function of the symmetric stable variable is

$$\ln \phi_n(t) = -|t|^{\alpha}\,. \tag{4-59}$$

Bergstrom presented the series expansions which can be used to approximate functions as given by Eq. 4-59. For $\alpha > 1$, his results yield the series

$$f(u) = \frac{1}{\pi\alpha}\sum_{k=0}^{\infty}(-1)^k\,\frac{\Gamma\left(\frac{2k+1}{\alpha}\right)}{(2k)!}\,u^{2k}\,. \tag{4-60}$$

For $\alpha > 1$, he also provided a limit series for $u > 0$, namely

$$f(u) = -\frac{1}{\pi}\sum_{k=1}^{n}\frac{(-1)^k}{k!}\,\frac{\Gamma(\alpha k+1)}{u^{\alpha k+1}}\sin\left(\frac{k\pi\alpha}{2}\right) + R(u)\,, \tag{4-61}$$

with the remainder term

$$R(u) = O\left[u^{-\alpha(n+1)-1}\right] \tag{4-62}$$

Term by term integrations of Eqs. 4-60 and 4-61 yield series for the cumulative distribution function $F(u)$ of u with $\alpha > 1$. Using the cumulative distribution function, Fama and Roll [10] tabulated the values of this function for twelve different values of α in the interval $1 \leq \alpha \leq 2$.

With the help of the tabulated values of $F(u)$, an estimate of the parameter c can be obtained from the sample fractiles. The 0.72 fractile of the symmetric stable distribution with $\delta = 0$ and $c = 1$, is located at the interval 0.827 ± 0.003 for $1 \leq \alpha \leq 2$, which is a minimum error value of 0.003 among all the fractiles [10], given a random sample of size N. Therefore, as a special case an estimate of c is

$$\hat{c} = \frac{1}{2(0.827)}\left(\hat{x}_{0.72} - \hat{x}_{0.28}\right)\,, \tag{4-63}$$

with $x_{0.72}$ and $x_{0.28}$ referring to the (0.72)(N+1)-st and (0.28)(N+1)-st order statistics, respectively. To estimate $x_{0.72}$ and $x_{0.28}$, the 0.72 and 0.28 fractiles of the distribution of x are used. This estimate has an asymptotic bias less than 0.4 percent [11].

The characteristic exponent α can be estimated from the sample by

$$\hat{u}_f = \frac{\hat{x}_f - \hat{x}_{1-f}}{2\hat{c}} = 0.872\,\frac{x_f - x_{1-f}}{\hat{x}_{0.72} - \hat{x}_{0.28}}\,, \tag{4-64}$$

with the suggested values for f from 0.92 to 0.99. Since x has a symmetric stable distribution with the characteristic exponent α, the scale parameter $\gamma = c^{\alpha}$, and the location δ, $\hat{u}_f$ is an estimate of the f-fractile of the symmetric stable distributions with the characteristic exponent α, of the scale parameter $\gamma = c = 1$, and the location parameter $\delta = 0$. Therefore, an estimate of α can be obtained by searching from the tables of $F(u)$ for the value, $\hat{\alpha}_f$, with its f-fractile most closely matching u_f. Since different f values will give different $\hat{\alpha}_f$, Fama and Roll [11] suggested using the equation

$$\hat{\alpha} = \left(\hat{x}_{0.93} + \hat{\alpha}_{0.99}\right)/2\,, \tag{4-65}$$

Finally, the location parameter δ can be estimated by the sample mean of x under the condition $1 < \alpha \leq 2$.

b. Asymmetric stable distributions. Press [26] proposed a procedure to estimate the parameters α,

β, δ, and γ of the asymmetric stable distributions by using the method of moments approach on the characteristic function. The procedure is as follows. From Eq. 4-50 for all α

$$\ln |\phi x(t)| = -\gamma |t|^{\alpha}. \tag{4-66}$$

Replacing $\phi_x(t)$ by the sample characteristic function,

$$\hat{\phi}_x(t) = \frac{1}{N} \sum_{i=1}^{N} e^{itx_i} , \tag{4-67}$$

after selecting the two nonzero values of t (like t_1 and t_2, with $t_1 = t_2$). For $\alpha \neq 1$, $\gamma |t_1|^{\alpha} = -\ln |\hat{\phi}_x(t_1)|$ and $\gamma |t_2|^{\alpha} = -\ln |\phi_x(t_2)|$. By solving these two equations simultaneously for α and γ, then

$$\hat{\alpha} = \frac{\ln \left| \frac{\ln|\hat{\phi}_x(t_1)|}{|\hat{\phi}_x(t_2)|} \right|}{\ln \frac{t_1}{t_2}} , \tag{4-68}$$

and

$$\ln\hat{\gamma} = \frac{\ln|t_1|\ln[-\ln|\hat{\phi}_x(t_2)|]-\ln|t_2|[-\ln|\hat{\phi}_x(t_1)|]}{\ln\left|\frac{t_1}{t_2}\right|} . \tag{4-69}$$

The imaginary part of logarithms of the characteristic function of Eq. 4-50 is defined by $Z(t) = \delta t - \gamma|t|^{\alpha} \beta w(t,\alpha)$. The nonzero values of t are again selected as t_3 and t_4, with $t_3 \neq t_4$; for $\alpha \neq 1$ and,

$$\delta t_k - \gamma \beta |t_k|^{\alpha} \tan \frac{\pi\alpha}{2} = Z(t_k), \quad k=3,4 . \tag{4-70}$$

Since $\hat{\phi}_x(x) = \frac{1}{N} \sum_{j=1}^{N} \cos t x_j + i \frac{1}{N} \sum_{j=1}^{N} \sin t\, x_j$, which in polar coordinates is $\hat{\phi}_x(t) \equiv \rho(t) \exp[i\theta(t)]$, where

$$\rho^2(t) = \left(\frac{1}{N} \sum_{j=1}^{N} \cos t x_j\right)^2 + \left(\frac{1}{N} \sum_{j=1}^{N} \sin t\, x_j\right)^2 ,$$

and

$$\tan[\theta(t)] = \left(\sum_{j=1}^{N} \sin t\, x_j\right) / \left(\sum_{j=1}^{N} \cos t x_j\right) .$$

Hence $\ln\hat{\phi}_x(t) = \rho(t) + i\theta(t)$, and

$$\hat{Z}(t) = \tan^{-1}\left[\left(\sum_{j=1}^{N} \sin t\, x_j\right) / \left(\sum_{j=1}^{N} \cos t x_j\right)\right]. \tag{4-71}$$

Replacing $Z(t_k)$ in Eq. 4-70 by its estimated value from Eq. 4-71, and solving the two implied equations simultaneously for β and δ, the estimates are

$$\hat{\beta} = \frac{\left| \frac{\hat{Z}(t_3)}{t_3} - \frac{\hat{Z}(t_4)}{t_4} \right|}{\left| |t_4|^{\alpha-1} - |t_3|^{\hat{\alpha}-1} \right| \hat{\gamma} \tan \frac{\pi\hat{\alpha}}{2}} \tag{4-72}$$

and

$$\hat{\delta} = \frac{|t_4|^{\hat{\alpha}-1} \frac{\hat{Z}(t_3)}{t_3} - |t_3|^{\hat{\alpha}-1} \frac{\hat{Z}(t_4)}{t_4}}{|t_4|^{x-1} - |t_3|^{\alpha-1}} . \tag{4-73}$$

The above estimates are the consistent estimates, i.e., estimates that converge to correct as $n \to \infty$, since both are based on $\hat{\phi}_x(t)$, which is a consistent estimate for $\phi_x(t)$. However, the rate of convergence to the population parameters varies depending on the selected values of t_1 through t_4.

The above equations used for estimating parameters are applied to the independent stochastic variables, but with different sets of t_1, t_2, t_3 and t_4. Thus, the complete different values of estimates for $\hat{\alpha}$, $\hat{\gamma}$, $\hat{\beta}$, and $\hat{\delta}$ are obtained. Since the probability density functions of asymmetric stable distributions are not available, the optimal choice of t_1, t_2, t_3 and t_4 is not defined.

4-2 Test for Goodness of Fit of Frequency Distributions.

Several methods can be used for testing the goodness of fit of a probability distribution function to frequency distributions. Such methods are the chi-square test, the likelihood ratio test (which is equivalent to chi-square test [35]), and the Kolmogorov-Smirnov distribution free test. The chi-square test is selected for this study because this test is well known and frequently applied both in statistics and hydrology.

The basic properties of the chi-square test are summarized as follows [20]. The total range of sample observations is divided into k mutually exclusive and exhaustive class intervals, each having the observed class probability O_j and the corresponding expected class probability $E_j (j = 1,2,\ldots,k)$. The expected value E_j is used as the norm of any class interval and the quantity $(O_j-E_j)^2$ is used as a measure of departure from the norm. The $(O_j-E_j)^2$ values cannot be compared from one class to another if the scale of each class interval is not nearly proportional to the expected value E_j. Therefore, a more suitable measure is obtained by using $(O_j-E_j)^2/E_j$. The measure of total discrepancy, χ^2, between the observations and the expectations becomes

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} . \tag{4-74}$$

This statistic is asymptotically distributed as the chi-square distribution with $k-1$ degree of freedom for the case of population parameters are estimated from the sample data, the number of degrees of freedom is further decreased by the number of estimated parameters. For m parameters, the total number of degrees of freedom is

$$f = k-1-m \ . \qquad (4\text{-}75)$$

The number of class intervals k has to be first selected for the application of chi-square test. If too many classes are used, the obtained frequency distribution will be very irregular. If there are too few classes, with large portions of frequencies falling in one or two classes, much information is lost. Sturges [31] gives the empirical expression for the number of class intervals k as

$$k = 1+3.3 \ln N \ , \qquad (4\text{-}76)$$

with $\ln N$ the natural logarithm of the sample size. For $N = 14,600$, $k = 33$ for 40 years of daily values. For $N = 480$, $k = 22$ for 40 years of monthly values.

Since Eq. 4-75 is empirical, no generally accepted method for determination of the number of class intervals exists. The number of class intervals selected for this study is 30, because it is convenient and lies between 22 and 33 for the cases of monthly and daily values, respectively.

Equal probability class intervals are used for the chi-square test in this study. The probability of each class interval is then determined by:

$$P_j = \frac{1}{k} \ , \text{ with } \ j=1,2,\ldots,k. \qquad (4\text{-}77)$$

With these equal probabilities, the corresponding lengths of class intervals are obtained from the c.d.f. The percentages of the chi-square distribution with $(k-1-m)$ degrees of freedom at $F(\chi^2) = 0.95$, 0.99 and 0.995 are summarized in Table 4-1.

Table 4-1 PERCENTAGES OF CHI-SQUARE DISTRIBUTION.

$F(\chi^2)$	Degrees of Freedom								
	30	29	28	27	26	25	24	23	22
0.95	43.8	42.6	41.3	40.1	38.9	37.7	36.4	35.2	33.9
0.99	50.9	49.6	48.3	47.0	45.6	44.3	43.0	41.6	40.3
0.995	53.7	52.3	51.0	49.6	48.3	46.9	45.6	44.2	47.8

4-3 Confidence Limits to Test for Departures from Exponentiality in Frequency Distributions.

Ordinarily it is difficult to select the theoretical probability density functions that will suitably model the heavy tails and the center parts of some frequency distributions of hydrologic random variables. Since the number of observations in the tails of these frequency distributions is usually small, the chi-square test may not adequately distinguish the goodness-of-fit for a selected probability density function. Because the cumulative probability distribution functions have values of zero and one at the tails, the Kolmogorov-Smirnov test also fails to determine how good this fit is [5]. Failure to find the proper theoretical density function which fits the frequency distribution of independent stochastic component in its tails results in a failure to preserve the properties of its extreme values. Consequently, it becomes difficult to generate new samples by the Monte Carlo method which possess the same or similar characteristics of extremes as the historical sample.

As reviewed in Chapter 2, Bryson [5] gives the definition of heavy-tailed probability distributions as distributions that converge to zero much less rapidly than an exponential function. Using the mean residual life time theorem, and the likelihood ratio approach, he derived the expression for the T-statistic in terms of exponentially distributed random variables. The expression for the T-statistic may be used for the test of hypothesis of how heavy the tails are in a distribution from a set of samples of random variables, at the given significance level. This expression is

$$T = \frac{\bar{x}\, x_{(N)}}{(N-1)\prod_{i=1}^{N}\left(x_i + \frac{x_{(N)}}{N-1}\right)^{2/N}} \ , \qquad (4\text{-}78)$$

with x_i the exponentially distributed random variable, $\bar{x}$ the mean of x_i, N the sample size, and $x_{(N)}$ the largest observed value of x_i's.

Unfortunately, the distribution of the T-statistics cannot be found in an explicit form. The 10, 5, and 1 percent critical values can be estimated only by simulating many T values, and taking the 90-th, 95-th, and 99-th percentiles of the frequency distribution of simulated values of T. Each T value is obtained by generating N exponentially distributed random variables x_i and substituting them into Eq. 4-78. These critical values of percentiles cannot be accurate without simulating a large sample of T values. Using both a large number of T values and a large N, this method then becomes applicable.

Because of the need for generating large samples, a simpler method of testing the heavy tails was obtained by Tao [33] by deriving the tolerance limits for the tail from an exponential function.

The exponential cumulative distribution function (c.d.f.) is of the form

$$F_e(x) = 1-e^{-\lambda x}. \qquad (4\text{-}79)$$

Plotting $1-F_e(x)$ against x on a semi-logarithmic paper produces a straight line from the origin with the slope $-\lambda$. For a heavy-tailed c.d.f., the same plotting technique will not show the linearity; instead, the curve will be concave upward. In the opposite case, for the light-tailed c.d.f., the plotted graph will be concave downward. These tails are illustrated in Fig. 4-1, with the light-tailed distribution in the case of normal c.d.f., heavy-tailed distribution in the class of stable c.d.f., and the tail of the exponential c.d.f.

The tolerance limits for the tail of an empirical frequency distribution are derived as follows: assume the empirical distribution denoted as $F_\lambda(x)$ is exponential with the scale parameter λ, then the

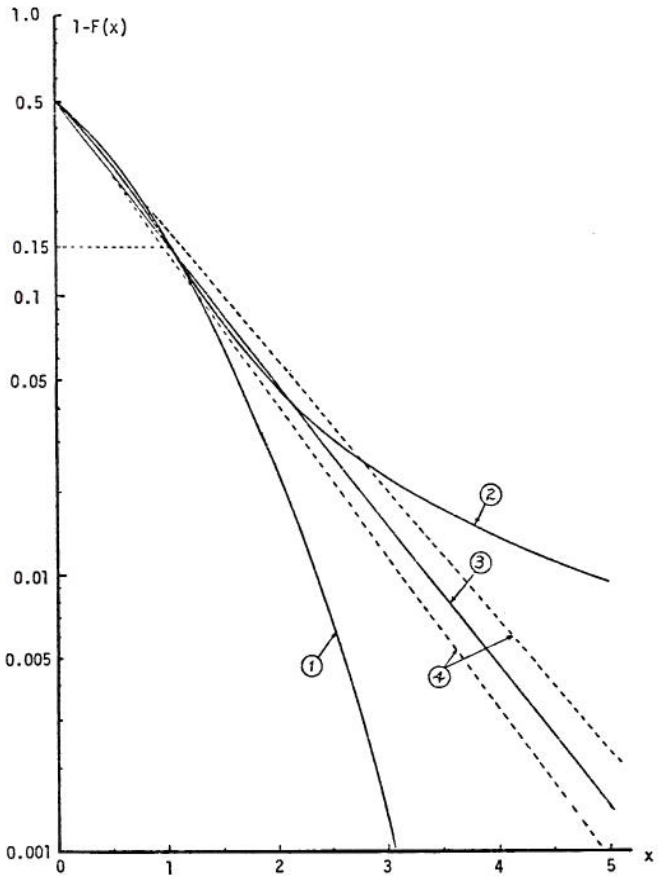

Fig. 4-2 Graphical representation of distributions with different types of tails: (1) standard normal c.d.f., light tail; (2) symmetrical stable c.d.f., with $\alpha = 1.5$, $\gamma = 0.5$, and $\delta = 0$, heavy tails; (3) exponential c.d.f., $F(x) = 1 - \frac{1}{2} e^{-1.155x}$; and (4) 90 percent tolerance limits for the exponential tail in case of sample size 500.

relationship between $\ln[1-F_\lambda(x)]$ and x is linear, the probability P that $\ln[1-F_\lambda(x)]$ is within certain limits is

$$P = Pr\left\{\theta_1(x) \le \ln[1-F_\lambda(x)] \le \theta_2(x)\right\} , \qquad (4\text{-}80)$$

with θ_1 and θ_2 the functions of x and the lower and the upper tolerance limits, respectively. Substituting Eq. 4-79 into Eq. 4-80 for $F_\lambda(x)$ and rearranging,

$$P = Pr\left[\frac{-\theta_1(x)}{x} \ge \lambda \ge \frac{-\theta_2(x)}{x}\right] , \qquad (4\text{-}81)$$

with $x > 0$. The parameter λ is estimated from the random variables x_i with sample size N by

$$\hat{\lambda} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_i} \qquad (4\text{-}82)$$

Substituting Eq. 4-82 into Eq. 4-81 and rearranging it,

$$P = Pr\left[\frac{-Nx}{\theta_1(x)} \le \sum_{i=1}^{N} Xi \le \frac{-Nx}{\theta_2(x)}\right] . \qquad (4\text{-}83)$$

With the estimate $\hat{\lambda}$ and the random variable x_i, the probability P of Eq. 4-83 is equal to or smaller than 1. If x_i is exponentially distributed, $\sum_{i=1}^{N} x_i$ has a gamma distribution with the shape parameter N and the scale parameter $\hat{\lambda}$. For a large shape parameter, the gamma distribution converges to a normal distribution with the mean $N/\hat{\lambda}$, and the variance $N/\hat{\lambda}^2$ [24]. The application to Eq. 4-83 at the 90 percent tolerance level results in

$$P = 0.90 = Pr\left[-\frac{Nx}{\theta_1(x)} \le \sum_{i=1}^{N} x_i \le \frac{-Nx}{\theta_2(x)}\right]$$

$$= Pr\left[\frac{N}{\hat{\lambda}} - 1.645\sqrt{\frac{N}{\hat{\lambda}^2}} \le \sum_{i=1}^{N} x_i \le \frac{N}{\hat{\lambda}} + 1.645\sqrt{\frac{N}{\hat{\lambda}^2}}\right] . \qquad (4\text{-}84)$$

After rearranging, the tolerance limits for $\ln[1-F_e(x)]$ at 90 percent level become, for the lower and upper limits are

$$\theta_1(x) = \frac{-\hat{\lambda}x}{1 - \frac{1.645}{\sqrt{N}}} , \text{ and } \theta_2(x) = \frac{-\hat{\lambda}x}{1 + \frac{1.645}{\sqrt{N}}} , \qquad (4\text{-}85)$$

where $\hat{\lambda}$ is estimated by Eq. 4-82 and N is the sample size.

With these tolerance limits for the exponential tail and a given significance level, the test of the hypothesis that the probability distribution of a variable has the heavy tail can then be made. Fig. 4-1 gives three types of tails: for the standard normal c.d.f., curve (1); for the symmetric stable c.d.f., curve (2); and the exponential c.d.f., curve (3). Curve (4) gives the tolerance limits for the exponential tail at the 90 percent level, and the sample $N = 500$. To make the tails comparable, two conditions are designed for c.d.f.'s:

a. All three c.d.f.'s should satisfy $F(0) = 0.5$ and $F(\infty) = 1$; and

b. Since the variance of the stable distribution does not exist, it is not possible to compare these three c.d.f.'s with equal variance. Therefore, all three c.d.f.'s should intersect at an arbitrarily selected point, like the 85 percent percentile. For the first condition, the exponential c.d.f. should be in the form of $F_e(x) = 1 - \frac{1}{2} e^{-\lambda x}$. For the second condition, the scale parameter of the exponential c.d.f. should be $\lambda = 1.155$, while the parameters of the symmetric stable c.d.f. should be: location, $\delta = 0$; scale, $\gamma = 1/2$, (note that for standard normal function, $\alpha = 2$, $\delta = 0$ and $\gamma = 0.5$), and the characteristic exponent α equal approximately 1.5.

With the above conditions for plotting the c.d.f., Fig. 4-1 shows that the tail of the normal function converges to zero much faster and is significantly higher than that of the exponential function for $x \ge 1.2$. The tail of the stable function is much heavier than that of exponential function, for $x \ge 2.8$. The tolerance limits are rather narrow because

$\theta_1(x)$ and $\theta_2(x)$ are estimated under the assumption that the random variable x are exponentially distributed and $\theta_1(x)$ and $\theta_2(x)$ are good for every x but not simultaneously for all x. It is easy to use the tolerance limits $\theta_1(x)$ and $\theta_2(x)$, derived for the exponential tail, to test the empirical tail distribution of the independent stochastic variables in order to show how rapidly this tail converges to zero in comparison with the corresponding exponential tail.

4-4 Use of Gnedenko's F-Criterion Statistic to Test the Nature of the Tails of Frequency Distributions.

The generalized exponential cumulative distribution function is of the form

$$F(x) = \begin{cases} 1 - e^{-\lambda(x-a)} & \text{for } x > a \\ 0 & \text{for } x \le a \end{cases} . \quad (4\text{-}86)$$

Here, λ and a are constants. The hypothesis that a sample of data is part of an exponential population against an alternative hypothesis that λ is not a constant but a function of x could be tested using the F-criterion test suggested by Gnedenko [13].

If x_i, $i = 1,2,\ldots,n$, form an ordered sample of data, with x_1 as the lowest, a new variable S_i may be defined as

$$S_i = (n-i+1)(x_i - x_{i-1}),\ x_0 = 0,\ i = 1,2,\ldots,n.$$

After dividing the S_i into two groups of length r and $n-r$, the following random variable, known as Gnedenko's statistic, is obtained:

$$Q(r,n-r) = \sum_{i=1}^{r} (S_i/r) \Big/ \sum_{i=r+1}^{n} (S_i/(n-r)). \quad (4\text{-}87)$$

This statistic has the F-distribution with $2r$ and $2(n-r)$ degrees of freedom under the null hypothesis, H_o, which is that x is exponentially distributed. This means that, if α is the level of significance,

$$Q(r,n-r) < f_{(1-\alpha/2)}(r,n-r) \quad (4\text{-}88)$$

and

$$Q(r,n-r)^{-1} < f_{(1-\alpha/2)}(n-r,r) . \quad (4\text{-}89)$$

The alternative hypothesis is that λ is either an increasing or a decreasing function of x.

If either constraint does not hold the null hypothesis is rejected at the level of significance α and the given division of the sample into segments of length r and $n-r$. It is obvious that in every case at least one constraint is satisfied. If constraint of Eq. 4-88 does not hold, one concludes that λ is a decreasing function of x and that the tail is light. On the other hand if Eq. 4-89 does not hold the conclusion is that λ is an increasing function of x and that the tail is heavy. In the terminology of failure rates [Fercho and Ringer, 13], the first of these nonconforming conditions signify that the failure rate decreases and the second that the failure rate increases. The null hypothesis corresponds to a constant failure rate.

Chapter 5
EMPIRICAL RESULTS AND THEIR DISCUSSION

Seventeen daily runoff series are used as the basic research data in this study. Patterns of these runoff series vary depending on the geographic location and climatic conditions of the river basin. Figures 5-1 and 5-2 show two selected years of daily flows, the daily flow means and daily flow standard deviations for the Tioga and Boise Rivers. It is obvious from the curves of the Tioga River that highly fluctuating runoff series result in highly fluctuating daily means and daily standard deviations. The smoother daily runoff series of the Boise River result in smoother daily means and standard deviation curves. This pattern should be expected taking into account the sampling variations.

Removal of periodicities from the daily means and daily standard deviations, followed by removing the dependence from the remaining series, produce the independent stochastic component. Frequency distributions of independent stochastic variables of the daily runoff series are of interest in this study. Frequency distributions of these seventeen daily series are plotted in Fig. 5-3. A general pattern of these curves is that the peaks are high and sharp at the center while the tails are long. Daily runoff series and the 3-day, 7-day, 13-day and monthly average runoff series are processed by using the same technique.

5-1 Procedure Used in Producing the Independent Stochastic Components and Their Properties.

The procedure used. The procedure and equations used to obtain the independent stochastic components from the observed series are summarized in the form of the flow chart.

The independent stochastic component obtained by the above procedure is denoted as the ξ-variable.

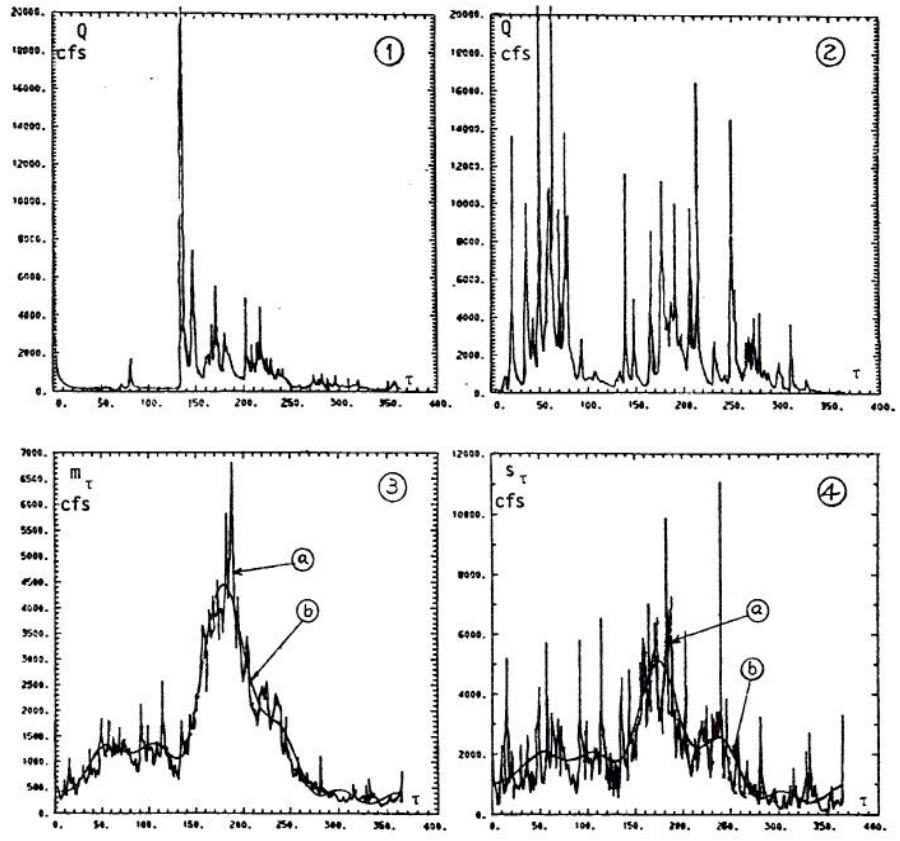

Fig. 5-1 (1) Daily flow series for the dry year, 1926; (2) daily flow series for the wet year, 1929; (3) daily means; (4) daily standard deviations, with (a) the computed values, and (b) the fitted periodic function, for the Tioga River (1921-1960).

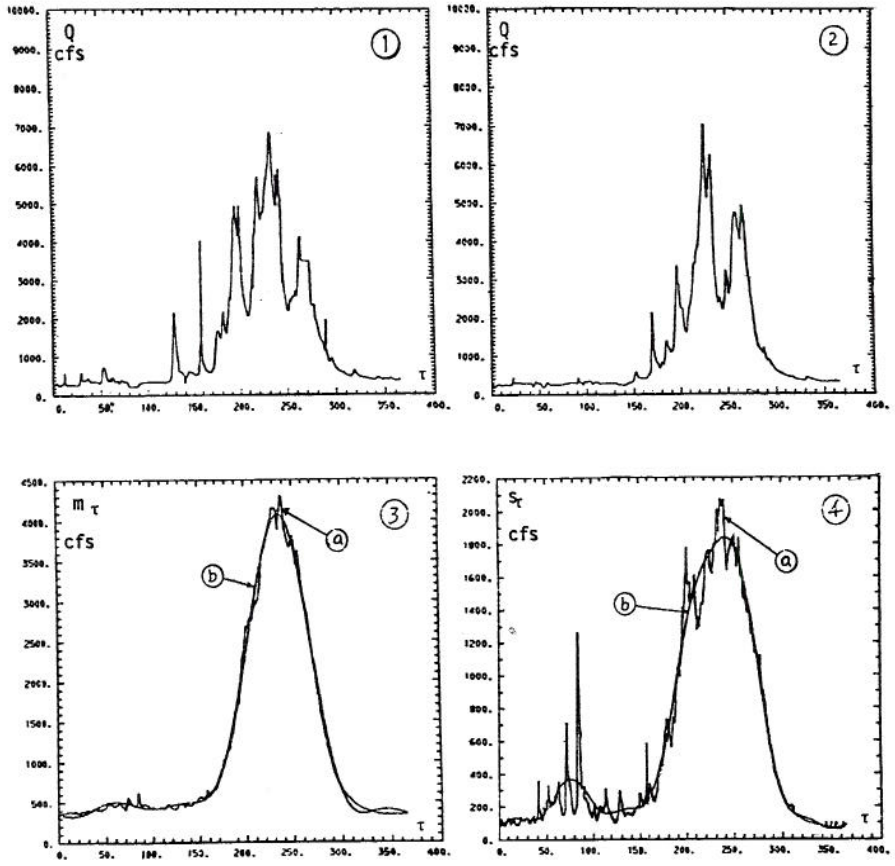

Fig. 5-2 (1) Daily flow series for the year 1926; (2) daily flow series for the year 1933; (3) daily means; (4) daily standard deviations, with (a) the computed values, and (b) the fitted periodic function, for the Boise River (1921-1960).

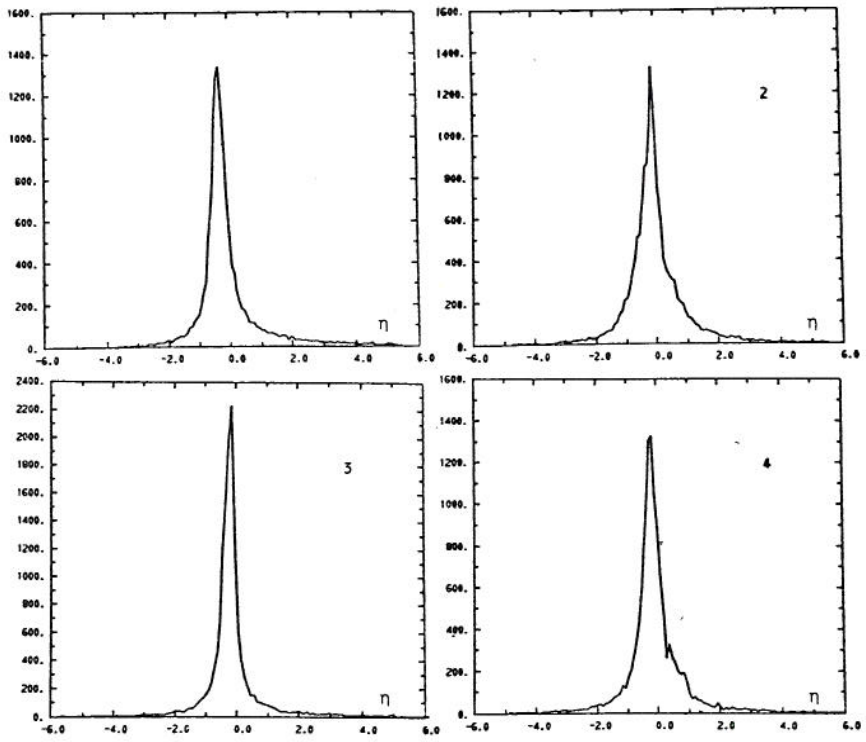

Fig. 5-3 Frequency distributions of daily independent stochastic components for 120 equal size class intervals (ordinates of curves are the absolute frequencies) (to continue).

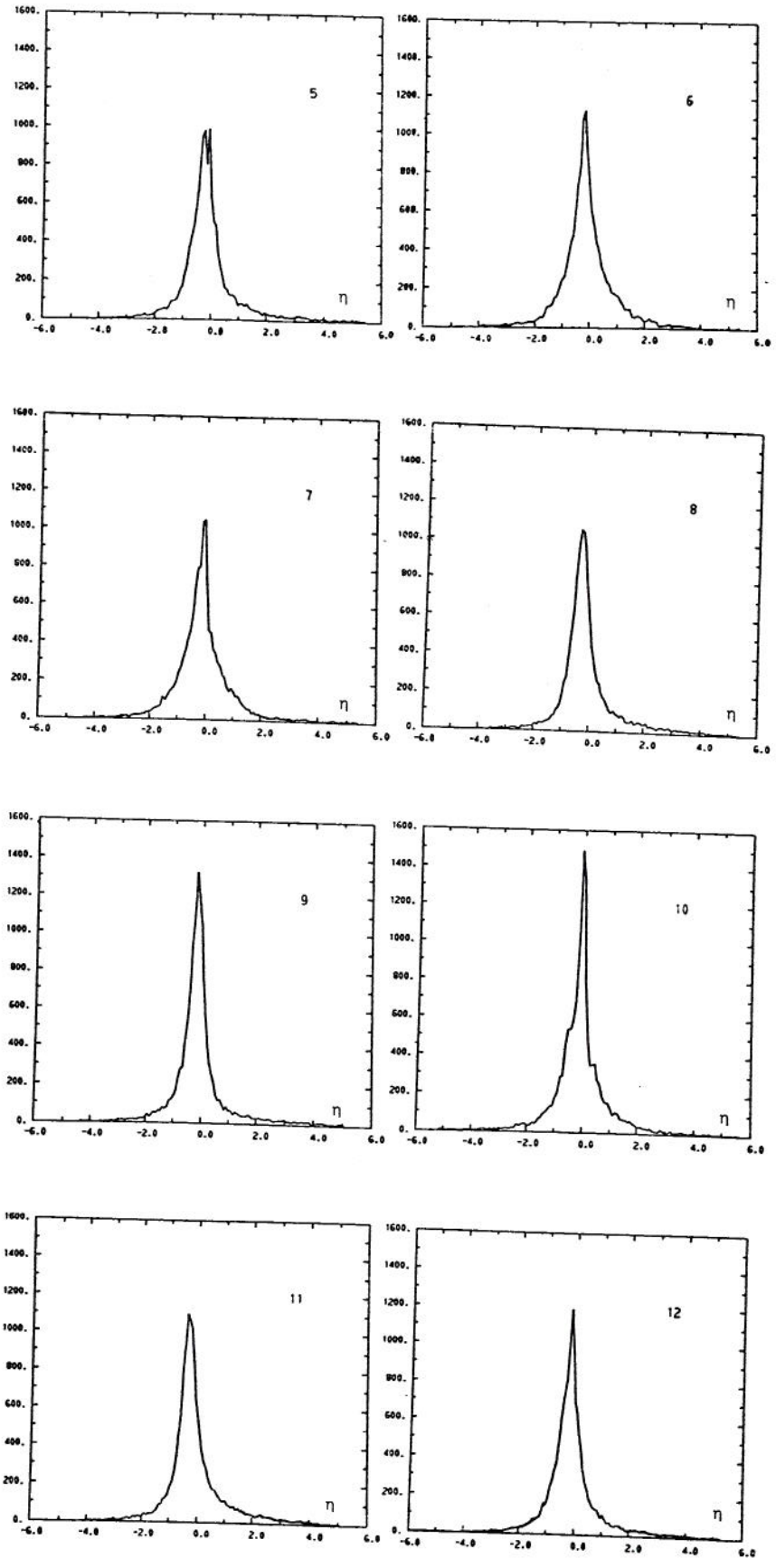

Fig. 5-3 Frequency distributions of daily independent stochastic components for 120 equal size class intervals (ordinates of curves are the absolute frequencies) (to continue).

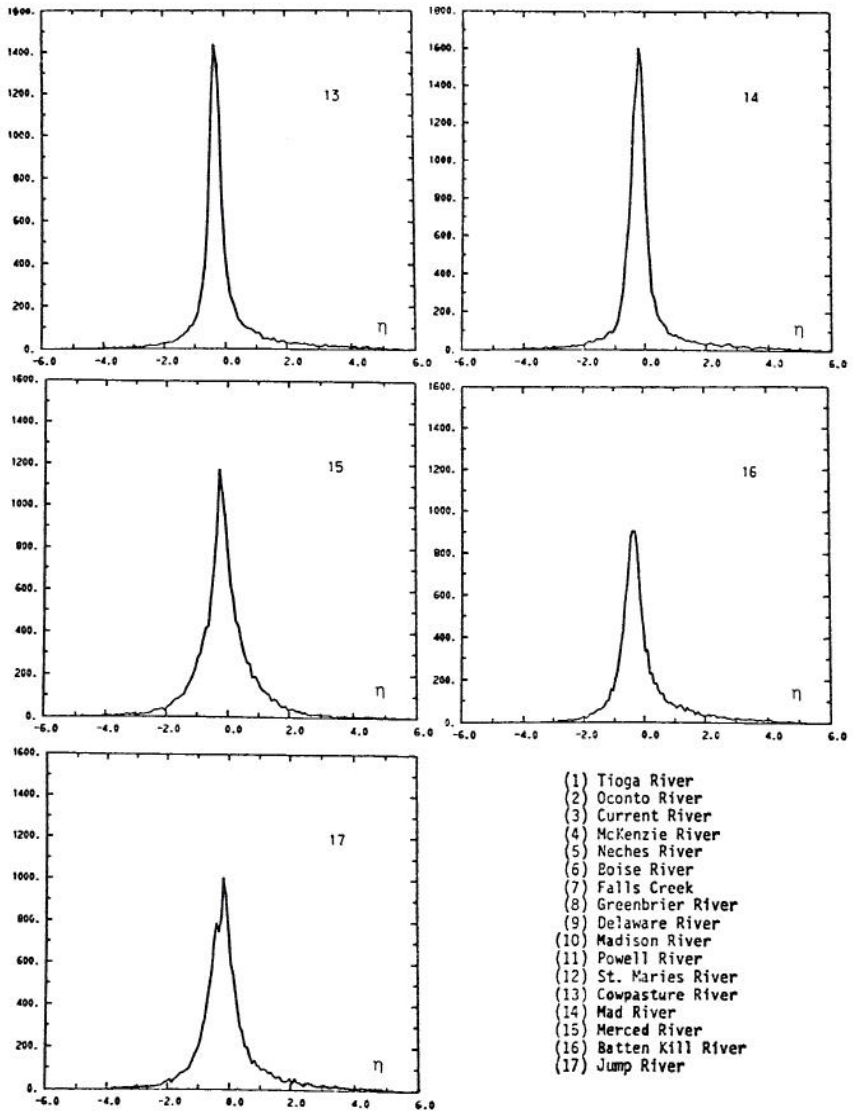

Fig. 5-3 Frequency distributions of daily independent stochastic components for 120 equal size class intervals (ordinates of curves are the absolute frequencies).

STEP 1 Prepare the series $Q_{p,\tau}$, $p = 1,2,\ldots,n$, $\tau = 1,2,\ldots,\omega$.

STEP 2 Compute means Q_τ, standard deviations s_τ, by Eqs. 3-2 and 3-3.

STEP 3 Substitute Q_τ series into Eq. 3-6, and compute the first six harmonics; determine the significant harmonics by using the empirical test method; compute the periodic component in Q_τ by using Eq. 3-15. Apply the same procedure to s_τ series.

STEP 4 Use Eq. 3-17 to remove periodicities in the mean and standard deviation.

STEP 5 Use Eq. 3-18 to standardize the remaining series of Eq. 3-17.

STEP 6 Compute serial correlation coefficients $r_{1,\tau}$, $r_{2,\tau}$ and $r_{3,\tau}$ by Eq. 3-36.

STEP 7 Apply the harmonic analysis to $r_{1,\tau}$, $r_{2,\tau}$, and $r_{3,\tau}$ series, following the procedure of STEP 3.

STEP 8 Compute the determination coefficients D_1, D_2, and D_3 by Eqs. 3-37 through 3-39, and determine the order of the autoregressive linear model.

STEP 9 Estimate the parameters of the autoregressive linear model by using Eqs. 3-27 through 3-35.

STEP 10 Compute the independent stochastic component by using Eq. 3-25.

In case the logarithmic transformation is used in the form of:

$$y'_{p,\tau} = \ln (y_{p,\tau} - \gamma) \qquad (5\text{-}1)$$

with $y_{p,\tau}$ obtained in STEP 4, and γ the lower boundary of $y_{p,\tau}$, a new STEP 4a is included with all other steps remaining unchanged. The independent stochastic component obtained by the logarithmic transformation of Eq. 5-1 is denoted as the ξ-variable. Since there are no negative observations in $Q_{p,\tau}$, it follows that $\min(Q_{p,\tau}) = 0$. Consequently,

$$\gamma = \min (y_{p,\tau}) = \min \left(\frac{Q_{p,\tau} - \mu_\tau}{\sigma_\tau} \right) = - \max \left(\frac{\mu_\tau}{\sigma_\tau} \right) .$$

This transformation is approximately equivalent to substituting the autoregressive linear model of Eq. 3-25 by a dependence model of the form

$$\varepsilon_{p,\tau} = \prod_{i=1}^{p} \left[(\varepsilon_{p,\tau-1})^{\alpha_{i,\tau}} \right] e^{\sigma_{\xi,\tau} \xi_{p,\tau}} . \qquad (5\text{-}2)$$

In the application of the logarithmic transformation to $Q_{p,\tau}$ in STEP 1, for $Q_{p,\tau} > 0$, $Q_{p,\tau}$ in STEP 1 is replaced by $\ln(Q_{p,\tau})$. For $Q_{p,\tau} = 0$, any small value such as 0.01, 0.001, or 0.0001 may be assigned to $Q_{p,\tau}$ in order to permit the logarithmic transformation, all other steps remaining unchanged. By this approach the resulting independent stochastic component is denoted as the η-variable.

Characteristics of independent stochastic components of seventeen daily flow series. The basic statistical characteristics of the ξ, ζ, and η variables of the selected daily flow series, each with approximately 40 years of data, are given in Tables 5-1 through 5-3, respectively.

Since the independent stochastic components are derived from the standardized dependent stochastic variables, $\varepsilon_{p,\tau}$, as demonstrated by STEP 5, the mean and the variance of these independent stochastic components should be zero and one, respectively. Tables 5-1 through 5-3 show that the means of ξ, ζ, and η variables for all 17 daily runoff series are practically zero. However, the variances of the η-variable are closest to unity. In addition, the η-variable has the smallest skewness and kurtosis coefficients, and the absolute values of extremes. The ξ-variable has the largest parameters.

When the standard deviation is proportional to the mean, the use of logarithmically transformed variables leads to the more consistent results for the independent stochastic component than the use of the original variables, because the transformed variable has a constant standard deviation over the τ positions. Although the application of logarithmic transformation does not convert the standard deviations to constants for all the 17 daily runoff series, this transformation still produces a substantial improvement. In addition, the logarithmic transformation has the advantage of not generating negative values when the new samples are generated.

The correlogram of independent stochastic components of 17 daily flow series up to 200 lags are given in Fig. 5-4. In order to test the independence of these variables, the tolerance limits of r_1, given by Anderson and computed by using Eq. 3-40, and -0.1629 and +0.01615 for samples of 40 years, at the 95 percent significance level. With the same sample size and the same significance level, the tolerance limits of r_k, given by Siddiqui, are computed by using Eq. 3-44. These results are practically the same as the Anderson's. The tolerance limits of the Fisher's z-transformation of r_k, computed by Eq. 3-46 are ± 0.01622.

The tolerance intervals of these three tests were found to be too narrow for practical use. Since r_k is asymptotically normally distributed with mean and variance given by Eqs. 3-42 and 3-43, respectively, an alternative test of r_k may be formulated by testing the frequency distribution of r_k. If the hypothesis that the frequency distribution of r_k is normal with the mean and variance given by Eqs. 3-42 and 3-43 is accepted, then the hypothesis of the independence of the tested stochastic component is also accepted.

The chi-square statistic may be used to test the goodness-of-fit of the theoretical normal function to the frequency distribution of r_k. Three stations passed this test at 99 percent significance level (the Tioga, Mckenzie, and Falls Creek Rivers) and two stations passed this test at 99.5 percent significance level (the St. Maries, and Merced Rivers). If this test is applied to the theoretical normal function for fitting the z-transform of r_k, five stations passed the test at 99 percent significance level (the Tioga, Mckenzie, Falls Creek, Greenbrier, and Powell Rivers) and three stations passed the test at 99.5 percent significance level (the St. Maries, Cowpasture, and Merced Rivers).

The models for the periodic-deterministic components. The three main sources of stream runoff are groundwater effluence, rainfall, and snowmelt. In areas where runoff is produced predominately by rainfall, the runoff is highly irregular because of randomness in rainfall. On the other hand, when the groundwater or snowmelt have large influences on runoff, the runoff is more regular because of the water storage of these two factors.

The daily runoff series of the Tioga River is used as an example to demonstrate the above points.

Table 5-1 BASIC CHARACTERISTICS OF THE DAILY ξ-VARIABLE FOR SEVENTEEN RUNOFF STATIONS IN USA.

River	Mean	Variance	Skewness Coefficient	Kurtosis Coefficient	Observed Min. (ξ)	Second Min. (ξ) Observed	Observed Max. (ξ)	Second Max. (ξ) Observed
1 Tioga	-0.00005	1.0000	12.47	292.10	-13.32	-10.22	32.45	30.42
2 Oconto	0.00013	0.9999	36.63	4130.09	-46.10	-26.37	86.18	14.33
3 Current	-0.00003	1.0000	29.15	2329.47	-36.53	- 7.67	74.48	20.71
4 Mckenzie	-0.00013	1.0000	4.96	203.57	-29.38	-12.66	31.33	19.58
5 Neches	0.00005	0.9999	5.53	322.69	-32.62	-17.73	36.26	23.18
6 Boise	-0.00021	1.0000	3.82	210.34	-25.65	-18.53	30.25	26.78
7 Fall Creek	-0.00001	1.0000	5.89	196.08	-17.39	-14.22	35.93	16.37
8 Greenbrier	-0.00001	1.0000	8.26	156.35	-18.41	- 7.33	32.04	19.15
9 Delaware	-0.00001	1.0000	8.10	115.59	-10.55	- 9.94	20.23	17.69
10 Madison	-0.00009	1.0000	0.43	160.40	-28.64	-21.04	27.23	19.50
11 Powell	0.00007	1.0000	4.68	80.32	-17.72	-10.20	20.53	18.18
12 St. Maries	0.00000	1.0000	47.62	4490.51	-34.83	-16.30	88.16	21.23
13 Cowpasture	-0.00001	1.0000	8.66	140.80	-10.27	- 9.75	26.45	20.16
14 Mad	-0.00003	1.0000	8.47	149.67	-12.91	-10.10	28.12	24.02
15 Merced	0.00000	1.0000	21.31	555.57	-14.15	- 2.43	38.29	31.76
16 Batten Kill	0.00004	1.0000	18.26	772.26	-21.66	-11.32	49.77	42.73
17 Jump	-0.00000	1.0000	10.44	211.37	-11.24	- 9.86	29.70	23.04

Table 5-2 BASIC CHARACTERISTICS OF THE DAILY ζ-VARIABLE FOR SEVENTEEN RUNOFF STATIONS IN USA.

River		Mean	Variance	Skewness Coeff.	Kurtosis Coeff.	Observed Min. (ζ)	Second Min. (ζ) Observed	Observed Max. (ζ)	Second Max. (ζ) Observed
1	Tioga	-0.0022	1.340	1.51	49.99	-5.67	-5.62	26.8	8.93
2	Octonto	0.0004	1.087	0.96	14.03	-5.20	-5.17	10.0	12.84
3	Current	-0.0007	1.578	-0.01	45.24	-7.76	-7.69	17.8	11.08
4	Mckenzie	-0.0005	1.240	0.92	20.75	-5.95	-5.90	14.7	12.09
5	Neches	0.0003	1.094	1.45	13.19	-4.33	-4.31	8.3	7.91
6	Boise	-0.0005	1.189	0.36	35.14	-4.98	-4.96	18.9	11.86
7	Falls Creek	-0.0013	1.457	1.97	61.32	-6.65	-6.61	30.4	14.39
8	Greenbrier	0.0089	1.354	1.78	19.93	-4.94	-4.93	16.4	11.62
9	Delaware	-0.0039	1.336	1.16	28.92	-6.59	-6.48	12.8	11.56
10	Madison	-0.0003	1.114	0.41	15.57	-5.30	-5.27	11.6	10.93
11	Powell	0.0058	1.401	0.67	21.26	-6.00	-5.94	13.4	12.06
12	St. Maries	0.0019	1.142	1.46	15.00	-4.83	-4.79	11.2	8.52
13	Cowpasture	0.0048	1.550	1.06	23.25	-6.67	-6.61	15.3	12.82
14	Mad	-0.0018	1.279	2.03	28.87	-5.70	-5.64	17.7	11.93
15	Merced	-0.0000	1.357	-1.51	56.19	-5.81	-5.74	14.6	12.22
16	Batten Kill	-0.0013	1.166	1.85	15.69	-4.11	-4.07	13.5	11.39
17	Jump	0.0053	1.476	0.93	33.48	-6.41	-6.35	17.4	17.08

In Fig. 5-1 the daily series of two typical years, dry and wet, (1) and (2), are given together with the 365 daily means, (3), and the 365 daily standard deviations, (4). Periodic functions used for daily means and standard deviations are the deterministic periodic components of the series. Because of large variations in the runoff process, daily means and its inferred periodic function depart significantly from the series of individual years. Under this high fluctuation, the estimated periodic function of parameters must deviate highly from the true population periodic component.

The opposite example is the daily flow series of the Boise River. The daily series of two years, and daily means and standard deviations are shown in Fig. 5-2. Fluctuations are much less than for the Tioga River. The daily flow series of individual years of the Boise River have similar general patterns

Table 5-3 BASIC CHARACTERISTICS OF THE DAILY η-VARIABLE FOR SEVENTEEN RUNOFF STATIONS IN USA.

	River	Mean	Variance	Skewness Coeff.	Kurtosis Coeff.	Observed Min. (η)	Second Min. (η) Observed	Observed Max. (η)	Second Max. (η) Observed
1	Tioga	-0.2011	1.104	2.09	13.74	-5.75	-5.69	15.22	7.26
2	Oconto	0.0003	1.066	0.83	11.92	-7.56	-6.82	7.49	11.92
3	Current	-0.0001	1.214	1.47	19.32	-12.25	-8.75	14.42	8.71
4	Mckenzie	-0.0006	1.150	1.11	14.11	-8.46	-7.63	10.84	8.71
5	Neches	-0.0000	1.068	1.43	10.40	-6.77	-6.11	6.59	6.45
6	Boise	-0.0008	1.080	0.74	11.19	-9.05	-8.78	8.62	7.58
7	Fall Creek	0.0006	1.072	0.92	10.14	-7.32	-6.99	7.00	6.76
8	Greenbrier	0.0057	1.113	1.67	9.50	-6.53	-5.53	7.27	6.52
9	Delaware	0.0000	1.080	1.30	10.76	-6.87	-6.21	6.74	6.63
10	Madison	-0.0009	1.092	0.30	13.69	-12.59	-10.66	8.69	8.24
11	Powell	0.0041	1.146	1.38	10.54	-7.88	-6.23	13.12	7.88
12	St. Maries	0.0005	1.078	1.38	11.28	-7.37	-5.79	11.21	8.11
13	Cowpasture	-0.0020	1.138	1.66	13.08	-11.74	9.86	7.45	7.17
14	Mad	-0.0004	1.100	1.52	18.66	-17.32	-6.85	10.58	8.37
15	Merced	0.0000	1.110	0.63	12.63	-9.85	-8.62	8.26	7.65
16	Batten Kill	-0.0009	1.077	1.86	9.15	-4.81	-4.34	7.55	7.01
17	Jump	-0.0003	1.072	1.29	9.20	-6.53	-6.22	7.01	6.86

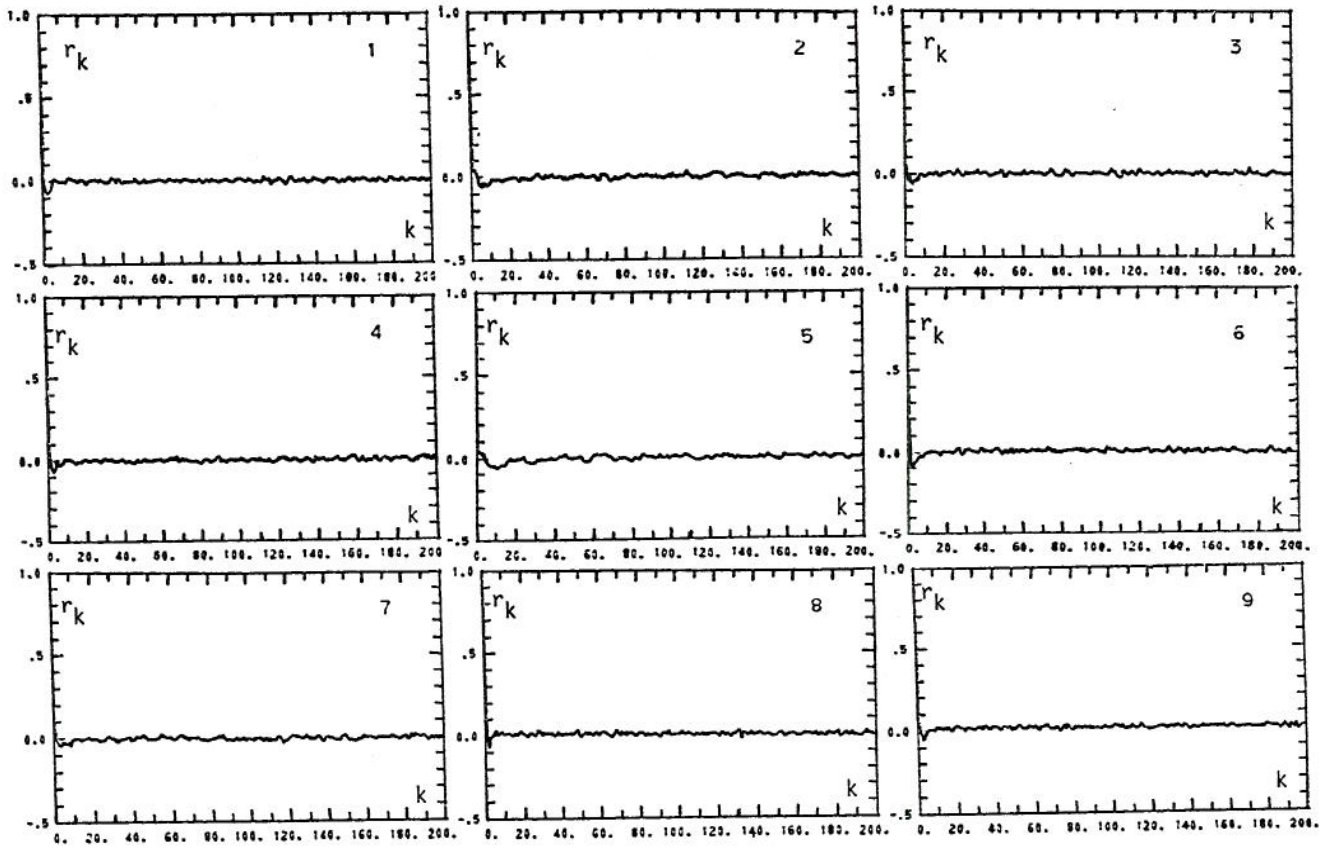

Fig. 5-4 Correlograms of the η-variable of daily runoff series (to continue).

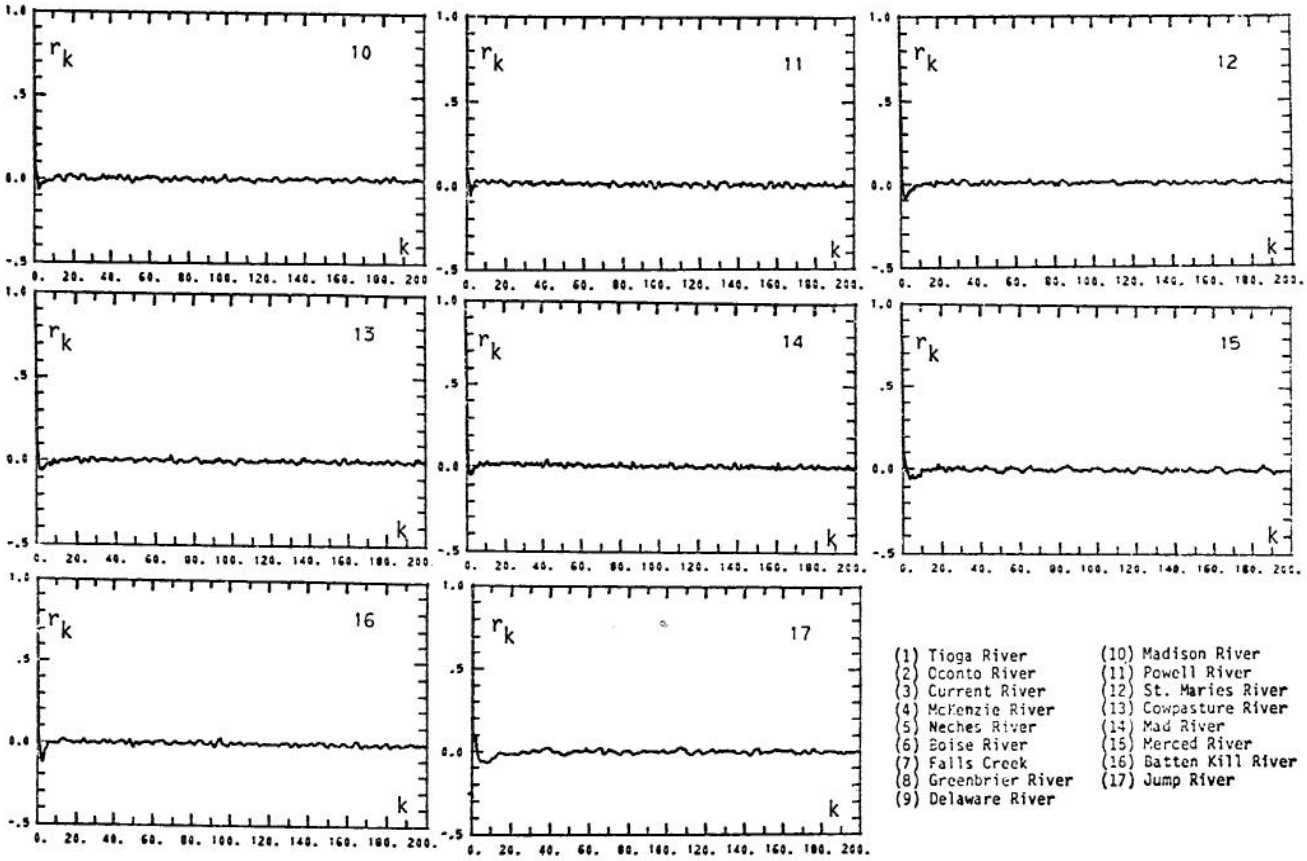

Fig. 5-4 Correlograms of the η-variable of daily runoff series.

as the means and standard deviations. The periodic mean and standard deviation are, therefore, subject to smaller sampling errors than for the Tioga River.

5-2 Periodicities in Serial Correlation Coefficients.

The serial correlation coefficients, $r_{k,\tau}$, with $k = 1$, 2, and 3, computed by Eq. 3-36 are independent of the removal of periodicities in the mean and standard deviation from the original variable, $Q_{p,\tau}$. These coefficients indicate the correlation between the τ-th and the $(\tau+k)$-th values for the available sample series. The periodicity may exist in $r_{k,\tau}$. Figure 5-5 shows periodicities in the 13-day, 7-day, and 3-day values of $r_{1,\tau}$. However, $r_{1,\tau}$ of daily series has a high fluctuation so that the periodicity must be inferred by an objective test, and not concluded by a visual inspection. Figures 5-6 and 5-7 show the $r_{2,\tau}$ and $r_{3,\tau}$ series, for the Tioga River, respectively, and Figs. 5-8 through 5-10 the $r_{1,k}$, $r_{2,k}$ and $r_{3,k}$ series for the Neches River.

The fluctuations of a daily runoff series is somewhat decreased by averaging the flows of Δt consecutive days. This may be the reason why the periodicity in the serial correlation coefficient becomes more obvious for larger values of Δt.

For a better insight into the periodic parameters, the cumulative periodograms of $r_{k,\tau}$ are plotted in Fig. 5-11 for the daily runoff series of 17 stations. Figure 5-11 shows the sums of explained variance of the first m harmonics versus m for $r_{1,\tau}$, $r_{2,\tau}$ and $r_{3,\tau}$, with m the sequential index of harmonics. The frequency of the m-th harmonic is m/ω, and for the daily series, $\omega = 365$. The shape of $P = f(m)$ is convex upward as shown for all 51 curves plotted in Fig. 5-11. A sudden rise of the cumulative periodogram of $r_{k,\tau}$, for a few harmonics with lowest frequencies indicates periodicities in autocorrelation coefficients. In the application of empirical tests to determine the significant harmonics in $r_{k,\tau}$, $k = 1,2,3$, the critical value, $P_{min} = 0.071$, as given by Eq. 3-12, with $c = 2$, $a = 0.033$, and $n = 40$. Only three stations show the sum of the explained variances in $r_{1,\tau}$ by the first six harmonics to be smaller than the critical value P_{min}: the Oconto (2), Neches (5), and Mad (14) Rivers. In general, the rivers with the runoff predominately produced by rainfall demonstrate less periodicity in serial correlation coefficients than rivers with runoff produced by both rainfall and snow accumulation and melt.

The $P = f(m)$ curves for $r_{1,\tau}$ are below the curves for $r_{2,\tau}$ and $r_{3,\tau}$, while the curve for $r_{2,\tau}$ is below that for $r_{3,\tau}$. These are the general patterns for all 17 daily series. One of the reasons for this pattern may be that the autocorrelation of successive values ($\varepsilon_{p,\tau}$ and $\varepsilon_{p,\tau+1}$) is affected by more sampling variation than the autocorrelation for the lags 2 and 3 ($\varepsilon_{p,\tau}$ and $\varepsilon_{p,\tau+2}$; $\varepsilon_{p,\tau}$ and $\varepsilon_{p,\tau+3}$).

For the daily series of the Tioga River, the total explained variance of the first six harmonics of

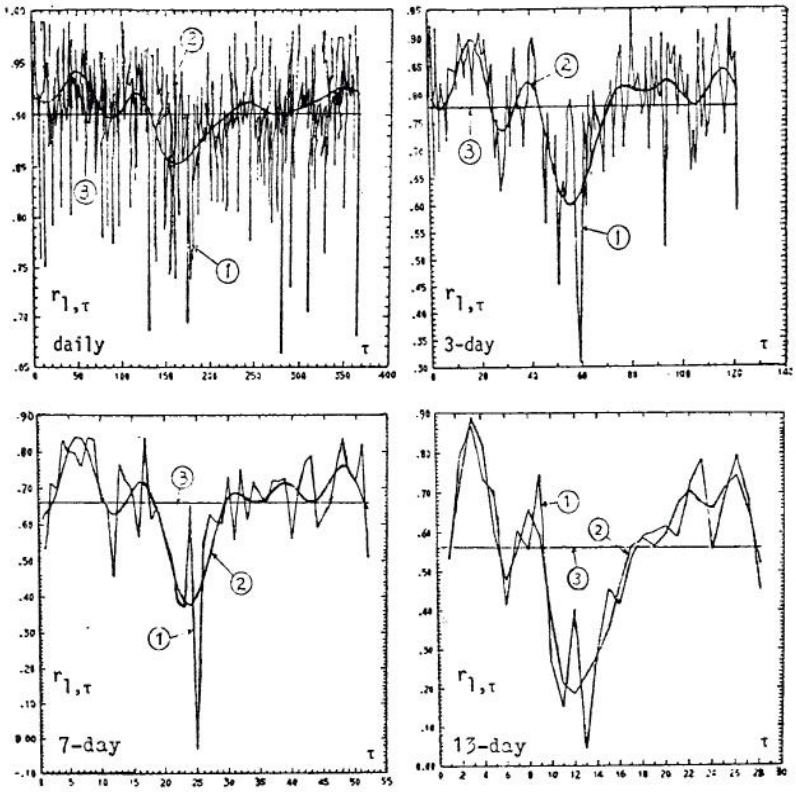

Fig. 5-5 The $r_{1,\tau}$ series of daily, 3-day, 7-day, and 13-day flow series for the logarithmically transformed discharges of the Tioga River: (1) Computed $r_{1,\tau}$; (2) $r_{1,\tau}$ series composed of six harmonics; and (3) the mean of $r_{1,\tau}$.

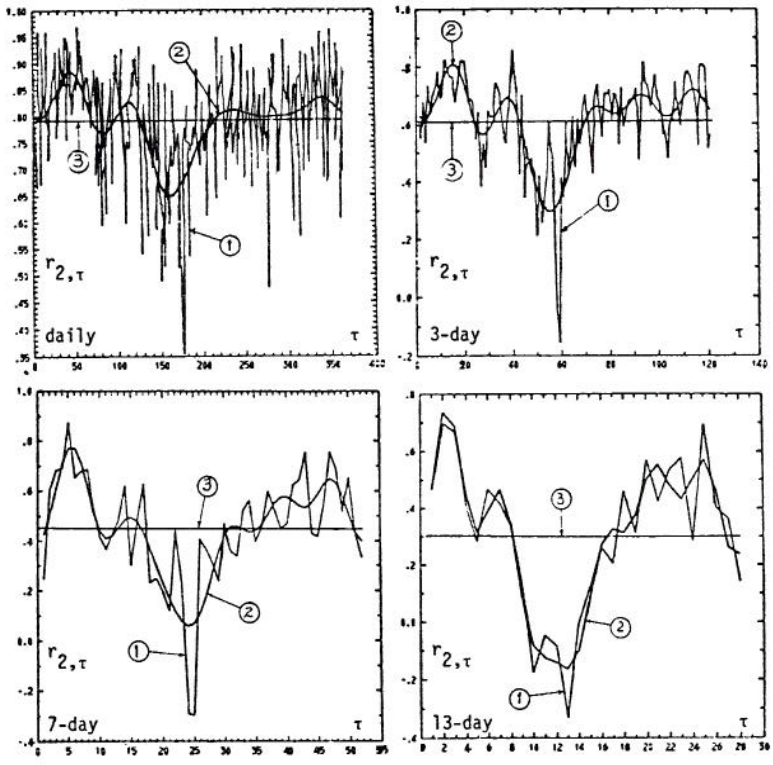

Fig. 5-6 The $r_{2,\tau}$ series of daily, 3-day, 7-day, and 13-day flow series for the logarithmically transformed discharges of the Tioga River: (1) Computed $r_{2,\tau}$; (2) $r_{2,\tau}$ series composed of six harmonics; and (3) the mean of $r_{2,\tau}$.

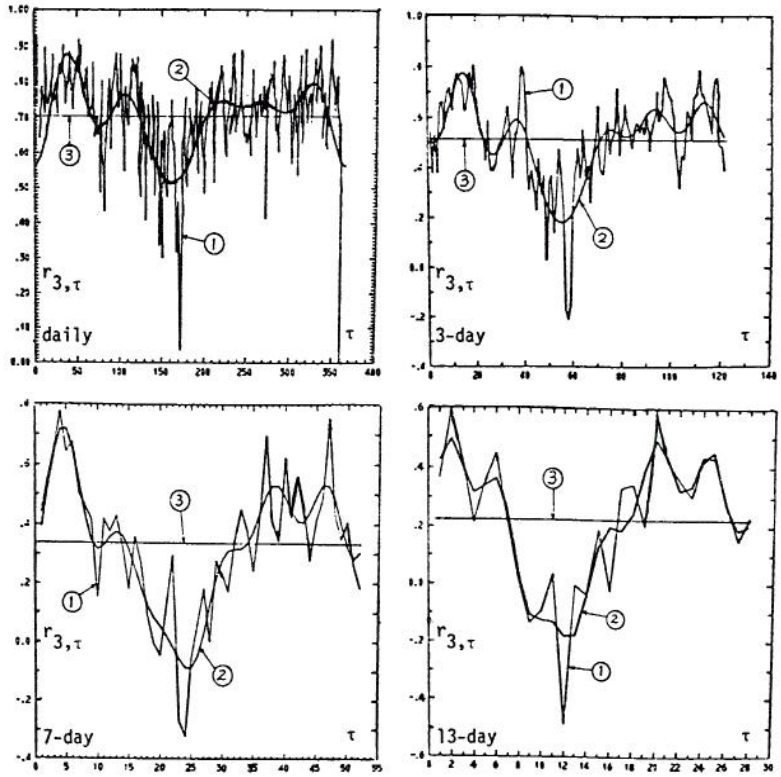

Fig. 5-7 The $r_{3,\tau}$ series of daily, 3-day, 7-day, and 13-day flow series for the logarithmically transformed discharges of the Tioga River: (1) Computed $r_{3,\tau}$; (2) $r_{3,\tau}$ series composed of six harmonics; and (3) the mean of $r_{3,\tau}$.

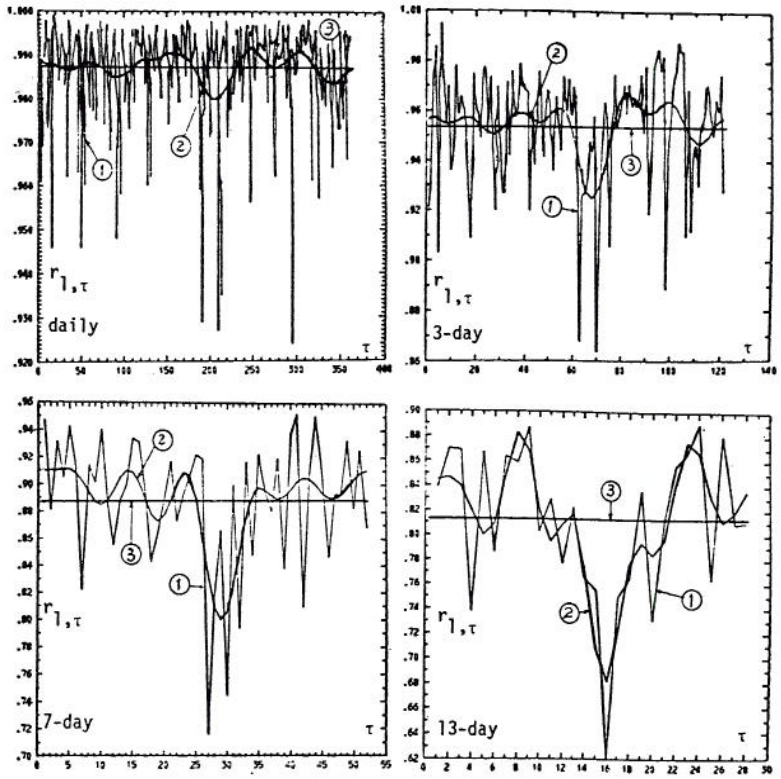

Fig. 5-8 The $r_{1,\tau}$ series of daily, 3-day, 7-day, and 13-day flow series for the logarithmically transformed discharges of the Neches River: (1) Computed $r_{1,\tau}$; (2) $r_{1,\tau}$ series composed of six harmonics, and (3) the mean of $r_{1,\tau}$.

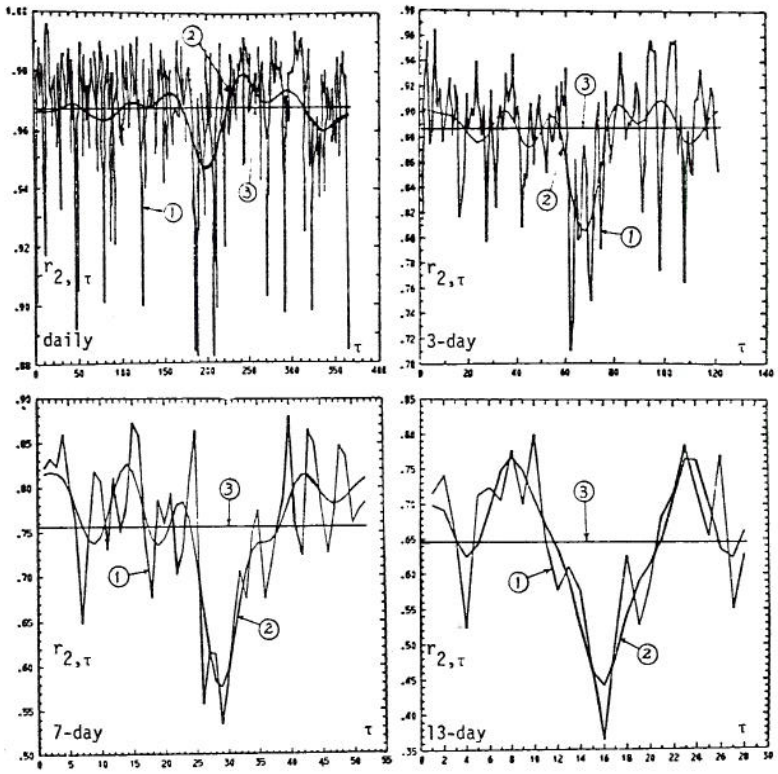

Fig. 5-9 The $r_{2,\tau}$ series of daily, 3-day, 7-day, and 13-day flow series for the logarithmically transformed discharges of the Neches River: (1) Computed $r_{2,\tau}$; (2) $r_{2,\tau}$ series composed of six harmonics; and (3) the mean of $r_{2,\tau}$.

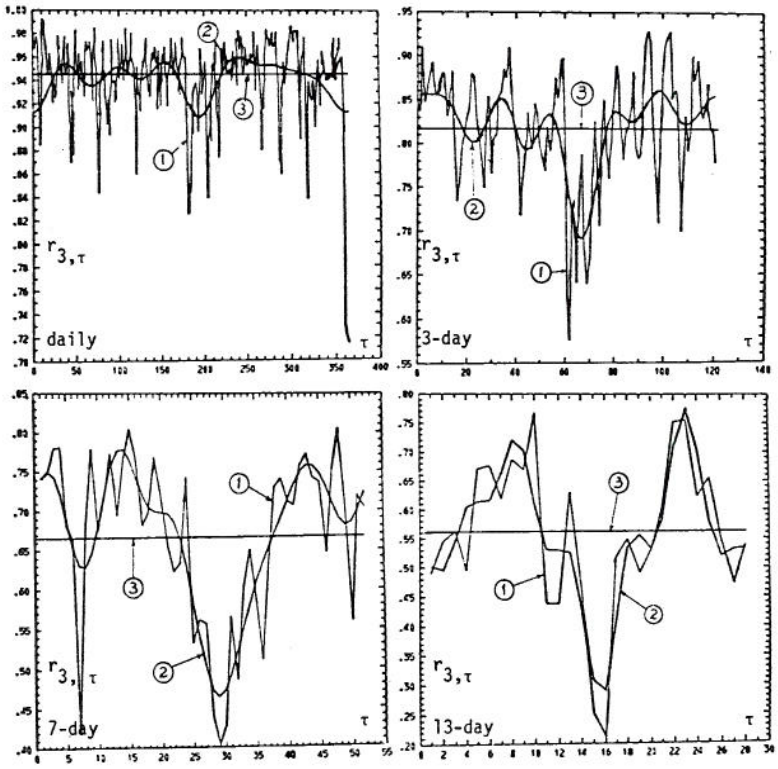

Fig. 5-10 The $r_{3,\tau}$ series of daily, 3-day, 7-day, and 13-day flow series for the logarithmically transformed discharges of the Neches River: (1) Computed $r_{3,\tau}$; (2) $r_{3,\tau}$ series composed of six harmonics; and (3) the mean of $r_{3,\tau}$.

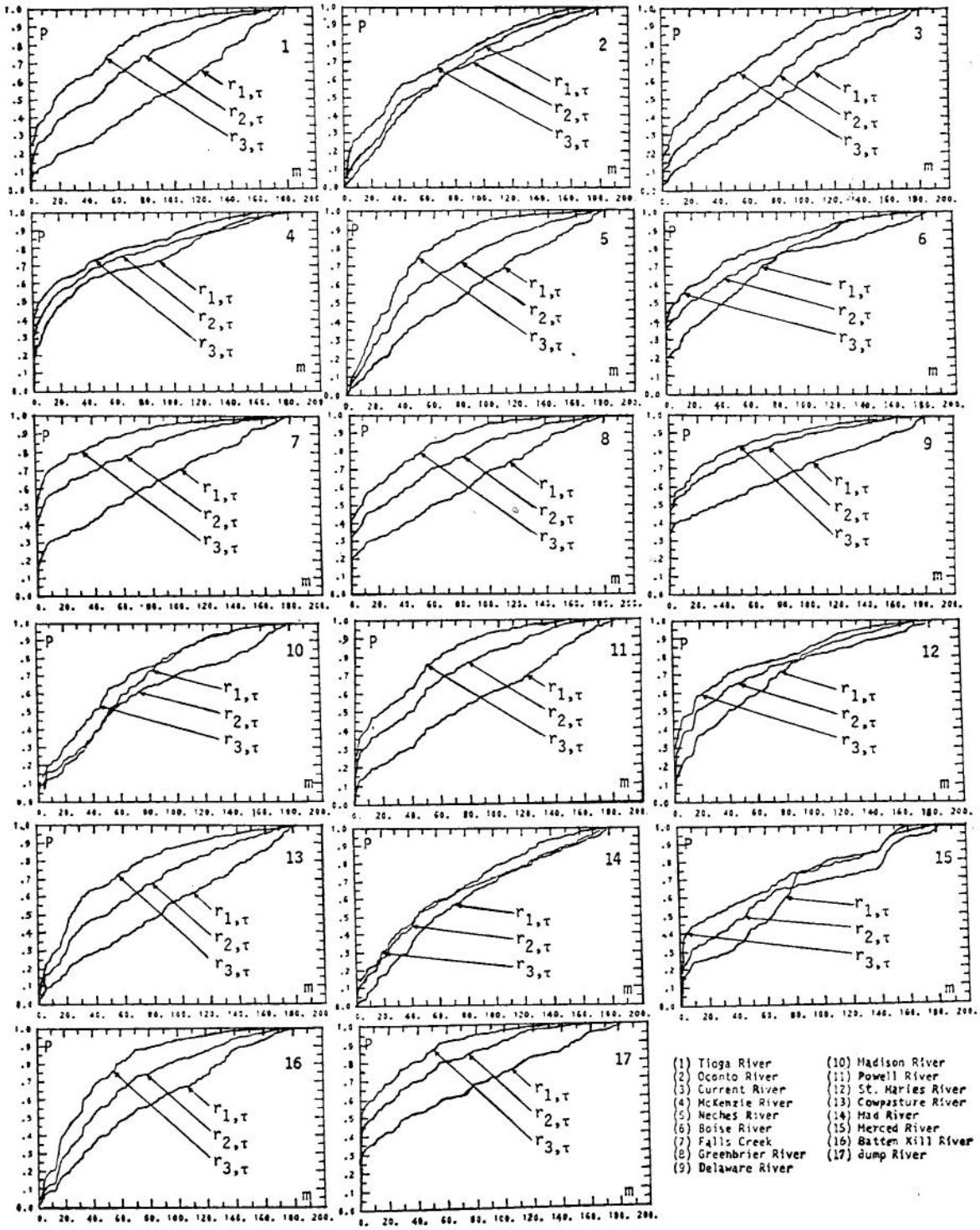

Fig. 5-11 Explained variances, P, by m harmonics in $r_{1,\tau}$, $r_{2,\tau}$, and $r_{3,\tau}$, of the logarithmically transformed 17 daily flow series, with m ranging from 1 to 182.

$r_{1,\tau}$, as shown in Fig. 5-5, is about 12 percent, which is greater than P_{min} = 7.1 percent.

It seems desirable to use the mean daily values of $r_{1,\tau}$ because of its high fluctuation over 365 values of τ. Figure 5-11 shows for most cases that the first six harmonics explain only a small portion of variances of $r_{1,\tau}$, $r_{2,\tau}$ and $r_{3,\tau}$. Hence, the use of mean values (r_1, r_2, r_3) is justified. If these means replace $r_{k,\tau}$ of Eqs. 3-27 through 3-35, the constant values of α_k should also be used to replace the periodic autoregressive coefficients $\alpha_{k,\tau}$, and the constant value of σ_ξ should also be used to replace the periodic standard deviation $\sigma_{\xi,\tau}$.

The effect of using the constant autoregressive coefficients may be tested by finding whether they still yield satisfactory results in removing the

dependence from stochastic variables, $\varepsilon_{p,\tau}$. However, $\sigma_{\xi,\tau}$ is highly sensitive to variations in $r_{k,\tau}$. Taking the Tioga River as an example, the first-order autoregressive linear model is found appropriate to describe the dependence of the stochastic component with the maximum and minimum of $r_{1,\tau}$ for the fitted periodic function are approximately 0.95 and 0.85, with the mean of 0.90. For the values 0.85, 0.90, and 0.95 of $r_{1,\tau}$, the $\sigma_{\xi,\tau}$ values of Eq. 3-28 are 0.316, 0.447, and 0.530, respectively. For 365 values of $r_{1,\tau}$ approximated by its mean value, the error in $\sigma_{\xi,\tau}$ is greater than ±20 percent. This error may affect significantly the computation of the independent stochastic component by Eq. 3-25.

5-3 Tolerance Limit Test of Tails of Frequency Distributions of Independent Stochastic Components.

The basic hypothesis in this text is that both tails are well approximated by simple exponential functions. Then the tails plot as straight lines in graphs with semilogarithmic scales. Instead of using the semilog graph paper for the independent stochastic ξ-component, ordinary graph paper can be used for the logarithmic transformation of the original variable $x_{p,\tau}$, with the resulting independent stochastic η-component, as described in the previous text. When testing the tails of distributions of the ξ- and η-variables, both cases should lead to the same inference for a particular series except that differences may originate from the estimates of coefficients of harmonics of periodic parameters because of the use of $x_{p,\tau}$ and $\log\ x_{p,\tau}$, respectively. However, the effects of differences in estimates by using $x_{p,\tau}$ on the final conclusions for the tails should be negligible. Therefore, the two approaches are applied in this study: (i) the tests of tails to be exponential by using the tolerance limits for the η-component, and (ii) the tests by the Gnedenko statistic for tails to be exponential for the ξ-component. The first approach is the subject of this section, and the second approach is the subject of the next section of this chapter.

Since either the right or the left tail of probability distributions of independent stochastic components is of interest, only the parts of tails with large absolute values are tested to determine if the tails are heave or not. The procedure for testing the right tail to be exponential is as follows:

i. The largest 500 values of the η-component of daily series are selected (approximately 3.5 percent of the total sample);

ii. The mean and the lower boundary of η are estimated by $\bar{\eta} = \sum_{i=1}^{500} \eta_i$ and $\eta_o = \min(\eta_i)$, respectively;

iii. Use of Eq. 4-85, with λ estimated by $\hat{\lambda} = 1/(\bar{\eta} - \eta_o)$, and $\theta_1(\eta)$ and $\theta_2(\eta)$ computed as the 90 percent tolerance limits for the tails of the exponential function;

iv. The 500 extreme values are sorted into 30 class intervals of equal length, with the relative frequencies of these intervals denoted by 0_η, $k = 1,2,\ldots,30$, and their class marks denoted by η_k, $k = 1,2,\ldots,30$; and

v. The tail distribution, based on the η_i values, is tested only by using these large η_k values; for $\theta_1(\eta_k) \le \ln[1 - \sum_{j=1}^{k} 0_j] \le \theta_2(\eta_k)$, the distribution has an exponential tail; for $\ln[1 - \sum_{j=1}^{k} 0_j] \ge \theta_2(\eta_k)$ the distribution has a heavy tail; and for $\theta_1(\eta_k) \ge \ln[1 - \sum_{j=1}^{k} 0_j]$ the distribution has a light tail.

The same procedure is applied to investigate the left tail by using the smallest 500 values of η, selected from the entire sample.

As noted before, this is a test of the hypothesis that the η_i's are exponentially distributed, with the shape parameter $\lambda = 1/(\bar{\eta} - \eta_o)$. The confidence limits θ_1 and θ_2 depend on η and λ, with λ estimated by using only the η_i values. Therefore, the confidence limits are subject to sampling errors in η_i.

The graphical representation of results of this investigation for the 17 daily series of η are shown in Fig. 5-12. The basic results, are:

1. For the right-tail test of the η-frequency distributions of 17 stations, nine fall clearly into the exponential tail category, while eight cross the lower limit into the light-tail region. For the right tails which fall into the exponential tail category, only the right tails of the Madison and Merced Rivers (10, 15) closely follow the upper confidence limit.

2. For the left-tail test, the tails tend to be close to the confidence limit at the heavy tail side. Only the Tioga, Current, and Boise River (1, 3, 6) show the left tails clearly crossing the upper confidence limit into the heavy tail region, but for a limited range.

3. There is sufficient evidence in Fig. 5-12 to conclude that the frequency distributions of the independent stochastic components possess the exponential tails.

4. The left tails are shorter than the right tails. Therefore, the left tails seem to be heavier and the right tails lighter. This fact may be explained as follows: (i) Theoretically, there is always a lower boundary for the stochastic component of a runoff series whereas the higher limit is unbounded, and (ii) The positive extreme values of stochastic components result from floods and the negative values result from low flows. The variation of floods is greater than of low flows.

5-4 Tests by Using the Gnedenko Statistic for the Tails of the ξ-Frequency Distributions.

The relationship $[1-F(\xi)]$ against $[\xi-a]$ of Eq. 4-86 is approximately linear, for $\xi > a$, when plotted on semilogarithmic paper, if the distribution $F(\xi)$ of the independent stochastic component is exponential. Therefore, in the first instance, it was decided to regraph the cumulative distribution of

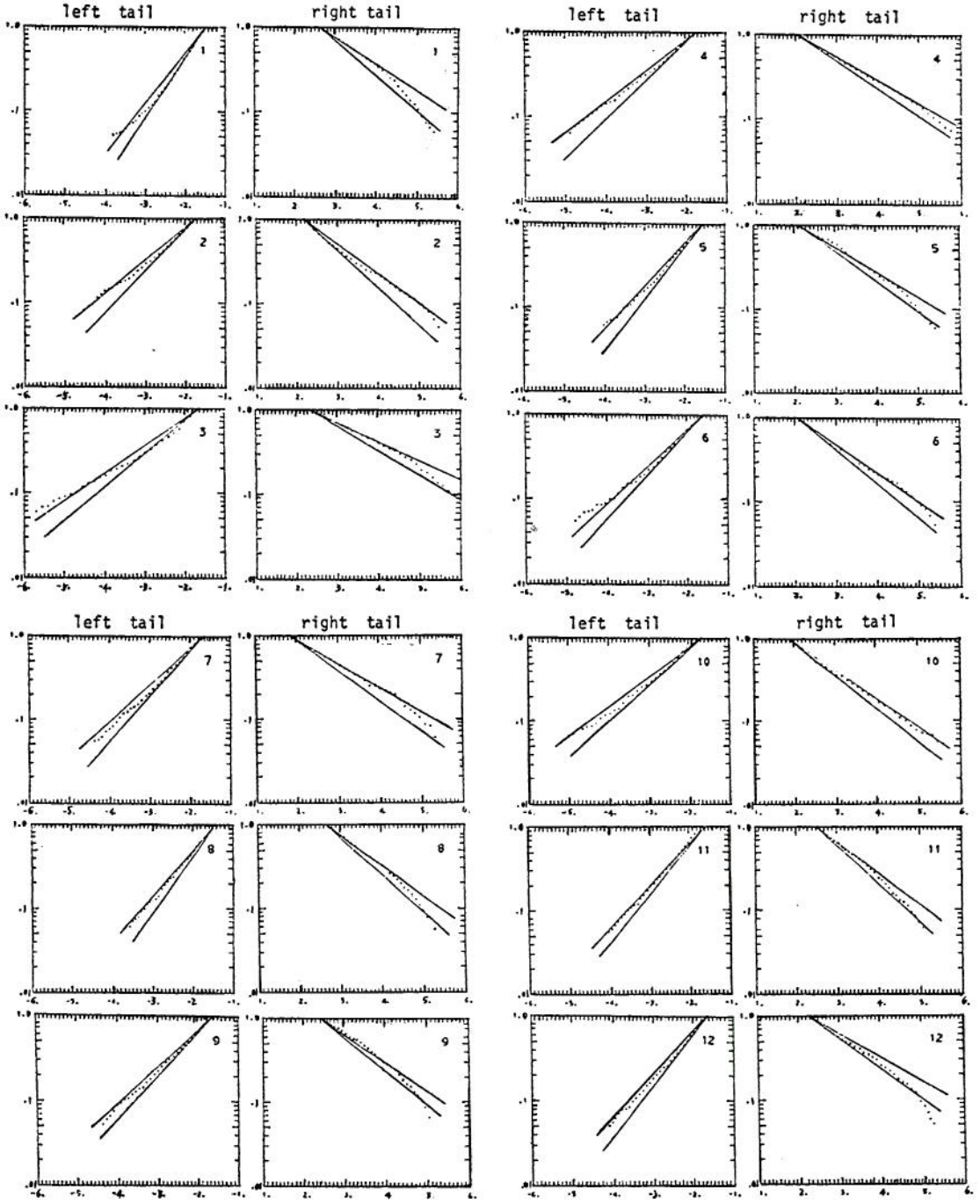

Fig. 5-12 Test of the left and right tail of the independent stochastic η-components: dotted line, the tails of the frequency distributions; solid lines, the 90 percent confidence limits; with η the abscissa and $F(\eta)$ the ordinates of the left tail, and $1-F(\eta)$ the ordinates of the right tail (to continue).

tails of the frequency distribution of standardized, independent stochastic components of the 17 daily series with more details than in Section 5-3 with the η-component.

Three uncertainties are involved in tests based on the Gnedenko statistic. The first is related to the particular application of this paper. It results from the choice of n, the length of sample or, in other words, the selected value a at which the tail begins. In an attempt to resolve this uncertainty, the standardized independent ξ-variable of each series was divided into six unequal class intervals (a total of 102 intervals for the 17 series). Contributions by values in each class interval to the total coefficients of skewness and kurtosis of the data are tabulated for all class intervals and all series. The abridged results are given in Table 5-4. Within a certain range of values of standardized variables, such as between -5.5 and 5.5, it is seen that, if somehow the values beyond these two limits did not exist, the distribution would be nearly normal for almost all 17 series. Admittedly, the truncated data for the range -5.5 to +5.5 may conform with some other type of symmetrical distributions, such as Pearson's Type IV, Johnson's S_u, [Pearson, 25]. However, with these range limits, leaving on the left tail about 0.1 percent and on the right tail about 1.0 percent

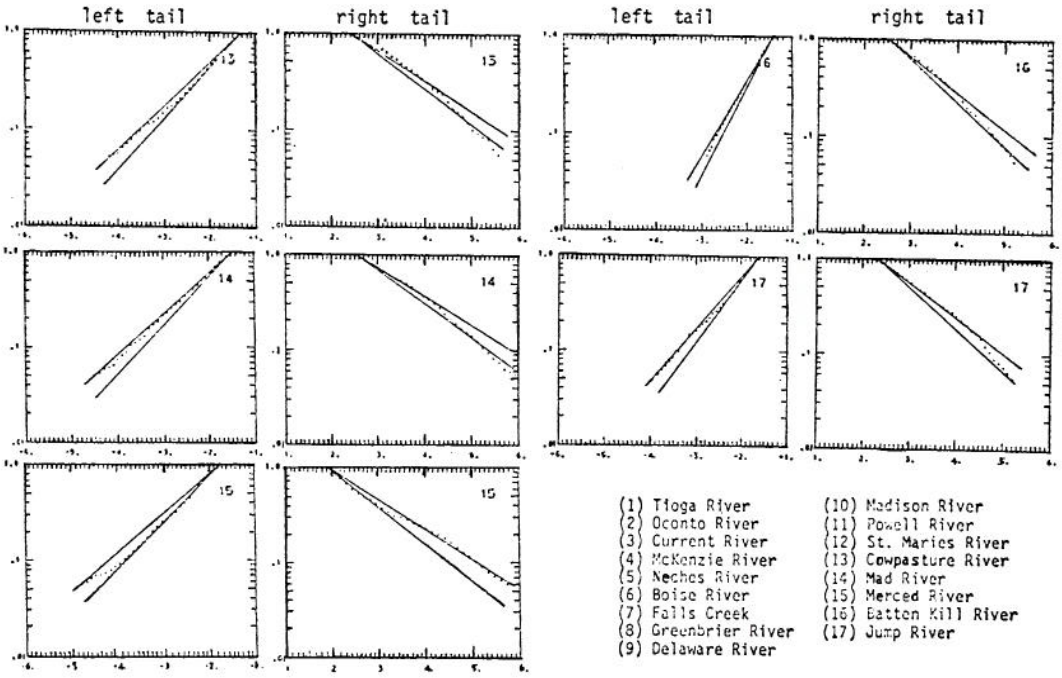

Fig. 5-12 Test of the left and right tail of the independent stochastic η-components: dotted line, the tails of the frequency distributions; solid lines, the 90 percent confidence limits; with η the abscissa and $F(\eta)$ the ordinates of the left tail, and $1-F(\eta)$ the ordinates of the right tail.

of the total of about 14,000 daily ξ-values, it was thought that a study of the 0.1 and 1 percent parts should form an important basis for the investigation. In addition, because this choice of limits is too restrictive, extended tails formed from 5 percent of the values were also investigated.

The second uncertainty comes from the division of the sample into groups of length r and $n-r$, which is arbitrary. However, if large values of r say equal to $n-1$ or $n-2$ are chosen, the Gnedenko statistic may have a large dependence on observations at the extremity [Bryson, 5]. Accordingly, it was decided to examine values of $Q(1,n-1)$, $Q(2,n-2)$ and reciprocals of the same. Nevertheless, in the program r was decreased in steps of 1 from $n-1$ to $n-16$, with limitations according to sample length, but the results of the tests were not significantly different.

Thirdly, one has to choose a level of significance α. For this two-sided test for exponentiality in distribution, 95 percent confidence limits or $\alpha = 0.025$ may be considered. Because the decision rules given by constraints of Eqs. 4-88 and 4-89 depend on the choice of α, values of $\alpha/2$ around which the constraints are reversed should be known. These are obtained from tables of the F-distribution and are presented in Tables 5-5 and 5-6. Tables correspond to constraints of Eqs. 4-88 and 4-89, respectively.

As discussed in connection with Fig. 4-2 light tail functions would be concave downwards and heavy tail functions would be concave upwards when $[1-F(\xi)]$ is plotted against ξ on semilogarithmic paper. Figures 5-13 to 5-18 show the empirical cumulative probability distributions or step functions of the data from the tails of the frequency functions of standardized independent ξ-residuals of daily flow data. The curvatures corresponding to the cases where there are significant heavy or light tails as noted in the tables are clearly evident in most of the cases.

From Table 5-6 it is seen that the extreme 0.1 percent of the samples do not exhibit any significance regarding heavy tails. For the stations 2, 3, 5, 12, 15 and 16, it appears that the tails containing 1 percent or 5 percent of the total sample are heavy. There is sufficient reason to suspect that this behavior may arise from errors in the estimates of coefficients of harmonics used to form the standardized ξ-residuals. It is noted that in the time series for these 6 stations the periodicities in the standard deviations need 2 to 5 significant harmonics for representation.

In a separate study to be published, problems arising from the estimation and removal of harmonics in the standard deviation are presented. Part of the same study is on generated autoregressive processes with normally distributed random components and added periodicities in the mean and standard deviation. When the periodicities and linear dependence are estimated and removed, the independent ξ-residuals exhibit often the non-normal behavior particularly with respect to the standardized third and fourth moments, if the periodicity in the standard deviation is initially incorporated through two or more harmonics. The departure from normality increases with an increase of the number of these harmonics. Similar distortions arise in application when the autoregressive structure is assumed to be linear but actually fluctuates in a manner related to the periodicity in the standard deviation.

Therefore, as a matter of interest, the tests were repeated using a nonparametric analysis, without harmonics, for the removal of periodicities. Results in Table 5-5, within brackets, show that the tails are now significantly light and not heavy. This apparent reversal in the nature of the tails shows the strong effects which might arise from harmonics which are not fully representative.

Table 5-4 COEFFICIENTS OF KURTOSIS AND SKEWNESS OF THE INDEPENDENT STOCHASTIC ξ COMPONENT OF DAILY DATA WITH CONTRIBUTIONS FROM VALUES WITHIN DIFFERENT CLASS INTERVALS.

Station Number	Description										
1	Range	-13.32		-11.00		-5.50	5.50		11.00	32.45	Total
	Number in interval		1		4	14,517		63	15		14,600
	Coefficient of skewness		-0.16		-0.20	0.57		1.75	10.42		12.47
	Coefficient of kurtosis		2.20		1.80	3.20		13.60	271.30		292.10
2	Range	-46.10		-11.00		-8.00	8.00		11.00	86.18	Total
	Number in interval		2		1	14,590		4	3		14,600
	Coefficient of skewness		-7.97		-0.06	0.33		0.48	44.18		36.63
	Coefficient of kurtosis		342.50		0.70	2.40		1.20	3,783.30		4,130.10
3	Range	-36.53		-11.00		-6.00	6.00		11.00	74.48	Total
	Number in interval		1		3	14,187		39	7		14,237
	Coefficient of skewness		-3.42		-0.10	0.53		1.60	30.54		29.15
	Coefficient of kurtosis		125.00		0.70	2.60		14.30	2,186.90		2,329.50
4	Range	-29.38		-11.00		-5.50	5.50		11.00	31.33	Total
	Number in interval		3		15	13,426		51	14		13,509
	Coefficient of skewness		-2.61		-0.51	0.36		1.89	5.38		4.96
	Coefficient of kurtosis		58.70		4.60	3.30		16.30	121.00		203.60
5	Range	-32.62		-11.00		-5.50	5.50		11.00	36.26	Total
	Number in interval		5		14	13,447		30	18		13,514
	Coefficient of skewness		-3.49		-0.45	0.26		0.97	8.24		5.53
	Coefficient of kurtosis		98.10		3.60	2.60		8.10	210.30		322.70
6	Range	-25.65		-11.00		-5.00	5.00		11.00	30.25	Total
	Number in interval		8		8	14,324		40	11		14,391
	Coefficient of skewness		-3.01		-0.09	0.22		0.90	5.90		3.82
	Coefficient of kurtosis		60.00		1.50	2.60		6.80	139.40		210.30
7	Range	-17.39		-11.00		-6.50	6.50		11.00	35.93	Total
	Number in interval		6		21	13,758		61	8		13,856
	Coefficient of skewness		-1.08		-0.96	-0.22		3.64	4.51		5.89
	Coefficient of kurtosis		15.50		8.00	2.90		33.70	136.00		196.10
8	Range	-18.41		-11.00		-5.00	5.00		11.00	32.04	Total
	Number in interval		1		5	14,489		87	18		14,600
	Coefficient of skewness		-0.43		-0.07	0.59		2.37	5.82		8.26
	Coefficient of kurtosis		7.90		0.40	3.00		19.10	126.00		156.40
9	Range			-10.55		-5.50	5.50		11.00	20.23	Total
	Number in interval		0		12	13,786		64	28		13,890
	Coefficient of skewness		0.0		-0.45	0.36		2.26	5.93		8.10
	Coefficient of kurtosis		0.0		3.90	3.00		19.40	89.30		115.60
10	Range	-28.64		-11.00		-5.50	5.50		11.00	27.23	Total
	Number in interval		6		15	13,187		32	6		13,246
	Coefficient of skewness		-3.29		-0.56	0.19		1.18	2.91		0.43
	Coefficient of kurtosis		77.50		4.50	3.20		10.10	65.10		160.40
11	Range	-17.72		-11.00		-4.50	4.50		11.00	20.53	Total
	Number in interval		1		31	14,457		93	15		14,597
	Coefficient of skewness		-0.38		-0.61	0.49		2.00	3.31		4.68
	Coefficient of kurtosis		6.80		4.30	3.00		15.70	50.50		80.30
12	Range	-34.83		-11.00		-7.50	7.50		11.00	88.16	Total
	Number in interval		3		0	13,860		8	2		13,873
	Coefficient of skewness		-3.49		0	0.63		0.38	50.01		47.62
	Coefficient of kurtosis		112.90		0	3.00		3.50	4,370.70		4,490.05
13	Range			-10.27		-5.00	5.00		11.00	26.45	Total
	Number in interval				8	12,674		24	21		12,727
	Coefficient of skewness				-0.42	0.50		2.01	6.57		8.66
	Coefficient of kurtosis				3.50	2.70		15.10	118.50		140.80
14	Range	-12.91		-11.00		-5.00	5.00		11.00	28.12	Total
	Number in interval		1		10	14,467		18	24		14,520
	Coefficient of skewness		-0.15		-0.29	0.43		1.90	6.78		8.47
	Coefficient of kurtosis		1.90		2.30	3.00		15.10	127.40		149.70
15	Range			-14.15		-11.00	11.00		38.29		Total
	Number in interval				1	14,562		37			14,600
	Coefficient of skewness				-0.19	0.33		15.11			21.31
	Coefficient of kurtosis				2.70	3.50		549.40			555.60
16	Range	-21.66		-11.00		-5.50	5.50		11.00	49.77	Total
	Number in interval		2		3	13,824		30	11		13,870
	Coefficient of skewness		-0.84		-0.12	0.65		0.94	17.63		18.26
	Coefficient of kurtosis		17.00		1.10	2.90		7.60	743.70		772.30
17	Range	-11.24		-11.00		-6.00	6.00		11.00	29.70	Total
	Number in interval		1		16	14,517		33	33		14,600
	Coefficient of skewness		-0.10		-0.60	-0.10		1.56	9.68		10.44
	Coefficient of kurtosis		1.10		4.80	3.00		14.70	187.80		211.40

It was also thought worthwhile to investigate the stability of the tails of the distributions. As noted before, the tails of stable distributions follow an asymptotic form of the Pareto Law given by Eq. 4-52

$$1-F(x) \to C_1 x^{-\alpha}, \quad x \to \infty$$

Figures 5-13 to 5-17 were redrawn using the logarithmic two-cycle paper. From these, the one percent tail values are presented in Figs. 5-19 and 5-20. If one ignores the effect of the last one or two values which are outliers and hence subject to high sampling bias, the slopes tend to some stable values asymptotically with α close to 2. However, on account of the finite sample sizes no useful inferences could be made.

Further investigations were made on whether the law which represents the three-parameter family of Weibull Distributions with its density function given by Eq. 5-2, holds instead of a simple exponential law. This is given by

$$1-F(x) = e^{-\lambda(x-a)^p}, \qquad (5\text{-}3)$$

where p is a constant which is equal to 1 in the exponential case. If this law holds, a plot of $[\log \log \frac{1}{1-F(x)}]$ against $(x-a)$ should show a linear relationship for $x > a$ provided that λ is constant. However, when applied to the tails of distributions, the graphs showed that this law is not applicable.

Table 5-5 SIGNIFICANCE LEVELS $\alpha/2$ FOR EXPONENTIALITY IN THE DISTRIBUTION OF THE TAILS OF THE INDEPENDENT STOCHASTIC ξ COMPONENT OF DAILY DATA, ON THE SIDE OF LIGHT TAILS. VALUES OF $\alpha/2 < 0.025$ ARE UNDERLINED. RESULTS FROM NONPARAMETRIC ANALYSIS ARE WITHIN BRACKETS.

Station Number	EXTREME 0.1%				EXTREME 1%				EXTREME 5%			
	LEFT TAIL		RIGHT TAIL		LEFT TAIL		RIGHT TAIL		LEFT TAIL		RIGHT TAIL	
	Q(1,n)	Q(2,n-1)	Q(1,n)	Q(2,n-1)	Q(1,n)	Q(2,n-1)	Q(1,n)	Q(2,n-1)	Q(1,n)	Q(2,n-1)	Q(1,n)	Q(2,n-1)
1	0.10 - 0.25	0.01 - 0.025	< 0.001	< 0.001	0.50 - 0.75	0.25 - 0.50	0.025 - 0.05	0.05 - 0.10	0.75 - 0.90	0.75 - 0.90	0.25 - 0.50	0.50 - 0.75
2	0.50 - 0.75 (0.01 - 0.025)	0.90 - 0.95 0.25 - 0.50	0.75 - 0.90 (0.005 - 0.01)	0.90 - 0.95 (0.01 - 0.025)	0.90 - 0.95 (0.10 - 0.25)	0.99 - 0.995 (0.75 - 0.90)	0.95 - 0.975 (0.05 - 0.10)	0.975 - 0.99 (0.10 - 0.25)	0.95 - 0.975 (0.25 - 0.50)	0.995 - 0.999 (0.90 - 0.95)	0.975 - 0.99 (0.25 - 0.50)	0.995 - 0.999 (0.50 - 0.75)
3	0.75 - 0.90 (0.01 - 0.025)	0.75 - 0.90 (0.005 - 0.01)	0.75 - 0.90 (< 0.001)	0.75 - 0.90 (< 0.001)	0.95 - 0.975 (0.05 - 0.10)	0.99 - 0.995 (0.05 - 0.10)	0.90 - 0.95 (< 0.001)	0.95 - 0.975 (< 0.001)	0.975 - 0.99 (0.25 - 0.50)	0.995 - 0.999 (0.25 - 0.50)	0.95 - 0.975 (0.025 - 0.05)	0.99 - 0.995 (< 0.001)
4	0.50 - 0.75	0.25 - 0.50	0.25 - 0.50	0.05 - 0.10	0.75 - 0.90	0.75 - 0.90	0.50 - 0.75	0.25 - 0.50	0.90 - 0.95	0.95 - 0.975	0.75 - 0.90	0.75 - 0.90
5	0.25 - 0.50	0.25 - 0.50	0.25 - 0.50	0.10 - 0.25	0.75 - 0.90	0.75 - 0.90	0.50 - 0.75	0.50 - 0.75	0.90 - 0.95	0.975 - 0.99	0.75 - 0.90	0.90 - 0.95
6	0.10 - 0.25	0.025 - 0.05	0.01 - 0.025	0.05 - 0.10	0.50 - 0.75	0.25 - 0.50	0.10 - 0.25	0.50 - 0.75	0.75 - 0.90	0.75 - 0.90	0.50 - 0.75	0.75 - 0.90
7	0.025 - 0.05	0.005 - 0.01	0.50 - 0.75	0.25 - 0.50	0.25 - 0.50	0.10 - 0.25	0.50 - 0.75	0.50 - 0.75	0.50 - 0.75	0.75 - 0.90	0.90 - 0.95	0.95 - 0.975
8	0.50 - 0.75	0.50 - 0.75	0.25 - 0.50	0.10 - 0.25	0.75 - 0.90	0.75 - 0.90	0.50 - 0.75	0.25 - 0.50	0.90 - 0.95	0.95 - 0.975	0.75 - 0.90	0.75 - 0.90
9	< 0.001	< 0.001	0.005 - 0.01	< 0.001	0.005 - 0.010	0.025 - 0.05	0.05 - 0.10	0.025 - 0.05	0.25 - 0.50	0.50 - 0.75	0.25 - 0.50	0.25 - 0.50
10	0.10 - 0.25	0.10 - 0.25	0.10 - 0.25	0.025 - 0.05	0.50 - 0.75	0.50 - 0.75	0.50 - 0.75	0.25 - 0.50	0.75 - 0.90	0.75 - 0.90	0.50 - 0.75	0.50 - 0.75
11	0.25 - 0.50	0.10 - 0.25	0.005 - 0.01	< 0.001	0.50 - 0.75	0.50 - 0.75	0.05 - 0.10	0.005 - 0.01	0.75 - 0.90	0.90 - 0.95	0.25 - 0.50	0.10 - 0.25
12	0.50 - 0.75 (0.10 - 0.25)	0.75 - 0.90 (0.01 - 0.025)	0.75 - 0.90 (0.01 - 0.025)	0.75 - 0.90 (< 0.001)	0.90 - 0.95 (0.25 - 0.50)	0.975 - 0.99 (0.10 - 0.25)	0.90 - 0.95 (0.05 - 0.10)	0.975 - 0.99 (0.005 - 0.01)	0.95 - 0.975 (0.50 - 0.75)	0.995 - 0.999 (0.50 - 0.75)	0.975 - 0.99 (0.25 - 0.50)	0.995 - 0.999 (0.10 - 0.25)
13	< 0.001 (0.001 - 0.005)	< 0.001 (< 0.001)	0.05 - 0.10 (< 0.001)	0.01 - 0.025 (< 0.001)	0.005 - 0.01 (0.01 - 0.025)	0.005 - 0.01 (< 0.001)	0.25 - 0.50 (< 0.001)	0.10 - 0.25 (< 0.001)	0.25 - 0.50 (0.10 - 0.25)	0.50 - 0.75 (0.025 - 0.05)	0.50 - 0.75 (0.005 - 0.01)	0.50 - 0.75 (< 0.001)
14	0.10 - 0.25	0.025 - 0.05	0.025 - 0.05	0.025 - 0.05	0.25 - 0.50	0.25 - 0.50	0.10 - 0.25	0.10 - 0.25	0.50 - 0.75	0.75 - 0.90	0.50 - 0.75	0.50 - 0.75
15	0.75 - 0.90 (0.01 - 0.025)	0.75 - 0.90 (0.025 - 0.05)	0.05 - 0.10 (< 0.001)	0.001 - 0.005 (< 0.001)	0.90 - 0.95 (0.10 - 0.25)	0.95 - 0.975 (0.25 - 0.50)	0.25 - 0.50 (0.01 - 0.025)	0.25 - 0.50 (< 0.001)	0.95 - 0.975 (0.25 - 0.50)	0.99 - 0.995 (0.50 - 0.75)	0.75 - 0.90 (0.10 - 0.25)	0.95 - 0.99 (0.01 - 0.025)
16	0.50 - 0.75	0.50 - 0.75	0.05 - 0.10	0.50 - 0.75	0.75 - 0.90	0.90 - 0.95	0.25 - 0.50	0.90 - 0.95	0.90 - 0.95	0.975 - 0.99	0.50 - 0.75	0.975 - 0.99
17	0.005 - 0.01	0.001 - 0.005	0.10 - 0.25	0.001 - 0.005	0.05 - 0.10	0.05 - 0.10	0.25 - 0.50	0.10 - 0.25	0.25 - 0.50	0.50 - 0.75	0.75 - 0.90	0.75 - 0.90
Normal Input	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 5-6 SIGNIFICANCE LEVELS $\alpha/2$ FOR EXPONENTIALITY IN THE DISTRIBUTION OF THE TAILS OF THE INDEPENDENT STOCHASTIC ξ COMPONENT OF DAILY DATA, ON THE SIDE OF HEAVY TAILS. VALUES OF $\alpha/2 < 0.025$ ARE UNDERLINED.

Station Number	EXTREME 0.1%				EXTREME 1%				EXTREME 5%			
	LEFT TAIL		RIGHT TAIL		LEFT TAIL		RIGHT TAIL		LEFT TAIL		RIGHT TAIL	
	$Q(1,n)^{-1}$	$Q(2,n-1)^{-1}$	$Q(1,n)^{-1}$	$Q(2,n-1)^{-1}$	$Q(1,n)^{-1}$	$Q(2,n-1)^{-1}$	$Q(1,n)^{-1}$	$Q(2,n-1)^{-1}$	$Q(1,n)^{-1}$	$Q(2,n-1)^{-1}$	$Q(1,n)^{-1}$	$Q(2,n-1)^{-1}$
1	0.75 - 0.90	0.975 - 0.99	> 0.999	> 0.999	0.25 - 0.50	0.50 - 0.75	0.95 - 0.975	0.90 - 0.95	0.10 - 0.25	0.10 - 0.25	0.50 - 0.75	0.25 - 0.50
2	0.25 - 0.50 (0.975 - 0.99)	0.05 - 0.10 (0.50 - 0.75)	0.10 - 0.25 (0.99 - 0.995)	0.05 - 0.10 (0.975 - 0.99)	0.05 - 0.10 (0.75 - 0.90)	0.005 - 0.01 (0.10 - 0.25)	0.025 - 0.05 (0.90 - 0.95)	0.01 - 0.025 (0.75 - 0.90)	0.025 - 0.05 (0.50 - 0.75)	0.001 - 0.005 (0.05 - 0.10)	0.01 - 0.025 (0.50 - 0.75)	0.001 - 0.005 (0.25 - 0.75)
3	0.10 - 0.25 (0.975 - 0.99)	0.10 - 0.25 (0.99 - 0.995)	0.10 - 0.25 (> 0.999)	0.10 - 0.25 (> 0.999)	0.025 - 0.05 (0.90 - 0.95)	0.005 - 0.01 (0.90 - 0.95)	0.05 - 0.10 (> 0.999)	0.025 - 0.05 (> 0.999)	0.01 - 0.025 (0.50 - 0.75)	0.001 - 0.005 (0.50 - 0.75)	0.025 - 0.05 (0.95 - 0.975)	0.005 - 0.01 (> 0.999)
4	0.25 - 0.50	0.50 - 0.75	0.50 - 0.75	0.90 - 0.95	0.10 - 0.25	0.10 - 0.25	0.25 - 0.50	0.50 - 0.75	0.05 - 0.10	0.025 - 0.05	0.10 - 0.25	0.10 - 0.25
5	0.50 - 0.75	0.50 - 0.75	0.50 - 0.75	0.75 - 0.90	0.10 - 0.25	0.10 - 0.25	0.25 - 0.50	0.25 - 0.50	0.05 - 0.10	0.01 - 0.025	0.10 - 0.25	0.05 - 0.10
6	0.75 - 0.90	0.95 - 0.975	0.975 - 0.99	0.90 - 0.95	0.25 - 0.50	0.50 - 0.75	0.75 - 0.90	0.25 - 0.50	0.10 - 0.25	0.10 - 0.25	0.25 - 0.50	0.10 - 0.25
7	0.95 - 0.975	0.99 - 0.995	0.25 - 0.50	0.50 - 0.75	0.50 - 0.75	0.75 - 0.90	0.25 - 0.50	0.25 - 0.50	0.25 - 0.50	0.10 - 0.25	0.05 - 0.10	0.025 - 0.05
8	0.25 - 0.50	0.25 - 0.50	0.50 - 0.75	0.75 - 0.90	0.10 - 0.25	0.10 - 0.25	0.25 - 0.50	0.50 - 0.75	0.05 - 0.10	0.025 - 0.05	0.10 - 0.25	0.10 - 0.25
9	> 0.999	> 0.999	0.99 - 0.995	> 0.999	0.99 - 0.995	0.95 - 0.975	0.90 - 0.95	0.95 - 0.975	0.50 - 0.75	0.25 - 0.50	0.50 - 0.75	0.50 - 0.75
10	0.75 - 0.90	0.75 - 0.90	0.75 - 0.90	0.95 - 0.975	0.25 - 0.50	0.25 - 0.50	0.25 - 0.50	0.50 - 0.75	0.10 - 0.25	0.10 - 0.25	0.25 - 0.50	0.25 - 0.50
11	0.50 - 0.75	0.75 - 0.90	0.99 - 0.995	> 0.999	0.25 - 0.50	0.25 - 0.50	0.90 - 0.95	0.99 - 0.995	0.10 - 0.25	0.05 - 0.10	0.50 - 0.75	0.75 - 0.90
12	0.25 - 0.50 (0.75 - 0.90)	0.10 - 0.25 (0.975 - 0.99)	0.10 - 0.25 (0.975 - 0.99)	0.10 - 0.25 (> 0.999)	0.05 - 0.10 (0.50 - 0.75)	0.01 - 0.025 (0.75 - 0.90)	0.05 - 0.10 (0.90 - 0.95)	0.01 - 0.025 (0.99 - 0.995)	0.025 - 0.05 (0.25 - 0.50)	0.001 - 0.005 (0.25 - 0.50)	0.01 - 0.025 (0.50 - 0.75)	0.001 - 0.005 (0.75 - 0.90)
13	> 0.999 (0.995 - 0.999)	> 0.999 (> 0.999)	0.90 - 0.95 (> 0.999)	0.975 - 0.99 (> 0.999)	0.99 - 0.995 (0.975 - 0.99)	0.99 - 0.995 (> 0.999)	0.50 - 0.75 (> 0.999)	0.75 - 0.90 (> 0.999)	0.50 - 0.75 (0.75 - 0.90)	0.25 - 0.50 (0.95 - 0.975)	0.25 - 0.50 (0.99 - 0.995)	0.25 - 0.50 (> 0.999)
14	0.75 - 0.90	0.95 - 0.975	0.95 - 0.975	0.95 - 0.975	0.50 - 0.75	0.50 - 0.75	0.75 - 0.90	0.75 - 0.90	0.25 - 0.50	0.10 - 0.25	0.25 - 0.50	0.25 - 0.50
15	0.10 - 0.25 (0.975 - 0.99)	0.10 - 0.25 (0.95 - 0.975)	0.90 - 0.95 (> 0.999)	0.995 - 0.999 (> 0.999)	0.05 - 0.10 (0.75 - 0.90)	0.025 - 0.05 (0.50 - 0.75)	0.50 - 0.75 (0.975 - 0.99)	0.50 - 0.75 (> 0.999)	0.025 - 0.05 (0.50 - 0.75)	0.005 - 0.01 (0.25 - 0.50)	0.10 - 0.25 (0.75 - 0.90)	0.10 - 0.05 (0.975 - 0.99)
16	0.25 - 0.50	0.25 - 0.50	0.90 - 0.95	0.25 - 0.50	0.10 - 0.25	0.05 - 0.10	0.50 - 0.75	0.05 - 0.10	0.05 - 0.10	0.01 - 0.025	0.25 - 0.50	0.01 - 0.025
17	0.99 - 0.995	0.995 - 0.999	0.75 - 0.90	0.995 - 0.999	0.90 - 0.95	0.90 - 0.95	0.50 - 0.75	0.75 - 0.90	0.50 - 0.75	0.25 - 0.50	0.10 - 0.25	0.10 - 0.25
Normal Input	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999	> 0.999

The conclusions which would be reached from the results of this part of the study is that, as in the previous section, there is insufficient evidence to indicate that the tails of distributions of the independent stochastic component differ from the exponential type on the side of heavy tails. Distributions do not seem to conform with those of the Weibull family.

5-5 Probability Distributions of Independent Stochastic Components.

The seven groups of probability distribution functions described in Chapter IV were used in fitting the frequency distributions of the independent stochastic components of daily flows. The group of stable distributions possesses special characteristics different from those of the other six groups. The results of application of stable distributions are presented and discussed in Section 5-6.

The goodness-of-fit of a theoretical p.d.f. to the frequency distributions of a set of random variables is determined by comparing the chi-square value obtained by Eq. 4-74 with the critical chi-square value for the given significance level. If the chi-square value is smaller than the critical chi-square value, this p.d.f. is accepted.

The frequency distribution of each set of random variables is fitted using all the six p.d.f. groups. In applying the chi-square test for goodness-of-fit to all the p.d.f.'s, it is desirable to investigate several p.d.f.'s., compare their characteristics and then to select one which best fits the empirical frequency distribution. The p.d.f. which possesses the minimum probability of the chi-square statistic should be selected as the function of best fit to the frequency distribution. Since the chi-square probability density function with given degrees of freedom is well defined, the chi-square probability was computed by integrating the chi-square p.d.f., with the lower and upper integration limits being zero and the chi-square value, respectively.

Based on the chi-square test, the acceptance of each p.d.f. used to fit the frequency distributions of the independent stochastic components is shown in

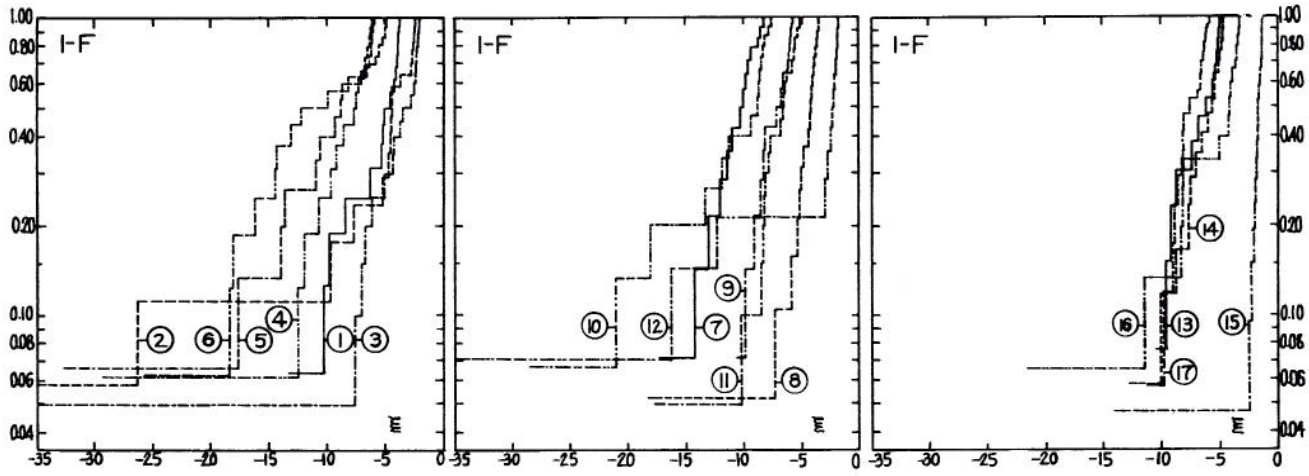

Fig. 5-13 Empirical distribution, Log (1-F) versus ξ of the left extreme 0.1 percent of the distribution of independent stochastic components.

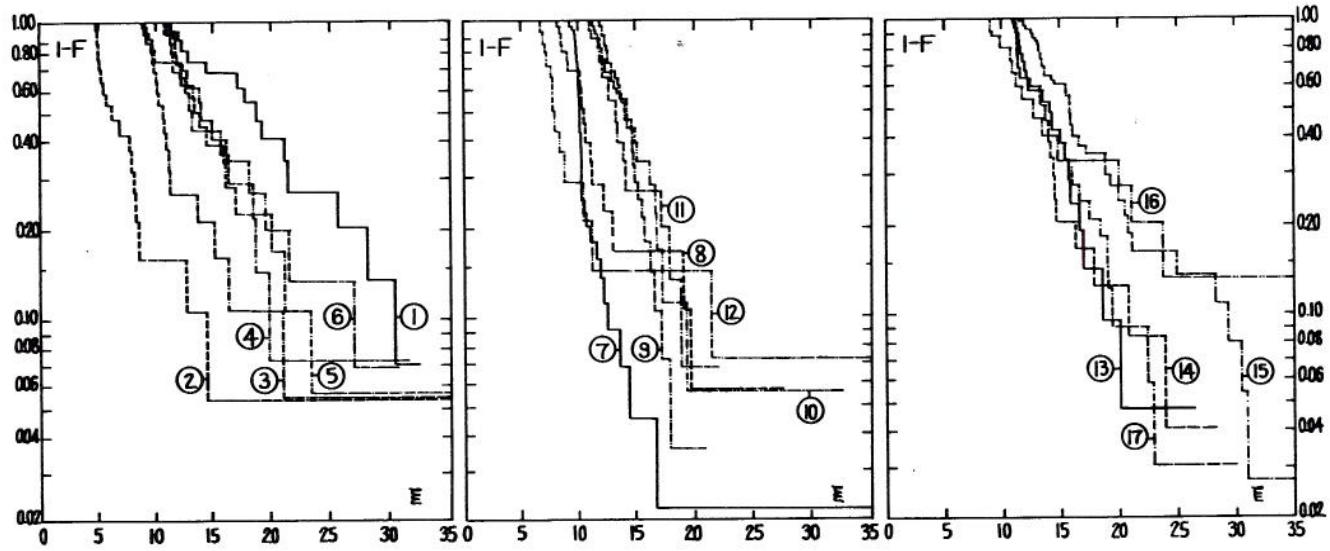

Fig. 5-14 Empirical distribution, Log (1-F) versus ξ of the right extreme 0.1 percent of the distribution of independent stochastic components.

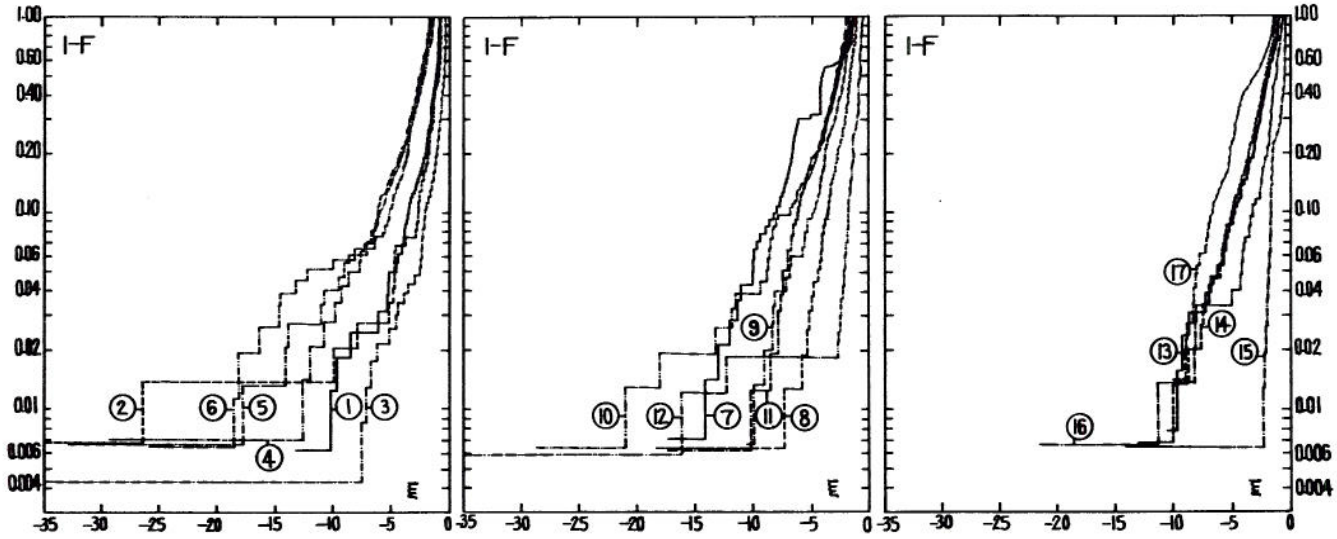

Fig. 5-15 Empirical distribution, Log (1-F) versus ξ of the left extreme 1 percent of the distribution of independent stochastic components.

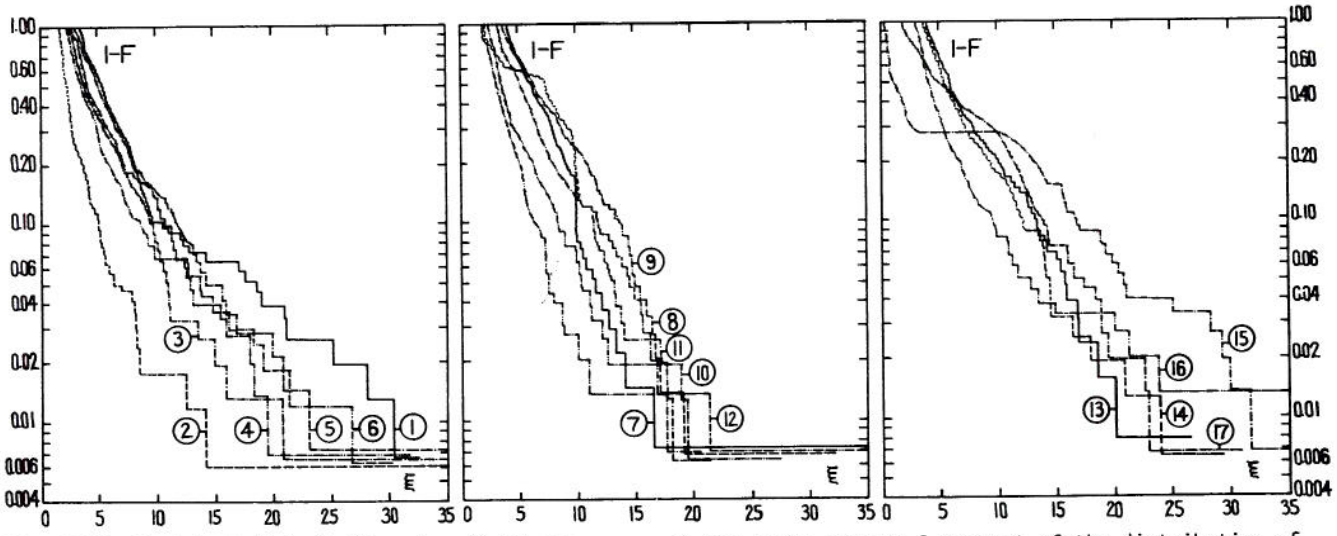

Fig. 5-16 Empirical distribution, Log (1-F) versus ξ of the right extreme 1 percent of the distribution of independent stochastic components.

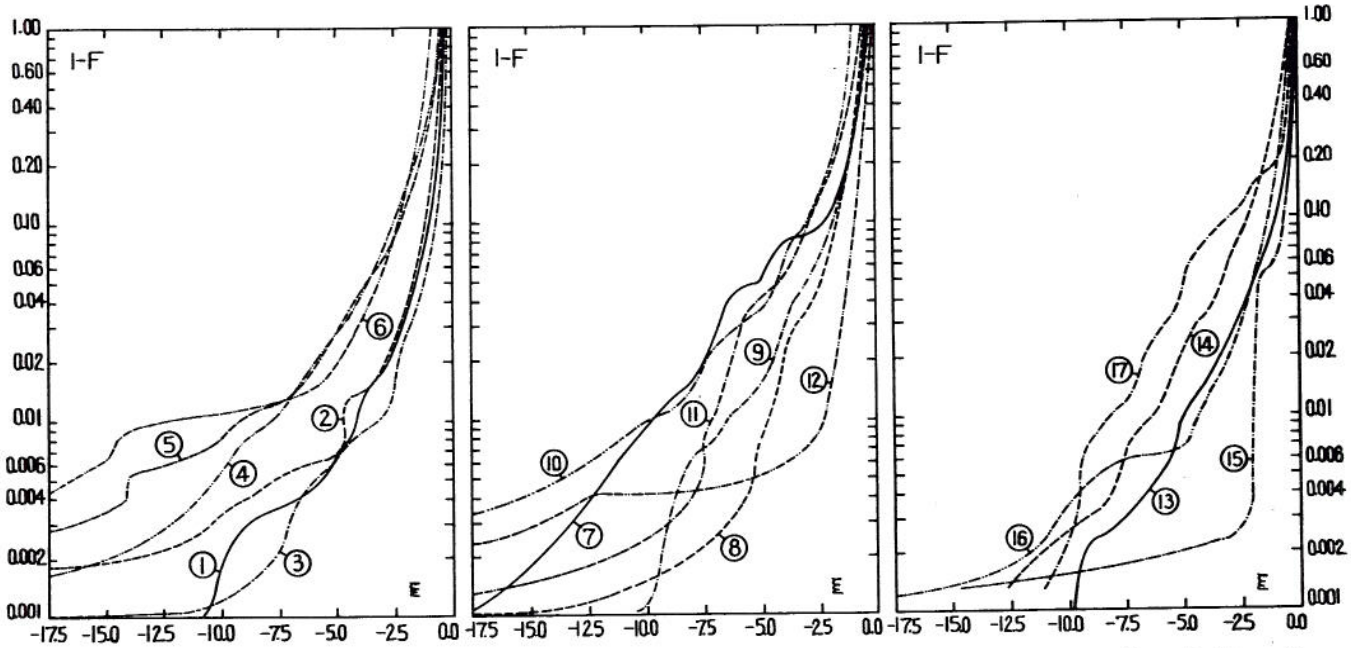

Fig. 5-17 Empirical distribution, Log (1-F) versus ξ of the left extreme 5 percent of the distribution of independent stochastic components.

Tables 5-7 through 5-12. The number of times of successful and unsuccessful fits are given in these tables. If a p.d.f. is applicable for all the 17 sets of series, the sum of numbers of successful and unsuccessful should be 17; however, for the p.d.f.'s such as Pearson's family functions, the good fit is determined only by certain criteria; therefore, the sum of the number of successful and unsuccessful fits is less than or equal to 17. By examining Tables 5-7 through 5-12, the results can be summarized as follows:

1. Lognormal and gamma functions best fit the frequency distributions of monthly series. The normal function modified by the three- or four-term Hermite polynomials also gives a good fit.

2. For the 13-day series, the frequency distributions of the η- and ζ-series are fitted well by the lognormal, gamma, and normal, modified by three- and four-term Hermite polynomials, as shown in Tables 5-7 and 5-8. For the ξ-series, the double-branch gamma and the mixture of Pearson's Type VII and gamma are shown in Table 5-11 to be best applicable.

3. For the 7-day series, the frequency distributions of the ζ- and ξ-series are difficult to fit by any of the p.d.f. studied except the double-branch gamma and the mixture of Pearson's Type VII and gamma functions, as shown in Tables 5-9 and 5-11. In the latter case, the fit is good in 9 out of 17 series. For the frequency distributions of the η-series, the lognormal, the normal modified by three- or four-term Hermite polynomials, and the gamma modified by three-term Laguerre polynomials, fit well about one third of all 17 cases, while the double-branch gamma function fits well 10 series, and the mixture of the normal and gamma functions applies to 8 out of 9 cases.

4. For the 3-day values, the double-branch gamma function fits well 9 out of the 17 η-series; however,

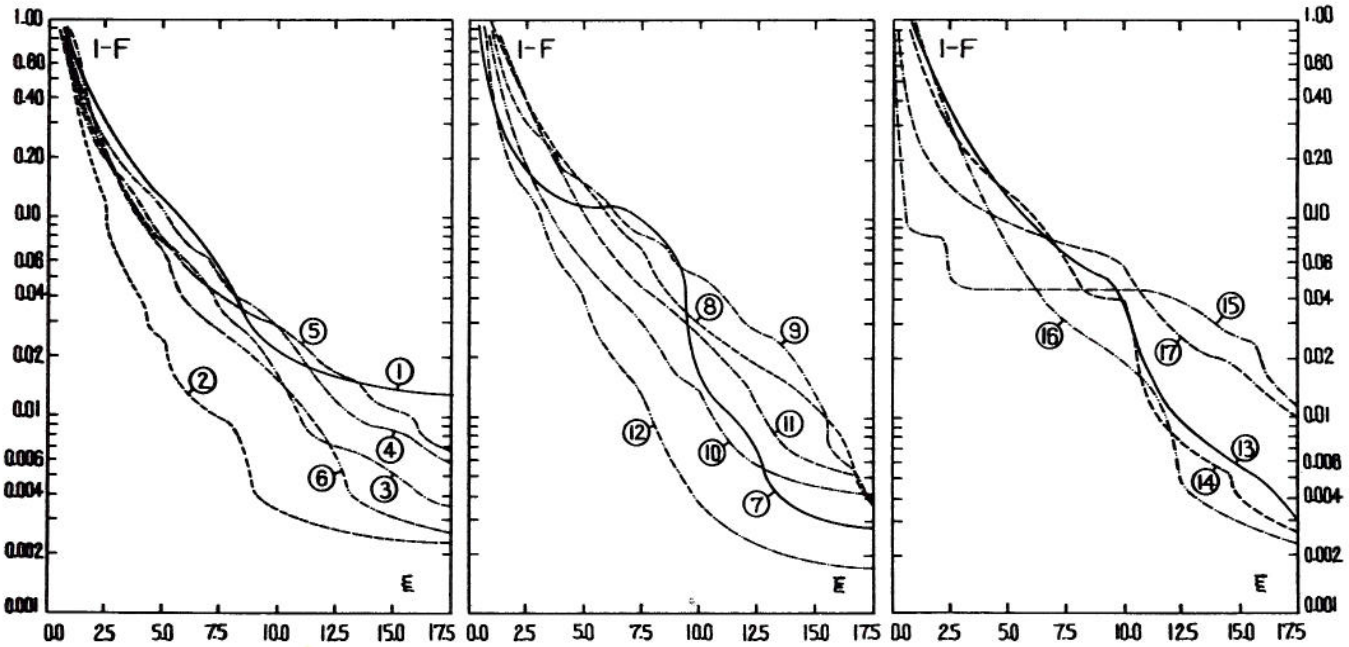

Fig. 5-18 Empirical distribution, Log (1-F) versus ξ of the right extreme 5 percent of the distribution of independent stochastic components.

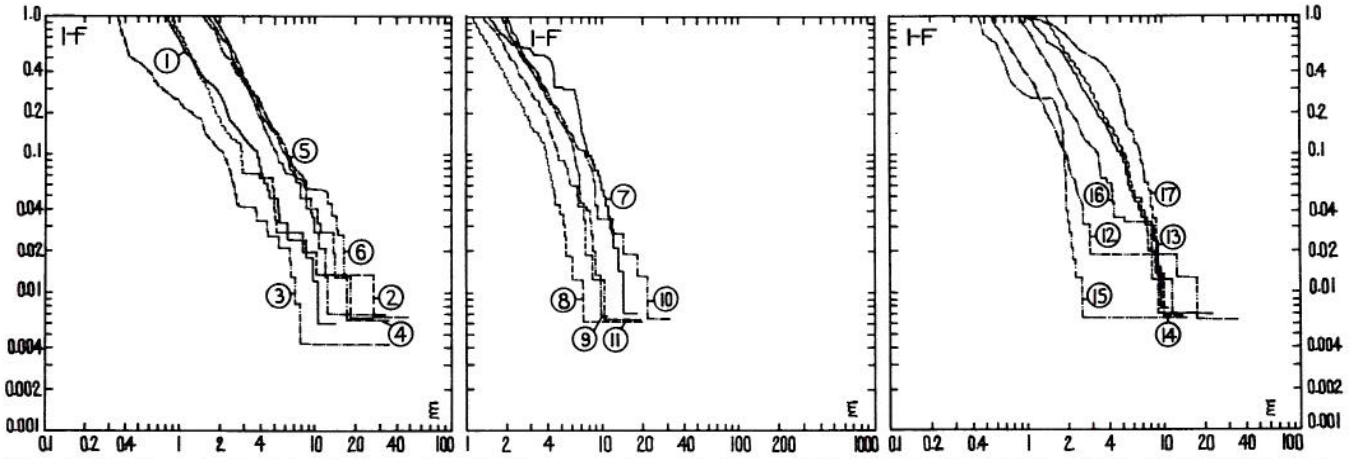

Fig. 5-19 Empirical distribution, Log (1-F) versus Log ξ of the left extreme 1 percent of the distribution of independent stochastic components.

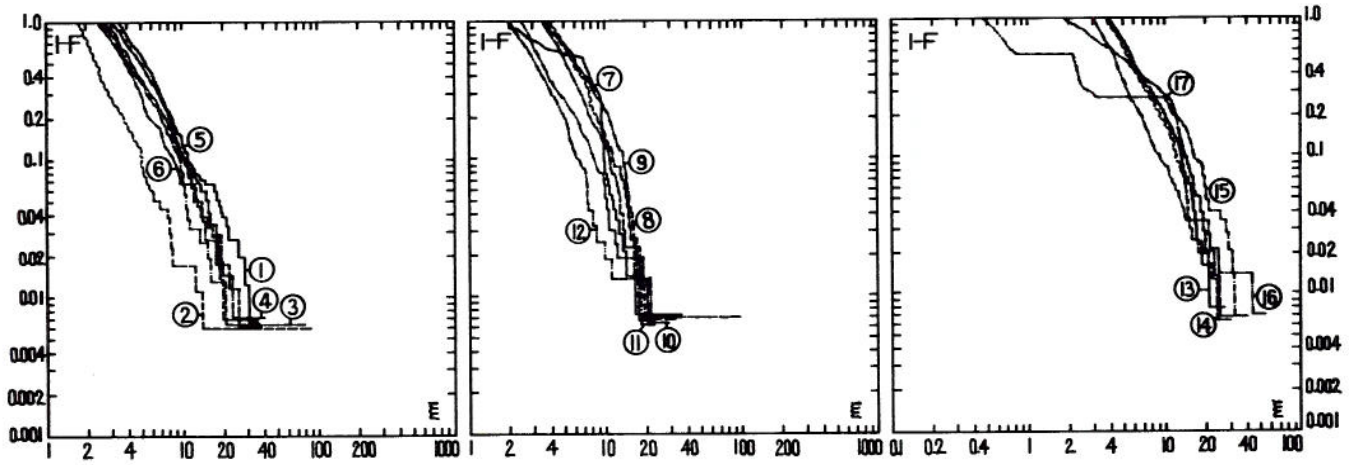

Fig. 5-20 Empirical distribution, Log (1-F) versus Log ξ of the right extreme 1 percent of the distribution of independent stochastic components.

Table 5-7 RESULTS OF FITTING VARIOUS PDF TO FREQUENCY DISTRIBUTIONS OF THE η-SERIES, WITH THE LEVEL OF SIGNIFICANCE 99 PERCENT; S = SUCCESSFUL FITS, F = UNSUCCESSFUL FITS.

Function	Monthly series		13-day series		7-day series		3-day series		Daily series	
	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F
Normal	16	1	3	14	0	17	0	17	0	17
Normal with Hermite-3 terms	17	0	15	2	1	16	0	17	0	17
Normal with Hermite-4 terms	17	0	16	1	7	10	0	17	0	17
Lognormal	17	0	16	1	6	11	0	17	0	17
Type IV					0	2	0	6	0	12
Type VI									0	1
Gamma	17	0	15	2	3	14	0	16	0	13
Gamma with Laguerre-3 terms	17	0	16	1	7	10	0	16	0	13
Gamma with Laguerre-4 terms					0	17	0	16	0	13
Weibull	15	2	4	13	0	17	0	16	0	13
Double-Branch Gamma	1	16	2	15	10	7	9	8	0	17
Mixtures of Normal and Gamma					8	1				
Mixtures of Type VII and Gamma			11	1	4	4	1	7	0	17

no p.d.f. studied fits well the frequency distributions of the ζ- and ξ-variables.

5. When the six p.d.f. groups are applied to frequency distributions of the daily values of η, ζ, and ξ variables, with their sample sizes approximately 14,600, none passes the chi-square test with the critical chi-square value 43.0 at the 99 percent significance level. However, the double-branch gamma functions have the smallest chi-squares. For the frequency distribution of daily values of the ζ- and ξ-variables, the double-branch gamma functions have the chi-square values in the range of 200 to 3000, having the smallest chi-squares for the η-series. As an example, the frequency distribution of η of the Falls Creek and the Merced Rivers were fitted by the double-branch gamma functions with chi-squares of 79.1 and 79.8, respectively. However, even with these lowest chi-squares the fit is rejected. Frequency distributions and the fitted double-branch gamma probability density and cumulative functions of the daily η-series of the Falls Creek and the Merced Rivers are shown in Figs. 5-21 and 5-22, respectively. Frequency distributions of η for all 17 series are plotted in Fig. 5-3. The daily series has a large sample size. Statistical parameters estimated from a sample of large size should be close to population values, with the tolerance interval of these parameters inversely proportional to the sample size. Given the narrow tolerance interval, the goodness of fit tests fail even at a high level of significance.

Since no probability density function can adequately fit the frequency distributions of independent stochastic components of daily flow series, an empirical method was attempted and is presented here.

All values were sorted into certain class intervals of equal length and frequency densities of each class interval computed. When these relative frequencies became highly irregular, a moving average was used to smooth them. The tails of frequency distributions were approximated by two exponential density functions. Parameters of the two exponential density functions were estimated from the values as the respective tails.

An example of this approach is given in Fig. 5-23. The maximum and minimum 500 values of the series for the Neches River were fitted by the exponential density functions. The center part of the frequency distribution was divided into 45 equal length class intervals, with the relative frequencies computed and smoothed.

5-6 Fitting of Symmetric Stable Distributions.

The symmetric stable distributions were fitted to frequency distributions of independent stochastic components of daily series of the Madison and Batten Kill Rivers, for the purpose of demonstrating their use in comparison with the use of the other distribution functions. Parameters of symmetric stable distributions were estimated by techniques described in Chapter IV, with the transformation of the original series into the u-variable by Eq. 4-58. Cumulative frequency distributions of u, and the fitted symmetric stable distributions are shown in Fig. 5-24 for the Madison River and Fig. 5-25 for the Batten Kill River. For the Madison River, the parameters of the stable distribution, estimated by percentiles, are $c = 0.337$, $\alpha = 1.264$, $\gamma = 0.253$ and $\delta = 0.0067$. For the Batten Kill River, parameters are: $c = 0.338$, $\alpha = 1.255$, $\gamma = 0.257$ and $\delta = 0.0425$. Using the transformation by Eq. 4-58, the u-variable has $c = \gamma = 1$, $\delta = 0$, and α the same as for the untransformed series. For the Kolmogorov-Smirnov test used for the goodness-of-fit test, the symmetric stable distribution fails to fit the frequency distribution of the u-variable even at the 99 percent confidence level, with the critical value of 0.0136.

Several factors limit the fit of stable distributions to frequency distributions of independent stochastic components of daily runoff series:

1. Density functions of stable distributions are not available in closed forms;

Table 5-8 PROBABILITY DISTRIBUTION FUNCTIONS OF BEST FIT FOR FREQUENCY DISTRIBUTIONS OF THE η-SERIES.

RIVER	Monthly series	13-day series	7-day series	3-day series	Daily series
Tioga	GL3	GL3	MNG	MPG	
Oconto	LN	LN	MPG	DBG	
Current	LN	MNG			
McKenzie	GL3	LN	DBG		
Neches	GL3	G	MPG		
Boise	GL3	LN	MPG	DBG	
Falls Creek	LN	LN	MNG	DBG	
Greenbrier	GL3	GL3	MNG	MPG	
Delaware	GL3	LN	DBG	DBG	
Madison	NH3	DBG	DBG	DBG	
Powell	GL3	G	MNG		
St. Maries	GL3	LN	MNG		
Cowpasture	GL3	LN	MNG		
Mad	LN	LN	DBG		
Merced	GL3	NH4	DBG	DBG	
Batten Kill	GL3	GL3	MNG	DBG	
Jump	GL3	LN	NH4		

Abbreviations: LN, lognormal; G, gamma; NH3, normal modified by 3 terms Hermite polynomials; NH4, normal modified by 4 terms Hermite polynomials; DBG, double-branch gamma; MNG, mixture of normal and gamma; MPG, mixture of Pearson's Type VII and gamma; PIV, Pearson's Type IV; GL3, gamma modified by 3 terms Laguerre polynomials.

Table 5-10 PROBABILITY DISTRIBUTION FUNCTIONS OF BEST FIT TO FREQUENCY DISTRIBUTIONS OF THE ζ-SERIES.

RIVER	13-day series	7-day series	3-day series	Daily Series
Tioga	NH4			
Oconto	NH4	MPG		
Current	MPG			
McKenzie	MPG	DBG		
Neches	DBG	DBG		
Boise	DBG	DBG	DBG	
Falls Creek	DBG	DBG		
Greenbrier	DBG			
Dalaware	DBG			
Madison	LN	MPG		
Powell	MNG	DBG		
St. Maries	DBG	DBG		
Cowpasture	DBG			
Mad	MNG	MPG		
Merced	DBG	DBG		
Batten Kill	MNG	DBG		
Jump	DBG			

Abbreviations: LN, lognormal; NH4, normal modified by 4 terms Hermite polynomials; MNG, mixture of normal and gamma; MPG, mixture of Pearson's Type VII and gamma; DBG, double-branch gamma.

Table 5-9 RESULTS OF FITTING VARIOUS PDF TO FREQUENCY DISTRIBUTIONS OF THE ζ-SERIES, WITH THE LEVEL OF SIGNIFICANCE 99 PERCENT; S = SUCCESSFUL FITS AND F = UNSUCCESSFUL FITS.

Function	13-day series		7-day series		3-day series		Daily series	
	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F
Normal	1	16	0	17	0	17	0	17
Normal with Hermite-3 terms	6	11	0	17	0	17	0	17
Normal with Hermite-4 terms	10	7	0	17	0	17	0	17
Lognormal	8	9	0	17	0	17	0	17
Type IV	0	1	0	3	0	8	0	12
Type VI							0	4
Gamma	6	11	0	17	0	17	0	17
Gamma with Laguerre-3 terms	6	11	0	17	0	17	0	17
Gamma with Laguerre-4 terms			0	17	0	17	0	17
Weibull	0	17	0	17	0	17	0	17
Double-Branch Gamma	14	3	9	8	1	16	0	17
Mixtures of Normal and Gamma	5	0	0	1				
Mixtures of Type VII and Gamma	5	3	7	5	0	17	0	8

Table 5-11 RESULTS OF FITTING VARIOUS PDF TO FREQUENCY DISTRIBUTIONS OF THE ξ-SERIES, WITH THE LEVEL OF SIGNIFICANCE 99 PERCENT: S = SUCCESSFUL FITS AND F = UNSUCCESSFUL FITS.

Function	Monthly series		13-day series		7-day series		3-day series		Daily series	
	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F	No. of S	No. of F
Normal	1	16	1	16	0	17	0	17	0	17
Normal with Hermite-3 terms	15	2	1	16	0	17	0	17	0	17
Normal with Hermite-4 terms	11	6	1	16	0	17	0	17	0	17
Lognormal	16	1	2	15	0	17	0	17	0	17
Type IV			1	0	0	1	0	10	0	8
Type VI					0	2			0	3
Gamma	14	0	1	11	0	17	0	17	0	17
Gamma with Laguerre-3 terms	14	0	2	10	0	17	0	17	0	17
Gamma with Laguerre-4 terms			0	17	0	17	0	17	0	17
Weibull			0	12	0	17	0	17	0	17
Double-Branch Gamma			12	5	3	14	0	17	0	17
Mixtures of Normal and Gamma			3	1						
Mixtures of Type VII and Gamma			7	1	9	5	0	14	0	15

2. Parameters of symmetric stable distributions must be estimated by percentiles; no method is yet available for a successful estimation of asymmetric cases;

3. Distributions of independent stochastic components do not possess heavy tails, particularly the right tails, while the stable distributions may fit well only when such heavy tails exist;

4. When $1 \leq \alpha < 2$, the second and higher-order moment of stable distributions do not exist; for $\alpha < 1$, the moments do not exist, while generally the independent stochastic components are standardized with the mean of zero and the variance of unity.

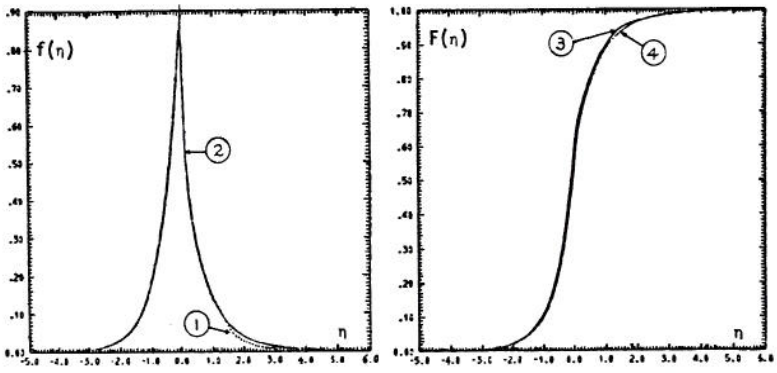

Fig. 5-21 Fitting the frequency distribution of the η-series of the Falls Creek River by the double-branch gamma density function: (1) frequency distribution, (2) fitted double-branch gamma density function, (3) cumulative frequency distribution and (4) fitted cumulative double-branch gamma distribution function.

Table 5-12 PROBABILITY DISTRIBUTION FUNCTIONS OF BEST FIT TO FREQUENCY DISTRIBUTIONS OF THE ξ-SERIES.

RIVER	Monthly series	13-day series	7-day series	3-day series	Daily series
Tioga	LN	MPG	MPG		
Oconto	LN	MPG	MPG		
Current	LN	MPG	DBG		
McKenzie	G	MPG			
Neches	LN	DBG	DBG		
Boise	LN	DBG			
Falls Creek	NH4	DBG	MPG		
Greenbrier	LN	DBG			
Dalaware	MPG				
Madison	LN	MNG	MPG		
Powell	LN	MPG	MPG		
St. Maries	LN	DBG	DBG		
Cowpasture	LN	MNG	MPG		
Mad	LN	MPG	MPG		
Merced	NH4	DBG			
Batten Kill	LN	MPG	MPG		
Jump	LN				

Abbreviations: LN, lognormal; G, gamma; NH3, normal modified by 3 terms Hermite polynomials; NH4, normal modified by 4 terms Hermite polynomials; DBG, double-branch gamma; MNG, mixture of normal and gamma; MPG, mixture of Pearson's Type VII and gamma; PIV, Pearson's Type IV; GL3, gamma modified by 3 terms Laguerre polynomials.

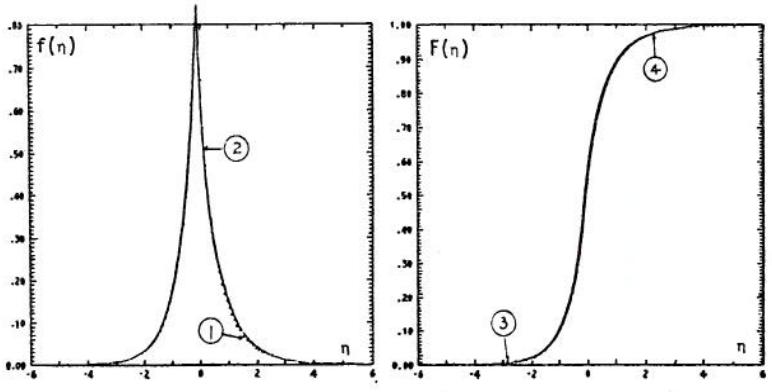

Fig. 5-22 Fitting the frequency distribution of the η-series of the Merced River by the double-branch gamma density function: (1) frequency distribution, (2) fitted double-branch gamma density function used, (3) cumulative frequency distribution and (4) fitted cumulative double-branch gamma distribution function used.

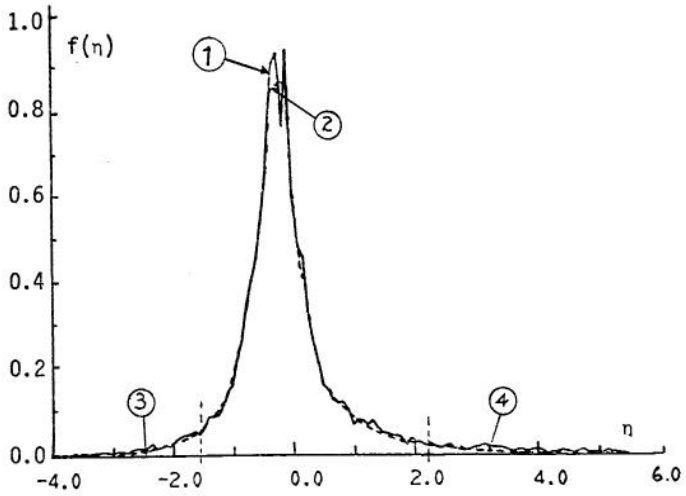

Fig. 5-23 (1) Empirical frequency density curves; (2) smoothed frequency density curves; (3) $f(\eta) = 0.0370\, e^{-1.333(1.538-\eta)}\, I_{(-\infty,-1.538)}$ fitted to the negative tail, and (4) $f(\eta) = 0.0370\, e^{-0.755(\eta-2.172)}\, I_{(2.172,\infty)}$, fitted to the positive tail of the η-series of the Neches River.

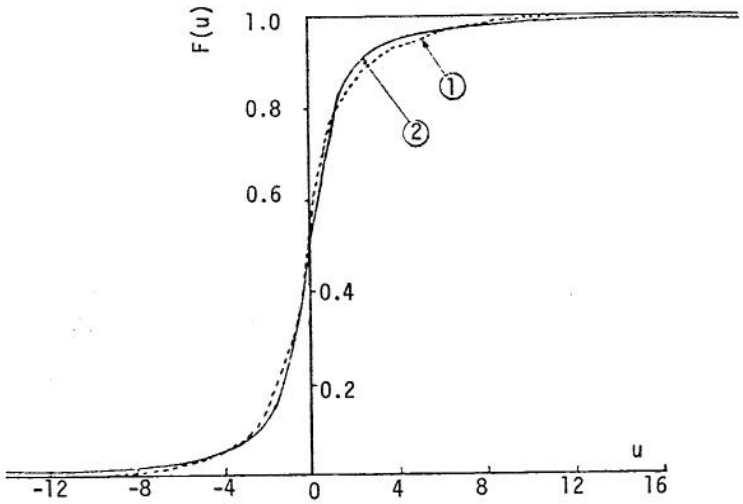

Fig. 5-24 Fitting the cumulative frequency distribution of u-series of the Madison River by the symmetric stable distribution: (1) cumulative frequency distribution and (2) stable distribution fitted, with $\alpha = 1.264$, $\delta = 0$, and $\gamma = 1$.

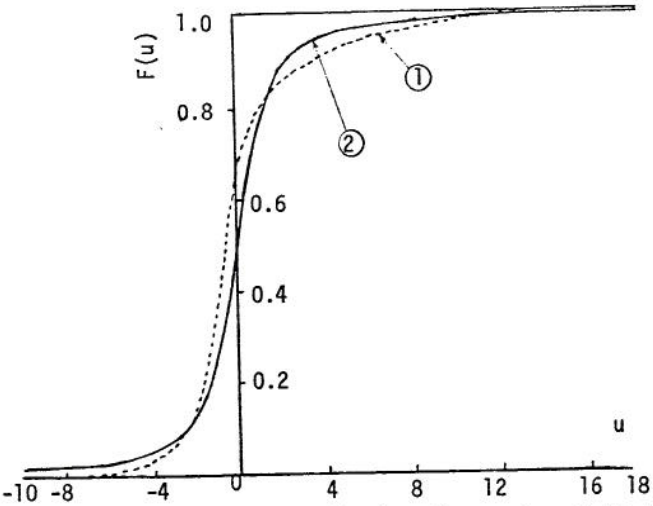

Fig. 5-25 Fitting the cumulative frequency distribution of u-series of the Batten Kill River by the symmetric stable distribution: (1) cumulative frequency distribution and (2) stable distribution fitted, with $\alpha = 1.255$, $\delta = 0$, and $\gamma = 1$.

Chapter 6
CONCLUSIONS

The following conclusions were drawn from the results of this study, namely:

1. To infer the periodicity in parameters from a series, harmonic analysis is used and the significant harmonics are identified. Errors in determining the number of significant harmonics and errors in estimating their coefficients greatly affect the accuracy of inferred periodic functions. For example, if a harmonic explains only one percent of the total variance of a parameter of its estimates over the discrete value of the basic period, and is incorrectly either accepted or rejected by the usual tests of significance, the maximum error in the inferred periodic function may be as high as 14 percent of the standard deviation of the periodic parameters.

2. Although high sampling fluctuations always exist in the estimated serial correlation coefficients, the periodicity in these series should be tested. Since the standard deviation of the independent stochastic component, as the residuals of the linear autoregressive models for stochastic variables, are highly sensitive to errors in the estimates of serial correlation coefficients, a careful investigation of the periodicities in these coefficients is necessary for more realistic models.

3. The logarithmic transformation provided improvements by assigning different weights to values (by decreasing the weights of high values in comparison with the weights of low values) and by reducing flow fluctuations in comparison with the original series. Consequently, the transformed data produced somewhat better results than the original series in fitting the independent stochastic component by selected probability distribution functions.

4. Distributions of independent stochastic components tend to have very long tails especially in series with small time units such as daily series. However, the evidence in this study suggests that the tails are not heavy. Exponential functions are found to be good approximations for the tails of frequency distributions in a large majority of cases.

5. Frequency distributions of the independent stochastic components cannot be adequately fitted by stable distributions with heavy tails and an infinite variance.

6. Independent stochastic components obtained for the logarithmically transformed monthly runoff series were found to be approximately normally distributed, while the frequency distributions were found to be skewed but bell-shaped for the 13-day series. The normal function, modified by Hermite polynomials, the lognormal, and the gamma probability distribution functions are found to fit well these frequency distributions. As the discrete time interval (in which the year is divided) becomes smaller, such as the 7-day and 3-day series, the well-rounded, bell-shaped distributions of independent stochastic components change to highly skewed distributions with a sharper peak and the longer tails. Mixed Pearson's Type VII and gamma functions and the double-branch gamma function are more suited for modeling this kind of distributions.

7. Frequency distributions of independent stochastic components for different intervals Δt indicate that the distributions become closer to normal when Δt increases. The central limit theorem is then useful in modeling probability distributions of hydrologic independent stochastic components.

8. For the independent stochastic components of the daily flow series, none of the probability distribution functions studied for fitting the frequency distributions could pass the chi-square test, even with the significance criterion of 99 percent. The double-branch gamma function had the smallest chi-square values. The Kolmogorov-Smirnov test also rejects the hypothesis of good fit for any probability distribution function used. This difficulty results from very large samples with the resulting very narrow tolerance limits. This problem requires a special attention in future studies.

REFERENCES

1. Anderson, R. L., 1941, Distribution of the Serial Correlation Coefficient: Annals of Mathematical Statistics, Vol. 13, pp. 1-13.

2. Barnes, F. B., 1954, Storage Required for a City Water Supply: J. Inst. Eng., 26, Australia, 198.

3. Bonne, J., 1971, Stochastic Simulation of Monthly Streamflow by a Multiple Regression Model Utilizing Precipitation Data: Journal of Hydrology, Vol. 12, pp. 285-310.

4. Box, G. E. P., and G. M. Jenkins, 1972, Time Series Analysis Forecasting and Control: San Francisco, California, Holden-Day Inc., 542 p.

5. Bryson, M. C., 1974, Heavy-Tailed Distributions, Properties and Tests: Technometrics, Vol. 16, No. 1, pp. 61-67.

6. Chow, V. T., and S. Rameseshen, 1965, Sequential Generation of Rainfall and Runoff Data: J. Hydraul. Div., Proc. Soc. Civil Eng., 4416, pp. 205-223.

7. Cohen, A. C., 1967, Estimation in Mixtures of Two Normal Distributions: Technometrics, Vol. 9, No. 1, pp. 15-28.

8. Elderton, W. P., 1953, Frequency Curves and Correlation: Washington, D.C., the Harren Press, 230 p.

9. Fama, E. F., 1965, The Behavior of Stock-Market Prices: The Journal of Business, Vol. 36, pp. 34-105.

10. Fama, E. F., and R. Roll, 1968, Some Properties of Symmetric Stable Distributions: Journal of the American Statistical Association, Vol. 63, pp. 817-836.

11. Fama, E. F., and R. Roll, 1971, Parameter Estimates for Symmetric Stable Distributions: J. Amer. Stat. Assoc., Vol. 66, No. 334, pp. 331-338.

12. Feller, W., 1966, An Introduction to Probability Theory and Its Applications: New York, John Wiley and Sons, Vol. 2, 610 p.

13. Fercho, W. W., and L. J. Ringer, 1972, Small Sample Power of Some Tests of Constant Failure Rates: Technometrics, 14, pp. 713-724.

14. Gnedenko, B., et al, 1969, Mathematical Theory of Reliability: Academic Press.

15. Hasselblad, V., 1966, Estimation of Parameters for a Mixture of Normal Distributions: Technometrics, Vol. 8, No. 3, pp. 431-444.

16. Johnson, N. L., and S. Kotz, 1970, Continuous Univariate Distribution-1: Boston, Houghton Mifflin Company, Vol. 1, 298 p.

17. Kneese, A. V., and S. C. Smith, 1966, Water Research: Baltimore, Maryland, Johns Hopkins Press, 375 p.

18. Llamas, J., and M. M. Siddiqui, 1969, Runs of Precipitation Series: Colorado State University Hydrology Paper No. 33, Fort Collins, Colo., 16 p.

19. Maass, A., et al, 1962, The Design of Water Resources Systems: Harvard University Press, Cambridge, Massachusetts, 620 p.

20. Markovic, R. D., 1965, Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff: Colorado State University Hydrology Paper No. 8, Fort Collins, Colorado, 33 p.

21. Mandelbrot, B., 1963, The Variation of Certain Speculative Prices: The Journal of Business, Vol. 36, pp. 394-419.

22. Matalas, N. C., 1967, Mathematical Assessment of Synthetic Hydrology: Water Resources Research, Vol. 3, No. 4, pp. 937-945.

23. Ord, J. K., 1970, Families of Frequency Distribution: No. 30 of Griffin's Statistical Monographs and Curves Edited by Alan Stuart.

24. Parzen, E., 1962, Stochastic Processes: San Francisco, California, Holden-Day Inc., 306 p.

25. Pearson, E. S., 1965, Tables of Percentage Points of $\sqrt{G_1}$ and b_2 in Normal Samples; A Rounding Off: Biometrika, Vol. 52, pp. 282-285.

26. Press, J. P., 1972, Estimation in Univariate and Multivariate Stable Distributions: Journal of the American Statistical Association, Vol. 67, pp. 842-846.

27. Quenouille, M. H., 1949, A Large Sample Test for the Goodness of Fit in Autoregressive Schemes: Journal Royal Statistical Society, Vol. 110, pp. 123-239.

28. Quimpo, R. G., 1966, Stochastic Analysis of Daily River Flows: Ph.D. Dissertation, Colorado State University, 96 p.

29. Quimpo, R. G., 1967, Stochastic Model of Daily River Flow Sequences: Colorado State University Hydrology Paper No. 18, Fort Collins, Colorado, 30 p.

30. Siddiqui, M. M., 1960, Test for Regression Coefficients When Errors are Correlated: An. Math. Statistics, Vol. 31, pp. 931-932.

31. Sturges, H. A., 1926, The Choice of a Class Interval: Journal of American Statistical Association, Vol. 21, p. 65-66.

32. Sudler, C. E., 1927, Storage Required for the Regulation of Streamflow: Trans. Amer. Soc. Civil Eng., Vol. 91, pp. 622-660.

33. Tao, Pen-Chih, 1973, Distribution of Hydrologic Independent Stochastic Components: Ph.D. Dissertation, Colorado State University, Fort Collins, Colorado.

34. Yevjevich, V., 1964, Fluctuations of Wet and Dry Years, Part II, Analysis by Serial Correlation: Colorado State University Hydrology Paper No. 4, Fort Collins, Colorado, 50 p.

35. Yevjevich, V., 1972, Probability and Statistics in Hydrology: Water Resources Publications, Colorado State University, Fort Collins, Colorado.

36. Yevjevich, V., 1972, Structural Analysis of Hydrologic Time Series: Colorado State University Hydrology Paper No. 56, Fort Collins, Colorado, 59 p.

37. Yevjevich, V., 1972, Stochastic Processes in Hydrology: Water Resources Publications, Colorado State University, Fort Collins, Colorado.

KEY WORDS: Time series, structural analysis, periodicities, distribution functions for runoff series, distribution tail characteristics:

ABSTRACT: Structural analysis and mathematical modeling used in evaluation and removal of periodicity and dependence from hydrologic time series are first reviewed, summarized and discussed. Seventeen river daily series in USA are used as basic research data. Periodicities in serial correlation coefficients of stochastic components are found not to be negligible. Independent stochastic components result from removing periodicity and dependence in parameters. Methods of testing distributions of tails of frequency distributions of these independent residuals are developed. Tails have been shown not to belong to the class of heavy tails, but rather the exponential tails. Seven groups of probability distribution functions, namely classical, Pearson's family,

KEY WORDS: Time series, structural analysis, periodicities, distribution functions for runoff series, distribution tail characteristics:

ABSTRACT: Structural analysis and mathematical modeling used in evaluation and removal of periodicity and dependence from hydrologic time series are first reviewed, summarized and discussed. Seventeen river daily series in USA are used as basic research data. Periodicities in serial correlation coefficients of stochastic components are found not to be negligible. Independent stochastic components result from removing periodicity and dependence in parameters. Methods of testing distributions of tails of frequency distributions of these independent residuals are developed. Tails have been shown not to belong to the class of heavy tails, but rather the exponential tails. Seven groups of probability distribution functions, namely classical, Pearson's family,

KEY WORDS: Time series, structural analysis, periodicities, distribution functions for runoff series, distribution tail characteristics:

ABSTRACT: Structural analysis and mathematical modeling used in evaluation and removal of periodicity and dependence from hydrologic time series are first reviewed, summarized and discussed. Seventeen river daily series in USA are used as basic research data. Periodicities in serial correlation coefficients of stochastic components are found not to be negligible. Independent stochastic components result from removing periodicity and dependence in parameters. Methods of testing distributions of tails of frequency distributions of these independent residuals are developed. Tails have been shown not to belong to the class of heavy tails, but rather the exponential tails. Seven groups of probability distribution functions, namely classical, Pearson's family,

KEY WORDS: Time series, structural analysis, periodicities, distribution functions for runoff series, distribution tail characteristics:

ABSTRACT: Structural analysis and mathematical modeling used in evaluation and removal of periodicity and dependence from hydrologic time series are first reviewed, summarized and discussed. Seventeen river daily series in USA are used as basic research data. Periodicities in serial correlation coefficients of stochastic components are found not to be negligible. Independent stochastic components result from removing periodicity and dependence in parameters. Methods of testing distributions of tails of frequency distributions of these independent residuals are developed. Tails have been shown not to belong to the class of heavy tails, but rather the exponential tails. Seven groups of probability distribution functions, namely classical, Pearson's family,

those modified by polynomials, Weibull, double-branch gamma, mixture of functions, and family of stable distribution functions, were applied to fit frequency distributions of independent stochastic components. The same techniques were applied to the 3-day, 7-day, 13-day and 30-day monthly series. It was found that the 3-parameter lognormal function fits well the frequency distributions of monthly independent stochastic components. Since frequency distributions for small time intervals were skeded, with shapr peaks and long tails, probability distribution functions with more parameters must be used to fit these distributions.

Reference: Tao, Pen-chih, V. Yevjevich & N. Kottegoda, Colorado State University, Hydrology Paper No. 82, January (1976), Distributions of Hydrologic Independent Stochastic Components.

those modified by polynomials, Weibull, double-branch gamma, mixture of functions, and family of stable distribution functions, were applied to fit frequency distributions of independent stochastic components. The same techniques were applied to the 3-day, 7-day, 13-day and 30-day monthly series. It was found that the 3-parameter lognormal function fits well the frequency distributions of monthly independent stochastic components. Since frequency distributions for small time intervals were skeded, with shapr peaks and long tails, probability distribution functions with more parameters must be used to fit these distributions.

Reference: Tao, Pen-chih, V. Yevjevich & N. Kottegoda, Colorado State University, Hydrology Paper No. 82, January (1976), Distributions of Hydrologic Independent Stochastic Components.

those modified by polynomials, Weibull, double-branch gamma, mixture of functions, and family of stable distribution functions, were applied to fit frequency distributions of independent stochastic components. The same techniques were applied to the 3-day, 7-day, 13-day and 30-day monthly series. It was found that the 3-parameter lognormal function fits well the frequency distributions of monthly independent stochastic components. Since frequency distributions for small time intervals were skeded, with shapr peaks and long tails, probability distribution functions with more parameters must be used to fit these distributions.

Reference: Tao, Pen-chih, V. Yevjevich & N. Kottegoda, Colorado State University, Hydrology Paper No. 82, January (1976), Distributions of Hydrologic Independent Stochastic Components.

those modified by polynomials, Weibull, double-branch gamma, mixture of functions, and family of stable distribution functions, were applied to fit frequency distributions of independent stochastic components. The same techniques were applied to the 3-day, 7-day, 13-day and 30-day monthly series. It was found that the 3-parameter lognormal function fits well the frequency distributions of monthly independent stochastic components. Since frequency distributions for small time intervals were skeded, with shapr peaks and long tails, probability distribution functions with more parameters must be used to fit these distributions.

Reference: Tao, Pen-chih, V. Yevjevich & N. Kottegoda, Colorado State University, Hydrology Paper No. 82, January (1976), Distributions of Hydrologic Independent Stochastic Components.

LIST OF PREVIOUS 25 PAPERS

No. 56 Structural Analysis of Hydrologic Time Series, by Vujica Yevjevich, November 1972.

No. 57 Range Analysis for Storage Problems of Periodic-Stochastic Processes, by Jose Salas-La Cruz, November 1972.

No. 58 Applicability of Canonical Correlation in Hydrology, by Padoong Torranin, December 1972.

No. 59 Transposition of Storms, by Vijay Kumar Gupta, December 1972.

No. 60 Response of Karst Aquifers to Recharge, by Walter G. Knisel, December 1972.

No. 61 Drainage Design Based Upon Aeration by Harold R. Duke, June 1973.

No. 62 Techniques for Modeling Reservoir Salinity by John Hendrick, August 1973.

No. 63 Mechanics of Soil Erosion From Overland Flow Generated by Simulated Rainfall by Mustafa Kilinc and Everett V. Richardson, September 1973.

No. 64 Area-Time Structure of the Monthly Precipitation Process by V. Yevjevich and Alan K. Karplus, August 1973.

No. 65 Almost-Periodic, Stochastic Process of Long-Term Climatic Changes, by William Q. Chin and Vujica Yevjevich, March 1974.

No. 66 Hydrologic Effects of Patch Cutting of Lodgepole Pine, by Thomas L. Dietrich and James R. Meiman, April 1974.

No. 67 Economic Value of Sediment Discharge Data, by Sven Jacobi and Everett V. Richardson, April 1974.

No. 68 Stochastic Analysis of Groundwater Level Time Series in the Western United States, by Albert G. Law, May 1974.

No. 69 Efficient Sequential Optimization in Water Resources, by Thomas E. Croley II, September 1974.

No. 70 Regional Water Exchange for Drought Alleviation, by Kuniyoshi Takeuchi, November 1974.

No. 71 Determination of Urban Watershed Response Time, by E. F. Schulz, December, 1974.

No. 72 Generation of Hydrologic Samples, Case Study of the Great Lakes, by V. Yevjevich May, 1975.

No. 73 Extraction of Information on Inorganic Water Quality, by William L. Lane, August, 1975.

No. 74. Numerical Model of Flow in Stream-Aquifer System, by Catherine E. Kraeger Rovey, August, 1975.

No. 75. Dispersion of Mass in Open-Channel Flow, by William W. Sayre, August, 1975.

No. 76 Analysis and Synthesis of Flood Control Measures, by Kon Chin Tai, September, 1975.

No. 77 Methodology for the Selection and Timing of Water Resources Projects to Promote National Economic Development, by Wendim-Agegnehu Lemma, August 1975.

No. 78 Two-Dimensional Mass Dispersion in Rivers, by Forrest M. Holly, Jr., September 1975.

No 79. Range and Deficit Analysis Using Markov Chains, by Francisco Gomide, October. 1975.

No. 80 Analysis of Drought Characteristics by the Theory of Run, by Pedro Guerrero-Salazar and Vujica Yevjevich, October 1975.

No. 81 Influence of Simplifications in Watershed Geometry in Simulation of Surface Runoff, by L. J. Lane, D. A. Woolhiser and V. Yevjevich, January 1976.