US Army Corps of Engineers
Hydrologic Engineering Center

Stochastic Analysis of Drought Phenomena

July 1985

TD-25

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
July 1985	Training Document	

4. TITLE AND SUBTITLE
Stochastic Analysis of Drought Phenomena

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David Goldman

5d. PROJECT NUMBER

5e. TASK NUMBER

5F. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Corps of Engineers
Institute for Water Resources
Hydrologic Engineering Center (HEC)
609 Second Street
Davis, CA 95616-4687

8. PERFORMING ORGANIZATION REPORT NUMBER
TD-25

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/ MONITOR'S ACRONYM(S)

11. SPONSOR/ MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
.

14. ABSTRACT

Droughts are caused by both hydrologic and socioeconomic components. This document concentrates on stochastic models of streamflow as the hydrologic component of drought. Stochastic models of drought are presented to the water resource engineer as an extension of the more commonly understood flood frequency analysis. The extension can be made because flood frequency analysis utilizes stochastic models of independent random variables whereas drought analysis utilizes stochastic models of dependent random variables. There are many different stochastic models that have been implemented to describe dependent random variables. A comparison of these models indicates that simple ones, such as the autoregressive model, are adequate for the water resource engineer's needs.

15. SUBJECT TERMS
stochastic hydrology, drought analysis, autoregressive model, crossing theory

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER
U	U	U	UU	154	

Stochastic Analysis of Drought Phenomena

July 1985

US Army Corps of Engineers
Institute for Water Resources
Hydrologic Engineering Center
609 Second Street
Davis, CA 95616

(530) 756-1104
(530) 756-8250 FAX
www.hec.usace.army.mil

TD-25

STOCHASTIC ANALYSIS OF DROUGHT PHENOMENA

CONTENTS

CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

This document would not have been possible without the support of the HEC staff. In particular, thanks are given to Bill Johnson, whose general guidance and support were invaluable to the completion of this document. Also, the support of Arlen Feldman (Chief of Research) and Darryl Davis (Chief of Planning) was also very valuable. Technical assistance was received from many members of the HEC staff. The comments of Mr. Harold Kubik were particularly valuable. The production of the manuscript was made possible by the efforts of Cathy Lewis, who performed the technical typing.

Special thanks are given to William L. Lane (United States Bureau of Reclamation) for his technical assistance. Dr. Lane's comments were invaluable in correcting errors and omissions in the original manuscript.

Preface

The study of extreme hydrologic events is of great importance because of their socio-economic impact. In fact, a great deal of time and effort has been invested in predicting the occurrence and quantifying the effects of hydrologic extremes. The effort expended in studying hydrologic extremes has been disproportionally focused on flood phenomena in comparison to the efforts made in studying droughts. However, the increasing demands on available water resources make the quantification and prediction of drought essential to water resources planning.

Although droughts have not been studied as extensively as floods, there is a growing body of knowledge on the subject. The purpose of this presentation is to discuss the current thinking on analyzing droughts, and to relate this analysis to the more frequently use and commonly understood flood frequency analysis. The presentation is divided into six sections. Section 1, Identification of Drought and Low-Flow, discusses the factors which can be used to identify these extreme events in the hydrologic record. Section 2, Stochastic Models Based on Introductory Probability Theory presents an introduction to the use of probability and statistics to model hydrologic phenomena. Section 3, Autoregressive Models for the Streamflow Process, a particular type of stochastic model is presented and example applications are given. Section 4, Drought Analysis, the stochastic models developed previously are applied to the drought analysis problem. Section 5, Evaluation of the Autoregressive Model, discusses the validity of the autoregressive model in view of some of

the research literature which has criticized its use. Section 6, Concluding Remarks, points out some of the advantages and disadvantages inherent in using stochastic hydrologic models.

This presentation focuses on the stochastic models of the hydrologic processes and avoids discussing the problems associated with modeling socio-economic demands on water resource projects which are an integral part of identifying drought. This approach is taken to simplify the general presentation of stochastic models and because the stochastic models of the socio-economic processes are of less interest to the hydrologist. Consequently, the general assumption is made that the demands on the water resource system are known and that the primary concern is with the stochastic modeling of the inputs (streamflow, rainfall, groundwater storage, etc.) to the water resource system.

Given that the discussion is focused on the stochastic model of the hydrologic process, the question is how do stochastic models differ from the models that the hydrologist usually employs in practice? As will be restated throughout the discussion, the probability models currently used to perform flood frequency analysis can be extended with a few additional concepts to develop models for the analysis of droughts.

Section 1

Drought Identification

1.1 Drought Definition

A major problem in analyzing droughts is separating their occurrence from the hydrologic record, i.e., defining their occurrence. The difficulty stems from the fact that drought occurrence depends on the interaction between the natural occurrence of water (hydrometeorologic factors) and the intended use of water (operational use).

As an example of this difficulty, consider the perception of drought from the viewpoints of the meteorologist, agriculturist and the hydrologist. The meteorologist views drought as below normal precipitation in a region; the agriculturist, as a soil moisture deficit during the growing season, the hydrologist, as below normal streamflow.

Even within each of these disciplines, the perception of drought varies. Consider the regional variability of meteorologic drought. Dracup et. al. (1980) report drought periods are considered to occur after six rainless days in Bali, and after two rainless years in Libya. The soil moisture deficit which corresponds to agricultural drought is a function of crop type as well as meteorologic conditions. The intended use of the water is a critical factor in hydrologic drought. As Beard and Kubik (1972) point out, streamflows which are considerably below normal for short periods (intense droughts of short duration) may be very significant in areas where demand is a small fraction of the normal supply but of little significance where ample storage

is present. On the other hand, long periods of slightly below average streamflow (long duration of low intensity) may be significant to uses which depend on storage but of little significance to small fraction users.

Thus, drought definition depends strongly on the particular focus of the analysis. A single characterization of this phenomenon is not possible.

1.2 Drought Analysis Tasks

In view of the different perceptions of drought, it is probably contentious to propose a general set of tasks to be followed in drought analysis. However, as a general point of discussion, the tasks proposed by Dracup et. al. (1980) are general enough to be used by all the disciplines mentioned and a good starting point. The following is a summary of the major points described in their article.

Drought analysis is divided among four tasks. The first task is to determine the nature of the water deficit. The water deficit refers to the choice of analyzing either precipitation, soil moisture or streamflow. Of course, a combined approach in drought analysis could be taken where all these phenomena are considered. In this presentation, a distinction is made between the cause of drought (precipitation) and the impacts due to drought (soil moisture or streamflow). For hydrologists/planners, the primary interest is in impacts. For that reason, the analysis is restricted to an individual deficit such as streamflow. However, either the combined or individual approach is valid.

The second task is to identify the integral period of time for the

analysis. The integral period of time is the time increment; hour, day, month, season, year, etc., over which the hydrologic data is averaged in the drought analysis, and is one of the two factors which determine the number of drought events in the hydrologic record (the truncation level is the other factor). An obvious effect of increasing the integral period length is the corresponding loss of information about the hydrologic process. For example, seasonal flows which are successively lower or higher than normal are not necessarily recognized when employing an annual integral period in the analysis.

The choice of the integral period distinguishes between the generally accepted definitions of extreme streamflow values, high-flows and floods on the higher end of the streamflow spectrum and droughts and low-flows on the lower end of the spectrum.

Low-flows and floods, are generally considered to be instantaneous measures of streamflow. For example, a flood is described in terms of a peak discharge, say the 100-year flood. The 100-year flood's peak discharge has a one percent chance of being equaled or exceeded in any given year. The term "one percent chance" is a probabilistic term which will be fully discussed in Section 2. Low flows are usually averaged over a number of days. Even though this is not technically an instantaneous measure, low-flows were grouped with floods since they are both analyzed in a statistically similar manner. For example, a common indicator of a low-flow event is the Q_7^{10} (the seven-day ten-year, low-flow). By definition, there is a ten percent chance that the mean daily flow volume for seven consecutive days will be less than the Q_7^{10} in any one year. On the other hand, high flows and droughts are measures of streamflow volume which are recorded on a time interval of months or years.

A third task is to establish the truncation level which is employed to distinguish droughts from other events in the hydrologic record. The truncation level reflects the socio-economic demands on the available water supply. For example, the mean annual streamflow (or some fraction of the mean) might be used to represent the expected demand of a muncipality on the available streamflow (see Figure 1.1). However, the demand need not be constant and can be represented by some time varying truncation level (See Figure 1.1). The assumption is made for the remainder of the discussion (unless it is stated to the contrary) that the truncation level is known, and for the sake of simplifying the discussion, is a constant value.

As can be seen from Figure 1.1, periods of flow below the truncation level are identified as drought or low-flow periods and flows above the truncation level as periods of high flow or flood periods. In fact, the separation (and symmetry) between the definitions of low-flow, drought, high-flow and flood can be seen quite readily in Figure 1.2 by combining the concepts of integral period and truncation level.

As a final task, a regional analysis approach to the problem is selected. Limiting the analysis to a single site is generally not feasible since the hydrologic record at a single site is too short to provide adequate estimates of drought statistics. The local hydrologic record can be extended in a regional analysis by considering the interrelationship between records covering a broad topographical area. The delineation of the study area is based on either <u>geomorphologic</u> or <u>statistical</u> homogeneity factors. Geomorphologic factors which delineate an area include topography (mountain ranges are an obvious factor), local storage (lakes) and soil properties. In the statistical approach, sites are grouped based on similar statistics of the hydrologic

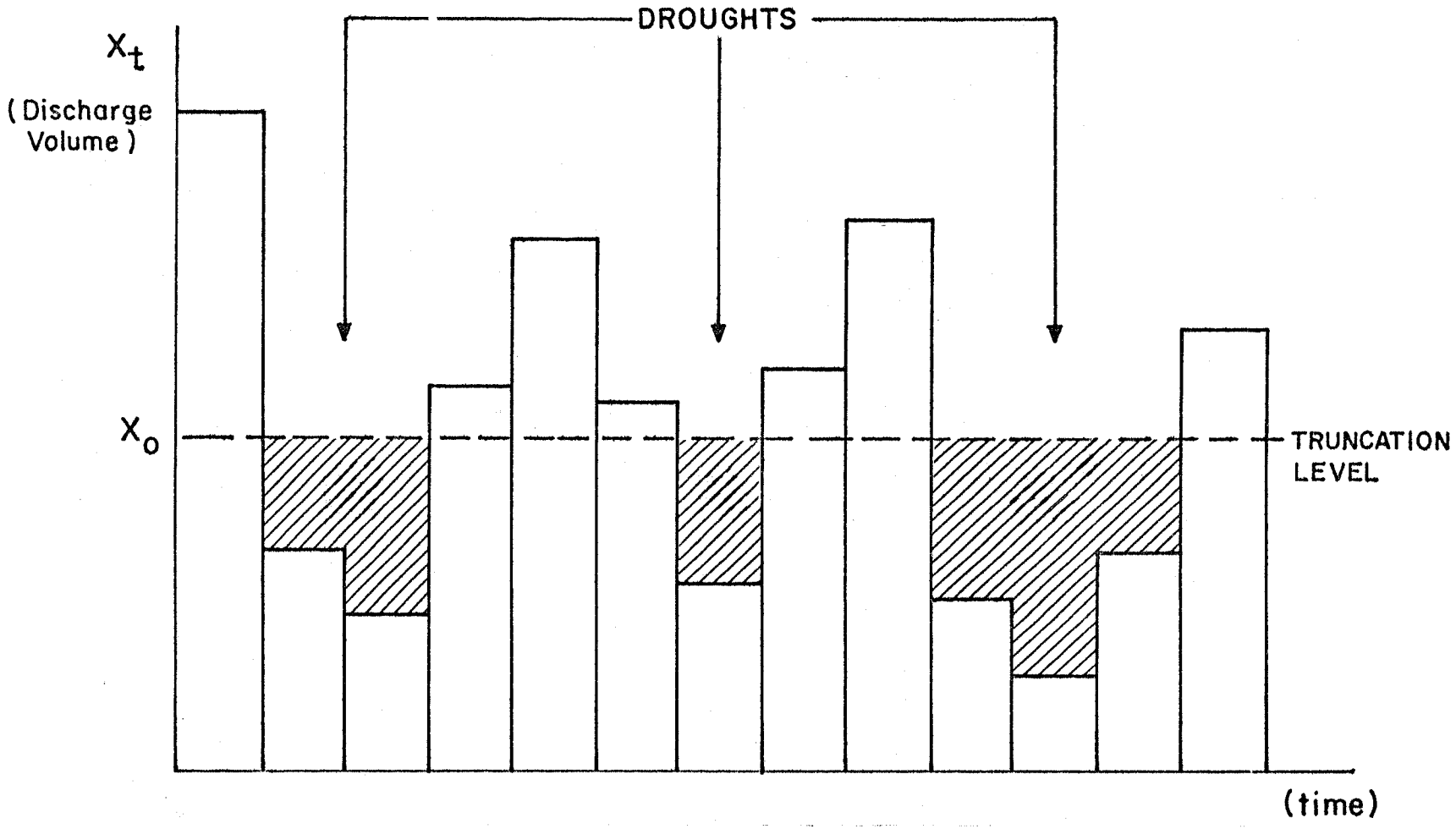


Figure 1.1 THE TRUNCATION LEVEL

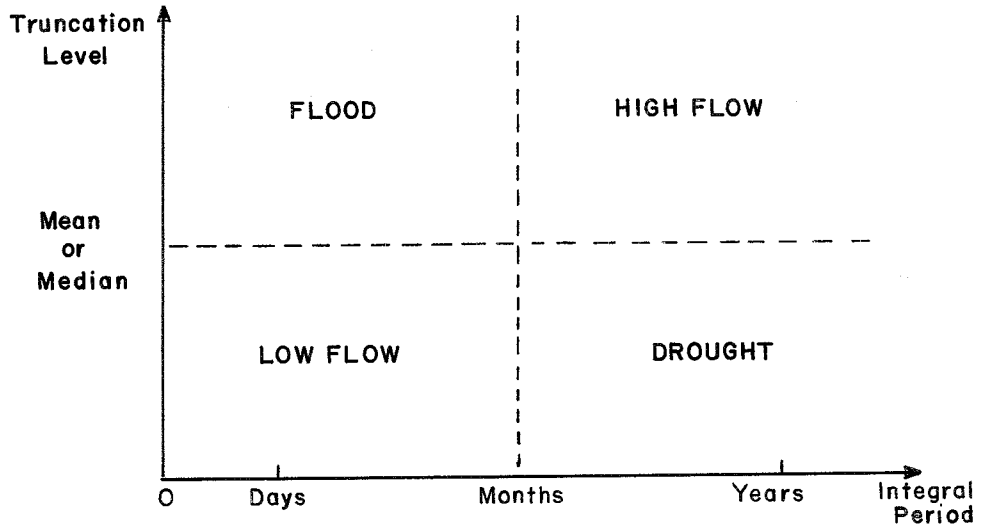


Figure 1.2 HIGH FLOWS, LOW FLOWS, DROUGHTS, FLOODS, AND THE INTEGRAL PERIOD. (after Dracup et al 1980)

record. At this time, delineating the areal extent of a drought based on these methods has not been adequately investigated and is an area of needed research.

The tasks defined by Dracup et. al. seem reasonable. The most difficult to apply is regional analysis. Regional analysis in flood studies has been applied successfully; however, the same cannot be claimed for drought analysis. In Section 2.5, the problems involved in performing a regional analysis in drought studies are explored.

In conlcusion, a drought can be identified from a hydrologic record based on the analysis tasks described. The major problem to be addressed is, therefore, to build a model of the hydrologic process, say a streamflow model, to predict, or at least estimate, the potential severity of future droughts, assuming that the truncation level and integral period are specified. Thus, the building of a stochastic model of the streamflow process is the desired end product of the subsequent discussion.

1.3 Summary

Drought occurrence is a function of socio-economic and hydrometeorologic factors. The combination of these factors make the identification and quantification of drought phenomena a difficult problem. The work done by Dracup et. al. (1980) has been referenced as a starting point for the identification of drought. Based on this work, a drought is identified in the hydrologic record based on four analysis tasks. The tasks determine the nature of the water deficit, the integral period, the truncation level for the analysis and whether or not regional analysis is to be applied. Consequently,

recognition of the fact that drought is defined in conjunction with socio-economic considerations, as embedded in these tasks, is a key concept.

Consideration of socio-economic factors are extremely important in accomplishing the above tasks. However, the inclusion of socio-economic variables into drought analysis is beyond the scope of this discussion. Thus, the discussion of stochastic analysis will presume that the first three tasks have been completed.

Section 2

Stochastic Models Based on Introductory Probability Theory

2.1 Introduction

The purpose of this section is to introduce stochastic models of the streamflow process by relating them to the probability models used in flood or low-flow frequency analysis. The advantage to this approach is that hyrologists often use frequency analysis and it is described in most introductory hydrology texts.

The discussion begins by delineating the difference between stochastic and probability models on the one hand and deterministic models on the other hand. Again, the hydrologist (and the engineer in general) is much more familiar and comfortable with deterministic models than with stochastic models. The comparison of these two categories of models is thus useful since it develops a framework in which the discussion can lead the hydrologist from models which are more commonly used, deterministic models to models which are less commonly used, probabilistic models used in frequency analysis, and finally; to models which are not well understood and used sparingly, stochastic models in drought analysis.

The probability models in low-flow and flood frequency analysis are related to streamflow stochastic models by recognizing that the former are models of independent random variables and the latter are models of dependent random variables. The relationship exists because even though the mathematical theory for independent random variables is much simpler to understand than that for dependent random variables, concepts are involved which are common to both

theories. For example, the concepts of probability or exceedance frequency, estimation and probability distributions are necesary in describing either independent or dependent random variables. Consequently, a detailed discussion of probability models for independent random variables is included as a stepping stone to the description of stochastic models.

However, the major difficulty in modeling dependent random variables and, in turn, developing streamflow stochastic models, is incorporating dependence between random variables into the mathematical theory. The extension of the technique for building mathematical models for dependent random variables by including dependence between random variables is theoretically simple, but leads to a very difficult estimation problem.

The discussion ends by describing how "time series" analysis uses regression techniques to solve the difficult estimation problem associated with modeling dependent random variables. The final step of including dependence into the relationship between random variables by regression methods is the essential concept necessary to build a stochastic streamflow model.

2.2 Stochastic vs. Deterministic Models

Stochastic and deterministic models are used extensively in water resources engineering. Although our primary interest focuses on stochastic modeling, it is instructive to examine how the two approaches differ by means of an example. Consider the problem of designing a system of reservoirs that is needed to meet the water supply requirements of a growing city. In order to estimate the required storage capacity of the reservoir system, the estimates of the future inflows to and demands on the reservoir system have to be estimated for the system's operating life.

Obviously, estimating the future inflows and demands for the reservoir system is a rather difficult problem. Our primary interest focuses on being able to predict the likely inflows. As previously explained in Section 1.3 the socio-economic aspects are beyond the scope of this discussion.

A possible means of determining future inflows (only theoretically possible) would be to create a mathematical model, based on the fundamental laws of classical physics, which simulates future weather conditions. The results of the model prediction coupled with a model which simulates the movement of precipitation through the earth's hydrologic cycle (a watershed model) is then used to predict future streamflows. Unfortunately, the present day technology does not exist to produce accurate long-term weather projections because of the complexity of the earth's atmospheric processes. Currently the best physical models of the atmosphere can make predictions on the order of a few days. However, if this type of model existed, then the meteorologic conditions and thus the inflows, to the reservoir over its economic life would be predicted a-priori (i.e., predicted before it is observed). This type of reservoir inflow model is <u>deterministic</u>. A deterministic model attempts to predict the value of some variable, in this case streamflow, before the variable can be observed.

Deterministic models of reservoir operations or watershed dynamics are commonly applied in water resources engineering. For example, given the future inflows to a reservoir system and the operating characteristics of the system, the resulting reservoir outflows can be predicted a-priori with a reservoir simulation model.

Since prediction of future streamflows by deterministic methods is an

extremely complex task, simplifying assumptions must be made to estimate the potential inflows to a reservoir system. A common approach is to presume that the future inflows are identical to the past inflows. A basic difficulty with this approach is that it is highly unlikely that the sequence of observed flows will be repeated in the future. An alternative to this approach is to assume that the past record flows are observations of a random or stochastic process. A random process is one in which the value of future occurrences (lets say streamflow) cannot be predicted with certainty. If the underlying probability laws governing the random process can be identified, then the probable inflows to the reservoir system might be estimated (a more in depth discussion of random variables and their corresponding probability laws is given in subsequent sections). This approach has the advantage over the more traditional approach in that the future sequences of inflow to the reservoir are not assumed to be identical to the historic flow sequence; and also, has the advantage of being a great deal simpler than the deterministic model alternative.

Of course, a price has been paid in viewing the streamflow process as a random process. First, a means for inferring the underlying probability law, or equivalently, developing a stochastic model, governing the streamflow process must be developed. Second, the stochastic model of the streamflow process is not able to predict future streamflow, but only the relative likelihood that future streamflows will take on certain values.

Consequently, the difference between deterministic and stochastic models is that the predictions of a deterministic model are in terms of a single value (e.g., the streamflow volume next year will be a 1000 acre-feet) whereas prediction of a stochastic model are in terms of the relative likelihood that

streamflow will take on certain valves (e.g.,there is a ninety percent chance that next years streamflow volume will exceed 1000 acre-feet).

The engineer is much more familiar with the deterministic than the stochastic approach. This may lead to the misconception that the deterministic approach is superior. This certainly is not true in general. For example, the accepted view of nature in the science of quantum mechanics is decidedly stochastic. In the water resource sciences, there are advocates of both approaches.

Stochastic models in water resources engineering are used to simulate processes which can be categorized as independent random variables or dependent random variables. In the case of streamflow analysis, annual floods and low-flows are usually assumed to be independent random variables. Processes represented by independent random variables are independent of any other process. For example, if the probability that the peak streamflow equals or exceed a certain value in any given year is independent of conditions of the previous years, or any other factor related to streamflow behavior, then the peak annual streamflow can be considered an independent random variable.

Processes that are represented by dependent random variables may be related to a number of factors. For example, if in a previous month the total streamflow volume is below normal then there is a good chance that the current month's streamflow will also be less than normal. The reason for this is that the available groundwater storage is a major factor in maintaining streamflow. Consequently, if the groundwater levels are depressed causing below normal streamflow in a previous month, it is quite likely that these groundwater levels will not recover in time to produce normal streamflow in the current

month. Consequently, monthly streamflows might be characterized by a random variable whose value is dependent (or conditional) on the previous month's value. Of course, this type of dependence is extremely important because successive monthly volumes below normal or below the truncation level can cause a drought.

2.3 Independent Random Variables

2.3.1 Probability

The concept of probability is thoroughly discussed in numerous books on probability and statistics (for example, Benjamin and Cornell, 1970). For the purpose of this discussion, probability is associated with observation frequency. For example, consider that peak annual streamflows are the obeservations of an independent random variable. After an extremely long period (longer than would be normally available from historic streamflow records), an estimate is made of the frequency with which streamflow peaks have certain values. In particular, lets say, that fifty percent of the observed streamflow peaks were greater than 1,000 cfs. Thus, there is the temptation to claim that the probability is 0.5 that an observed peak annual streamflow will be greater than 1,000 cfs. Equivalent statements would be that a flow of 1,000 cfs has an exceedance frequency of 50 percent, or that on the average one out of every two peak annual flows will exceed 1,000 cfs.

By convention the probability that an observation of a random variable will take on a value between its maximum and minimum values is one. Thus the probability that a random variable, X, is greater than a certain value, x, is equal to one minus that value. The mathematical notation for this is:

$$P[X > x] = 1 - P[X \leq x] \tag{2.1}$$

where: $P[X > x]$ = exceedance probability

$P[X \leq x]$ = nonexceedance probability

Technically speaking, unless the streamflow record is infinitely long, the observation frequency is only an estimate of the true probability. The estimation of probabilities associated with the values of random variables is a significant problem for the water resources engineer. (Unfortunately, the classic statistical techniques used to determine the reliability of probability estimates are of little use to the water resources engineer because hydrologic records are relatively short, on the order of 50 years.) Consequently, observations frequencies estimated from these short records may not give very good probability estimates. The problem of estimation is thus extremely important and will be continually emphasized throughout this discussion.

2.3.2 Probability Model Inference

A probability model (whether or not it pertains to independent or dependent random variables) defines the probability that a random variable will be observed with values between certain limits. Probability laws are usually described by a mathematical function which in this discussion is referred to as a probability or stochastic model. A major step in the analysis of random processes is to select the appropriate probability model.

There are two major tasks involved in selecting a probability model for an independent random variable. The first task is to estimate probabilities based on observation. The second task is to determine the probability model's

functional form based on the probabilities estimated in task one.

The methodology used for probability estimation is best described by an example. Consider the observations of annual streamflow volumes of the West Branch of the Oswegatchie River near Harrisville, New York shown in Table 2.1. The volumes are grouped in increasing intervals of 50 acre-feet, graphically represented as a histogram (Figure 2.1). The fraction of the total number of observations within each interval is an estimate of the random variable's occurrence frequency. The estimation methodology associates the occurrence frequency with the probability that an observation of the random variable, streamflow volume, will occur in any interval. In other words, the probability that a random variable occurs in a given interval is equal to the ratio of the expected number of observations in the interval to the total number of observations. Thus, probabilities are estimated from the observed occurrence frequencies. In the example, there are 7 volumes out of the total 65 between 250 and 300 acre-feet, giving an observation frequency or an estimate of the probability as 0.107 (number of observations/total number of observations). Another means of expressing this estimate is that there is a 10.7 percent estimated probability (or chance) that an observation occurs between 250 and 300 acre-feet.

Another convenient representation of the observed frequencies is the cumulative frequency distribution. This distribution is calculated by successively adding the frequency distribution values from the lowest interval to the interval of interest. In the example, the occurrence frequencies 0.046, 0.107, 0.276 add to the cumulative frequency of .429 at 350 acre-feet. An alternative expression for this estimate is that there is an estimated 42.9 percent chance that an observation will be less than 350 acre-feet.

TABLE 2.1
ANNUAL STREAMFLOW VOLUMES
of the West Branch of the
Oswegatchie River, Harrisville, N.Y
NY.

Year	Volume (acre ft)	Year	Volume (acre ft)
1917	338.1	1950	336.7
1918	392.3	1951	392.4
1919	406.2	1952	307.1
1920	350.7	1953	325.8
1921	361.3	1954	442.3
1922	414.1	1955	406.2
1923	255.6	1956	333.3
1924	409.4	1957	300.4
1925	400.3	1958	363.4
1926	449.6	1959	353.3
1927	348.3	1960	413.8
1928	534.3	1961	286.7
1929	463.3	1962	354.7
1930	453.2	1963	319.2
1931	249.8	1964	270.7
1932	415.3	1965	246.9
1933	354.7	1966	320.8
1934	261.3	1967	299.8
1935	363.4	1968	307.1
1936	326.7	1969	409.7
1937	421.3	1970	310.6
1938	401.1	1971	406.2
1939	313.5	1972	396.3
1940	288.2	1973	451.0
1941	241.1	1974	427.8
1942	335.9	1975	372.8
1943	432.3	1976	564.1
1944	315.8	1977	442.3
1945	368.5	1978	463.3
1946	380.1	1979	411.3
1947	604.5	1980	336.8
1948	341.2	1981	520.5
1949	334.5		

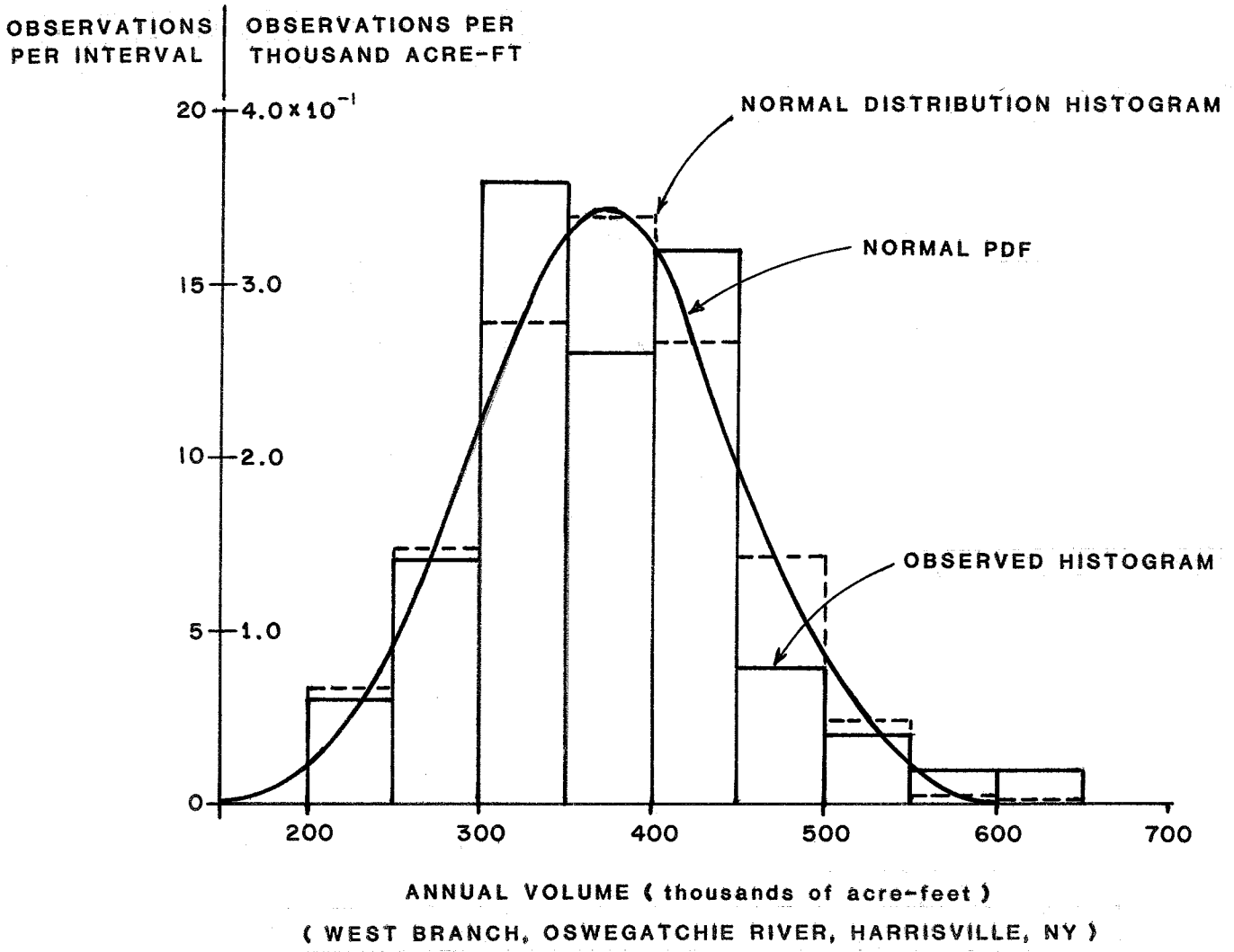


Figure 2.1 COMPARISON OF ANNUAL STREAMFLOW VOLUME HISTOGRAM AND NORMAL DISTRIBUTION

Estimating probabilities in this manner has some drawbacks. First, the choice of intervals (in this example 50 acre-feet) tends to be arbitrary. Second, the method constrains the occurrence probability of the random variable between the highest and lowest observations. This is an unfortunate constraint for the hydrologist who can never be sure that future observations will not exceed historical observations. This can be seen by inspection of the cumulative frequency distribution. By definition, the probability that the event occurs in an interval ranges from zero (no observation of the event) to one (absolute certainty of an observation). Consequently, the method estimates there is one hundred percent probability that an observation is between 200 and 650 acre-feet.

To avoid the interval problem, plotting positions are assigned to each observation (see Haan, 1977, pg. 133). The plotting positions are calculated by first arranging the flows from highest to lowest and assigning a rank to each observation (see Table 2.2). An estimate of the cumulative probability at each point is calculated using a plotting position formula, such as the Weibull formula:

$$P\,[X < x_i] = \frac{m}{N+1} \qquad (2.2)$$

where: x_i = observed event

X = random variable

m = rank

N = number of events

The factor N+1 is employed to allow for a finite probability that a flow occurs outside the observed flows.

Table 2.2
Plotting Position for
Annual Flows of the
West Branch of the Oswegatchie River,
Harrisville, N.Y.

Rank	Year	Annual Flow	% Exceedence Frequency	Rank	Year	Annual Flow	% Exceedence Frequency
1	1947	604.5	1.5	34	1921	361.3	51.5
2	1976	564.1	3.0	35	1962	354.7	53.0
3	1928	534.3	4.6	36	1933	354.7	54.6
4	1981	520.5	6.1	37	1959	353.3	56.1
5	1978	463.3	7.6	38	1920	350.7	57.6
6	1929	463.3	9.1	39	1927	348.3	59.1
7	1930	453.2	10.6	40	1948	341.2	60.1
8	1973	451.0	12.1	41	1917	338.1	62.1
9	1926	449.6	13.6	42	1980	336.8	63.6
10	1954	442.3	15.2	43	1950	336.7	65.2
11	1977	442.3	16.7	44	1942	335.9	66.7
12	1943	432.3	18.2	45	1949	334.5	68.2
13	1974	427.8	19.7	46	1956	333.3	69.7
14	1937	421.3	21.2	47	1936	326.7	71.2
15	1932	415.3	22.7	48	1953	325.8	72.7
16	1922	414.1	24.2	49	1966	320.8	74.2
17	1960	413.8	25.7	50	1963	319.2	75.7
18	1979	411.3	27.3	51	1944	315.8	77.3
19	1969	409.7	28.8	52	1939	313.5	78.8
20	1924	409.4	30.3	53	1970	310.6	80.3
21	1955	406.2	31.8	54	1968	307.1	81.8
22	1971	406.2	33.3	55	1952	307.1	83.3
23	1919	406.2	34.9	56	1957	300.4	84.9
24	1938	401.1	36.4	57	1967	299.8	86.4
25	1925	400.3	37.9	58	1940	288.2	87.9
26	1972	496.3	39.4	59	1961	286.7	89.4
27	1951	392.4	40.9	60	1964	270.7	90.1
28	1918	392.3	42.4	61	1934	261.3	92.4
29	1946	380.1	43.9	62	1923	255.6	93.9
30	1975	372.8	45.5	63	1931	249.8	95.5
31	1945	368.5	47.0	64	1965	246.9	97.0
32	1935	363.4	48.5	65	1941	241.1	98.5
33	1958	363.4	50.0				

The plotting position method estimates a cumulative probability for each data point. The theoretical justification for the plotting position approach is derived from the theory of order statistics, which is beyond the scope of this presentation (see Gumbel, 1958, for further reading).

The second task is to choose a probability model that corresponds to the probabilities estimated from the observed data. Probability models are generally represented in either of two functional forms. One form is the probability density function (PDF) (see Benjamin and Cornell, pg. 70, 1977). The PDF is the model proposed for comparison with the observed histogram. A second form is the cumulative distribution function (CDF) for both the case of discrete and continuous functions. The CDF is the model proposed for comparison with the observed cumulative frequency distribution. Although hydrologists generally deal with discrete data, the continuous PDF and CDF are most often used as probability models since streamflow or rainfall is thought of as a continuous process. The CDF is related to the cumulative area under the PDF, analagous to the relationship between the histogram and cumulative frequency distribution. Mathematically, this is expressed by the integral relationship:

$$F_X(x) = P\,[X \leq x_i] = \int_{-\infty}^{x_i} f_X(x)dx \qquad (2.3)$$

where $f_X(x)$ is the PDF and $F_X(x)$ the CDF, and minus infinity $-\infty$ is taken as the lowest bound for the random variable. This relationship is graphically demonstrated in Figure 2.2, and $P\,[X \leq x_i]$ is read as the probability that the random variable X is less than or equal to x_i, the upper bound on the integral.

The total probability of observing a random variable between its maximum and minimum limits is, by convention, equal to one. Taking plus and minus

infinity as the limits of the random variable, results in the integral relationship:

$$P\ [-\infty < x < \infty] = 1 = \int_{-\infty}^{\infty} f_X(x)dx$$

thus the area under the PDF is always unity. Note that in Figure 2.2 the cumulative probabilities approach zero and one, leaving a small but non-zero probability for any value of the random variable.

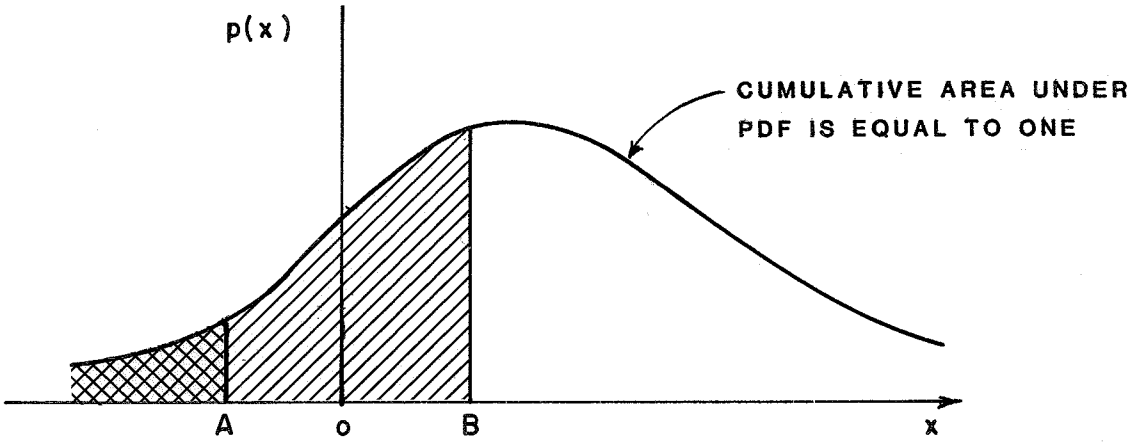
The method used to compare the probability model and the observed data is best illustrated by an example. Assume that the data for the West Branch of the Oswegatchie River is to be modeled by the normal distribution which is given by:

$$f_X(x) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left\{ \frac{(-1/2)(x-\mu_X)^2}{\sigma_X^2} \right\} \qquad (2.4)$$

where: μ_X = mean or average value of PDF

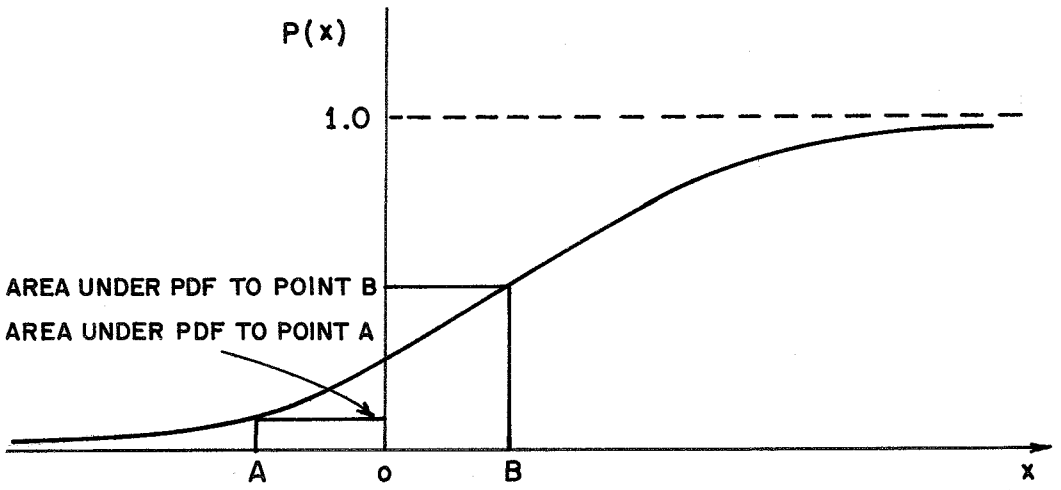
σ_X^2 = variance of the PDF

which are parameters of the distribution. (The normal distribution is the most well known distribution in the statistical and physical sciences. Tabulated values of the normal PDF and CDF may be found in most statistical texts, including the references already mentioned.) A comparison of the proposed model and the observed data is commonly made in either of two ways. One way is to compare the observed histogram and the theoretical or model histogram predicted by the PDF. The model histogram frequencies were calculated for each interval by computing the integral:



PROBABILITY DENSITY FUNCTION (PDF)

P(x) = EXCEEDANCE PROBABILITY



CUMULATIVE DISTRIBUTION FUNCTION

Figure 2.2 RELATIONSHIP BETWEEN PROBABILITY DENSITY FUNCTION AND CUMULATIVE DISTRIBUTION FUNCTION

$$P\,[x_i \leq X \leq x_{i+1}] = \int_{x_i}^{x_{i+1}} f_X(x)dx \qquad (2.5)$$

The observed probabilities or frequencies are then normalized by dividing by the interval length (50 acre-feet) so that the area under the histogram is equal to one. From the comparison, a judgment can be made as to the goodness of fit of the model and observed histograms (Figure 2.1).

Although the above approach is viable, it is cumbersome, and also suffers from the interval problem mentioned earlier. A second more convenient approach is to compare the CDF with the observed cumulative frequency distribution (Figure 2.3).

The comparison is facilitated by use of probability paper, which is specific to a particular CDF (see Haan, pg. 128, 1977). The example data are plotted on normal probability paper in Figure 2.4, for demonstration purposes. If the proposed probability model fits the data, then the data will lie close to a straight line on the probability paper.

The comparisons made in Figures 2.1, 2.3 and 2.4 indicate that the proposed model fits the data reasonably well in the central portion of the distribution but deviates in the "tails" of the distribution (e.g., the regions in the extreme portions of the distribution, 200 to 300 and 550 to 650 acre-feet). These deviations from the observed data probably indicate that the underlying distribution is skewed. A skewed distribution having a preponderant tail, is not symmetrical like the normal distribution.

Since most streamflows have a lower bound of zero, their frequency

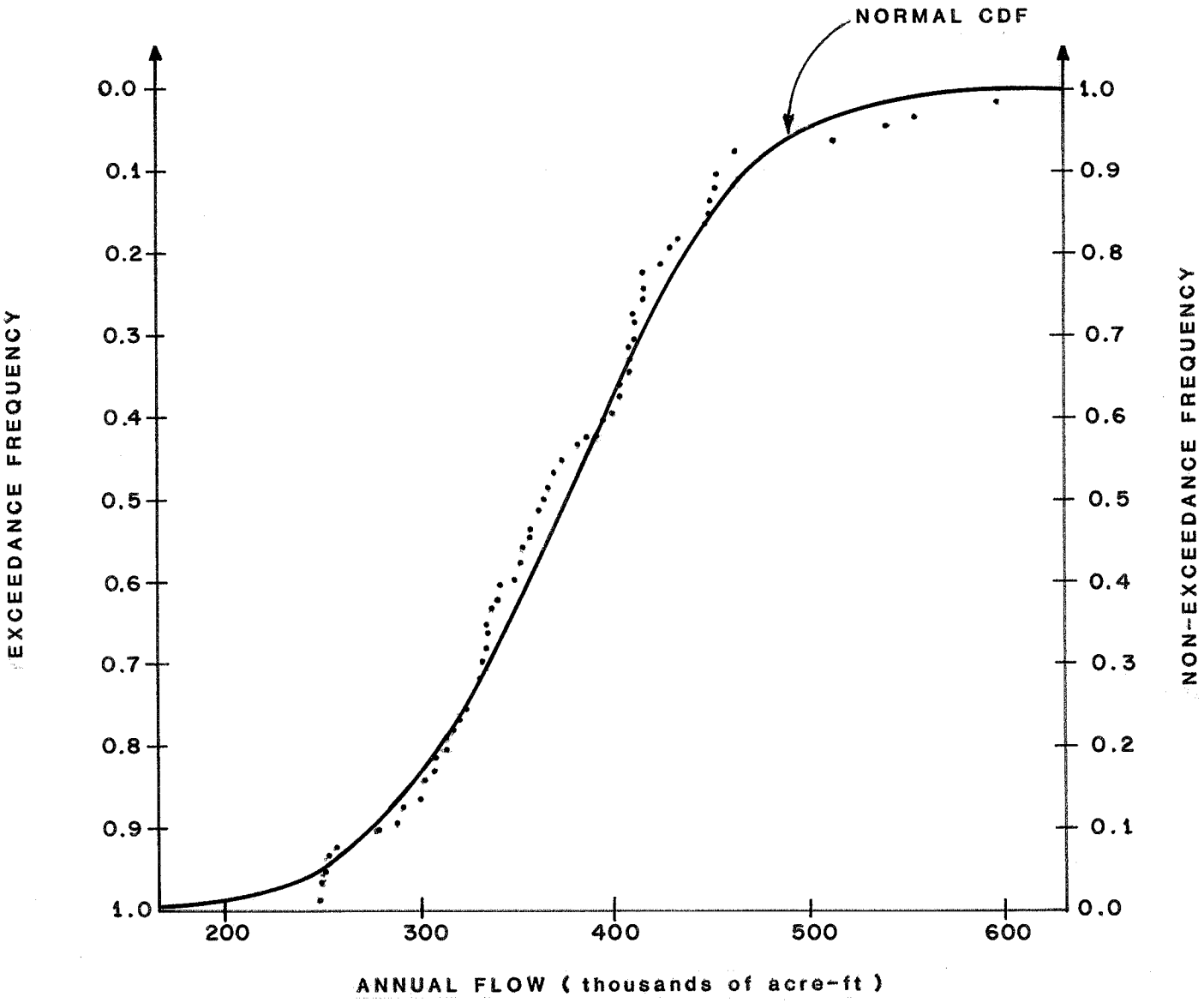


Figure 2.3 COMPARISON OF ANNUAL STREAMFLOW VOLUME, CUMULATIVE FREQUENCY DISTRIBUTION, AND NORMAL CDF.

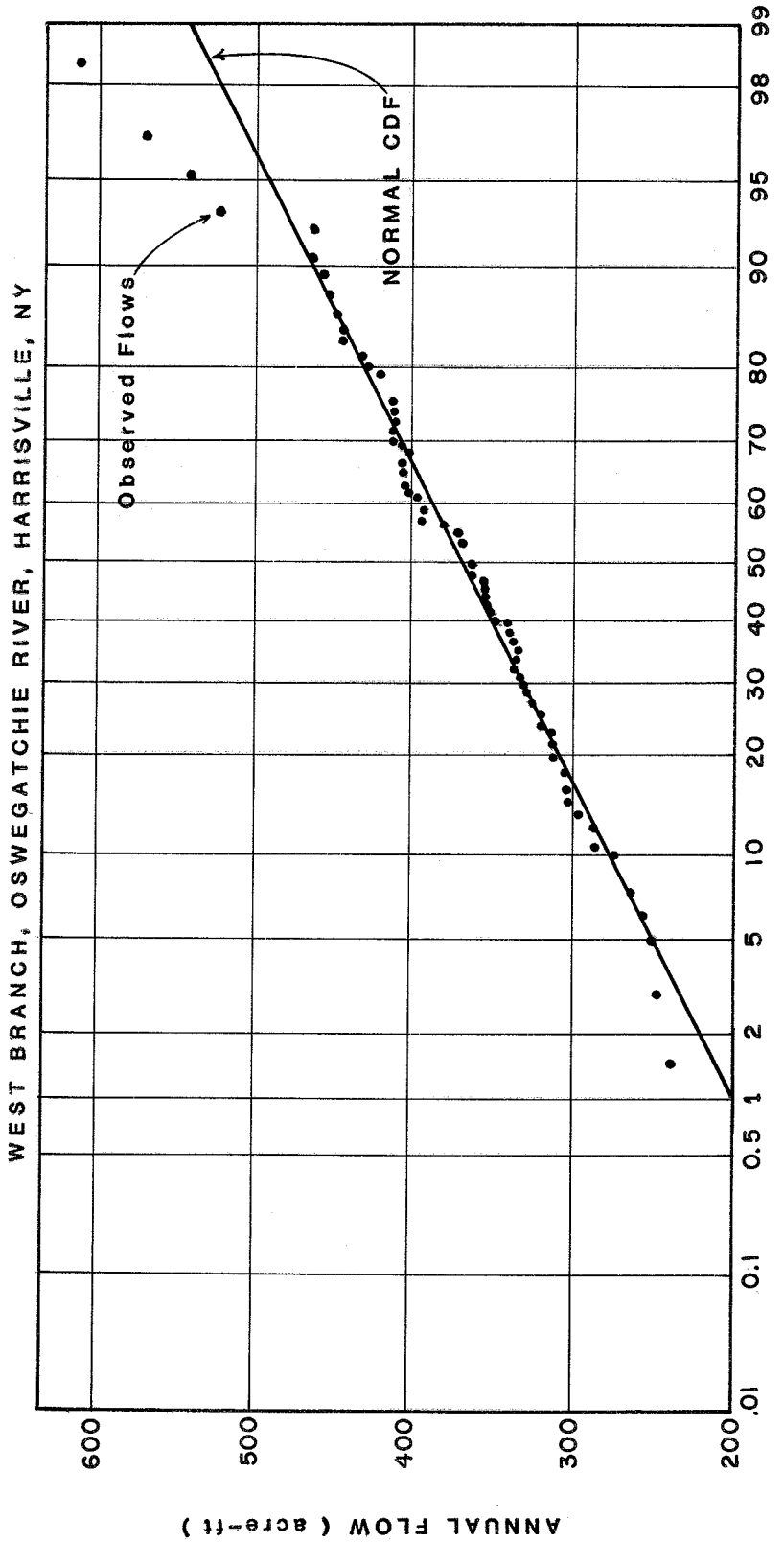


Figure 2.4 ANNUAL STREAMFLOW

distribution are necessarily asymmetrical. In some cases, streams are best treated as having a non-zero lower limit due to the existence of channel losses or external sources (e.g., spring flow). The considerations involved in selecting a frequency distribution appropriate for analyzing streamflows in drought analysis are discussed in Section 3.2., "Selection of the Marginal Distribution."

In the above discussion, terms such as "acceptable difference" or "close" were subjectively offered as criteria for accepting or rejecting the proposed model. The method can be made more objective by employing statistical "goodness of fit tests" (see Haan, pg. 174, 1977). However, there are problems with these tests when hydrologic data is involved. Criteria that might be more appropriately used for analysis of droughts are discussed in Section 3.

2.3.3 Probability Model Moments

The probability models shape indicates important properties of the random process. For example, the interest might focus on the central tendency or spread of values that can be expected. A means of characterizing these properties are the moments of the PDF. The moments which are of greatest interest are the mean, variance and skew coefficients.

The mean value, also referred to as expected value, the average value or the first moment, measures the central tendency value of the random variable X. The variance (the second central moment of the PDF) measures the width or the spread of squared values about the mean. The square root of the variance is the standard deviation. The skew coefficient (proportional to the third central moment) is a measure of the asymmetry of the PDF about the mean value.

The normal distribution has a skew coefficient of zero, being symmetrical about the mean. A distribution which has a pronounced tail to the right of the mean has a positive skew and to the left a negative skew (see Figure 2.5).

Each of these moments may be calculated from the PDF as follows:

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x) dx \tag{2.6}$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx \tag{2.7}$$

$$\gamma_X = \int_{-\infty}^{\infty} (x - \mu_X)^3 f_X(x) dx \,/(\sigma_X)^3 \tag{2.8}$$

where; μ_X = mean

σ_X = standard deviation

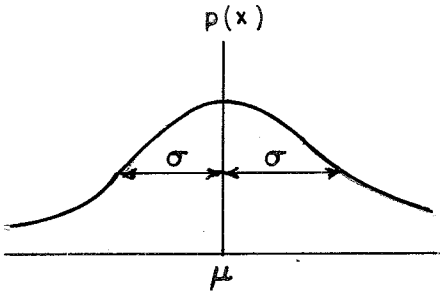
σ_X^2 = variance

γ_X = skew

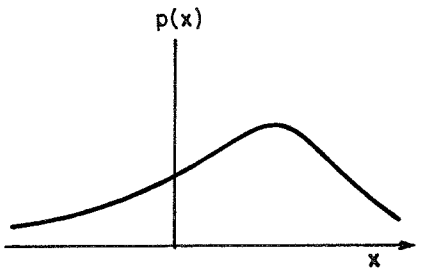
In the most general case, the moments of the PDF can vary with time. However, this type of model leads to some rather difficult estimation problems. To simplify this problem and for practical considerations, the moments are assumed constant for frequency analysis. For a further discussion of this point, see Section 2.4.1 on time series analysis. Given that the moments are constant, the problem is to estimate these parameters from observations of the random variable (i.e., the data, as discussed in the next section).

2.3.4 Moment Estimators

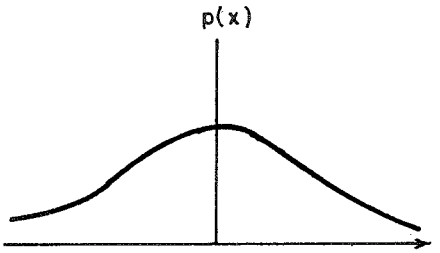
A large body of statistical theory is devoted to the estimation of model parameters from observations. In this theory, a great deal of effort is spent defining the "best" estimators. Our discussion is very limited in this regard



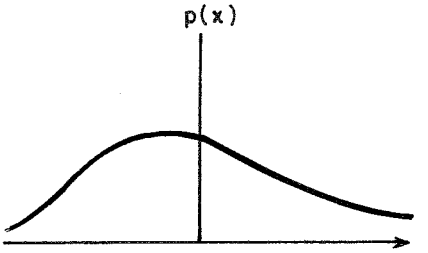
MEAN AND VARIANCE



NEGATIVE SKEW



ZERO SKEW



POSITIVE SKEW

Figure 2.5 SKEWED PROBABILITY FUNCTIONS

and presents only formulas for the sample estimates of distribution moments. Note that there are a number of methods available for estimation, but this approach is the most prevalent because it is easy to apply.

The moment estimating formulas for the estimation of the mean, variance and skew coefficient (in contrast to the true population values which were discussed previously) are:

$$m_X = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.9}$$

$$s_X = \sum_{i=1}^{N} \left[(x_i - m_X)^2 / (N - 1) \right]^{1/2} \tag{2.10}$$

$$g_X = N \sum_{i=1}^{N} (x_i - m_X)^3 / (N - 1)(N - 2) S_X^3 \tag{2.11}$$

where: m_X = sample mean

s_X = sample standard deviation

g_X = sample skew coefficient

X = random variable

x_i = ith observation

N = number of observation

The formulas given are unbiased, i.e., the expected value of the estimating equation is equal to the "true" value of the moments.

The numbers of observations play an important role in evaluating the reliability of the sample estimates. Consider for example the effect of the number of observations, N on the sample mean. The sample mean is a sum of random variables, x_i and thus is a random variable, with its own mean, standard deviation, skew and other moments. The standard deviation of the

sample mean is related to the standard deviation of the observations as (remembering that X is an independently distributed random variable):

$$\sigma_{m_X} = \sigma_X/\sqrt{N} \tag{2.12}$$

where: σ_{m_X} = standard deviation of the mean

Since the larger the standard deviation the greater the spread in the PDF and the greater the uncertainty in evaluating the random variable, the uncertainty in evaluating the sample mean decreases as the inverse of the square root of the number of observations. In water resources, the record lengths at a single station range on the order of 20 to 100 years. A rough calculation demonstrating the percent improvement in the estimate of the sample mean over this range is approximately:

$$\frac{\sigma_{m_{X\ N=100}}}{\sigma_{m_{X\ N=20}}} = \sqrt{\frac{20}{100}} = .45 \tag{2.13}$$

Consequently, if the estimates based on only twenty years are viewed with skepticism, then 100 years of data improves our view of the estimate by 45 percent.

2.4 Dependent Random Variables

2.4.1 Time Series Analysis

As mentioned previously, sequences of streamflow volumes (monthly, annual, etc.) may be modeled as a stochastic process. The inference of the probability model for stochastic processes is accomplished by techniques available in time series analysis. The term time series is an apt description for streamflow which is a sequence of observations in time. However, the term time series is

somewhat of a misnomer as a general description for stochastic processes. For example, the variation of some type of soil property, such as porosity, width distance may be modeled as a stochastic process. Techniques available in time series analysis can be used to characterize these observations even though they are a sequence in space rather than time.

At first glance, an attempt to analyze a time series might seem hopeless due to its chaotic nature. To simplify this analysis, we can take an operational view of the time series. The operational view assumes that the time series can be separated into deterministic and random components. The deterministic components are trends, periodicities and spurious events (see Figure 2.6). These components may result from identifiable physical phenomena. The random component is subtracted from the original time series. This random component may reflect an inherent property of the process or in the limitations of our physical model.

A trend is manifested in a long-term change in a property of the time series. A physical basis for a hydrologic trend is an identifiable long-term climatic change. An example of a hydrologic trend is a consistent increase with time of a stream's mean annual flow. Another example would be the onset of an ice age, possibly caused by a long-term decrease in average global temperature or increase in precipitation or a combination of both.

Trends in streamflow are difficult to identify since their occurrence may be due to the scale of observation being employed. For example, consider the trend free trace of a normal independent process generated by a numerical procedure (Figure 2.7). On the local scale shown, an argument might be made for the identification of a trend in the data. However, this conclusion based

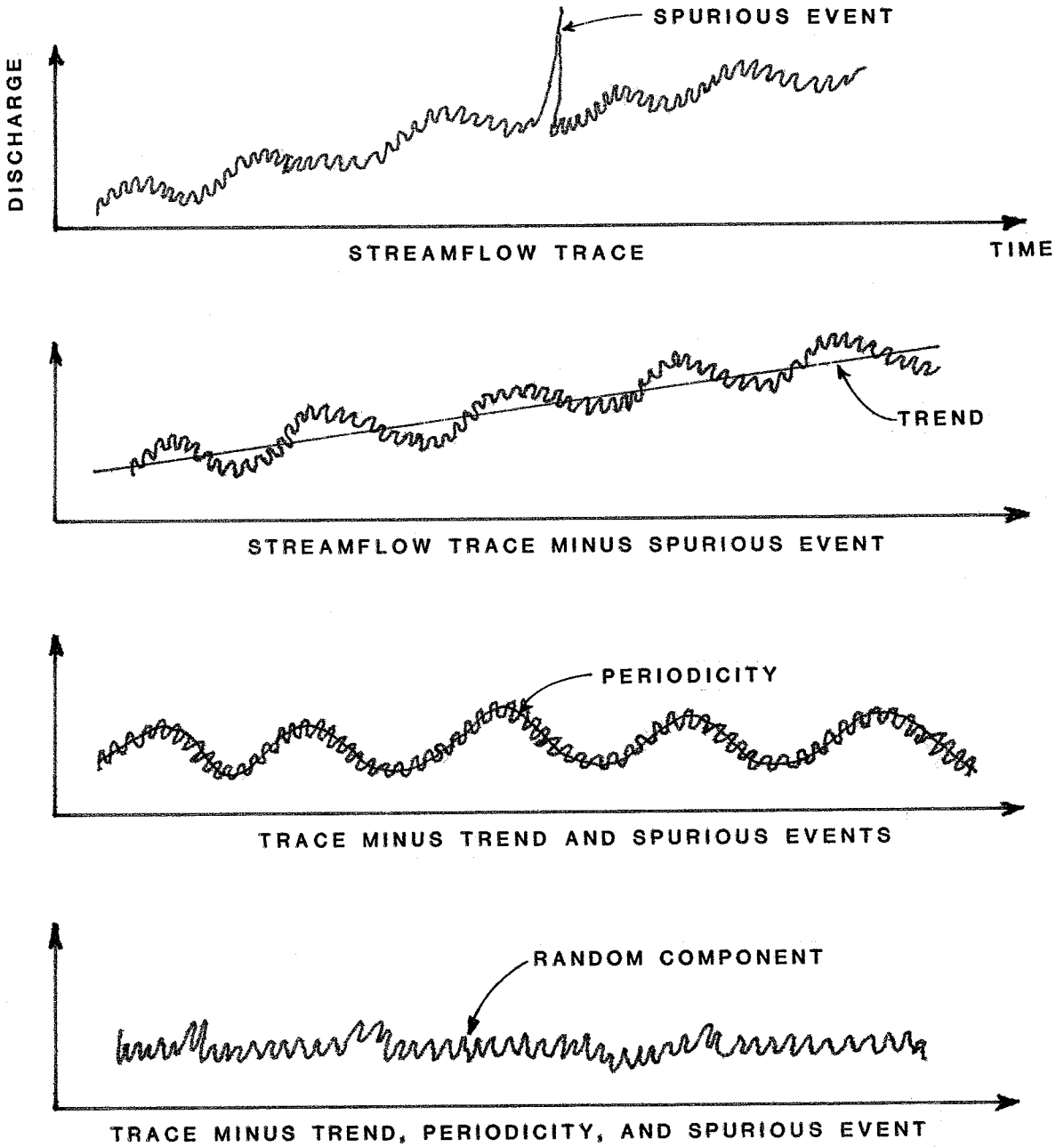


Figure 2.6 TIME SERIES, TRENDS, PERIODICITIES, SPURIOUS EVENTS, AND RANDOM PHENOMENA

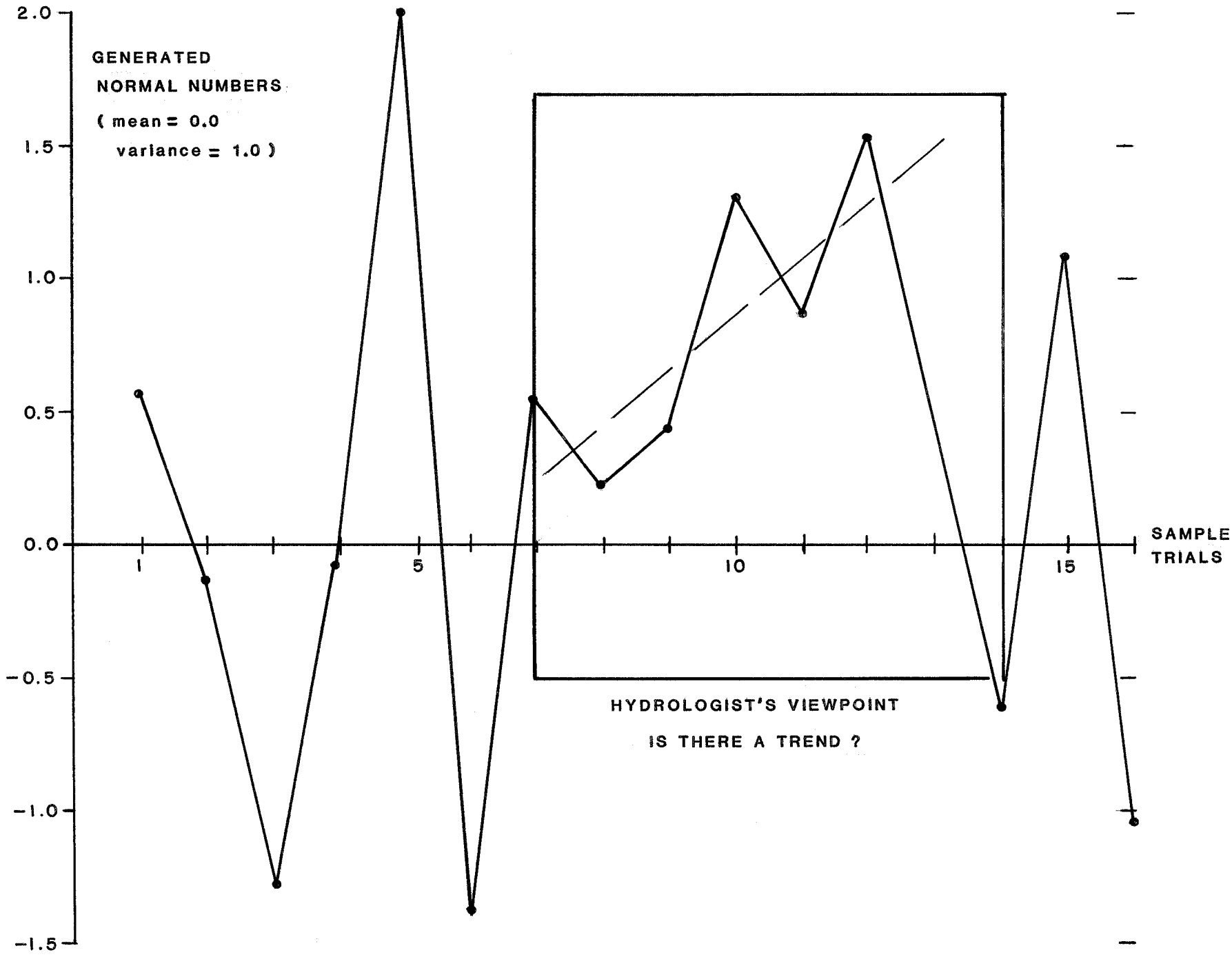


Figure 2.7 TRENDS AND THE NORMAL INDEPENDENT PROCESS

on the local scale of observation is incorrect since by construction a trend is not present.

Generally, trends are characterized by a linear or polynomial function (e.g., quadratic, cubic, etc.). There are statistical tests available for testing which of these models are adequate in describing the trend (ref. Draper and Smith, 1966). However, the water resources engineer's observation scale (50 years) is necessarily local due to the short-term nature of hydrologic records. Statistical tests only signify the reasonableness of the trend model on this scale. The extrapolation of an identified trend much beyond the observed record is difficult at best.

Periodicities are the identifiable cyclical aspects of a time series. Unlike trends, some periodicities are easily recognizable. For example, the annual cycle as it effects precipitation or streamflow (or weather patterns in general) are obvious periodic components in a natural time series.

The recognizable periodicities in a natural time series are readily added to stochastic models. For example, monthly or seasonal periodic fluctuations in mean streamflow volumes can be modeled (see Salas et.al. chapters 3 and 4), although inclusion of the periodicities makes inference of the model parameters more difficult. However, attempts to identify periodicities that exceed the annual cycle (e.g. identifying the twenty year drought) fall into the same difficulties as identifying long term trends. In general, the mathematical techniques in spectral analysis used to identify periodicities (Haan, 1977, or Jenkins and Watt, 1968) are beyond the scope of this presentation. The assumption is made that any periodicities identifiable in the hydrologic record are easily identifiable (annual cycle at most). The periodicities can

then be subtracted from the time series to simplify the analysis.

A spurious event or an outlier is a phenomena that is completely uncharacteristic of the time series record. A spurious event in the hydrologic record could be caused by a catastrophic event, such as a volcanic eruption. There has been speculation that the additional volcanic dust emitted by a volcano has a direct effect on atmospheric processes and thus on the hydrologic record.

In classic statistics, outliers are identified with measurement errors. Certainly, if a stream gage is not operating properly, then the engineer has good reason to discard data. Otherwise, the categorization of a datum as an outlier is risky (as well as controversial) business.

When all the deterministic components of a time series are removed; the trends, periodicities and spurious events, what remains is the random phenomena. In general, and in the case of streamflow volumes, the random phenomena usually demonstrates stochastic dependence. That is, the random phenomenda needs to be treated as a dependent random variable.

In general, time series analysis attempts to characterize a sequence of observations in two steps. First, the trends, periodicities and spurious events are identified and subtracted from the time series. Second, the probability model is postulated for the remaining random phenomena.

2.4.2 Stationarity and Ergodicity

Although the operational view of a time series is useful for analysis purposes, a more general view is to use a probability density function (PDF)

to model the time series. In the previous section, the moments of the PDF for an independent random variable were assumed to be constant. However, these moments could be allowed to vary with time. For example, a trend might be manifested in the increase with time of the mean annual flow (the first moment of the streamflow PDF).

Irrespective of the analysis point of view, long-term trends or periodicities are not included in stochastic models of streamflow. The reason for this is that extrapolation of trends or periodicities over the design life of a project (50 years) based on a record of 50 years involves too much uncertainty. Thus, for the purpose of predicting drought, or any long-term prediction involving streamflow, the assumption is made that streamflow sequences are free of trends or long-term periodicities (periodicities greater than the annual cycle).

The statistical equivalent to this view is to assume stationarity and ergodicity. Stationarity requires that all moments of the PDF are constant, such as the mean, standard deviation, skew etc. This assumption is actually too general for statistical analysis in water resources. The condition is usually relaxed to include only certain moments of the PDF (in our case only the first three moments). This is termed weak stationarity.

Ergodicity states that a property of a stochastic process derived from a single realization is the same as that derived from a number of realizations. As an example, consider three separate traces of hypothetical streamflows shown in Figure 2.8. Ergodicity requires that the mean for a single realization be the same as that determined from the observations across a group of realizations (A, B, C, etc). Of course, this property could be

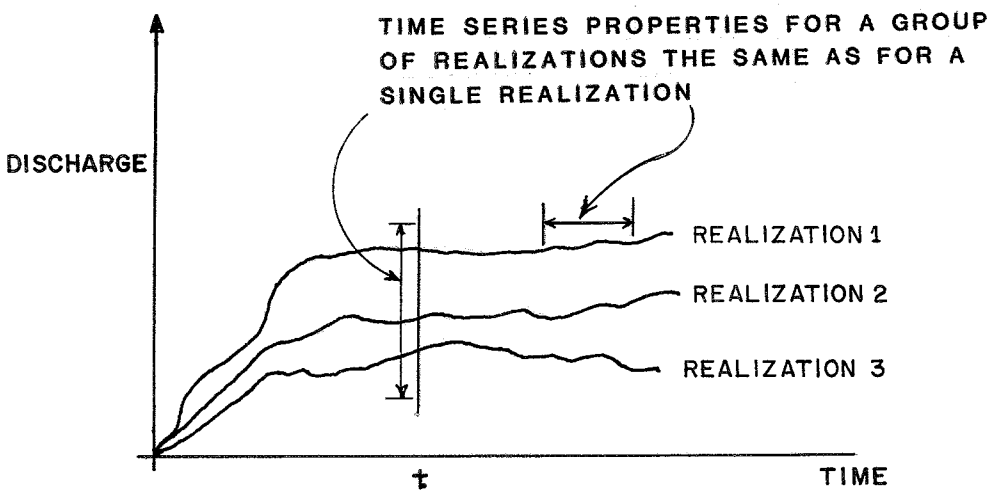


Figure 2.8 ERGODIC PROCESSES

required for all moments of the probability distribution, but as in the case of stationarity, the requirement is only for the first three moments.

Given the stationary and ergodic assumptions, the problem of inferring the probability model for a time series has been reduced to inferring a PDF with constant moments. The stationary and ergodic assumptions may appear severe. However, given the current state of knowledge, they are necessary simplifying assumptions.

2.4.3 Probability Models for Dependent Random Variables

In Section 2.3, methods were given for modeling independent random variables. However, to more completely describe a random process, the concept of dependence must be incorporated into models of random variables.

Two variables X and Y are dependent if the likelihood of X taking on values within a certain range are dependent on Y taking on values within a certain range, and vice-versa. X and Y may be observations of two completely different variables (streamflow and air temperature) or may relate current and previous observations of the same variables (the current and previous months streamflow volume). The dependence between observations of the same variable is termed serial dependence. In this regard, dependent and independent random variables differ in that if X and Y are independent then the likely values of one variable do not depend on likely values of the other variables. However, the independent and dependent cases are similar in that values of a random variable cannot be predicted prior to their observation and models of random variable are based on observation frequency.

Although the stationary and ergodic assumptions result in a major simplification in modeling random processes, as will be seen in later sections, the modeling of a dependent random variable is still too complicated a problem. The purpose of this section is to briefly review traditional concepts pertaining to dependent random variables (this material may be found in any texts on probability and statistics).

The joint behavior or the relationship between random variables is defined by a joint probability law, functionally expressed by a joint CDF or PDF. In the present discussion, the examination of the joint behavior or dependence of random variables is relevant to both single stream gage analysis and regional analysis. Streamflow observations at a single gage usually demonstrate serial dependence, e.g., there is a relationship between current and previously observed flows. In this case, the joint behavior of interest is between different observations of the same random variable.

The purpose of regional analysis is to establish a relationship between observations at different stream gages as well as the serial dependence at each gage. These relationships are then used to improve estimates of model parameters. The relationships needed are the joint behavior of streamflows at a number of gages, modeled as the joint behavior of a number of random variables.

The probability of joint occurrence of two random variables is given as (Benjamin and Cornell, pg. 91):

$$P\ [X \leq x_i \text{ and } Y \leq y_i] = \int_{-\infty}^{x_i} \int_{-\infty}^{y_i} f_{X,Y}(x,y)dx\ dy \qquad (2.14)$$

where: X,Y = random variable

$f_{X,Y}(x,y)$ = joint probability density function

The probability computation is equivalent to determining the area under a two-dimensional distribution between the limits specified. As in the case of a single random variable the total probability must be equal to one and the area under the PDF must be equal to one.

In some instances, the behavior of X is of interest over certain ranges of the variable Y (or vice-versa). As an example, consider the joint behavior of annual streamflow volume and well pumping rates. Annual streamflow volumes and pumping rates are related by the effect pumping rates have on aquifer levels. Aquifer levels in turn are the primary source of long-term baseflow. The pumping rates may vary randomly during the year in response to varying domestic and industrial demand. However, the pumping rates vary only between certain limits, constrained by the pump capacities. An average probable annual streamflow volume may be of interest over the full range of pumping rates.

This average behavior is determined by integrating the joint PDF over the required range of Y (pumping rates) to determine the probability distribution of X (annual streamflow volumes). The distribution of X calculated in this way is the marginal distribution which determines the behavior of X over the total range of Y (Benjamin and Cornell, pg. 92, 1977):

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \qquad (2.15)$$

where: $f_X(x)$ = marginal PDF of X

$$F_X(x_i) = P\,[X \leq x_i] = \int_{-\infty}^{x_i} f_X(x)dx \qquad (2.16)$$

where: $F_X(x)$ = cumulative marginal distribution X

The conditional distribution determines the probability of X having values between x and x + dx for values of Y < y. The conditional PDF is determined by dividing the joint PDF by the probability that Y < y_o over the total range of X (Benjamin and Cornell, pg. 92, 1977):

$$p = \int_{-\infty}^{\infty} \int_{-\infty}^{y} f_{XY}(x,y')dy'\ dx = \int_{-\infty}^{y} f_{y}(y')dy' \qquad (2.17)$$

$$f_{X|Y}(x,y) = \frac{f(x,y)}{p} \qquad p>0 \qquad (2.18)$$

where: $f_{X|Y}(x,y)$ = conditional PDF of x

$f_{Y}(y)$ = marginal PDF of Y

Division by the probability p is done to renormalize the total conditional probability to one. The conditional probability of X given a value of Y<y is thus:

$$F_{X|Y}(x_i y) = P\ [X \leq x_i \mid Y \leq y] = \int_{-\infty}^{x_i} f_{X|Y}(x,y)dx \qquad (2.19)$$

where: $F_{X|Y}(x,y)$ = conditional CDF

The cumulative, conditional and marginal CDF are the n related using a formula analagous to equation 2.18 as:

$$\text{(conditional)}\quad P[X \leq x \mid Y \leq y] = \frac{P[X \leq x \text{ and } Y \leq y]}{P[Y \leq y]} \quad \begin{matrix}\text{(cumulative)}\\ \text{(marginal)}\end{matrix} \qquad (2.20)$$

Knowledge of the conditional or joint PDF for the flow in a particular stream is of great value in drought analysis. For example, assume that the annual volume of streamflow is dependent on only the past year's streamflow. The probability that two consecutive year flows are less then the truncation level (the demand) is (using equation 2.20):

$$P\,[X_2 < x_o \text{ and } X_1 < x_o] = P\,[X_2 < x_o \mid X_1 < x_o]\; P\,[X_1 < x_o]$$

where: X_2, X_1 = annual flows in consecutive years

x_o = the truncation level

(Note the probability that the first years flow is less than the truncation level is calculated without knowledge of any previous condition in the stream). The means by which this calculation can be carried out in drought analysis are fully detailed in Section 4. The important point is to realize that once the PDF of streamflow process is known, the probability associated with a given drought can be calculated.

2.4.4 Dependence and Linear Regression Analysis

The inference of the joint PDF for two or more variables or the serial dependence for a single variable is an extremely difficult problem. The problem can be appreciated by considering the methodology described in Section 2.4.2 for identifying the PDF of an independent random variable and trying to extend this method to identifying multivariate distributions (i.e., joint distributions of two or more variables). Let's consider an example where paired observations of the random variables X and Y are available. This example is applicable to the univariate problem where X and Y may be observations of the same random variable if serial dependence is being investigated (e.g., X and Y are the current and previous months streamflow) or observations of two differenct random variables as in the case of regional analysis (e.g., X and Y are streamflows at two different gages).

In direct analogy to the method for independent random variables, a two-dimensional histogram is developed for X and Y and compared to a theoretical PDF. The agreement between the observed and theoretical distributions

determines whether or not the theoretical PDF is an appropriate model.

This is obviously a cumbersome procedure to make inferences on the PDF of a random process. Furthermore, if more than two variables are involved, the procedure is too cumbersome to be of practical use.

In practice, this methodology is not employed. Instead linear regression analysis is used to investigate the dependence between random variables. The purpose of this section is to introduce linear regression analysis and to describe its relationship to a particular type of joint PDF, the multivariate normal distribution.

The most common approach (particularly in considering stochastic processes in water resources engineering) is to presume that the relationship between random variables or transformations of the random variable is linear. A transformation is performed because the relationship between variables presents some special problems. To keep this discussion simple, a discussion of transformations is delayed until Section 3. For the remainder of this discussion, the relationship between variables is assumed linear.

A linear relationship between any two deterministic or random variables X and Y is expressed as (see Figure 2.9):

$$y_i = ax_i + b \tag{2.21}$$

where: y_i, x_i = observations of the variables Y and X

a = the slope of the straight line relationship

b = intercept of the straight line relationship

One method of determining the coefficients of the straight line is to require that the squared difference between the left and right hand sides of equation

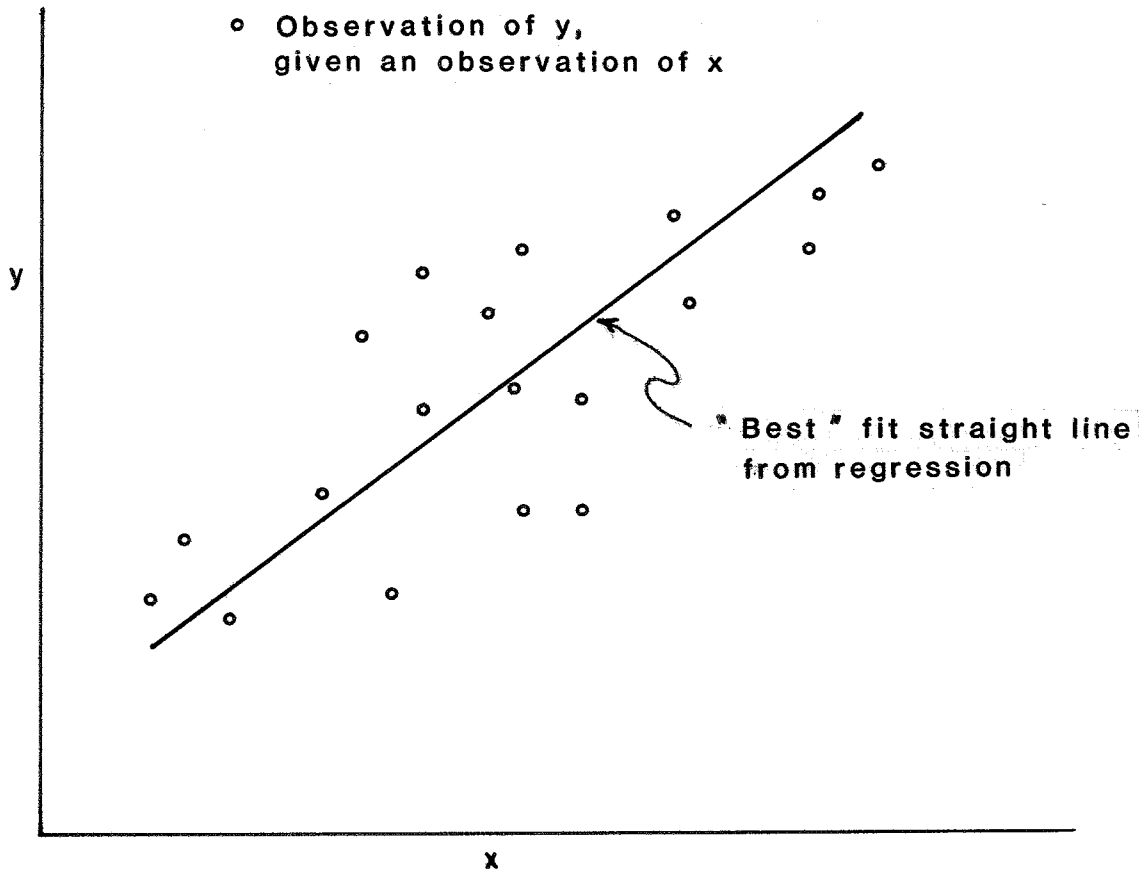


Figure 2.9 LINEAR REGRESSION

(2.21) be a minimum:

$$\min \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \qquad (2.22)$$

where: n = the total number of paired observations of X and Y

R_i^2 = the squared residual

It can be shown (Draper and Smith, 1966) that equation (2.23) results in the following values for the coefficient of the linear regression:

$$a = r \frac{s_Y}{s_X} \qquad (2.23)$$

$$b = m_Y - am_X \qquad (2.24)$$

where: m_Y, m_X = sample mean of Y and X

s_Y, s_X = sample standard deviation of Y and X

The sample correlation coefficient can be calculated by:

$$r = \frac{s_{X,Y}}{s_X s_Y} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - m_X)(y_i - m_Y)}{s_X s_Y} \qquad (2.25)$$

where: r = sample correlation coefficient

$s_{X,Y}$ = sample covariance

It can be shown that the sample correlation coefficient has the interesting property:

$$-1 < r < 1 \qquad (2.26)$$

Note that although the above relationships refer to "sample" estimates, X and Y need not be random variables to apply the equations of Section 2.3.4

The sample correlation coefficient indicates the degree of linear association between variables X and Y. This can be seen by noting that in equation (2.23) the slope of the straight line in Figure 2.9 is directly proportional to the value of r. Consequently, if r = 0 then knowledge of the value of X is of no help in estimating Y. As the value of r increases, estimates of value of Y can be made with greater confidence based on values of X. The correlation coefficient between streamflows should always be greater than or equal to zero. However, sample estimates may in fact be less than zero due to sampling error. Negative values should never be used in an analysis because it is not physically reasonable. Instead, a possible alternative is to substitute a value for r based on analysis of streams in the same region.

Regression analysis takes on added significance if the joint probability density function is multivariate normal. The bivariate normal form for two random variables X and Y:

$$f_{X,Y} = \frac{1}{2\pi\ \sigma_X\ \sigma_Y\ (1 - \rho^2)^{\frac{1}{2}}} \exp\left\{ \frac{-1}{2\ (1 - \rho^2)} \left[\left(\frac{x - \mu_X}{\sigma_X} \right)^2 - 2\rho\ \frac{(x - \mu_X)(y - \mu_Y)}{\sigma_X\ \sigma_Y} + \left(\frac{y - \mu_Y}{\sigma_Y} \right)^2 \right] \right\} \qquad (2.27)$$

where: μ_Y, μ_X = the mean of the marginal distribution of X or Y

σ_Y, σ_X = the variance of the marginal distribution of X or Y

ρ = the correlation coefficient

The correlation coefficient is defined by:

$$\rho = \frac{\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\ f_{X,Y}\ (x,y)\ dx\ dy}{\sigma_X\ \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \qquad (2.28)$$

where: σ_{XY} = the covariance of the variable X and Y

The correlation coefficient is an indicator of the dependence between random variables X and Y. Notice that if ρ = o then the joint normal PDF, equation (2.27) reduces to:

$$f_{X,Y} = \frac{1}{2\pi\ \sigma_X\ \sigma_Y} \exp\left\{ - 1/2 \left(\frac{x - \mu_X}{\sigma_X} \right)^2 \right\}$$

$$\exp\left\{ - 1/2 \left(\frac{y - \mu_Y}{\sigma_Y} \right)^2 \right\} = f_X(x)\ f_Y(y) \qquad (2.29)$$

where: $f_X(x)$, $f_Y(y)$ = PDF for independently distributed random variables X,Y

Thus ρ = o corresponds to the condition that X and Y are independently distributed.

As might be expected from the discussion of regression analysis, there is a relationship between the sample estimates of the regression and the parameters of the bivariate normal distribution. This can be seen by noting two relationships which can be derived from the joint normal PDF:

$$y_i = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x_i - \mu_X) + \sigma_Y \sqrt{1-\rho^2} \times e_i \qquad (2.30)$$

$$x_j = \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y_i - \mu_Y) + \sigma_X \sqrt{1-\rho^2} \times e_j \qquad (2.31)$$

where: e_i, e_j = independently distributed normal random variables, mean zero, variance one

y_i, x_j = joint normally distributed random variables

Taking the conditional expectations of both sides of these equations (noting that $E[e_i] = 0$, $E[e_j] = 0$):

$$E\,[Y|X = x_i] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x_i - \mu_X) \tag{2.32}$$

$$E\,[X|Y = y_i] = \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y_i - \mu_Y) \tag{2.33}$$

which are linear relations between two variables similar to those shown in equation (2.21). The relationship between these equations can be specifically seen by noting the direct correspondence between sample estimates of the parameters in the regression ($\rho, \sigma_Y, \sigma_X, \mu_Y, \mu_X$) and the parameters of the bivariate normal distribution (r, s_X, s_Y, m_X, m_Y). There are many sophisticated tests available in time series analysis to determine if a set of random variables has a joint normal dependency. However, these methods measure dependence in terms of r, the correlation coefficient of a linear regresson. This coefficient is a poor indicator of the general stochastic dependence between two random variables. This can be readily seen by constructing a relationship between Y and X as quadratic and then trying to fit a linear relationship between two variables. In this case, the correlation coefficient certainly would not be equal to one, yet, by construction, there is perfect quadratic dependence between the two variables. Consequently, caution must be used in interpreting the correlation coefficient as a measure of dependence.

In summary, the only possible means of correctly specifying the joint dependence of random variables is by determining the joint PDF. However, inference of the joint PDF in general is cumbersome if not impossible and thus regression analysis is used as the primary technique to investigate dependence, even though it has limitations.

2.4.5 System Memory, Serial Dependence and the Correlogram

As briefly mentioned in Section 2.3, the physical justification for modeling streamflow as a dependent random variables is aquifer storage. To explain this more fully, consider the means by which precipitation results in streamflow. Water from precipitation appears in streams via two major paths, either over the land surface as overland flow or beneath the land surface as base flow. (Note that sometimes the distinction is made between various modes of water travel beneath the surface. For the purpose of this discussion only the distinction between overland and base flow is made). The water particles traveling along these separate paths require different travel times to reach the stream. This difference results in a significant lag between the time that overland and base flow is observed in the stream.

Base flow is the direct result of storage in aquifers. It is this storage and slow release of water by aquifers which allows some streams to continue to flow in extended periods of no precipitation. Thus, current observations of streamflow at a given location are effected not only by the current meteorologic events but also by meteorologic conditions in the past.

The extent of time needed for streamflow to be unaffected by previous meteorologic events is usually referred to as the "system memory" of the stream basin. Both watershed physical characteristics and atmospheric processes are possible factors effecting the magnitude of system memory. Certainly, aquifer storage has a direct effect on system memory. It has also been suggested that long-term memory exists in atmospheric processes. This would then be manifested in streamflow records. The actual atmospheric processes which are responsible for this effect have not been detailed.

Consequently, system memory is an indicator of how great the dependence of a currently observed flow is on the past. When a random variable's value is a function of its own past, the variable is said to exhibit serial dependence. Ideally, this dependence is described by a joint PDF. However, as explained in the previous section, regression analysis is generally used, since it is very difficult to ascertain the joint PDF.

To explore streamflow serial dependence, consider a plot of current streamflow versus a preceeding time period streamflow (the time or integral period may be weeks, months, years, etc.), Figure 2.10. Assuming a linear relationship between concurrent flows the following regression relation can be developed (Jackson and Fiering, pg. 50, 1971).

$$(x_i - \mu_X) = \rho_1(x_{i-1} - \mu_X) \tag{2.34}$$

where: x_i, x_{i-1} = streamflows in the current and preceding years

μ_X = the mean annual flow

ρ_1 = lag one serial correlation coefficient

The term lag one refers to the fact that the observations have been lagged one integral period in the comparison (for example lagged one year).

The memory of streamflow systems is determined by calculating correlation coefficients for increasing lags. A plot of the correlation coefficient versus lag is termed a correlogram (Figure 2.11). Ideally, the correlogram approaches zero as a smoothly decaying function. Correlograms which approach zero at relatively long lags indicate relatively greater system memory.

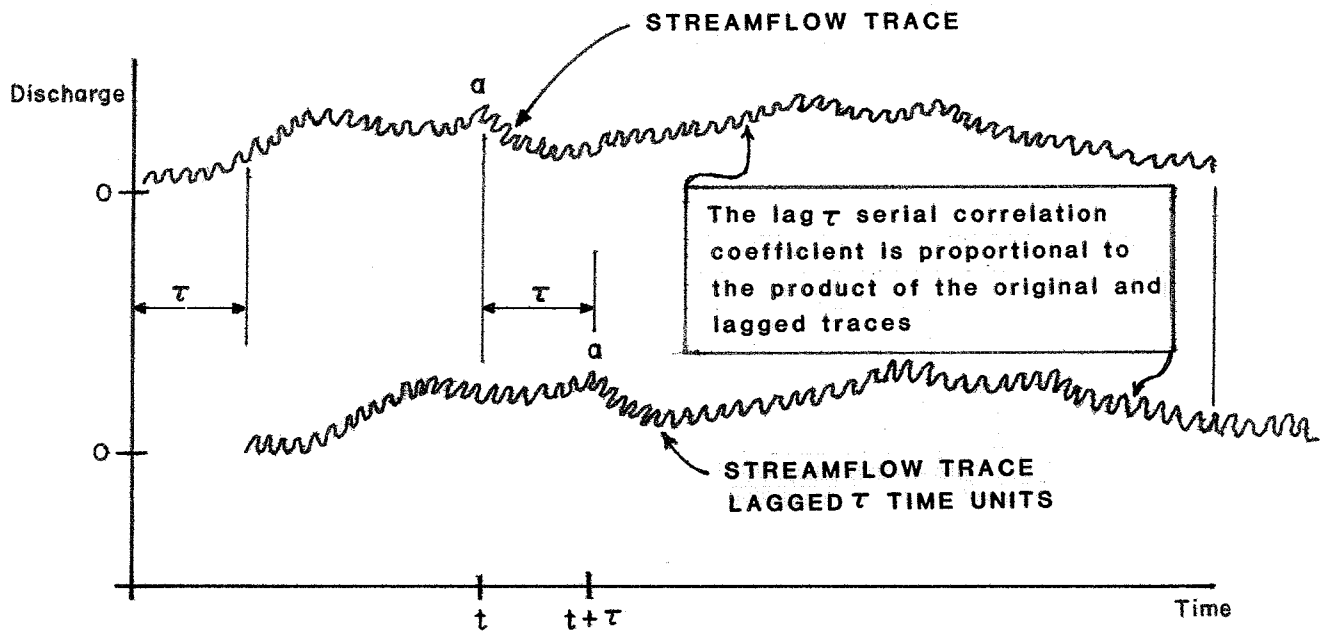


Figure 2.10 STREAMFLOW AUTOCORRELATION

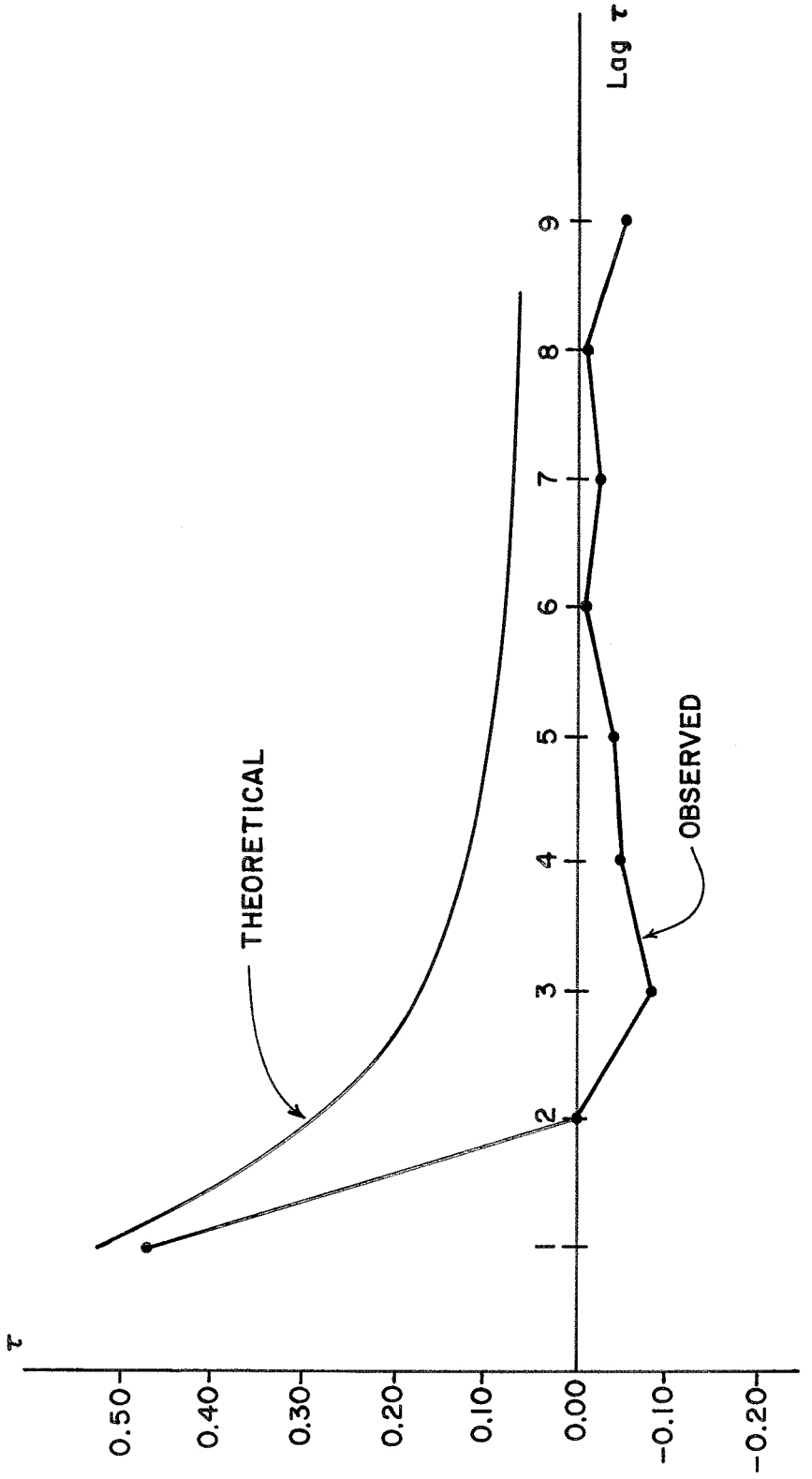


Figure 2.11 CORRELOGRAM

Estimating the number of lags needed to adequately describe the serial dependence in a stochastic model is a difficult and controversial problem. The reader interested in a discussion of the statistical techniques used to identify the appropriate number of lags from correlogram analysis should see Jackson and Fiering, 1971, pg.67 or Salas et. al. 1980, chapter 4. However, caution must be used when these statistical techniques indicate a need for an excessive number of lags (more than <u>two</u>). Caution is needed because there is some question as to whether or not the "persistence" (or long term memory) that is indicated by multiple lag models is justifiable or necessary. A discussion of this controversial point is delayed until Section 5, since the discussion of lag-one models are only necessary for introductory purposes of Section 3.

The serial correlation, can be estimated from observed data using the following formula (Jackson and Fiering, pg. 30, 1971)

$$r_k = \frac{\sum_{i=1}^{N-k} x_i\, x_{i+k} - \frac{1}{N-k}\left(\sum_{i=1}^{N-k} x_i\right)\left(\sum_{i=k+1}^{N} x_i\right)}{\left[\sum_{i=1}^{N-k} x_i^2 - \frac{1}{N-k}\left(\sum_{i=1}^{N-k} x_i\right)^2\right]^{0.5}\left[\sum_{i=k+1}^{N} x_i^2 - \frac{1}{N-k}\left(\sum_{i=k+1}^{N} x_i\right)^2\right]^{0.5}} \qquad (2.35)$$

where: r_k = the sample lag k correlation coefficient

As in the estimation of moments, the sample serial correlation coefficient is a random variable which is subject to sampling variability. The problem of sampling variability increases with increasing lag (see Figure 2.10). Because of this, the observed correlograms are not smooth and may go negative. As the lag increases there are fewer data to cross compare, thus decreasing the accuracy of the estimate. For a further discussion of this point see Jackson and Fiering, pg. 67.

Thus, a method is now available for specifying in general the joint dependence between sucessive observations. Presuming that the process is stationary and multivariate normal, then knowledge of the sample mean, standard deviation and correlogram allows the estimation of the joint PDF. However, as mentioned in previous sections, this type of probability model may not be wholly appropriate for modeling streamflows. The reason for this difficulty is discussed in Sections 3 and 5.

2.5 Regional Analysis

Commonly, streamflow records are either non-existent or of insufficient length at a location of interest. Regional analysis involves the use of streamflow records from nearby gaging stations to either extend existing records or estimate flows where records are non-existent. A major difficulty is to determine which stations to include in the analysis.

This discussion focuses on the use of regional analysis to extend existing streamflow records. For a more general discussion of the regional analysis problem see Haan, pg. 229, 1977.

As discussed in Section 1, the number of stations (or the areal extent) to include in the analysis may be determined by geomorphologic or statistical similarity criteria. Traditionally, geomorphologic similarities considered included watershed physical characteristics such as stream length, stream slope and drainage area. Riggs (1968) points out that low flows are more generally affected by subsurface characteristics rather than the surface characteristics used in traditional regional analysis. Because subsurface characteristics are difficult to characterize over a wide area, regional

analysis based on geomorphologic criteria has not yet proved to be effective. This is an area of ongoing research (Task Committee on Low-Flow Evaluation, 1980).

Extrapolation of Riggs' low-flow analysis conclusions to drought analysis may be somewhat misleading. As pointed out earlier, low-flow analysis implies a shorter integral period than drought analysis. Thus, differences between aquifer characteristics, although important in both low-flow and drought analysis, probably play a more important role in affecting the "instantaneous" measure of low-flow, such as the Q_7^{10} (the seven-day ten-year low flow, see section 1.2), than in the longer term averages of interest in drought analysis. However, Riggs' comments are well worth noting. A general understanding of watershed characteristics, including those altered by man's activities, is of paramount importance in selection of stations to be included in a regional analysis.

An alternative approach is to utilize statistical similarity criteria in a regional analysis. However, statistical similarity criteria have not been proposed in the literature. One might suspect that stations might be selected based on length of record, and on the individual statistics of each record, such as the mean, standard deviation, skew and correlations between stations. Further research in this area is needed.

Given that criteria were available for selecting stations to include in a regional analysis, records at these stations could be used to reconstitute (fill in missing records) at a particular station by employing some type of mathematical interpolation procedure. The method of interpolation most commonly used is linear regression (Draper and Smith, 1966). The technique is

analagous to the method used to determine the serial correlation of an individual streamflow record. Assuming that only a single additional station is used, the relationship between two stations can be expressed.

$$y_i = r_{X,Y} \frac{s_Y}{s_X} x_i \tag{2.37}$$

where: y_i = streamflow value to be reconstituted

x_i = streamflow record chosen in the analysis

i = streamflow period under consideration

s_X, s_Y = sample standard deviations of X and Y

$r_{X,Y}$ = sample cross correlation

The above relationship can be generalized to include more stations in the regression relation, as follows:

$$y_i = b_1 x_{1,i} + b_2 x_{2,i} + \ldots\ldots + b_n x_{n,i} + t_i \tag{2.37}$$

where: $x_{j,i}$ = streamflows at station j (j=1, 2,...,n) during period i

b_j = functions of the cross correlation between stations, determined by regression

n = the number of stations

t_i = random error component

For a discussion of the number of stations to include in the regression see Haan, pg. 230, (1977).

Equation 2.38 can be used to improve estimates of the mean and standard deviation of the record y_i (Matalas and Jacobs (1964), or Fiering (1963)). More importantly, the linear regression for multiple stations is a model that can be used to formulate the conditional probabilities of interest in drought

analysis (a logical extension of Section 2.3.5). The relationships developed to express serial dependence and the regional relationships in streamflow records are used in Section 4 to determine these probabilities.

2.6 Summary

The purpose of this section was to relate probability models currently used in flood and low-flow frequency analysis (models for independent random variables) to probability models used in modeling streamflow for drought analysis (models for stochastic processes). This approach was taken because the hydrologist is familiar with frequency analysis but not familiar with the techniques or language used in modeling stochastic processes.

Independent random variables and stochastic processes were related by considering the common problem of inferring a probability model, the probability density function (PDF), from observation of a random process. As it turned out, the inference of the independent random variables PDF is conceptually straight forward if the moments of the PDF are assumed constant with time. The inference procedure involved comparing the closeness of fit between the observed histogram and the histogram of the proposed PDF.

The only practical drawback to this procedure is that streamflow records are short. Consequently, the accuracy of the probability estimates for rare hydrologic events of interest to the hydrologist are not very reliable (or at least there is a lack of confidence in these estimates). However, the advantage of dealing with independent random variables, is that the mathematics are simple and the identification of the PDF is possible if the number of observations are plentiful.

The same cannot be said for the stochastic models used for drought analysis. Not only is there the lack of data problem inherent in frequency analysis but the mathematics become much more difficult. This of course leads to a double problem in identifying the PDF for a stochastic process.

The inference of the PDF for a stochastic process is done using techniques in time series analysis. In time series analysis, the general description of the stochastic process requires the use of a PDF whose parameters could vary with time. For example, the mean annual streamflow might be modeled as increasing with time. However, this type of model was found to be very difficult to infer from the data and not appropriate for use in predicting likely streamflows for the hydrologists planning horizon based on available data.

The problem was simplified by making the assumption that the stochastic process is stationary and ergodic (at least weakly stationary and ergodic). This assumption presumes that certain moments of the PDF are constant with time. Thus, the stationary and ergodic assumptions are made to simplify the models used in frequency and drought analysis.

The assumptions of stationarity and ergodicity reduce the problem of modeling streamflow as a stochastic process to that of modeling a <u>dependent random</u> variable. However, inference of the PDF for the jointly distributed random variables is still a very difficult problem. The methods used to infer the PDF of an independent random variable might be applied to dependent random variables. However, this methodology for even two variables is cumbersome if not impossible to apply in practice.

The inference problem for dependent random variable is solved by using linear regression to determine the dependence between random variables. In this methodology dependence is indicated by a correlation coefficient. However, the caveat was made that the correlation coefficient indicates the linear dependence between variables. The correlation coefficient is a perfect indicator of dependence only when the random variables are joint normally distributed. In general, the inference of the PDF for dependent random variables is not a problem that is easily solved in practice.

Correlation can be used to model serial dependence and joint dependence between random variables. Serial dependence is used to model the relationship between current and past observation of streamflow. The number of time periods into the past that should be included into the serial dependence between streamflows depends on the stream system memory. The system memory can be deduced from constructing a correlogram which is a plot of the serial correlation coefficient versus lag.

Correlation is also used to model streamflow dependence at different locations or gaging stations in a regional analysis. Unfortunately, regional analysis is probably not as applicable to drought analysis as it has been to flood frequency analysis. The reason for this is that droughts are likely to be more affected by watershed subsurface characteristics which are difficult to ascertain on a regional scale.

In conclusion, the problem of modeling a streamflow process has been equated to the modeling of a dependent random variable. In this section, the normal distribution has been used as the probability model of dependent random variables for example purposes. In subsequent sections, probability models which are more appropriate for modeling streamflow are presented.

Section 3

Autoregressive Model for the Streamflow Process

3.1 Introduction

The purpose of this section is to describe a particular type of stochastic streamflow model, the autoregressive mode. This model was chosen not only because of its simplicity, but also because it demonstrates some of the general difficulties in modeling streamflow as dependent random variable. There are many models that are more sophisticated than the autoregressive model. A discussion of the reason why these models might be more desirable than the autoregressive model is given in Section 5. However, a general description of all the possible stochastic models described in the literature is beyond the scope of this discussion. For those interested in persuing the topic further, see either Salas et. al. (1980) or Kottegoda (1980).

In the previous section, the examples used the normal distribution as the probability model. However, the normal distribution is not generally recognized as being appropriate because histograms of observed streamflow are usually skewed.

Therefore, the goal is to choose a probability model that is more appropriate than the normal distribution for the streamflow process. However, a dilemma is reached at this point because the mathematical tool for inferring dependence between random variables, and thus a probability model, is linear regression analysis. Yet, the inference of the stochastic dependence between random variables for a non-normal joint PDF is, generally, not represented by the linear regression coefficient.

A solution to the problem is to determine the marginal distribution of the random variables (see Section 2.5.4). In other words, given that X and Y are dependent random variables (say streamflow at two different gages) determine the PDF of X over all possible values of Y (i.e. the marginal distribution of X). Once the marginal distributions of all the random variables being modeled are known then transformations of these variables to a group of variables with normal marginal distributions can be performed. After the transformation is accomplished, regression analysis is performed to determine the dependence between the transformed random variables. Thus, the problem of finding a probability model for the streamflow process that has the desired characteristics has been reduced to finding the appropriate marginal distribution for dependent random variables and the correct transformation of these variables to obtain a set of normally distributed variables.

Consequently, there are four steps to be performed in identifying the appropriate PDF for the streamflow process. First, the marginal distribution of the random variables must be found. This is analogous to finding the PDF of an independent random variable. Second, a transformation of the original data to a normally distributed set of data is performed so that the dependence identified by the regression analysis can be related to a joint normally distributed set of data. The third step is to perform a regression analysis of the transformed data. The probability model resulting from the regression analysis of the transformed data is referred to as an autoregressive model. The "auto" descriptor indicates that serial dependence is involved, i.e., regression between lagged observations of a given random variable such as monthly streamflows. The fourth and last step is to perform an inverse transformation to determine the probability model for the streamflow process.

In the following sections, the issues involved in selecting a marginal distribution and the procedure used to construct and employ an autoregressive model are discussed. In addition, a currently available computer model, HEC-4 "Monthly Streamflow Generator" (Corps, 1971) is described and an example application is given.

3.2 Selection of the Marginal Distribution

The inference of the marginal distribution for a stochastic model of annual or monthly streamflows (a dependent random variable) is essentially the same in low-flow or flood frequency analysis (an independent random variable). The inference procedure relies on the acceptance of the fit between the proposed PDF and observed histogram. The criteria for the acceptance of fit is a controversial subject. Because the short length of hydrologic records does not allow confident estimate of rare event probabilities (i.e., estimates of the distribution's tails). Yet, these are often the probabilities of most interest.

For a better understanding of the problem, return to the example of Section 2.3.1, where the observed frequency distribution of the West Branch of the Oswegatchie River was developed. Consider the comparison made between the normal distribution histogram and the observed frequency distribution in Figure 2.3. Note that in the last two intervals between 550 and 650 acre-feet, there are only two observations available. The comparison of the model and observations over this interval is of most interest to the hydrologist. Yet this interval is where the estimates of probabilities/occurrence frequencies are the poorest.

Statistical "goodness of fit" tests can be used to try and make an objective decision regarding the fit of the data. However, there are problems with using such tests. The problems are exemplified by considering the widely employed chi-square goodness of fit test. Basically, this test models the number of observations occurring in any interval of the observed frequency histogram as a random variable. The method requires that the difference between the expected and observed frequency of occurrence be summed for all intervals of the histogram. If the total deviation does not exceed an "expected" deviation given by the chi-square test then the fit of the probability distribution is accepted (for a more thorough explanation see most elementary statistics, texts, for example Benjamin and Cornell, pg. 459, 1970, and Haan, pg. 174, 1977).

The problem with this test is that in the tails of the distribution, the occurrence frequency of a random variable (the number of observations per interval) is being estimated by very little data. Consequently, the computation of the deviation between the expected and observed frequency is not nearly as reliable for the tails as towards the center part of the distribution. Again, the problem is with a lack of data, a problem statistical tests cannot resolve. These conclusions are supported by the comments of Haan (pg. 178, 1977), Fiering and Jackson (pg. 69, 1971), Lane (1979), and Riggs (1968). In particular, consider the comments of Riggs on low-flow frequency analysis (pg. 3),

> "Because particular basin characteristics fix the shape of a frequency curve, no one theoretical distribution is generally applicable and no theoretical frequency distribution will adequately describe the low-flow frequency curve of certain streams,..."

Further, he feels that the effects of sampling error and basin characteristics

on the shape of the frequency curve are much greater than an error made by "hand fitting a curve." Annual and monthly volume frequency curves tend to be smoother than shorter duration frequency curves. However, the implication of these comments would seem to be that statistical criteria used to select among theoretical probability models do not have much advantage over the graphical method described in Section 2.3.2.

Obviously, classical statistics is not going to be of much help in selecting a probability distribution. As an example of possible selection criteria that might be more suited to the water resources engineer's need, consider the study performed by Matalas (1963). Matalas proposed two criteria to fit probability models to low flow data. First, the probability model must predict a low flow at least as severe as the most severe observed low flow while still remaining non-negative (i.e., a lower bound of zero flow is required). The second criteria involved choosing probability models which had an explicit relationship between the skew and kurtosis (the kurtosis is proportional to the fourth central moment of the probability distribution). The criteria then involved calculating the kurtosis by two methods, one by the method of moments, the second by using the theoretical relationship between skew and kurtosis (having calculated the skew by the method of moments). Best fit was based on the consistency of the two calculated values of kurtosis. Based on these criteria, Matalas found that of the four distributions tested, the Gumbel Extreme Value and Pearson Type III were superior to the three parameter Log Normal and Pearson Type V.

The important point to note is that Matalas constrained the fit to estimate a low-flow at least as severe as that estimated from the observed data. The water resources engineer constrained by social or political

requirements may wish to use this type of criteria in the drought analysis.

In Matalas' study, an important point is made concerning the effect of large flows (floods) on the calculation of the probability distribution skew and the occurrence frequency of low-flows. In Figure 3.1, a typical cumulative probability plot of the data in his study and the fitted Pearson Type III and Gumbel distributions are shown. The important aspect of this plot to note is the location of the outlying high-flow point in the data. Matalas points out that this type of point has an extreme effect on the distribution skew and resulting calculation of flow probabilities. In this particular example, if the largest flow is neglected the skew is approximately zero, whereas, inclusion of this point results in a skew of 1.85. If this point is included, the lowest observed flow of 1,070 cfs is underestimated by the Gumbel and Pearson Type III distributions which indicate 1,370 cfs at the same exceedance level. Neglecting this point yields a much better fit of the lower tail of the frequency curve. This is an example of an extreme point which is responsible for rejection of a probability distribution based on selection criteria (see Section 2.4.1 on spurious events). Since the study is concerned with fitting the lowest flow, possibly the extreme point should be disregarded as an outlier.

The above examples in Matalas's study point out the decision problems facing the engineer in selecting a probability distribution. What are the criteria? Should an extreme point be excluded as an outlier? These questions cannot be answered by statistical methods alone. The decision must be made with consideration of the socio-economic realities of the project for which the analysis is being done. For example, the consequences of failure of the project such as a reservoir (failure to meet demand) may cause such a severe

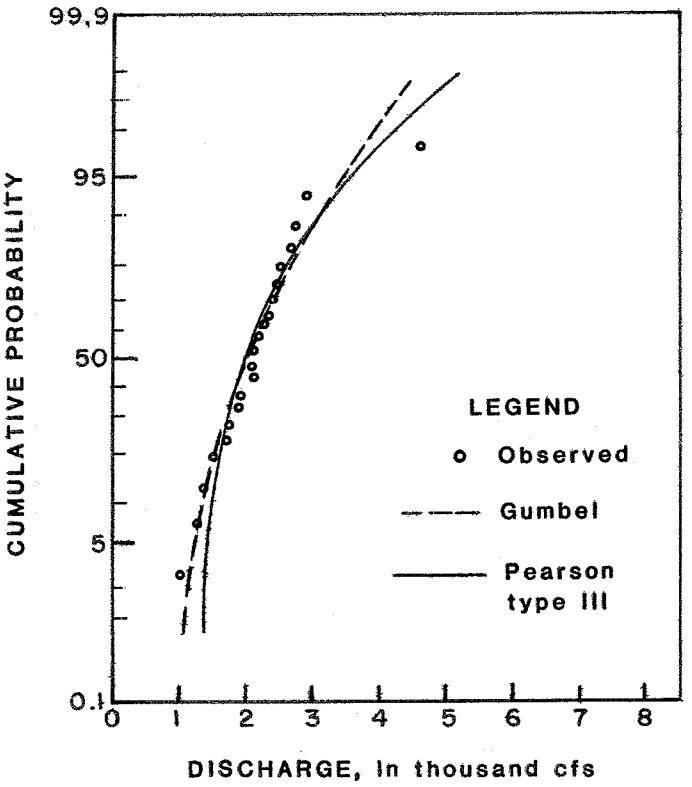


Figure 3.1 EFFECT OF EXTREME POINTS ON SKEWS

(Reproduced from Matalas, 1963)

economic hardship, that the engineer is constrained to meet the most severe drought on record. Consequently, the probability model is constrained to predict as severe a drought as observed in the record, analagous to the criteria which Matalas used in low-flow analysis.

In summary, there are significant difficulties involved in choosing the "best" marginal PDF out of the infinite number possible. The rejection or acceptance of any distribution, including the normal distribution, always will involve some uncertainty.

3.3 Autoregressive Model Formulation for Annual Flows

The mathematical technique for performing a regression analysis has been described in Section 2.4.5. However, it has only been described as a mathematical technique for incorporating dependence in a probability model. In this section, the regression model is developed based on an operational view of a stochastic process. This should lead to better understanding of the autoregressive model.

In Section 2, a stochastic process was defined as a purely random process. However, an operational definition is that a stochastic process is composed of a deterministic component and a probabilistic component. Fiering and Jackson (1969) expressed these components mathematically as:

$$(q_i - \bar{q}) = d_i + t_i \tag{3.1}$$

where: q_i = value of the stochastic process at the ith time step (or integral period) from an initial known state

$\bar{q}$ = mean value of q_i

d_i = deterministic component

t_i = probabilistic component

A proposed form of the deterministic portion of the model is based on the concept of system memory or equivalently serial correlation. Assume that the process is stationary and ergodic and that there is a linear dependence between the current annual flow, q_i and the previously observed flow, q_{i-1}. From equation (3.1), d_i is expressed as:

$$d_i = \rho_1 (q_{i-1} - \bar{q}) \tag{3.2}$$

where: ρ_1 = lag one correlation coefficient

A more general form considers multiple lags:

$$q_i = b_o + b_1 q_{i-1} + \ldots + b_n q_{i-n} \tag{3.3}$$

where: q_{i-n} = the annual flow lagged n years

b = coefficients of linear regression (functions of the lagged serial correlation coefficient)

Practically speaking, the number of lags is difficult to determine because of sampling errors in estimating ρ. For the introductory purposes of this discussion, only lag one models are considered. Those interested in a further discussion of multi-lag models should see Fiering and Jackson, page 67, 1971 or Salas et. al. 1980, Chapter 4.

The probabilistic component t_i, represents the randomness present in the annual streamflows, e.g., the portion of the process that cannot be explained with the adopted deterministic component. t_i is chosen to preserve the underlying probability model of q_i. For example, assume q_i has a normal distribution. t_i is chosen such that the mean and standard deviation of q_i

is preserved. There are two mathematical constraints used to determine t_i. Requiring the mean to be preserved gives the following results.

$$E[q_i - \bar{q}] = E[\rho_1(q_{i-1} - \bar{q}) + t_i] \tag{3.4}$$

where: $E[\cdot]$ = expected or mean value of the term within the brackets

Since the expected value of q_i and q_{i-1} is equal to $\bar{q}$ then it follows:

$$E[t_i] = 0$$

Consequently if we assume that the t_i are normally distributed with mean zero then the first condition is satisfied. The requirement that the standard deviation of q_i is preserved results in the final form of t_i (see Fiering and Jackson, pg. 50, 1971).

$$t_i = e_i \sigma \sqrt{1 - \rho_1^2} \tag{3.6}$$

where: σ = the standard deviation of q_i

e_i = the normally distributed random variable of unit variance and zero mean

In summary, if the annual streamflow records are lag one linear autoregressive and are normally distributed, then sample estimates of the mean, m, standard deviation s, and the lag one serial correlation coefficient ρ_1, are calculated from the streamflow record using formula given in Section 2. The lag one linear autoregressive model becomes:

$$(q_i - \bar{q}) = r_1(q_i - \bar{q}) + e_i s \sqrt{1 - r_1^2} \tag{3.7}$$

which is equivalent to the relations given in Section 2.4.5 for joint normally distributed random variables. This can be seen by examining, for example, equation 2.31. In the case of serial dependence, $x_i = q_{i-1}$, $y_i = q_i$, and

$r_{X,Y} = r_1$. Thus if sample estimates r_1 and s are used for ρ_1 and σ, then upon substitution, equation (2.31) is equivalent to equation (3.7). Note the equivalence between the marginal bivariate normal distributions of equations (2.30) and 2.31) and the autoregressive model containing a normally distributed error term, equation (3.7).

The critical point to this development is embodied in the steps taken to derive the form for the error component t_i through equations (3.5) and (3.6). Due to the form of t_i, the sample statistics of q_i, the mean, standard deviation and lag one serial correlation coefficient are preserved by the autoregressive model. This property gives the model validity from a statistical point of view.

The difficulty in applying the autoregressive model is that the appropriate form for the error term, t_i, is not easy to derive in the case that the marginal distribution of q_i is skewed. This is why a transformation is applied to observed streamflows to obtain a set of data that has a marginal normal sample distribution. The autoregressive model is then utilized for the transformed data. In the next section, the means by which the autoregressive models can be used in conjunction with the appropriate distribution are discussed.

3.4 Monte Carlo Simulation

3.4.1 Methodology

The autoregressive model has been proposed for the streamflow process. The calculation of drought probabilities with this model require that the conditional probabilities given by equation (2.21) be determined. Although it

is possible to do this explicitly using equation (2.20) under special circumstances (an example is given in Section 4.3), in general it is not possible for an autoregressive model that involves multiple lags at a number of sites. However, the conditional probabilities of interest can be derived from the multi-lag autoregressive model using Monte Carlo simulation.

A Monte Carlo simulation is a method of sampling the values of a function at random. In this particular case, the interest is in sampling at random the values of a PDF. Thus, the Monte Carlo simulation can be viewed as creating observations of the random variable by artifically sampling the random variable's PDF.

The advantage of this method is that the integration shown in equation (2.20) does not have to be performed to calculate the drought probabilities. Instead, the artificial observations of the random variable can be used to create a sample frequency distribution as described in Section 2. The frequency distribution generated is then used to estimate occurrence probabilities. The probability estimates can be made as accurate as needed since as many observations as needed can be generated by the simulation.

The key to the simulation is to be able to generate random samples of the probability distribution. Random sampling is done by generating random numbers. There are many examples of random number generators that are encountered in games of chance. The rolling of a die generates random numbers from one to six. The roulette wheel spins out random numbers. A naturally occurring random number generator is the number of emissions of particles by radioactive substances in a given period.

Since the number of computations involved is large, practical considerations require the use of the computer which has standard routines available for generating "pseudo" random numbers. These routines cannot be used to generate true random numbers because the computer's finite memory dictates that the sequence of random numbers will eventually have to repeat (a good random number generator has a long period before repetition). However, the pseudo random numbers generated are generally considered to be adequate for practical purposes.

As an example, consider the simulation of the lag-one auto regressive process (reproduced from Fiering and Jackson (1971)), Table 3.1. In this process, random standard normal deviates are used to produce artificial or synthetic samples of annual flows. The procedure is simple, first the sample statistics are calculated from the observed data (in this case the flows are assumed to have a normal marginal distribution). The simulation is begun by assuming an initial flow value. A random normal deviate is generated and combined with the initial estimate to produce a streamflow value for the next year. This procedure is repeated to successively obtain flows. The simulation is terminated based on the accuracy needed for the simulated histogram. Note that the initial estimates for the simulation has little effect if the simulation is of sufficient length. Most often, several of the initial values are discarded to remove the effect of the assumed initial flow.

The synthetically generated flows can then be used to calculate frequency histograms to evaluate the conditional probabilities of equation (2.21). However, the synthetic sequences of flows can also be useful in simulation studies as is discussed in Section 3.6 or in deriving the distributions of certain drought statistics, as is discussed in Section 4.3.2.

Table 3.1

†Example Monte Carlo Simulation

* i	q_i	q_i-m	$r_1(q_i-m)$	$m+r_1(q_i-m)$	e_i	$e_i \cdot s\sqrt{1-r_1^2}$	q_{i+1}
0	588.80	0.00	0.00	588.80	-0.523	-83.60	505.20
1	505.20	-83.60	-31.62	557.18	0.611	97.66	654.85
2	654.85	66.05	24.98	613.78	-0.359	-57.38	556.40
3	556.40	-32.40	-12.26	570.54	-0.393	-62.82	513.73
4	513.73	-75.07	-28.39	560.41	0.084	13.43	573.83
5	573.83	-14.97	-5.66	583.14	-0.931	-148.81	434.33
6	434.33	-154.47	-58.42	530.38	-0.027	-4.32	526.06
7	526.06	-62.74	-23.73	565.07	0.798	127.55	692.65
8	692.63	103.83	39.27	628.07	1.672	267.26	895.32
9	895.32	306.52	115.92	704.72	-1.077	-172.15	532.57

* $q_{i+1} = m+r_1(q_i-m)+e_i \cdot s\sqrt{1-r_i^2}$

m = sample mean = 588.8 (cfs)

s = sample standard deviation = 172.667 (cfs)

r_1= lag-one serial correlation coefficient = .37819

q_i, q_{i+1} = generated flows in ith and i+1th years

† Reproduced from Jackson and Fiering, 1971, pg. 64.

3.4.2 Transformations

As pointed out previously, skewed distributions are likely to be used in streamflow analysis. Unfortunately, these distributions cannot be easily employed to meet the constraints needed to derive a relationship similar to equation (3.7). For example, assume that the observed flows are distributed log-gamma. If q_{i-1} and e_i have this same distribution, then the resulting q_i is not log-gamma distributed (the sum of log-gamma distributed flows is not necessarily distributed log-gamma). Thus, the constraints involved in deriving equation could not be met and the desired conditional probabilities could not be calculated.

These constraints can be fulfilled by creating a data set with a normal distribution from the original data set. This is accomplished by using a mathematical transformation. Equation (3.7) is then employed to generate transformed flows and an inverse transformation is then applied to these results to obtain an untransformed synthetic data set. This data set is then used to derive the conditional probabilities.

The log transform is often used on streamflow data. This transformation takes the form:

$$y_i = \log_e(x_i + b) \tag{3.8}$$

where: y_i = the transformed flow

b = a small percentage of the mean of X_i

The constant is added to the original flows to avoid the logarithm of zero, which is undefined. For example purposes, the log transformation was performed

on data shown in Table 2.1 and plotted on normal probability paper as shown in Figure 3.2. The data falls close to a straight line, indicating that the log of the transformed data is approximately normally distributed. This indicates that the original data is approximately log-normally distributed.

The calculation of the conditional probabilities by this method may only be approximate. For example, the log transformation produces an equation set:

$$(y_i - \mu_y) = (y_{i-1} - \mu_y) + \sigma_y \sqrt{1 - \rho_1^2} \quad e_i \qquad (3.9)$$

where μ_y, σ_y, and ρ_1 are the model parameters for the logrithms of the data.

Obviously, the conditional probabilities preserved by this equation will be for the logarithms of the data, not of the original data.

The accuracy of this approximate method for deriving the conditional probabilities depends on the severity of the transformation (for a further discussion of this point see Salas et. al., 1980, pg. 70). In some cases, the transformation will allow the preservation of the transformed and untransformed data. For example, the subtraction of a constant from the normally distributed data in Section 4.2 introduces no approximation. A general method for determining the severity of the transformation is to compare the model parameters calculated from the observed and generated data. If the mean, standard deviation, skew and serial correlation coefficients are not "significantly" different, then the transformations used and the calculated conditional probabilities are acceptable. Statistical tests are useful in determining if the differences between the observed and generated data are significant (see Haan, pg. 161, 1977). In Section 3.5.3, statistical tests are discussed which are employed by HEC-4 (Corps of Engineers, 1971) for this very purpose.

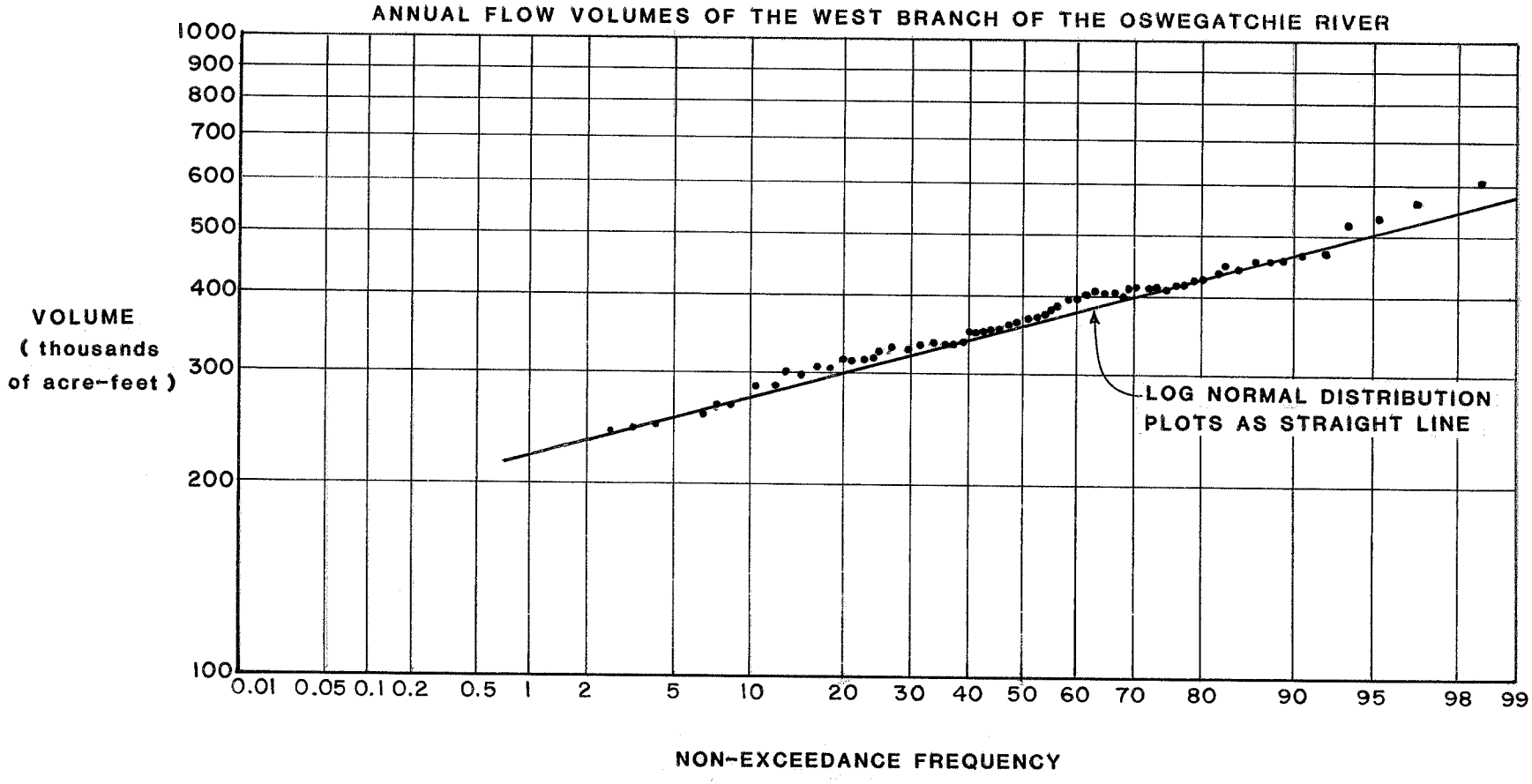


Figure 3.2 FLOW VOLUME FREQUENCIES ON LOG NORMAL PROBABILITY PAPER

3.5 HEC-4 Autoregressive Monthly Streamflow Generator

3.5.1 Basic Methodology

The HEC-4 computer model generates monthly flows at multiple stations in a manner analagous to the simple annual model. HEC-4 assumes that observed monthly flows have a log Pearson III distribution. The Wilson-Hilferty transforma- tion is performed on the observed data to obtain normally distributed variates.

As in the case of the simple annual model, the model is composed of a deterministic and random component (equation 3.1). The deterministic component accounts for the cross correlations between stations and the serial correlations between successive monthly flows (lag one serial correlations). The correlations are determined by a linear regression:

$$K_{i,j} = B_1 K_{i,1} + B_2 K_{i,2} + \ldots + B_{j-1} K_{i,j-1} + B_j K_{i-1,j} + \ldots + B_{j+1} K_{i-1,j+1} + \ldots + B_n K_{i-1,n} \quad (3.10)$$

where: B_i = parameters of the regression

$K_{i,j}$ = transformed flows of the ith month and the jth station

Note that this regression includes lagged serial correlations, regression with $K_{i-1,j}$, cross-correlations, regression with $K_{i,j-1}$, and lagged cross-correlations, regression with $K_{i-1,j+1}$, which is essentially a combination of the factors given in equations (2.36) and 2.37).

The random portion of the stochastic model is given by:

$$t_i = e_{i,j} \sqrt{1 - (R_{i,j})^2} \quad (3.11)$$

where: $(R_{i,j})^2$ = multiple coefficient of determination (see Draper and Smith, 1966)

$e_{i,j}$ = normally distributed random numbers with mean zero and variance one

The random component represents the uncertainty in predicting the values of $K_{i,j}$ with the linear regression.

As a simple example of the use of HEC-4, consider a two station model:

$$K_{i,1} = B_1 K_{i-1,1} + B_2 K_{i-1,2} + e_{i,1}\sqrt{1 - R_{i,1}^2} \qquad (3.12)$$

$$K_{i,1} = B_1' K_{i,1} + B_2' K_{i-1,2} + e_{i,2}\sqrt{1 - R_{i,2}^2} \qquad (3.13)$$

where $K_{i,1}$ is the generated flow at station one and $K_{i,2}$ is the generated flow at station two. Note that since HEC-4 is designed to utilize monthly flow data, there are twelve sets of the B_i and $R_{i,j}$ coefficients for each station j. In this case, the program computes 48 B_i coefficients and 24 $R_{i,j}$ coefficients. The equation is general and could be used to simulate flows for any integral period, such as seasonal or annual. At this time HEC-4 is designed to simulate only monthly flows.

The generation of flow values proceeds similarly to that of the single station annual model described in Section 3.4.1. For example, $K_{1,1,}$ the flow in the first month at the first station (say January), is related to flows one month earlier (December), $K_{12,1}$ and $K_{12,2,}$ and an artificially generated random number, $e_{1,1.}$ Note that $K_{i,1}$ is related to a lagged flow at station two, $K_{i-1,2}$ rather than a concurrent flow, $K_{i,2}$. This is necessary since the concurrent flow $K_{i,2}$ has yet to be generated. $K_{1,1}$ is then used with the lagged flow at station two, $K_{12,2}$ to generate the January flow at station two, $K_{1,2.}$ This is done recursively for all months in as many years as deemed necessary.

3.5.2 Transformation of Historical Data

The HEC-4 model assumes that streamflow statistics can be modeled by a log-gamma (Log-Pearson Type III) distribution. The advantage of this type of distribution is that it allows for a non-zero skew in the observed streamflow. However, log-gamma distributed variables cannot be used in equation (3.10) since the sum of log-gamma distributed variables is not necessarily distributed log-gamma. This fact prevents the preservation of the original data statistics when equation (3.10) is used with log-gamma variates.

HEC-4 overcomes this problem by transforming the original data from an assumed gamma to an approximately normal distribution using the Wilson-Hilferty transformation equation. The transformation steps used by HEC-4 are as follows:

1) Base 10 logarithm transformation

$$X_i = \log_{10}(q_i + b) \tag{3.14}$$

where: q_i = observed flow

b = one percent of $\bar{q}$

b is used to assure a non-zero flow, as the logarithm of zero is undefined.

2) Adjust to zero mean and unit standard deviation

$$Z_i = \frac{X_i - \bar{X}}{s_X} \tag{3.15}$$

where: X_i = logarithm of flows

$\bar{X}$ = mean of the X_i

s_X = sample standard deviation of the X_i

3) Wilson-Hilferty transformation from Pearson III to a normally distributed variate:

$$K_i = 6[(.5g_Z \cdot Z_i + 1)^{1/3} - 1]/g_Z + g_Z/6 \qquad (3.16)$$

where: g_Z = the sample skew of Z_i

The K_i result in an approximately normally distributed variable appropriate for use in evaluating the parameters of equation (3.10). Obtaining synthetic flow values from the generated K_i is a simple matter of applying the inverse of the transformations just described.

3.5.3 Statistical Analysis Performed by HEC-4

HEC-4 provides a number of statistics for evaluating the generated streamflows. As might be expected, information comparing observed, reconstituted and generated flow mean, standard deviation, skew and percentage volumes of yearly flow values are provided. In the latest version, statistical tests are performed to determine if the mean and standard deviations of the generated flows are "significantly" different than the historical (or reconstituted flows). As mentioned previously, the autoregressive generating scheme guarantees that the mean, standard deviation, cross and lag-one correlation coefficients are preserved for the transformed data and not the untransformed (i.e., actual streamflow observations) data. Thus, statistical tests have been provided to check if the transformation used by HEC-4 has resulted in a generated set of untransformed flows which are significantly different than the historic data. If there is a significant difference, then the generated sequence is not representative of the historical trace and the transformation has been too severe.

The statistical tests provided are only strictly applicable to normally

distributed variables. Consequently, these tests only give a rough guideline as to the acceptability of the generated sequence.

To test if q_1, representing the untransformed generated flow-mean, and q_2, representing the untransformed historical and reconstituted flow mean, come from the same normal population the test statistic formed is as follows:

$$z = (\bar{q}_1 - \bar{q}_2)/\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)} \qquad (3.17)$$

where: n_1, n_2 = sample sizes of observed and generated data

σ_1, σ_2 = standard deviations of observed and generated untransformed data

$\bar{q}_1$, $\bar{q}_2$ = mean values of observed and generated untransformed data

z can be shown to be a normally distributed random variate with mean zero and standard deviation one. As an example of the use of the statistic, determine the value of z such that the generated results are rejected at a significance level of 10% (i.e., an error is made in rejecting the generated results in 10% of the model applications). The confidence level is met if (see Figure 3.3):

$$P[z_1 < z < z_2] = 1 - \alpha \qquad (3.18)$$

where: z_1 = normal deviate at $\alpha/2$ significance

z_2 = normal deviate at $1-\alpha/2$ significance level

where α is the percent significance level. As a specific example, in a particular HEC-4 run for the Red River watershed, the maximum value of the z-statistic = 1.62 and $\alpha/2$ = 5.2%, for all months considered. Hence the generation results can be accepted at a 10% significance level when considering means.

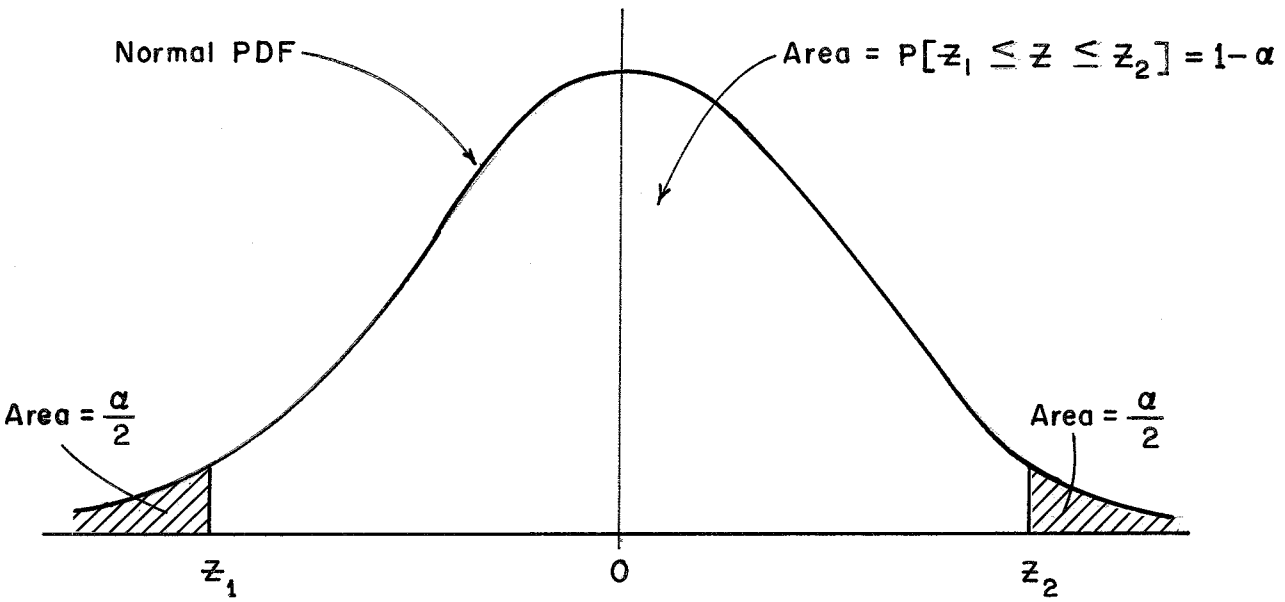


Figure 3.3 SIGNIFICANCE TEST WITH Z-STATISTIC

The generated standard deviations are tested by forming the f statistic as follows:

$$f = s_1^2 / s_2^2 \tag{3.19}$$

f can be shown to follow an F distribution. And as in the case of the z test, a significance level α can be attached to the probability:

$$P[f_1 < f < f_2] = 1 - \alpha/2 \tag{3.20}$$

where: f_1 = f-distributed deviate at $\alpha/2$ significance level

f_2 = f-distributed deviate at $1-\alpha/2$ significance level

s_1 = the larger of the untransformed observed or generated flow's standard deviation based on n_1 degrees of freedom

s_2 = the smaller of the untransformed observed or generated flow's standard deviation based on n_2 degrees of freedom

Based on this information, a judgement as to whether or not to accept the generated results can be made as in using the z test.

3.6 Example Application

For example purposes, HEC-4 was used to determine the probability associated with various reservoir capacities needed to satisfy the water supply needs for a community located on the Arroyo Seco River near Soledad, California. The procedure used to perform this analysis involved the following steps:

1. The historic record of 54 years was used to perform a mass curve analysis (see below for a description of mass curve analysis) based on annual inflows. The mass curve analysis assumes that the communities demand is 59420 acre-feet, one half of the mean annual streamflow.

2. HEC-4 was then used to "generate" one hundred, 50-year sequences of synthetic monthly streamflow. To supply the analysis, the assumption was made that no periodicities or trends were identifiable from the historic record.

3. 50-year sequences, each of the one hundred of monthly streamflows were totaled to produce one hundred, 50-year sequences of annual streamflow.

4. A mass curve analysis of each of the 50 sample sequences was performed to obtain a reservoir capcity for each sequence.

5. A histogram of the reservoir capacities to estimate the probable reservoir capacities.

Mass curve analysis is a well known technique for estimating reservoir capacity, given a period of inflow and demand. A brief description of this technique follows; however, for a more complete description consult Maass et. al. 1962, pg. 120.

Consider the mass curve (sometimes referred to as a Rippl diagram) shown in Figure 3.4. The figure shows the cumulative inflows to the reservoir, the cumulative draft and the cumulative departures from the demand. In this case, the draft was assumed to be one-half the mean flow. The cumulative draft is the draft multiplied by the number of years and the cumulative departure from the demand is the cumulative inflows minus the cumulative demand as shown at point B. The reservoir capacity is determined by first assuming that at the peak of the cumulative departures curve the reservoir is full. As the cumulative departures decrease from this point, there is a draft on the reservoir because the demand exceeds the inflow. Therefore, the total draft

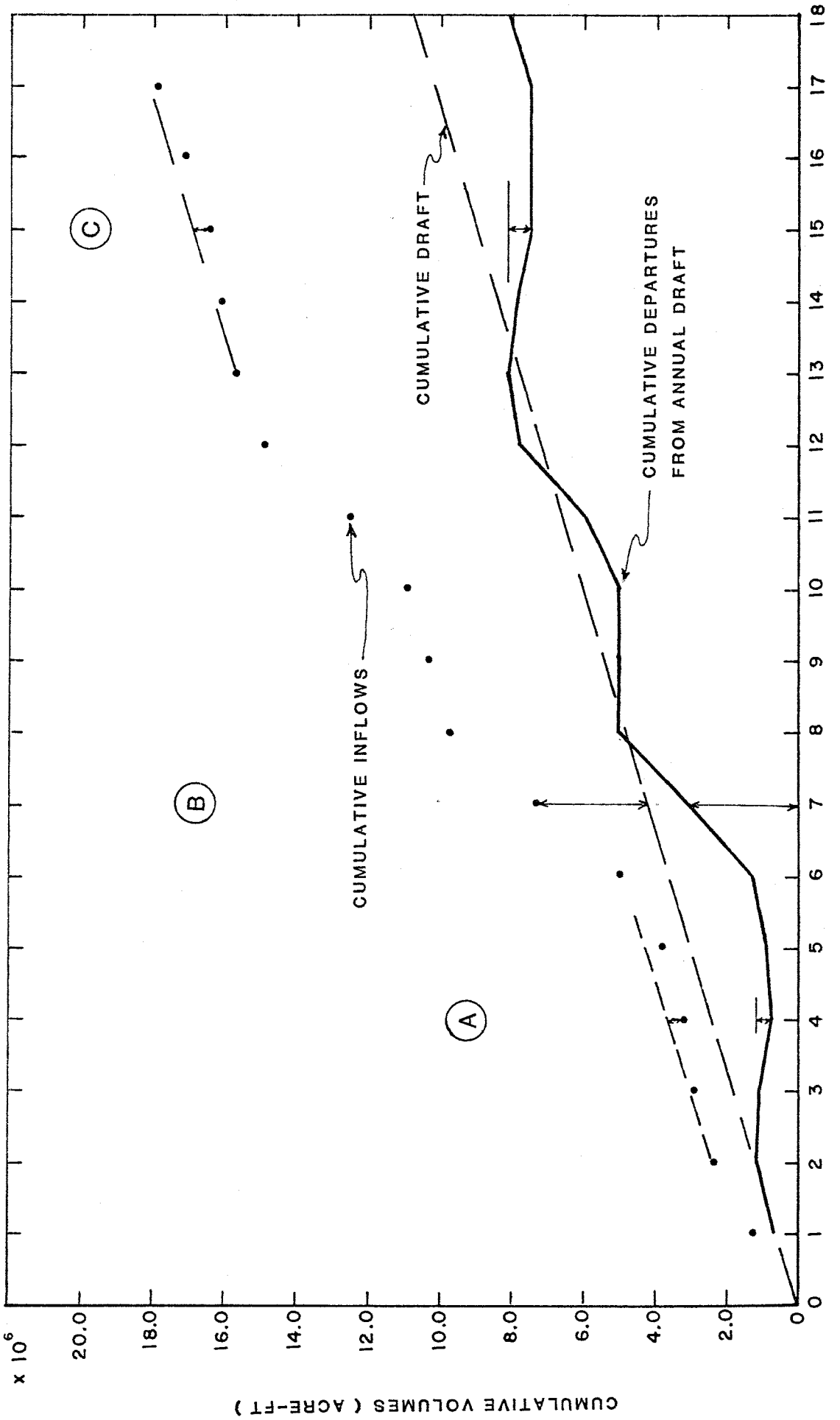


Figure 3.4 Example Mass Curve Analyses of Synthetic Streamflow Sequences of 50 Years

on the reservoir is measured as the distance from the peak to the trough on the cumulative departures curve and the maximum draft over the period is the required size of the reservoir (see points A and C). This is a fairly simple procedure but is complicated by the assumption of the starting storage in the reservoir (see Maass et. al. 1962, pg 120 for a complete explanation of this point). However, the starting condition of the reservoir is of no importance in this example since a comparison is being made between the storage required based on the historic record and the probability of required storages based on the HEC-4 lag-one autoregressive model.

In Figure 3.5, an example of the HEC-4 computer runs made to generate the synthetic inflows to the reservoir is shown. In Figure 3.6, the distribution of the storages as a function of sample size, i.e., the number of 50-year synthetic streamflow sequences, is displayed and compared to reservoir storage of 81,500 acre-feet.

The most striking feature of this result is the significant sampling variability of reservoir storage that results based on a lag-one autoregressive model. Note from Figure 3.6, that approximately 25 percent of the generated 50-year sequences, estimated probability of 0.25, require a reservoir storage that exceeds the storage required by the historic record.

The important conclusion to be drawn from this analysis, and probably one that is not too suprising, is that there is a great deal of uncertainty in trying to predict the future reservoir capacity requirements. Of course, the above results are a function of the probability model (the lag-one autoregressive model) used and the historic sample (in this case 54 years of streamflows for the Arroyo Seco River). However, irrespective of the probability model

Figure 3.5: Example HEC-4 Output

Job specification data

```
WORKSHOP WATER SUPPLY
ARROYO SECO RIVER
SOLEDAD CALIFORNIA

IYRA IMNTH IANAL MXRCS  NYRG NYMXG NPASS IPCHQ IPCHS  NSTA NCOMB NTNDM NCSTY IGNRL NPROJ IYRPJ MTHPJ LYRPJ
1906    10     1    50    50    50     1     0     0     1     0     0     0     0     0     0     0     0
```

Observed monthly flow data

RECORDED AND RECONSTITUTED FLOWS

STA	YEAR	10	11	12	1	2	3	4	5	6	7	8	9	TOTAL
520	1906	6	23	37	540	301	1366	556	297	180	53	18	7	3384
520	1907	14	33	511	1182	586	1956	544	198	109	46	19	19	5217
520	1908	43	43	111	243	433	199	92	52	24	4	0	0	1244
520	1909	6	14	37	1398	1587	444	334	119	54	24	13	11	4041
520	1910	18	28	216	357	142	362	180	59	22	6	2	3	1395
520	1911	7	15	26	1159	702	2335	310	141	72	29	15	13	4824
520	1912	16	23	34	67	37	173	142	71	34	10	2	2	611
520	1913	2	8	12	100	32	39	28	12	3	0	0	0	236
520	1914	0	31	293	2425	829	380	201	99	46	19	9	9	4341
520	1915	12	13	153	389	1611	565	308	313	111	53	23	19	3570
520	1916	24	29	135	1896	885	641	236	128	73	25	11	12	4095
520	1917	34	43	435	131	1481	362	130	74	35	9	6	5	2745
520	1918	8	14	23	27	251	670	113	56	20	2	1	24	1209
520	1919	20	63	87	73	482	271	92	31	11	2	0	0	1132
520	1920	1	6	205	27	69	251	225	62	17	7	1	0	871
520	1921	6	44	104	653	272	202	62	32	16	4	0	0	1395
520	1922	3	6	528	172	1840	436	195	85	34	13	6	4	3322
520	1923	10	283	395	297	199	86	691	106	29	8	3	3	2110
520	1924	5	8	12	59	26	79	75	8	1	0	0	0	273
520	1925	0	59	44	95	301	83	158	125	32	7	1	2	907
520	1926	7	10	15	65	1400	175	795	79	16	3	0	0	2565
520	1927	4	650	28	59	1438	331	282	102	43	17	2	2	2958
520	1928	5	40	64	80	319	415	250	49	16	3	1	0	1242
520	1929	1	20	120	39	315	244	82	31	17	2	0	0	871
520	1930	0	0	11	125	123	402	64	35	13	1	0	0	774
520	1931	0	8	10	89	36	37	15	5	1	0	0	0	201
520	1932	0	1	769	355	885	154	69	39	17	3	0	0	2292
520	1933	0	2	13	129	70	51	31	21	7	1	0	0	325
520	1934	0	2	292	398	328	193	57	26	17	2	0	0	1315
520	1935	1	33	28	373	83	223	618	112	37	12	5	4	1529
520	1936	4	9	23	190	1213	228	251	73	41	14	2	1	2049
520	1937	2	7	50	96	1045	862	297	103	46	13	3	2	2526
520	1938	3	9	422	153	2611	1770	304	141	65	29	12	10	5529
520	1939	13	20	31	56	104	121	38	17	4	0	0	0	404
520	1940	0	1	15	796	1198	653	298	97	45	12	5	4	3124
520	1941	8	14	456	754	1780	1543	1364	264	120	54	29	18	6404
520	1942	19	18	385	634	466	347	601	200	92	36	14	10	2822
520	1943	11	63	60	649	326	759	189	81	39	16	5	3	2201
520	1944	6	11	30	690	499	615	116	78	73	13	4	1	2136
520	1945	3	47	82	55	1025	316	174	72	32	7	1	1	1815
520	1946	22	37	626	179	110	112	144	52	20	5	0	0	1307
520	1947	2	146	54	39	124	82	64	21	8	0	0	0	540
520	1948	1	7	12	10	20	85	161	54	19	3	0	0	372
520	1949	0	2	40	52	90	540	90	35	10	1	0	0	860
520	1950	0	8	20	152	415	90	116	37	12	2	0	0	852
520	1951	7	578	365	195	102	117	53	47	16	4	0	0	1484
520	1952	1	27	374	1589	349	676	245	97	51	21	6	3	3439
520	1953	3	16	296	407	87	133	108	87	39	10	2	1	1189
520	1954	2	15	14	70	248	192	136	45	18	2	0	0	742
520	1955	0	14	117	150	111	78	97	82	23	6	0	0	678

Figure 3.5 (continued)

GENERATED FLOWS FOR PERIOD 1

50 years of monthly synthetic flows

STA	YEAR	10	11	12	1	2	3	4	5	6	7	8	9	TOTAL
520	1	14	24	38	40	32	398	252	43	19	7	1	1	869
520	2	4	40	35	135	957	171	99	77	34	14	4	1	1571
520	3	2	1	112	2893	671	188	202	111	54	27	7	11	4279
520	4	41	180	746	660	206	59	30	16	9	1	0	0	1948
520	5	1	368	86	386	387	1119	285	79	53	21	1	1	2787
520	6	10	3	7	58	148	205	142	57	28	8	5	2	673
520	7	2	18	40	110	46	440	295	52	31	5	0	0	1039
520	8	4	3	75	151	815	366	178	92	46	16	2	1	1749
520	9	13	92	89	226	180	221	43	78	42	13	4	1	1002
520	10	5	17	93	203	19	525	407	102	63	16	13	10	1473
520	11	28	27	200	96	735	293	151	31	9	0	0	1	1571
520	12	2	5	150	455	161	111	54	51	25	3	0	0	1017
520	13	3	23	57	85	86	200	158	79	42	10	0	1	744
520	14	3	41	87	59	6	21	41	27	11	1	1	1	299
520	15	6	13	57	147	328	405	185	53	9	1	0	0	1204
520	16	0	24	371	1552	4025	774	790	194	71	30	2	3	7836
520	17	15	71	90	40	165	148	162	175	69	33	23	17	1008
520	18	18	364	72	45	154	228	27	19	3	0	0	0	930
520	19	0	3	649	434	326	138	157	78	53	10	2	0	1850
520	20	7	157	74	67	487	60	70	45	17	3	1	1	989
520	21	9	167	43	1338	120	219	55	11	3	0	0	0	1965
520	22	2	2	9	26	37	39	132	40	16	3	0	0	306
520	23	0	7	6	32	608	293	32	21	8	2	0	0	1009
520	24	1	22	526	411	932	702	545	157	84	50	74	51	3555
520	25	33	166	200	167	570	434	188	182	101	61	27	29	2158
520	26	20	25	52	188	356	1443	237	42	14	2	0	0	2379
520	27	3	3	280	812	267	94	108	21	4	0	0	0	1592
520	28	0	17	159	647	734	146	316	137	48	13	0	1	2218
520	29	1	6	584	145	573	65	322	87	30	6	1	1	1821
520	30	1	7	19	46	335	217	75	15	4	0	0	0	719
520	31	0	11	3	44	44	90	74	15	2	0	0	0	283
520	32	0	1	19	216	272	121	45	22	4	0	0	0	700
520	33	1	14	209	290	192	307	165	97	26	10	1	1	1313
520	34	6	30	273	146	858	577	307	109	52	13	0	1	2372
520	35	4	34	362	889	410	565	121	56	24	6	0	0	2471
520	36	3	13	304	1374	717	132	283	157	58	12	8	10	3071
520	37	5	23	123	70	578	464	240	81	25	4	1	2	1616
520	38	6	19	20	55	250	829	572	251	125	45	19	5	2196
520	39	5	6	6	17	90	390	124	104	43	16	1	1	803
520	40	0	4	6	118	603	758	117	53	14	1	0	0	1674
520	41	0	2	15	25	232	435	273	89	55	10	1	1	1138
520	42	0	8	96	222	2289	283	281	41	21	3	0	0	3244
520	43	0	4	151	508	197	153	31	11	6	1	0	0	1062
520	44	0	4	148	352	195	119	93	68	34	12	2	1	1028
520	45	3	6	76	2268	1142	2173	1361	250	129	79	97	110	7694
520	46	118	36	237	700	2092	1191	533	212	67	13	1	1	5201
520	47	20	22	81	410	1657	649	349	218	91	25	9	1	3532
520	48	1	24	401	558	43	63	53	29	8	1	0	0	1181
520	49	0	54	15	342	1538	831	608	152	58	27	16	25	3666
520	50	39	59	57	103	1628	2001	308	79	38	10	2	2	4326

MAXIMUM VOLUMES FOR PERIOD 1 OF 50 YEARS OF SYNTHETIC FLOWS

STA	10	11	12	1	2	3	4	5	6	7	8	9	1-MO	6-MO	54-MO	AV MO
520	118	368	746	2893	4025	2173	1361	251	129	79	97	110	4025	7708	21138	169

MINIMUM VOLUMES

STA	10	11	12	1	2	3	4	5	6	7	8	9	1-MO	6-MO	54-MO	AV MO
520	0	1	3	17	6	21	27	11	2	0	0	0	0	4	3088	

Figure 3.5 (continued)

Observed monthly flow volumes

MAXIMUM VOLUMES OF RECORDED FLOWS

STA	10	11	12	1	2	3	4	5	6	7	8	9	1-MO	6-MO	54-MO	AV MO
520	43	650	769	2425	2611	2335	1364	313	180	54	29	24	2611	6161	18119	169

MINIMUM VOLUMES

STA	10	11	12	1	2	3	4	5	6	7	8	9	1-MO	6-MO	54-MO	AV MO
520	0	0	10	10	20	37	15	5	1	0	0	0	0	2	2708	

Statistics of transformed monthly flows

FREQUENCY STATISTICS

STA	ITEM	10	11	12	1	2	3	4	5	6	7	8	9
520	MEAN	0.379	1.254	1.897	2.300	2.499	2.446	2.207	1.804	1.400	0.719	-0.024	-0.095
	STD DEV	0.836	0.594	0.563	0.528	0.533	0.430	0.391	0.362	0.443	0.713	0.904	0.898
	SKEW	-0.661	0.216	0.170	0.181	-0.224	0.169	-0.020	-0.575	-0.818	-0.934	0.077	0.181
	INCRMT	0.10	0.52	1.64	3.98	5.80	4.49	2.36	0.85	0.38	0.12	0.10	0.10
	YEARS	50	50	50	50	50	50	50	50	50	50	50	50

Lag-one correlation coefficients for transformed monthly flows

RAW CORRELATION COEFFICIENTS FOR MONTH 10

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.808 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 11

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.543 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 12

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.267 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 1

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.393 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 2

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.416 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 3

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.634 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 4

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.616 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 5

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.848 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 6

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.956 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 7

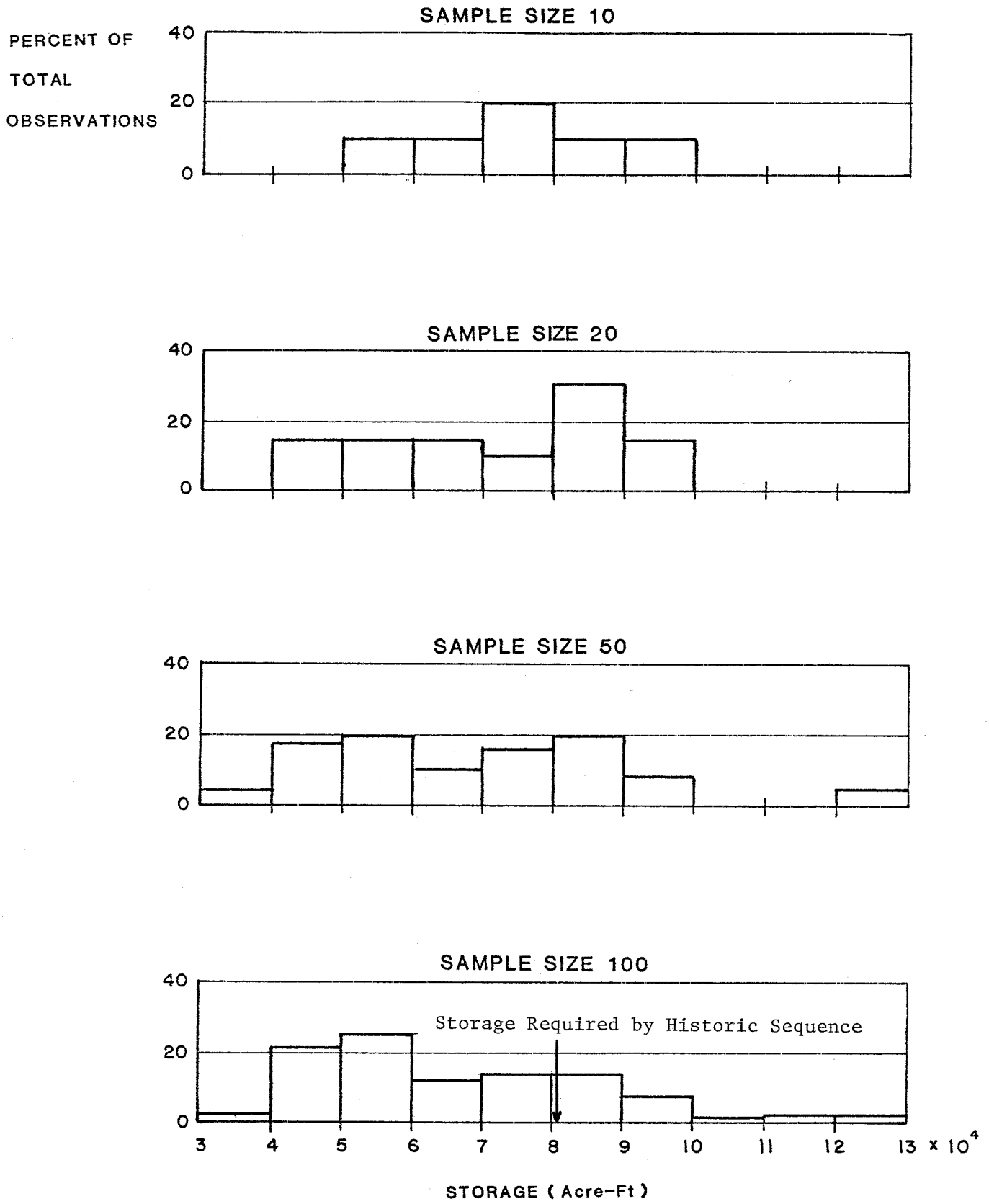
STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.959 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 8

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.872 (WITH PRECEDING MONTH AT ABOVE STATION)

RAW CORRELATION COEFFICIENTS FOR MONTH 9

STA	520
520	1.000 (WITH CURRENT MONTH)
520	0.941 (WITH PRECEDING MONTH AT ABOVE STATION)



FREQUENCY DISTRIBUTION OF RESERVOIR STORAGE

Figure 3.6 Distribution of Reservoir Storage for Synthetic Sequences of 50 Years

chosen, the qualitative results of the stochastic analysis would be the same; the prediction of future reservoir requirements is highly uncertain. Consequently, the reliability of the reservoir design should be insured based on some factor of safety.

3.7 Annual vs. Seasonal Autoregressive Models

The integral period, or equivalently, the computation interval for the drought analysis determines the computation interval for the autoregressive model (remember selection of the integral period is one of the four drought analysis tasks described in Section one). The selection of the integral period depends on two factors. One factor is the type of analysis to be performed. For example, if a reservoir is designed only to satisfy a seasonal demand (the annual summer drought) then a seasonal (say monthly) autoregressive model may be only necessary for the drought analysis. However, if reservoir storage must consider more than a single year drought, an over-year storage problem, then an annual model may be all that is necessary.

A second factor is the need to preserve the statistics of the annual flow series in the generated seasonal flow series. This need arises if the modeller identifies a degree of persistence (a long-term tendency for above or below normal flows to be followed by the same) in the observed annual flow series. Persistence is a very important issue which is discussed in detail in Section five.

This second factor poses a difficulty for the autoregressive model described in this section. Although the seasonal autoregressive model will preserve the statistics of the observed seasonal flows, it will not in general

guarantee that the corresponding observed annual flow statistics are preserved. Consequently, if there is persistance in the annual flow records, then another modeling scheme might be chosen. For example, a technique known as disaggregation can be used to preserve both annual and seasonal flow statistics in a stochastic streamflow model (see Salas et al. 1980, chapter 9). A computer program which currently utilizes disaggregation for autoregressive schemes and is currently available to the public is LAST (1979).

An alternative to using a more sophisticated model, if persistence is not an issue, is to compare the annual statistics (i.e., mean, variance, and lag-one autocorrelation) of the flows generated by a seasonal autoregressive model with the observed historical statistics. If the two are equal, then the seasonal model is probably adequate for the drought analysis. However, disaggregation is the preferred technique because it guarantees, at least approximately, that the observed annual and seasonal statistics are preserved by the generating procedure.

3.8 Simulation and Synthetic Streamflows

To this point, the discussion has focused on calculating the conditional probabilities of drought severity based soley on hydrologic factors. However, the probability of drought occurrence is also a function of the socio-economic factors mentioned in Section 1. Therefore, the probability model should account for these factors to arrive at a proper evaluation of drought probability.

As an example, consider again the problem of determining reservoir storage capacity. The probablity that the reservoir storage will not be sufficent is a

function of the inflow, the demand, the storage capacity (which is to be determined), the reservoir operational policy and a host of other factors. To include all these factors in a single probability statement would seem to be hopeless.

In synthetic hydrology, the problem is solved by utilizing hydrologic simulation models to calculate the probabilities. To perform the simulation a hydrologic model is created which is able to calculate reservoir storage levels based on reservoir design, inflows, demand, operating policy, etc. Inflows to the model are generated from the Monte Carlo simulation of the autoregressive model. The hydrologic model is then used to determine storage levels based on these "synthetic" inflows. An example of how this is done was given in a previous example.

The calculated storage levels are observations of a random variable (a variable which is a function of other random variables, inflow, demand and operating policy). Frequency distributions from the model simulation results could be constructed to determine the conditional probabilities of storage.

In practice, the storage level frequency distribution is not constructed. A more common approach is to generate the synthetic inflows and determine, by simulation, if the proposed reservoir design and operational policy are adequate (e.g., determine if the storage levels meet the required demand).

The important point to remember is that the Monte Carlo simulation produces no new information. The generation of "synthetic" streamflows is a means of calculating a conditional probability. In general, the explicit calculation of these probabilities is impossible when serial and cross

correlations are to be modeled in the streamflow record. The Monte Carlo simulation is the only practical means to evaluate these probabilities when a great many factors are involved.

3.9 Summary

In this section, the means by which a lag-one autoregressive model of the streamflow process is implemented was described in detail. The implementation of this type of probability model can be summarized in four steps:

1. The marginal distribution for the streamflows is selected using techniques described in Section 2.

2. The autoregressive model is only practically applicable to random variables which have a joint normal distribution. Consequently, a transformation is made of the observed streamflow data with underlying marginal distribution selected in step 1 to produce a data set with a marginal normal distribution.

3. The probable values of streamflow that are implied by the autoregressive model cannot be evaluated exactly in most cases of practical interest. Thus, a numerical procedure, Monte Carlo simulation, is performed to evaluate these probabilities. The results of the simulation are in terms of transformed synthetic streamflows.

4. An inverse transformation is performed to obtain synthetic streamflows.

The synthetic streamflows resulting from the autoregressive model may be used

to create streamflow histograms. These histograms can be used to estimate the probable streamflows implied by the autoregressive model. However, the major use of the synthetic streamflows is in simulation studies of water resource systems.

As an example of these studies, the HEC-4 computer model, which uses a monthly autoregressive model, was used to "generate" synthetic streamflows to estimate probable reservoir storage requirements. Since reservoir storage was the random variable of interest, the snythetic sequence of streamflow were used along with mass curve analyses to create histograms of reservoir storage. The histograms of reservoir storage were then used to estimate probable reservoir storage requirements.

Thus, synthetic streamflows are generated based on a probability model, The synthetic flows represent no new information. In fact, if the probabilities of interest could be determined by the direct methods outlined in Section 2, then simulation with synthetic flows would not be necessary. However, since water resource systems are complex, the probability models which necessarily describe these systems are also complex. Thus the direct methods of Section 2 cannot be employed, and the simulation approach which employs synthetic flows must be used.

Section 4

Drought Analysis

4.1 Introduction

In the previous sections the discussion has focused on the random variable streamflow. However, the main interest of this presentation is on the likely occurrence of drought. Recalling the discussion in Section 1, drought is defined by the nature of the water deficit (e.g., streamflow), the truncation level (e.g., the mean annual flow) and the integral period (e.g. one year). The purpose of this section is to combine the concepts used in defining drought with the streamflow probability models discussed in the previous section to derive distributions for statistical parameters of drought occurrence. In Section 4.2 on the theory of runs, statistical parameters for drought are defined, and in Section 4.3 stochastic streamflow models are used to derive the distribution of drought statistics.

4.2 Theory of Runs

In previous sections, streamflow was the process being modeled as a random variable. However, the probabilities associated with various levels of streamflow do not relate directly to drought probabilities. The level of streamflow must be related to a demand for water to assess drought potential. A convenient method to develop this relationship is to formulate a parameter or parameters which are a function of the difference between the level of streamflow and demand. Since streamflow and demand may be modeled as random variables, their difference or parameters of their difference are also random variables. Consequently, probability models of these parameters would give a

more direct measure of drought potential.

The theory of runs is an approach to analyzing a time series, such as streamflow, which develops parameters more suited to directly analyzing drought potential (Yevjevich, 1972). The theory of runs separates a time series into areas above or below a truncation level, x_o (see Figure 4.1). As described in Section 1.1, (see Figure 1.1), the truncation level (from the water resources engineer's point of view) is used to identify drought phenomena in the hydrologic record.

The parameters of interest are the run sum S (the cumulative deviation from x_o), the run intensity M (average deviation from x_o) and the run length D (time between successive crossings of x_o). The parameters may either indicate a positive run (an upcrossing at x_o) or a negative run (a down-crossing at x_o). In drought terminology, S_L is the severity, D_L is the duration and M_L is the intensity (Dracup el. al., 1980). The relationship between the parameters is:

$$S_L = M_L \cdot D_L \tag{4.1}$$

The truncation level x_o can be a constant, a function of time or a stochastic variable. A typical choice in drought analysis is some percentage of the mean flow, which could be considered as an average demand.

Consequently, the objective of drought analysis is to derive the likely values of these statistics. As will be shown, this objective can be met once a stochastic model for streamflows has been developed.

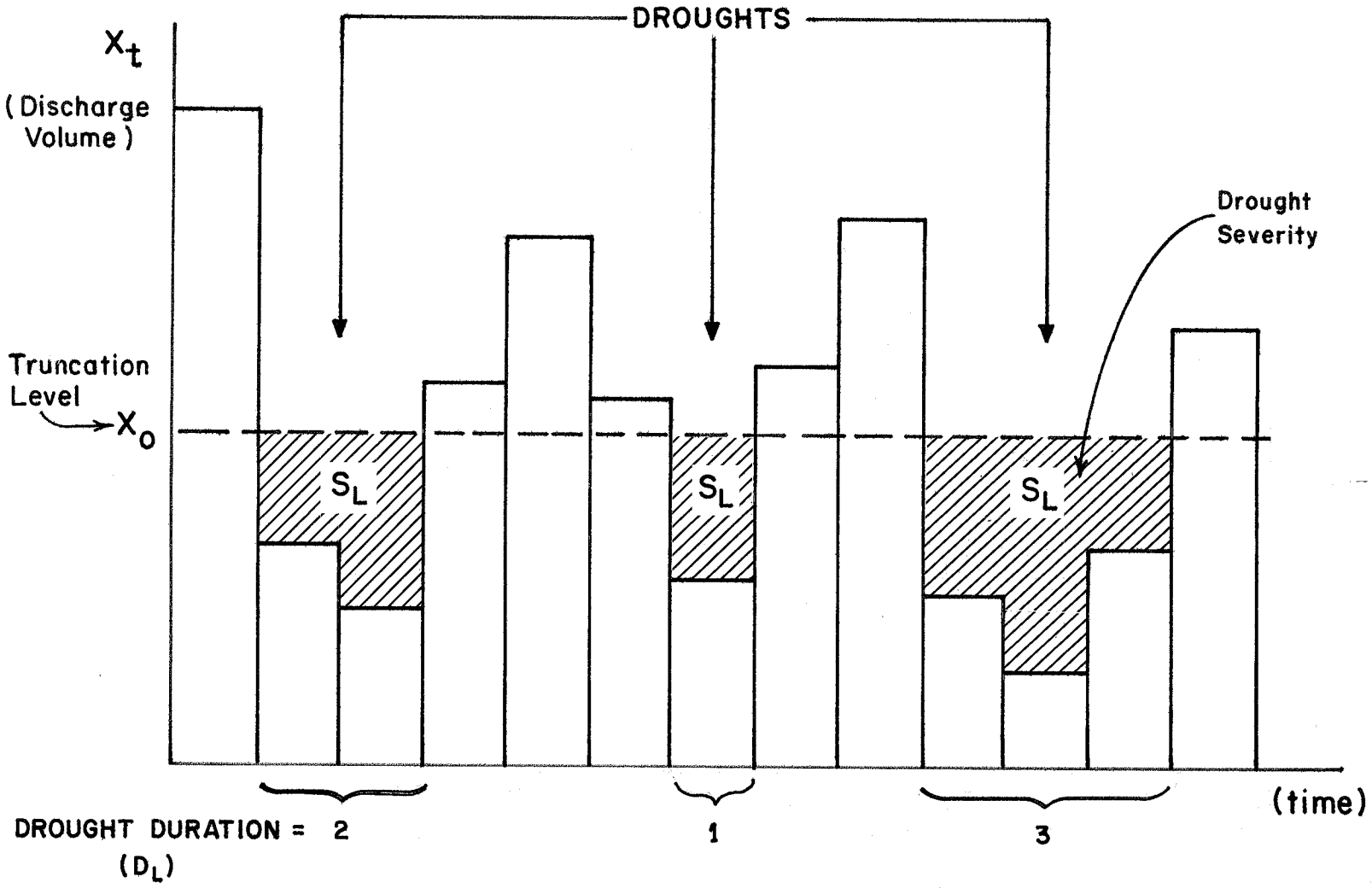


Figure 4.1 RUN PARAMETERS

4.3 Drought Duration Analysis

4.3.1 Exact Calculation of Probable Drought Duration

In this section, an exact expression for probable drought duration is derived for a simple problem by performing the integration of a joint PDF. The expression is "exact" in only the mathematical sense. The integration is performed without using any numerical approximations. This is generally not possible for most practical engineering problems because the PDF is usually not of a directly integrable form. However, the reader should remember that there are sampling errors involved in inferring the PDF from observed flow records. Consequently, any espression for probable drought duration is approximate because of the uncertainty in deriving the PDF.

The calculation of drought duration exceedance probability is equivalent to calculating the probability that successive flows are less than the truncation level. Assume for the moment that annual flows are being considered and no serial dependence exists (e.g., the flows are independently distributed random variables). The probability that N annual flows are less than the truncation level is the product of the probabilities that a flow in any one year is less than the truncation level:

$$P[D_L < N] = P[x_1 < x_o]\ P[x_2 < x_o] \ldots P[x_i < x_o]\ P[x_{i+1} < x_o] \ldots P[x_n < x_o] \tag{4.2}$$

where: x_i = the transformed flow in the ith year

$= (q_i - \bar{q})/\sigma_q$

$\bar{q}$ = mean annual flow (assuming sample equal to population value)

σ_q = standard deviation of flows

Note that the observed flows, q, have been transformed by subtracting the mean and dividing by the standard deviation. This results in a variate, x, with mean zero and variance one, which is convenient for calculation purposes.

The probable drought duration can be calculated by using a method described by Sen (1976). He notes that Feller (1957) developed an equation for the run lengths, positive or negative, in a series of infinite length as:

$$P[D_L \geq N] = P(N) + \sum_{M=1}^{\infty} P(M,N) \qquad (4.3)$$

where: $P[D_L \geq N]$ = the probability that the drought duration exceeds N years

$P(N)$ = probability that N successive streamflow volumes will be less than or equal to the truncation level

$P(M,N)$ = the joint probability that N successive streamflow volumes less than or equal to the truncation level will be followed by M values greater than the truncation level.

Since the flows are independent, the probability that a given flow is less than or greater than a particular value can be obtained by integrating the PDF, as shown in equation (2.3), to obtain the probabilities:

$$P[x \leq x_o] = p \qquad (4.4)$$

$$P[x > x_o] = n = 1-p \qquad (4.5)$$

where x_o, the mean annual flow, has been assumed to be the truncation level. The probability of an independent random variable having N successive values less than x_o is just the individual probability raised to the Nth power. Thus, the two terms in equation (4.3) can be calculated as:

$$P(N) = p^N \qquad (4.6)$$

$$P(M,N) = \sum_{M=1}^{\infty} p^N n^M \tag{4.7}$$

Equation (4.7) is the sum of a geometric series (see Kaplan, 1952, pg. 167) which simplifies to:

$$p^N \sum_{M=1}^{\infty} n^M = [(1/1-n) - 1]p^N = (1/p - 1)p^N \tag{4.8}$$

Substituting equation (4.8) and (4.6) into equation (4.3) the probable drought duration exceedance probability is obtained as:

$$P[D_L \geq N] = p^{N-1} \tag{4.9}$$

The probability that drought duration equals a given value is calculated as (Feller, 1957):

$$P[D_L = N] = P[D_L \geq N] - P[D_L \geq N+1] \tag{4.10}$$

Assuming that the probability of exceeding or being below x_o are equal, p=n=(1-p), then applying equation (4.9) to equation (4.10):

$$P[D_L = N] = p^{N-1} - p^N = p^{N-1}(1-p) = p^N \tag{4.11}$$

p can be calculated once the PDF of the flows is determined.

The above solution is not generally applicable because flow volumes usually demonstrate serial dependence. In this case, the flow's serial dependence is expressed by a joint PDF. The drought duration probability is then found by calculating the terms in equation (4.3) for a dependent random variable:

$$P(N) = \int_{-\infty}^{x_o} \cdots \int_{-\infty}^{x_o} f(x_1, x_2 \cdots x_N)\, dx_1\, dx_2 \cdots dx_N \tag{4.12}$$

where: $f(x_1, x_2 \cdots x_n)$ = joint PDF

P(N) = the probability that N successive flows are less than truncation level for a dependent random variable

As discussed in Section 2.4.5, deriving the joint PDF is a difficult task. This task is simplified by assuming that the flows can be modeled by an annual lag-one autoregressive process. In this case, the probability of a flow being less than the truncation level is dependent only on the previous year's flow, x_{i-1}:

$$P[x_i \leq x_o \mid x_{i-1} \leq x_o] = \int_{-\infty}^{x_o} f(x_i \mid x_{i-1})dx \tag{4.13}$$

where: $f(x_i \mid x_{i-1})$ = conditional probability for annual flows

The joint PDF of equation 4.4 becomes (see equation 2.18):

$$f(x_1,x_2 \ldots x_N) = f(x_1)\, f(x_2 \mid x_1)\, f(x_3 \mid x_2) \ldots f(x_{i+1} \mid x_i) \ldots f(x_N \mid x_{N-1}) \tag{4.14}$$

where: $f(x_i)$ = marginal distribution of the x_i

Substituting for the joint PDF in equation (4.4) P(N) becomes:

$$P(N) = \int_{-\infty}^{x_o} f(x_1)dx_1 \int_{-\infty}^{x_o} f(x_2 \mid x_1)dx_2 \ldots$$

$$\int_{-\infty}^{x_o} f(x_{i+1} \mid x_i)dx_{i+1} \ldots \int_{-\infty}^{x_o} f(x_N \mid x_{N-1})dx_N \tag{4.15}$$

and correspondingly:

$$P(M,N) = \int_{-\infty}^{x_o} f(x_1)dx_1 \ldots \int_{-\infty}^{x_o} f(x_N \mid x_{N-1})dx_N.$$

$$\int_{-\infty}^{x_o} f(x_{N+1} \mid x_N)dx_N \ldots \int_{-\infty}^{x_o} f(x_{N+M} \mid x_{N+M-1})dx_{N+M} \qquad (4.16)$$

where: P(M,N) = the probality that N successive flows less than or equal to the truncation level are followed by M successive flows greater than the truncation level for a dependent random variable

The conditional probability is calculated using equation (2.20):

$$P[x_i \leq x_o \mid x_{i-1} \leq x_o] = \frac{P(x_i \leq x_o,\ x_{i-1} \leq x_o)}{p} \qquad (4.17)$$

where: $P(x_i \leq x_o,\ x_{i-1} \leq x_o)$ = joint probability distribution between two successive annual flows

p = marginal probability that x_i is less than x_o

Sen (1976) calculated drought distribution probability by assuming that the joint probability between two successive streamflows in equation 4.4 is bivariate normal:

$$P(x_i \leq x_o,\ x_{i-1} \leq x_o) = \int_{-\infty}^{x_o}\int_{-\infty}^{x_o} \left(\frac{1}{2\pi(1-\rho_1^2)^{1/2}}\right) \exp\left(-\left(x_i^2 - 2\rho_1 x_i x_{i-1} + x_{i-1}^2\right)/2(1-\rho_1^2)\right)dx_i\, dx_{i-1} \qquad (4.18)$$

To solve equation 4.18 explicitly, Sen assumed that the mean adjusted flow is equal to the truncation level (note the mean of the adjusted flows is equal to zero; see equation 4.2). The drought probabilities were derived by Sen by integrating equation (4.18) exactly; and then using this result with equations (4.15), (4.16) and (4.17) to calculate the terms in equation (4.3) to obtain [equation 28, pg. 1509, ASCE, HY10, October, 1976):

$$P[D_L = N] = (1 - m)m^{N-1} \tag{4.19}$$

where: N = number of drought periods

ρ_1 = lag one serial correlation coefficient

$m = (0.5 + (1/\pi) \arcsin(\rho_1))^{N-1}$

$P[D_L = N]$ = Probability that N consecutive flows are less than the demand

This gives an explicit value for drought duration probability assuming a lag-one autoregressive process (a joint bivariate normal distribution between successive flows).

As an example of the use of this methodology, consider the annual flow volumes of the West Branch of the Oswegatchie River (Table 2.1), assuming that these flows demonstrate a joint normal serial dependence. The lag one serial correlation coefficient is then calculated using equation 2.36. The resulting drought duration probabilities calculated are shown in Table 4.1.

4.3.2 Probable Drought Occurrence Calculated by Monte Carlo Simulation

The purpose of this section is to demonstrate the equivalence between the exact solution given in the previous section and an approximate technique, Monte Carlo simulation, for probable drought duration based on a lag-one autoregressive model. Furthermore, the versatility of the simulation approach is demonstrated by calculating probable drought severity which is not easily obtained exactly.

The probable drought duration, D_L, and, severity, M_L, for an annual

Table 4.1

Comparison of Probable Drought Duration
Obtained by Exact and Monte Carlo Methods

	$P[D_L = N]$		
N (Years)	†Exact	*Monte Carlo	**Monte Carlo
1	.4456	.4532	.4583
2	.2470	.2426	.2500
5	.0421	.0414	0.0
10	.0022	.0019	0.0

† $P[D_L = N]$ = probability of having drought length of N years

$= (1-m)m^{N-1}$

where: $m = 1/2 + (1/\pi) \arcsin (r_1)$
r_1 = Sample Lag-one Serial Correlation Coefficient
= .17 (W. Br. Oswegatchie River)

* Estimate after 100000 simulations

** Estimate after 100 simulations

---

integral period were obtained by generating synthetic streamflows with the lag-one autoregressive model. The assumptions were made that, as in the previous section, the flows are joint normally distributed and that the demand is equal to mean annual streamflow. The sequence of steps taken to calculate these parameters are as follows:

(1) "Generate" synthetic streamflows using Monte Carlo simulation.

(2) Note the durations D_L, and volumes of the synthetic sequence, S_L, that are less than the truncation level (see Figure 4.1).

(3) Construct sample histograms of S_L and D_L for the synthetic sequence. This is done as in the previous examples, where histograms

were constructed (see Section 3.6) by grouping the number of observations of D_L and S_L into selected intervals.

(4) Repeat steps (1) through (3) for synthetic streamflow sequences of increasing length. The histograms for successive sequences are compared to determine if the estimated probabilities for S_L and D_L are approaching a final value.

The above steps were incorporated into a computer program and used to analyze the annual streamflow sequence given for the West Branch of the Oswegatchie River (the same data used as in the previous section). The computer program listing and sample output from the program is displayed in Appendix A. To demonstrate the validity of the synthetic flows obtained by Monte Carlo simulation, the sample mean, standard deviation and skew for the generated sequence of synthetic flows was calculated and is displayed in Table 4.2. Note that as the number of simulations increase, the sample statistics of the generated sequence approach that of the original data set which validates the simulation procedure.

The results of the analysis are summarized by the histograms of D_L and S_L given in Figures 4.2 and 4.3. To compare the exact mathematical solution for D_L given in the previous section and the Monte Carlo simulation, the estimated cumulative probabilities must be calculated from the histogram. This is done by the same procedure performed in Section 2 for flood frequency analysis to calculate sample cumulative probabilities. For example, consider the histogram based on 100 years of simulation for drought duration shown in Appendix A. The estimated probability that a drought is only one year in duration is equal to the total number of simulated droughts of one year in

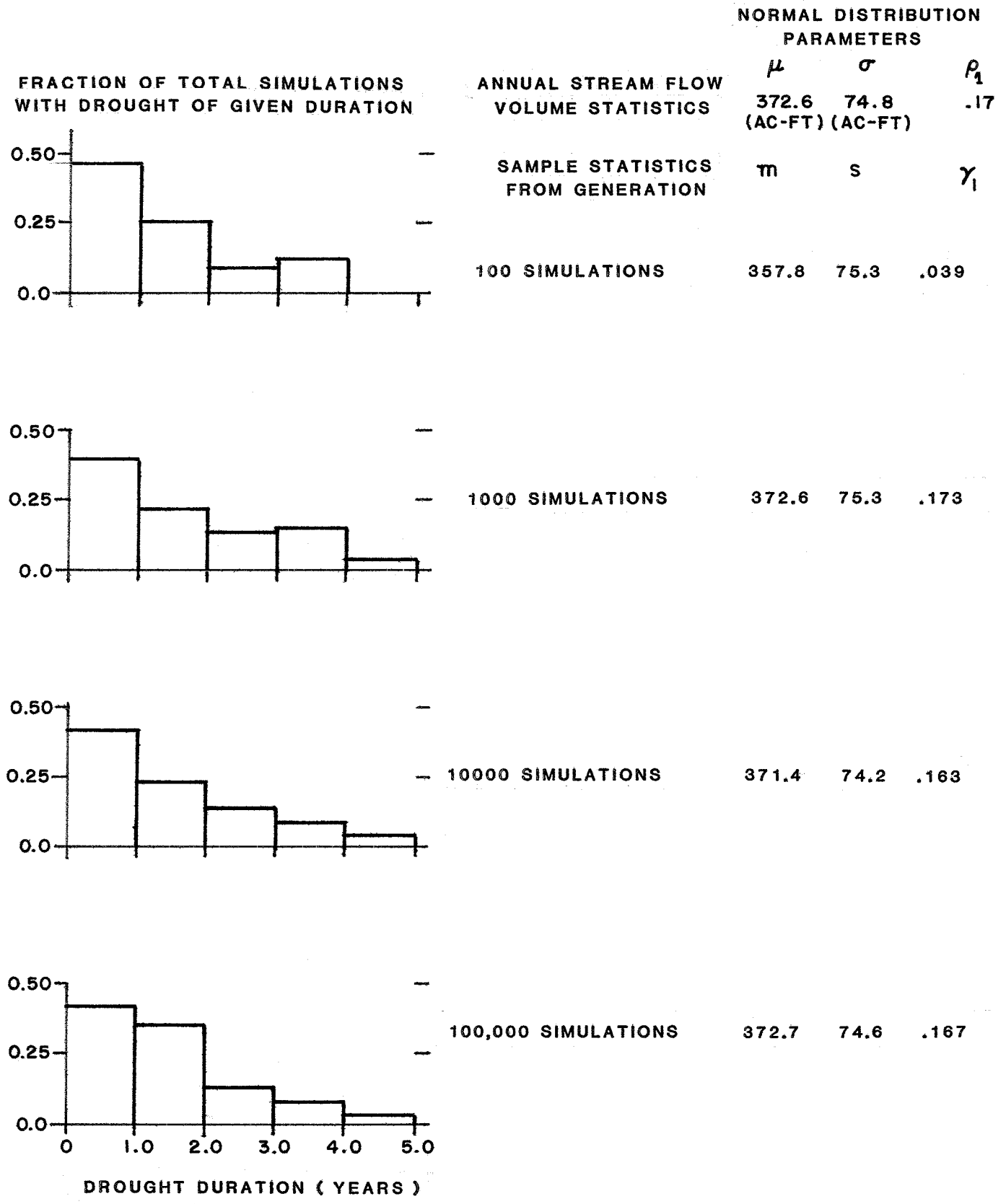
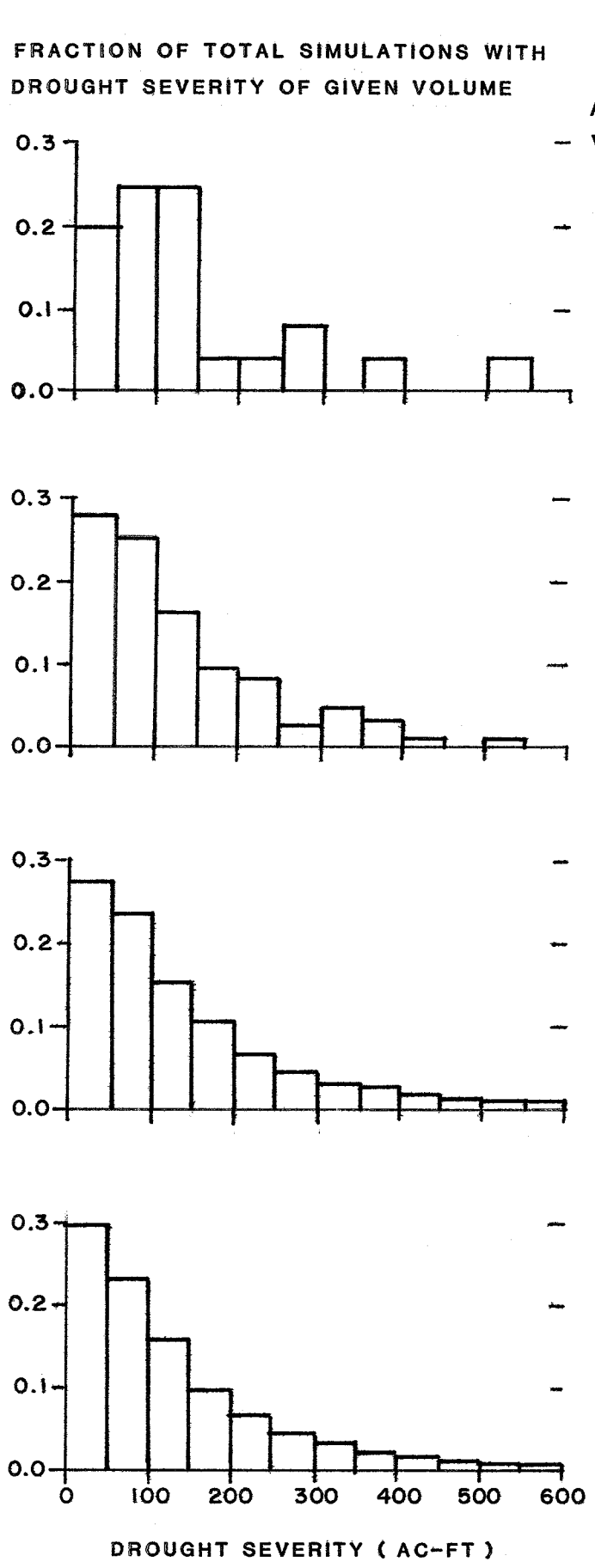


Figure 4.2 DROUGHT DURATION HISTOGRAMS DERIVED BY MONTE CARLO SIMULATION



	NORMAL DISTRIBUTION PARAMETERS		
ANNUAL STREAM FLOW VOLUME STATISTICS	μ	σ	ρ_1
	372.6 (AC-FT)	74.8 (AC-FT)	.17
SAMPLE STATISTICS FROM GENERATION	m	s	γ_1
100 SIMULATIONS	357.8	75.3	.039
1000 SIMULATIONS	372.6	75.3	.173
10000 SIMULATIONS	371.4	74.2	.163
100,000 SIMULATIONS	372.7	74.6	.167

Figure 4.3 DROUGHT SEVERITY HISTOGRAMS DERIVED BY MONTE CARLO SIMULATION

duration (equal to 11) divided by the total number of simulated droughts (24) resulting in an estimated probability $P[D_L = 1] = 11/24 = .4583$. This is shown as the drought duration fraction in Table 4.2. Comparison of the Monte Carlo estimates with the exact solution are shown in Table 4.1 for both 100 and 100000 years of simulation. Note that for the more common drought durations that both simulation lengths compare favorably to the exact solution. However, for the rare durations (5 or 10 years) a large number of simulations are needed to approximate the exact solutions.

Table 4.2

Sample Statistics for Monte Carlo Simulation

	Sample Estimate of Statistics at the End of N Simulations		
N	Mean (ac-ft)	Standard Deviation (ac-ft)	Lag-one Correlation
100	378.8	75.3	.039
1000	372.6	75.3	.173
10000	372.9	74.6	.166
50000	372.7	74.6	.163
100000	372.7	74.6	.170
	Specified Model Parameters		
	372.6	74.8	.170

Inspection of the histograms indicates that the central portions have not changed significantly in proceeding from (10,000) simulations to (100,000) simulations. The corresponding probable drought severity histograms for the maximum number of simulations are shown in Figure 4.3.

As a concluding note, the number of simulations needed to estimate the probabilities near the central portion of a distribution are necessarily less than those needed to estimate the probability at the tails of the distribution. This can be seen by inspecting Figures 4.2 and 4.3 and noting that the tails of the histograms do not converge to a constant value as quickly as the central portion. Again, this demonstrates the problem of trying to estimate rare events, i.e., the tails of the probability distribution. Relatively few observations (or Monte Carlo simulations) are needed to estimate the mean of the distribution (the central portion) but significanly more are needed to estimate the skew.

4.4 Summary

Drought was characterized by three statistics, drought severity, S_L, drought duration, D_L and drought intensity, M_L. The likely values of these statistics can be derived if the truncation level is defined and if a probability model for the streamflow process is properly inferred from observations.

Two methods for deriving the likely values of drought statistics were described. The first method, by Sen (1976), calculates exactly the likely drought durations by integrating a joint conditional probability distribution (i.e., assuming a joint normal distribution) for annual streamflow. The second method utilized Monte Carlo simulation, a numerical technique, to solve the same problem and to derive the likely values of drought severity. The equivalence between the methods was demonstrated.

Again, the fact that Monte Carlo simulation is a numerical technique for integrating complex probability density functions is emphasized. Although

this technique has been referred to in the past as streamflow "generation", the above example clearly demonstrates that the true interest is in evaluating a conditional probability with Monte Carlo simulation. The advantage in using Monte Carlo techniques is apparent when the engineering problem is complex and explicit evaluation of the conditional probabilities is not possible, as explained in Section 3.8.

SECTION 5

Evaluation of the Autoregressive Model

5.1 Introduction

The stochastic streamflow model discussed to this point, for example purposes, is the lag-one autoregressive model. In Section 3 the discussion focused on how to implement this model and the trade-offs in using annual vs. season models. In this section an evaluation is made of the autoregressive model in light of past criticism of the model, and more recently, support of this model.

The reason that doubt has been cast on this model is the observation of long-term persistence in streamflow records. For this reason, many stochastic models have been proposed which better account for this observation.

Consequently, the discussion first focuses on defining persistence and its implications for stochastic models. Secondly, the perceived inadequacy of the autoregressive model due to persistence is described. Finally, an evaluation of the autoregressive model is given, considering more recent research which supports its use in light of the parameter uncertainty issue.

5.2 Persistence

5.2.1 Introduction

Drought is perceived when streamflow volumes remain below some expected level, previously defined as the truncation level (e.g., the mean annual

streamflow). Thus the important question to be answered in modeling drought is given that the previous periods flow is below the truncation level, how likely is it that the current periods flow is going to be below the truncation level? This question can be answered if the serial dependence between successive periods of flow is known, i.e., if the correlogram is known. Remember that the correlogram is a plot of the serial correlation coefficient versus the period lag and indicates the degree of dependence of current flows on past flows.

The number of lags to include in a probability model, or the amount of serial dependence, may seem to be a simple matter based on analysis of the correlogram. To the contrary, this has turned out to be a controversial subject when modeling stochastic processes in hydrology.

The reason for this is the observation of persistence in hydrologic records. Persistence occurs when a great deal of serial dependence is observed in the hydrologic record. This essentially means the correlogram does not approach zero in a "reasonable" number of lags. In this subsection, the discussion focuses on how this phenomenon was discovered, its physical interpretation, and its impact on probability models for drought analysis.

5.2.2 Definition

The observation that hydrologic extremes are preceded or followed by the same is generally referred to as persistance in hydrologic time series, sometimes referred to as the Hurst phenomena (Hurst, 1951). The purpose of this section is to derive a parameter, known as the Hurst coefficient, which indicates the degree of persistence in streamflow. This parameter can then be

used to evaluate drought potential.

As an example, consider the calculation of the Hurst coefficient for the annual streamflow volumes of the West Branch of the Oswegatchie River (Table 2.1) shown in Table 5.1. The calculation is based on analyzing the mass curve (Rippl diagram) of demand and supply in Figure 5.1 (for an alternative approach see Wallis and Matalas, 1970). The mass curve analysis performed here differs from that described in Section 4. In this instance, the analysis determines the reservoir storage needed to supply water under the worst drought conditions and store water during the maximum flood without overtopping the dam. In Section 4, the reservoir storage determined by mass curve analysis only satisfied drought conditions.

To calculate the Hurst coefficient, assume that the demand is equal to the average yearly streamflow, define the cumulative departures from the mean flow or demand as:

$$S_k^* = S_k - k\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = S_k - k\bar{x} \qquad (5.1)$$

where: S_k^* = cumulative departures from mean flow

S_k = cumulative inflow at the end of the kth year

x = mean flow

x_i = observed annual flow

n = number of observation

Define the range R (Figure 5.2) as the difference between the maxium and minimum values of S_k. R then would be the maximum reservoir capacity needed to supply a constant outflow x without overtopping the dam, given the

historical inflow. (Note: in practice, reservoirs are not designed this large. The reference to reservoir storage is a convenient artifice for calculating the Hurst coefficients). Let the adjusted range be defined as:

$$R_n = R/s_n \tag{5.2}$$

where: R_n = the adjusted range

s_n = the standard deviation of the observed flows for a record length of n years

R = range, difference between the maximum and minimum values of S_k

Hurst then defined the relation:

$$R_n = (n/2)^H \tag{5.3}$$

where: H = the Hurst coefficient

The coefficient can be determined by taking the logarithm of equation (5.3):

$$\frac{\log (R_n)}{\log (n/2)} = H \tag{5.4}$$

The results of Table 5.1 give H = .76. The larger the value of the Hurst coefficient the greater the level of persistence in the streamflow record. Typically, the range of the Hurst coefficient is $0.5 \leq H \leq 0.9$. Values of H greater than 0.7 demonstrate a high level of persistence.

5.2.3 Physical Interpretation

The physical interpretation of long-term persistence has caused a great deal of discussion because of its relationship with system memory (see Section 2.4.6). One would reasonably expect that the greater the system memory the higher the level of observed persistence. Consequently, streamflow records which exhibit long-term dependence, correlograms which approach zero at

Table 5.1 Calculation of Hurst Coefficient (Volume in acre-feet)

YEAR	*CUMULATIVE STREAMFLOW	†CUMULATIVE DEPARTURE	MINIMUM DEPARTURE	MAXIMUM DEPARTURE	RANGE	MAXIMUM RANGE
1917	338.1	-34.5	-34.5	0.0	34.5	34.5
1918	730.4	-14.8	-34.5	0.0	34.5	34.5
1919	1136.6	18.8	-34.5	18.8	53.3	53.3
1920	1487.2	-3.2	-34.5	18.8	53.3	53.3
1921	1848.5	-14.5	-34.5	18.8	53.3	53.3
1922	2262.6	27.0	-34.5	27.0	61.5	61.5
1923	2518.2	-90.0	-90.0	27.0	117.0	117.0
1924	2927.6	-53.2	-90.0	27.0	117.0	117.0
1925	3327.9	-25.5	-90.0	27.0	117.0	117.0
1926	3777.5	51.5	-90.0	51.5	141.5	141.5
1927	4125.7	27.1	-90.0	51.5	141.5	141.5
1928	4660.1	188.9	-90.0	188.9	278.9	278.9
1929	5123.4	279.6	-90.0	279.6	369.6	369.6
1930	5576.6	360.2	-90.0	360.2	450.2	450.2
1931	5826.3	237.3	-90.0	360.2	450.2	450.2
1932	6241.6	280.0	-90.0	360.2	450.2	450.2
1933	6596.3	262.1	-90.0	360.2	450.2	450.2
1934	6857.7	150.8	-90.0	360.2	450.2	450.2
1935	7221.1	141.7	-90.0	360.2	450.2	450.2
1936	7547.7	95.7	-90.0	360.2	450.2	450.2
1937	7969.1	144.5	-90.0	360.2	450.2	450.2
1938	8370.2	172.9	-90.0	360.2	450.2	450.2
1939	8683.7	113.8	-90.0	360.2	450.2	450.2
1940	8971.8	29.4	-90.0	360.2	450.2	450.2
1941	9212.9	-102.1	-102.1	360.2	462.3	462.3
1942	9548.8	-138.8	-138.8	360.2	499.0	499.0
1943	9981.1	-79.2	-138.8	360.2	499.0	499.0
1944	10296.9	-135.9	-138.8	360.2	499.0	499.0
1945	10665.4	-140.0	-140.0	360.2	500.2	500.2
1946	11045.5	-132.5	-140.0	360.2	500.2	500.2
1947	11650.0	99.4	-140.0	360.2	500.2	500.2
1948	11991.2	67.9	-140.0	360.2	500.2	500.2
1949	12325.7	29.8	-140.0	360.2	500.2	500.2
1950	12662.3	-6.1	-140.0	360.2	500.2	500.2
1951	13054.7	13.7	-140.0	360.2	500.2	500.2
1952	13361.8	-51.8	-140.0	360.2	500.2	500.2
1953	13687.6	-98.7	-140.0	360.2	500.2	500.2
1954	14129.9	-28.9	-140.0	360.2	500.2	500.2
1955	14536.1	4.6	-140.0	360.2	500.2	500.2
1956	14869.3	-34.7	-140.0	360.2	500.2	500.2
1957	15169.7	-106.9	-140.0	360.2	500.2	500.2
1958	15533.2	-116.1	-140.0	360.2	500.2	500.2
1959	15886.5	-135.4	-140.0	360.2	500.2	500.2
1960	16300.3	-94.1	-140.0	360.2	500.2	500.2
1961	16587.0	-180.1	-180.1	360.2	540.2	540.2
1962	16941.7	-197.9	-197.9	360.2	558.1	558.1
1963	17261.0	-251.3	-251.3	360.2	611.4	611.4
1964	17531.7	-353.1	-353.1	360.2	713.3	713.3
1965	17778.7	-478.8	-478.8	360.2	839.0	839.0
1966	18099.4	-530.6	-530.6	360.2	890.8	890.8
1967	18399.2	-603.5	-603.5	360.2	963.7	963.7
1968	18706.2	-669.0	-669.0	360.2	1029.2	1029.2
1969	19116.0	-631.9	-669.0	360.2	1029.2	1029.2
1970	19426.6	-693.9	-693.9	360.2	1054.1	1054.1
1971	19832.7	-660.3	-693.9	360.2	1054.1	1054.1
1972	20229.1	-636.6	-693.9	360.2	1054.1	1054.1
1973	20680.1	-558.2	-693.9	360.2	1054.1	1054.1
1974	21107.9	-503.0	-693.9	360.2	1054.1	1054.1
1975	21480.7	-502.7	-693.9	360.2	1054.1	1054.1
1976	22044.8	-311.2	-693.9	360.2	1054.1	1054.1
1977	22487.2	-241.5	-693.9	360.2	1054.1	1054.1
1978	22950.5	-150.8	-693.9	360.2	1054.1	1054.1
1979	23361.7	-112.1	-693.9	360.2	1054.1	1054.1
1980	23698.6	-147.9	-693.9	360.2	1054.1	1054.1
1981	24219.1	0.0	-693.9	360.2	1054.1	1054.1

* Cumulative annual streamflow in acre-feet

† Departure from sample mean annual flow

$\bar{x}$=sample mean annual flow=372.6 s=sample standard deviation=74.8

H=Hurst coefficient=log(R)/log(N/2)=log(1054.1)/log(65/2)=.76

where R=Maximum Range, N=Number of Data Years

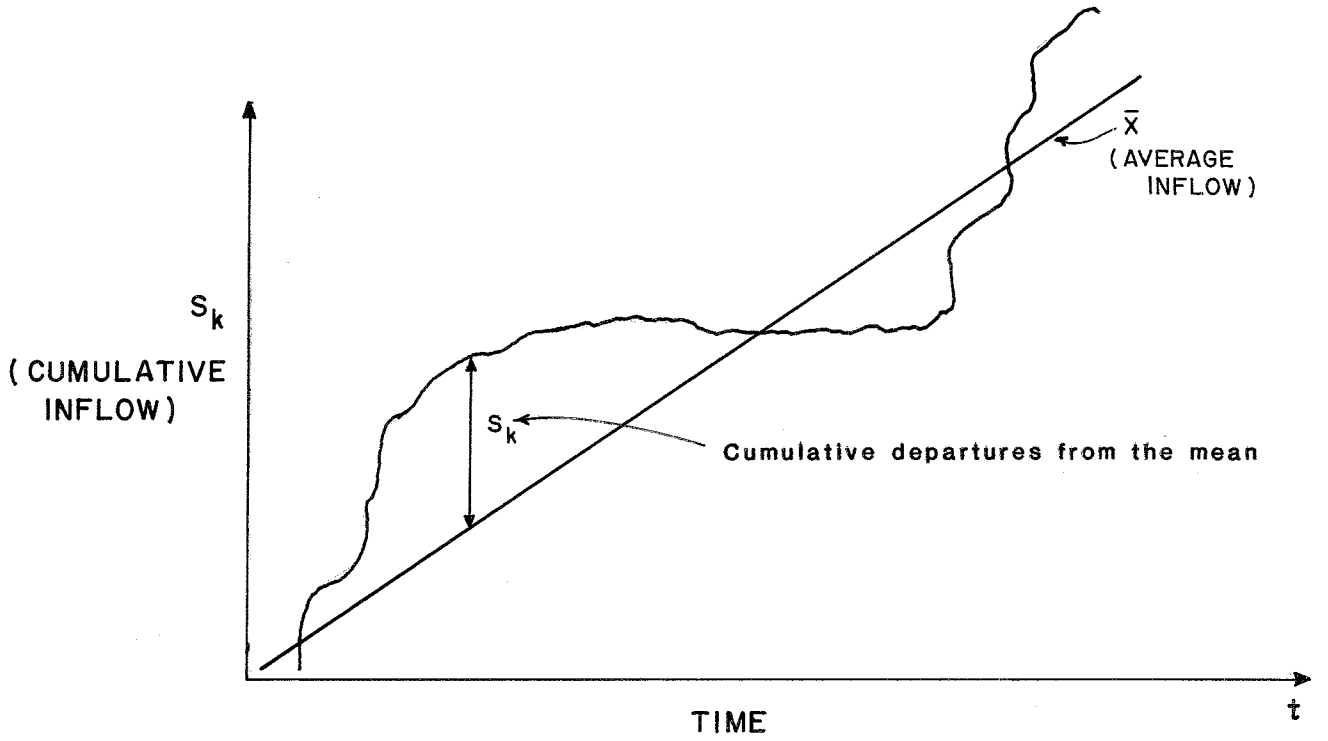


Figure 5.1 RIPPL DIAGRAM

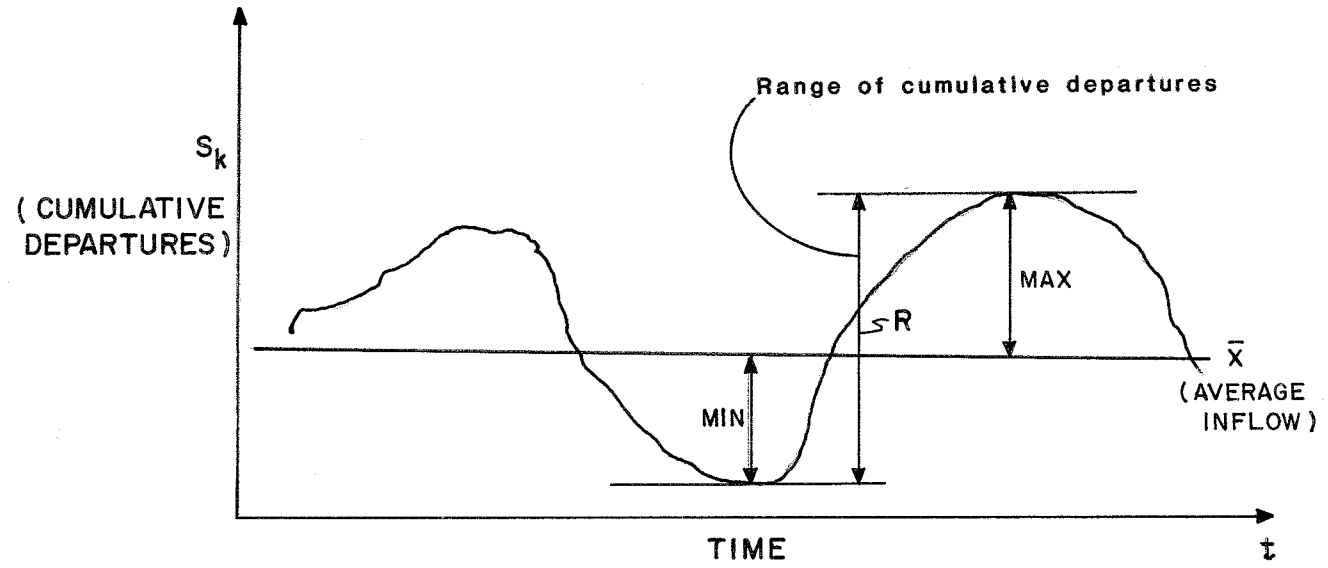


Figure 5.2 CUMULATIVE DEPARTURES FROM THE MEAN

relatively long lags, should demonstrate high levels of persistence. Thus, probabilistic models which use the observed correlogram as a means of calculating the conditional probability of a drought should correctly account for persistance in the streamflow record. Unfortunately, this is not true.

Mathematicians and statisticians have consistently shown that the Hurst coefficient has a value $H = 0.5$ for any stationary process with finite system memory (Klemes, 1974; Haan, 1977) as the number of observations approach infinity. Obviously, this is in direct contradiction of Hurst's findings that $H \geq 0.5$.

Further, Mandlebrot and Van Ness (1968) and Mandebrot and Wallis (1969) produced probabilistic models which produced $H \geq .5$; but required that the system have <u>infinite</u> memory. The implication is that the correlogram diverges even at infinite lag, i.e., the lagged serial correlation coefficient is always greater than zero!

There is a problem because there is no identifiable watershed process which is endowed with infinite memory. However, researchers have suggested that atmospheric processes may exhibit long-term memory. The high level of persistence seen in streamflow records may be due to this atmospheric effect. The actual mechanisms that would account for this phenomena has yet to be identified. Certainly, long-term groundwater storage, which is a physically plausible explanation for serial correlation, does not endow a watershed with infinite memory for past meteorologic events. Consequently, the argument for a process with infinite memory is less than conclusive.

Klemes (1974) clarified the situation by noting that the ability to fit a

model (probabilistic or deterministic) to data only mimics observed phenomena, but does not necessarily explain the nature of the processes which cause them. In the above reference, Klemes proposed a number of alternatives to the infinite memory explanation of the Hurst phenomena. In one of these, the stationarity constraint was relaxed and a system was assumed to have zero memory (no serial dependence) and a mean (average) value that fluctuated with time. Klemes demonstrated this type of model was capable of producing $H \geq 0.5$.

The above argument demonstrates the danger of inferring physical properties of a watershed from statistical analyses alone. The use of H in a statistical analysis of drought is important from an operational viewpoint because it indicates the potential severity of droughts. However, the "H" statistic does not explain the hydrometeorologic process behind the drought.

5.3 Model Comparisons

Criticism of the autoregressive model in the past has been that it is unable to model long-term persistence. The reason for this is that the autoregressive model is a "finite" memory model, that is there are a finite number of lagged correlation coefficients preserved in the model. Referring to the previous section a finite memory implies asymptotically (e.g., as the number of generated flows becomes large) that the Hurst coefficient, $H = 0.5$. Consequently, this model would underpredict the drought potential of a stream which exhibited long-term persistence, $H \geq 0.5$.

The inadequacy of the autoregressive model caused a great number of alternative models to be proposed which could simulate long-term persistence (see Kottegoda, 1980). However, these models present a disadvantage in that

they are much more difficult to use than the autoregressive model, particularly in multistation analysis.

More recently, research has indicated that the autoregressive model is more appropriate for use in water resources analysis in general, and in drought analysis in particular, than had previously been thought in the past. There are two reasons for this change of perspective. First, although the autoregressive model produces Hurst coefficient's, $H = 0.5$, asymptotically, for finite record lengths an $H \geq 0.5$ is produced. This is particularly true when employing an annual model with multiple lags. A seasonal autoregressive model will have greater difficulty in producing large H, unless the corresponding annual statistics are preserved through a disaggregation scheme (see Section 3.7 and Salas et. al, 1980, Chapter 9). Second the efficacy of using high powered mathematical models as alternatives to the autoregressive model is highly suspect when it is difficult to obtain reliable estimates of the model parameters such as the serial correlation coefficients and H.

Bowles et. al. (1980) investigated the range of applicability of the lag two autoregressive model (AR) and four other more advanced stochastic models (autoregressive moving average (ARMA), broken line (BKL), fast fractional Gaussian noise (FFGN) and the ARMA-markov model (AMAK)) based on a range of criteria. The applicability was determined based on a water supply study of four streams in Utah.

The models were evaluated based on the following range of criteria:

(1) Ability to preserve annual persistence statistics (the Hurst coefficient and serial correlation) and the run properties of the

seasonal statistics (such as number of droughts per year).

(2) Cost and ease of model use.

(3) Magnitude of economic regret associated with drought related losses (the economic damage associated with flows generated by each model were compared).

(4) A comparison of reservoir capacity and critical drought design parameters (reservoir design capacities were determined based on specified yield requirements from generated flows of each model. The critical drought parameter evaluated was the drought magnitude, M_L).

Criteria (1), (3) and (4) are related. The expectation is that the models that preserve long-term persistence should also produce the worst droughts and thus the largest economic regret (3) and the most conservative reservoir capacity (4).

The results of the study indicated the following: Criteria (2), the AR(2), ARMA and AMAK were much less costly to run then the FFGN and BKL. Implied in the conclusions is that the parameter estimation for the AR(2) was easiest. Criteria (1) indicated that all models except the AR(2) model are effective in preserving the Hurst coefficient, the ARMA model being most effective. All models seemed to be equally effective in preserving the expected number of droughts per year. Criteria (3) indicated that use of the BKL model minimized the expected agricultural damage. This is an interesting result since the BKL model was judged under criteria (1) to be inferior to the ARMA model in preserving long-term persistence. Based on this result, the

authors conclude that, pg. 59, "the objective of preserving the persistence statistics is not compatible with the objective of minimizing economic regret for the study streams." Criteria (4) indicates that the ARMA and AMAK were the most conservative in estimating the required reservoir capacity, followed by the AR(2) model and then the FFGN and BKL models. All models produced drought magnitudes, M_L, considerably larger than the worst historical drought.

As a result of this study, the authors developed a model choice strategy. The strategy is to use criteria (1) as the most important with (2) and (3) being secondary considerations. Criteria (3) is relegated to secondary consideration because, pg. 60, "economic regret will vary so much for different uses of generated sequences that is not possible to use it in a generalized choice strategy." Using criteria (1), the preservation of persistence parameters as the primary strategy, the authors proposed (see Figure 5.3) regions based on the lag-one serial correlation coefficient and the Hurst coefficient where various types of models are acceptable. Based on this figure, the autoregressive lag-one or lag-two models are acceptable for serial correlations less than 0.6 and Hurst coefficients less than 0.7.

Certainly the Bowles et. al. study demonstrates that an AR model is able to produce generated sequences with useful Hurst coefficients. However, caution should be used in interpreting these results. The study is based on only four streams in a relatively small region. More importantly, the study does not address the problem of parameter uncertainty on model choice.

Past studies have shown that there is a great deal of uncertainty in estimating the correlation and Hurst coefficients. Rodriguez-Iturbe (1969) examined the sampling variance of the mean, standard deviation and lag-one

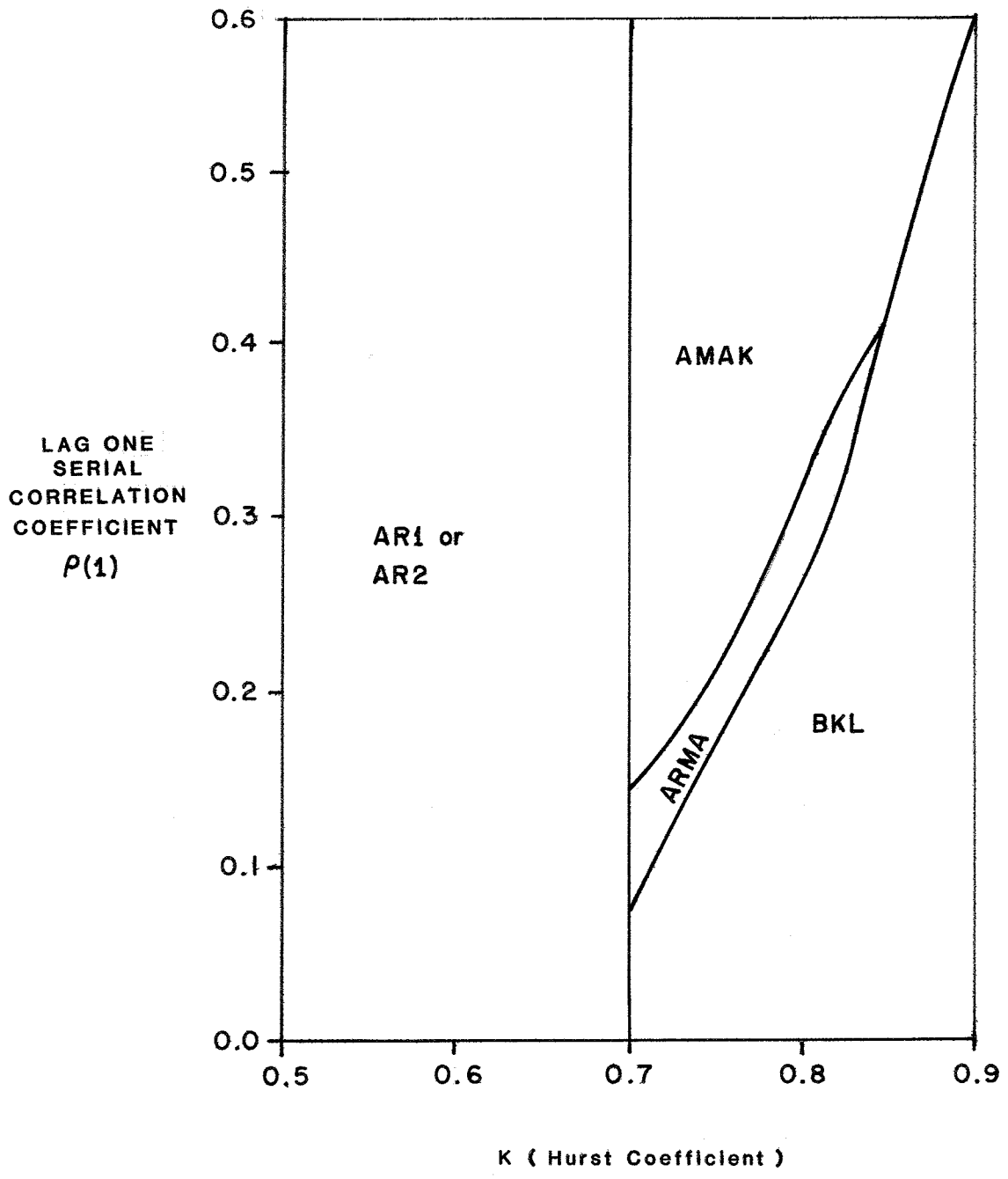


Figure 5.3 ANNUAL STREAMFLOW STOCHASTIC MODELS BASED ON LAG ONE SERIAL CORRELATION COEFFICIENT AND HURST COEFFICIENT RECOMMENDED BY ROWLES et. al. 1980

serial correlation coefficient assuming a linear, lag-one serial dependence between annual streamflows. He concluded that sampling errors are extremely high for all three variables for record lengths less than 40-60 years. Further he adds (page 1421),

"... that in respect to ρ_1, [the lag-one serial correlation coefficient] instability is the rule rather than the exception, even for records much longer than normally found."

The estimates of the mean and standard deviation tend to be significantly more reliable.

The sampling properties of the Hurst coefficient are also notoriously poor. Wallis and O'Connell (1973) conclude "For many regions of the world there is entirely insufficient hydrological data to make a reliable estimate of long-term persistence." In fact, these researchers claim that for normally available streamflow records it would be difficult to judge if persistence exists at all (e.g., $H = .5$ or $H \neq .5$). Considering these findings, it seems contradictory to judge a model based on its ability to preserve observed Hurst and correlation coefficients when these parameters are very difficult to estimate.

The importance of considering parameter and "intrinsic" uncertainty is choosing between stochastic models is addressed by Klemes et. al. (1981). They looked at parameter uncertainty by first comparing estimated reliability of reservoir performance based on long memory (ARMA) and short memory (AR) models. Second, reservoir reliability was investigated with a zero memory (the serial correlation set equal to zero) model. Only in this case, the parameters of the model, the streamflow mean and standard deviation, were varied based on their sampling variance (see equation 2.11). The conclusion

was that the difference between reservoir reliability* derived from a comparison of long and short memory flow models (about 3% greater with the long memory model) was insignificant compared to the 20% variability in the estimate of reservoir reliability due to model and parameter uncertainty. Consequently, parameter uncertainty tends to make the model choice academic.

Of more importance is the intrinsic uncertainty in defining future conditions. In the case of a reservoir, these researchers point out that social and economic changes during the project life of a reservoir are likely to introduce error into any estimates of reservoir reliability. This uncertainty has a greater effect on the estimate of reservoir performance reliability then does parameter uncertainty.

Consequently, from Klemes' et. al. point of view the advantage of using long memory flow models instead of an autoregressive models is marginal at best. Quoting, pg. 750, "To summarize, the replacement of a short-memory streamflow model with a long-memory model amounts to the incorporation of a small safety factor into the reservoir performance reliability. However, in most practical cases this factor will be much smaller than the accuracy with which the performance reliability can be assessed." Based on this conclusion, the model selection criteria given by Bowles et. al. seem to be to restrictive and that the AR model is more widely applicable than Figure 5.3 indicates.

*Three types of reservoir reliability were investigated. In this part of the analysis, quantity based reliability was investigated. This is the actual amount of water supplied to the consumer expressed as a percentage of the total demand over the period of simulation.

5.4 Summary

In the opinion of many researchers, the observation of long-term persistence in streamflow records invalidates the use of the autoregressive model. Long term persistence is characterized by a Hurst coefficient that exceeds 0.5. Researchers have been able to demonstrate that for any stationary probability model, including the autoregressive model, the derived Hurst coefficient is equal 0.5. Consequently, the autoregressive model was discarded in favor of more sophisticated models which are able to better model observed persistence.

However, more recently, the significance of long term persistence has been questioned because of the parameter uncertainty issue. Long term persistence can be equated with extremely long system memory, or equivalently, a correlogram that diverges (does not approach zero at long lags). Unfortunately, estimates of the lag-one serial correlation coefficient based on existing record lengths of 40 to 60 years has been shown to be highly uncertain. Consequently, estimation of serial correlations at longer lags, which is necessary to characterize long term persistence, must be even more uncertain. Thus it probably does not make sense to propose a model more sophisticated than an autoregressive model if its parameters are highly unreliable.

Another reason for some skepticism is the absence of a physical mechanism that endows a watershed with the "infinite memory" needed to account for long term perisistence. Although atmospheric processes have been suggested, the actual physical mechanism has not been identified.

The autoregressive model was found to be more valid than research has

indicated in the past. The reason for this is not only the effect of parameter uncertainty in choosing a model but the uncertainty intrinsic to the long term planning problem. The intrinsic uncertainty arises because the demands on a water resources project are very difficult to forecast over its lifetime (say 50-years). Consequently, the autoregressive model may very well be as sophisticated a model as needed given the intrinsic uncertainty involved in modeling demand.

Section 6

Concluding Remarks

There is an enormous volume of research devoted to the development and use of stochastic hydrologic models. If stochastic models include the probability models used in flood frequency analysis then at least some of this literature has been of some relevance to the practicing engineer. However, the application of stochastic models which produce synthetic streamflows has been of minor importance. This must be very disappointing to the research hydrologist considering the enormous amount of research effort devoted to stochastic model development.

Stochastic models do not receive widespread attention from the practicing hydrologist because these models are not well understood and there is some question as to how the models should be used. The reason that these models are not well understood is probably due to the jargon of time series analysis which permeates the literature on stochastic hydrology. However, this jargon can be dispensed with when the relationship between frequency analysis, which is understood by the practicing engineer, and stochastic streamflow models is recognized.

The relationship exists because in frequency analysis, streamflow peak discharge, for example, is modeled as an independent random variable and stochastic streamflow models represent streamflow volume as a dependent random variable. The link between the two is the incorporation of dependence to the description of random variables. Practically speaking, this is done by using linear regression to ascertain the degree of linear dependence between random variables. Thus, the combination of the probability models of independent

random variables used in flood frequency analysis and the modeling of dependence with linear regression results in a stochastic hydrologic model, namely the autoregressive model. Consequently, with addition of one simple concept, regression analysis, the engineer can see that stochastic models are not much different than the familiar techniques in frequency analysis.

The utilization of stochastic methods in practice is less easily addressed due to the nature of the prediction problem the hydrologist is trying to solve. The prediction problem generally involves estimating the likelihood of severe droughts or floods over the lifetime of a project given a relatively short historic record (i.e., a historic record that has length on the order of the useful life of the project being built). Obviously, there is a great deal of uncertainty in any prediction made under these conditions. This uncertainty is reflected in the uncertain estimates of the parameters in stochastic models which in turn leads to a small level of confidence in the estimates of likely levels of drought or high flow periods. Thus one might wonder what use, if any, that these stochastic models may have for the practicing hydrologist.

Stochastic models are useful because they point out the effect that variability in the hydrologic record can have on engineering design. Returning to the single reservoir design problem of Section 3.6, the variability in reservoir design capacity based on a probability model of the reservoir inflows (a lag-one autoregressive model) is extreme as can be seen from Figure 3.6. This exercise points out the need to modify a design based on the historic record by some type of safety factor. The magnitude of this safety factor may be based on engineering judgement or on simulations with stochastic model which produce droughts that are more severe than contained in the historic record. However, selection of a safety factor based on engineering judgement

is probably as valid as that derived by using a stochastic model given the uncertainties involved in the prediction problem.

The true value of stochastic methods results from the simulation approach to analyzing complex water resource systems. Since the stochastic models can be used to generate conditions as severe or more severe than the historic record, simulations using these models demonstrate the operational robustness of the water resource system under both severe wet and dry conditions. This is a convenient method of investigating the reliability of the system, considering that severe conditions generated by the stochastic model represent some safety factor.

List of References

Beard, L. R., and Kubik, H. E., 1972 Drought Severity and Water Supply Dependability, Technical Paper No. 30, Hydrologic Engineering Center, Corps of Engineers, U.S. Army, Davis, California.

Benjamin, J. R., and Cornell, C.A., 1970 Probability, Statistics and Decisions for Civil Engineers, McGraw-Hill, New York.

Bowles, S. D., Hughes, T. C., James, W. R., Jensen, D. T., Haws, F. W., 1980. Vulnerability of Water Supply Systems to Droughts, Utah Water Research Laboratory, College of Engineering, Utah State University.

Dracup, J. A., Lee, K. S. and Paulson, Jr., E. G., 1980. On the Definition of Droughts, Water Resources Research, April, V16(2), 297-302.

Fiering, M. B., 1963. Use of Correlation to Improve Estimates of the Mean and Variance, Geological Survey Professional Paper 434-C.

Haan, C. T., 1977. Statistical Methods in Hydrology, the Iowa State University Press, Ames, Iowa.

HEC-4, 1974. Monthly Streamflow Simulation (User Manual) Corps of Engineers, Hydrologic Engineering Center, Davis, California.

Hurst, H. E., 1951. Long-Term Storage Capacity of Reservoirs (with discussion), Trans. American Society of Civil Engineers, 116, paper 2447, 770-808.

Jackson, B. B. and Fiering, M. B., 1971. Synthetic Streamflows, Water Resources Monograph 1, American Geophysical Union, Washington, D. C.

Jenkins, G. M. and Watts, D. G., 1968. Spectral Analysis and Its Applications, Holden-Day.

Klemes, V., 1974. The Hurst Phenomenon: A Puzzle?, Water Resources Research, August, V10(4), 675-683.

Klemes, V., Srikanthan, R., McMahon, T. A., 1981. Long-Memory Flow Models in Reservoir Analysis: What is Their Practual Value?, Water Resources Research, V17(3), 737-751.

Kottegoda, N. T., 1980. Stochastic Water Resources Technology, Halsted Press, John Wiley and Sons, New York.

Last, 1979. Lane's Applied Stochastic Techniques (User Manual), Division of Planning Technical Services, Engineering and Research Center, Bureau of Reclamation, U.S. Department of the Interior.

Maass, A., Hufschmidt, MM., Dorfman, R., Thomas Jr., H.A., Marglin, S.A., Fair, G.M., 1962. Design of Water-Resource Systems, Harvard University Press, Cambridge, MA.

Mandelbrot, B. B., and Van Ness, J. W., 1968. Fractional Brownian Motions, Fractional Noises and Applications, Soc Ind Appl Rev., 10, 422-437.

Mandelbrot, B. B., and Wallis J. R., 1969. Computer Experiments with Fractional Gaussian Noises, Parts 1, 2 and 3, Water Resources Research, 5, 228-267. Correction WRR, 5, 1164.

Matalas, N. C., and Jacobs, B., 1964. A Correlation Procedure for Augmenting Hydrologic Data, Geological Survey Professional Paper, 434-E.

Matalas, N. C., 1963. Probability Distributions of Low Flows, Geological Survey Professional Paper 434-A.

Riggs, H. C., 1968: Low-Flow Investigations, Techniques of Water Resource Investigations of the United States Geological Survey.

Rodriguez-Iturbe, I., 1969. Estimation of Statistial Parameters for Annual River Flows, December, V5(6), 1418-1421.

Salas, J. D., Delleur, J. W., Yevjevich, V., and Lane W. L., 1980. Applied Modeling of Hydrologic Time Series, Water Resource Publications, Littleton, Colorado.

Sen, Z., 1976. Wet and Dry Periods of Annual Flow Series, October, American Society of Civil Engineers, HY10, 1503-1514.

Task Committee on Low-Flow Evaluation, Methods, and Needs of the Committee on Surface Water Hydrology of the Hydraulics Division, 1980. American Society of Civil Engineers, May, HY5, 717-731.

Wallis, J. R., and O'Connell, P. E., 1973. Firm Reservoir Yield How Reliable are Historic Hydrologic Records? Hydrologic Sciences Bulletin, 18(3), 347-365.

Wallis, J. R., and Matalas, N. C., 1979. Small Sample Properties of H and K Estimators of the Hurst Coefficient h, Water Resources Research, December, V6(6).

APPENDIX A

Computer Program for Drought Duration and Severity Calculation

Table A-1

Computer Program for Drought Duration and Severity Calculation

```
C
C     PROGRAM TO CALCULATE DROUGHT DURATION
C
      DIMENSION HI(2,51),VALUE(2,51),SAVE(10,3)
      INTEGER*6 ICON1,ICON2,ICON3,ISEED,JSEED
      COMMON/GEN/ ICON1,ICON2,ICON3,CON3
C
C     INITIALIZE DATA
C
      DATA ICON1,ICON2,CON3,ISEED /8388611,
     *703687441177663,.1421085472E-13,759821/
      DATA HI,VALUE /102*0,102*0.0/
C
C
      CALL ASGN(7,6HINPUT ,3H*0 ,-1)
      CALL ASGN(6,7HOUTPUT ,3H*3 ,1)
C
C
      IR=7
      IW=6
      IDIM=2
      JDIM=51
      ISAVE=0
      ICON3=ICON2+1
      NOBS=0
      ID=0
C
C     READ DATA
C
C       READ MEAN STANDARD DEVIATION AND LAG ONE SERIAL CORRELATION
C       COEFFICIENT
C
      READ(IR,*) XM,S,R
      WRITE(IW,10) XM,S,R
   10 FORMAT (/10X,"MEAN ANNUAL FLOW (AC-FT)=",F10.4/
     *10X,"FLOW STANDARD DEVIATION (AC-FT)=",F10.4/
     *10X,"LAG-ONE SERIAL CORRELATION=",F10.4/)
C
C     READ HISTOGRAM INTERVALS
C
      READ(IR,*) HD,HS,NHIST
      WRITE(IW,20) HD,HS,NHIST
   20 FORMAT (//10X,"HISTOGRAM INTERVALS"/
     *10X,"DURATION INTERVAL (YRS)=",F10.2/
     *10X,"SEVERITY INTERVAL (AC-FT)=",F10.0/
     *10X,"NUMBER OF CLASS INTERVALS=",I10)
C
C
      DO 25 I=1,(NHIST-1)
      VALUE(1,I)=FLOAT(I)*HD
      VALUE(2,I)=FLOAT(I)*HS
   25 CONTINUE
      VALUE(1,NHIST)=VALUE(1,(NHIST-1))
      VALUE(2,NHIST)=VALUE(2,(NHIST-1))
C
C     READ SIMULATION DATA
C
      READ(IR,*) NSIMUL,NINTER,JSEED
      IF (NINTER.EQ.0) NINTER=NSIMUL
      IF (JSEED.EQ.0) JSEED=ISEED
      WRITE(IW,30) NSIMUL,NINTER
   30 FORMAT(//10X,"SIMULATION DATA"/
     *10X,"# OF SIMULATIONS=",I10/
     *10X,"PRINTING INTERVALS=",I10)
C
C     INITIALIZE DATA
C
      XS=0
      XSUM=0
      X1=XM
      XCROSS=0
C
C GENERATE RANDOM NUMBER DISTRIBUTED NORMALLY WITH
C MEAN ZERO AND STANDARD DEVIATION ONE
C
      E=XNORM(JSEED)
      FACT=S*SQRT(1.0-R*R)
      DSUM=0
      QSUM=0
      X3=0
C
C     PERFORM SIMULATIONS
C     SAVE FIRST GENERATED FLOW
C
      X0=XM+R*(X1-XM)+E*FACT
      DO 500 I=1,NSIMUL
C
C     GENERATE FLOWS
C
      X2=XM+R*(X1-XM)+E*FACT
C
C     CALCULATE STATISTICS OF GENERATED FLOWS
C
      XSUM=XSUM+X2
      XS=XS+X2*X2
      XCROSS=XCROSS+X3*X2
C
C     DETERMINE DROUGHT CHARACTERISTICS
C
      IF (X2.LT.XM) THEN
          ID=1
```

Table A-1 (continued)

```
            DSUM=DSUM+1
            QSUM=QSUM+(XM-X2)
        ELSE
            IF (ID.EQ.1) THEN
                NOBS=NOBS+1
                ID=0
                J=1
                CALL HIST(J,DSUM,IDIM,JDIM,NHIST,VALUE,HI)
                J=2
                CALL HIST(J,QSUM,IDIM,JDIM,NHIST,VALUE,HI)
                DSUM=0
                QSUM=0
            END IF
        END IF
C
C       SET PARAMETERS
C
        X3=X2
        X1=X2
C
C GENERATE RANDOM NUMBER DISTRIBUTED NORMALLY WITH
C MEAN ZERO AND STANDARD DEVIATION ONE
C
        E=XNORM(JSEED)
        J=MOD(I,NINTER)
        IF(J.EQ.0)THEN
            N=I
            CALL STAT(N,XCROSS,XSUM,XS,X0,X2,SXM,DEV,SR)
            CALL RPRINT(DEV,HI,IDIM,NINTER,IW,JDIM,SR,SXM,NHIST,NOBS,N,
     *      VALUE)
            ISAVE=ISAVE+1
            SAVE(ISAVE,1)=SXM
            SAVE(ISAVE,2)=DEV
            SAVE(ISAVE,3)=SR
        END IF
  500 CONTINUE
        WRITE(IW,600)
  600   FORMAT(//10X,"SUMMARY OF ANNUAL FLOW STATISTICS"/)
        WRITE(IW,610)
  610   FORMAT(15X,"ITERATIONS",22X,"SAMPLE"/
       *38X,"MEAN",14X,"STD DEV",13X,"R LAG-ONE"/
       *36X,"(AC-FT)",13X,"(AC-FT)")
        DO 650 I=1,ISAVE
        N=I*NINTER
        WRITE(IW,640)N,SAVE(I,1),SAVE(I,2),SAVE(I,3)
  640   FORMAT(15X,I10,10X,F10.2,10X,F10.2,10X,F10.5)
  650   CONTINUE
        STOP
        END
        SUBROUTINE HIST(J,X,IDIM,JDIM,NHIST,VALUE,HI)
C
C       THIS SUBROUTINE SAVES DROUGHT PARAMETERS
C
        DIMENSION HI(IDIM,JDIM),VALUE(IDIM,JDIM)
        DO 100 I=1,(NHIST-1)
        Y=VALUE(J,I)
        IF (X.LE.Y) HI(J,I)=HI(J,I)+1
        IF (X.LE.Y) GO TO 200
  100 CONTINUE
        HI(J,NHIST)=HI(J,NHIST)+1.0
  200 RETURN
        END
        SUBROUTINE STAT(N,XCROSS,XSUM,XS,X0,X2,SXM,DEV,SR)
        DOUBLE PRECISION *12 TS,TERM1,TERM2,TERM3
C
C       THIS SUBROUTINE COMPUTES STATISTICS OF FLOWS
C       COMPUTE SAMPLE MEAN
C
        XN=FLOAT(N)
        XN1=XN-1.0
        SXM=XSUM/XN
C
C       COMPUTE SAMPLE STANDARD DEVIATION
C
        TERM2=XSUM*XSUM
        TS=XN*XS
        TS=TS-TERM2
        S=TS
        DEV=SQRT(S/(XN*XN1))
C
C       COMPUTE SAMPLE CORRELATION COEFFICIENT
C       SUM OF FLOWS FOR N-1 PERIODS
C
        TERM2=SXM*(XSUM-X2)
C
C       SUM OF FLOWS FOR N-1 LAGGED PERIODS
C
        TERM3=SXM*(XSUM-X0)
C
C
        TERM1=XCROSS
        TS=XN1*SXM*SXM
        TS=TERM1-TERM2-TERM3+TS
C
C       COMPUTE DIFFERENCE BETWEEN SUM OF SQUARES AND N*SAMPLE MEAN
C       SQUARED
C
         TERM1=XS
         TERM2=XN*SXM*SXM
         TERM1=TERM1-TERM2
C
C       SAMPLE LAG ONE CORRELATION COEFFICIENT
```

Table A-1 (continued)

```
C
      TS=TS/TERM1
      SR=TS
      RETURN
      END
      SUBROUTINE RPRINT(DEV,HI,IDIM,ITER,IW,JDIM,SR,SXM,NHIST,NOBS,N,
     *VALUE)
      DIMENSION VALUE(IDIM,JDIM),HI(IDIM,JDIM)
C PRINT STATISTICS
      WRITE(IW,100)N,SXM,DEV,SR
  100 FORMAT(//10X,"STATISICS AT SIMULATION #=",I10/
     *10X,"SAMPLE MEAN (AC-FT)=",F10.4/
     *10X,"SAMPLE STANDARD DEVIATION (AC-FT)=",F10.4/
     *10X,"SAMPLE LAG-ONE CORRELATION COEFFICIENT=",F10.4)
      WRITE(IW,150)
  150 FORMAT(//26X,"HISTOGRAMS FOR DROUGHT DURATION AND SEVERITY"/)
      WRITE(IW,160)
  160 FORMAT(//52X,"DROUGHT DURATION"/
     *16X,"INTERVAL",11X,"NUMBER OF",12X,"FRACTION",11X,"CUMULATIVE"/
     *16X,"(YEARS)",10X,"OBSERVATIONS",30X,"FRACTION")
      XN=FLOAT(NOBS)
      SUM=0
      EINT=0.
      DO 200 J=1,(NHIST-1)
      BINT=EINT
      EINT=VALUE(1,J)
      FRAC=HI(1,J)/XN
      SUM=SUM+FRAC
      WRITE(IW,170)BINT,EINT,HI(1,J),FRAC,SUM
  170 FORMAT(14X,2F6.0,11X,F7.0,2(13X,F6.4))
  200 CONTINUE
      WRITE(IW,210)HI(1,NHIST),VALUE(1,NHIST)
  210 FORMAT(/10X,"# OF OBSERVATIONS=",F10.0,
     *" WHICH EXCEED DURATION=",F10.1," YEARS")
      WRITE(IW,220)
  220 FORMAT(//42X,"DROUGHT SEVERITY"/
     *16X,"INTERVAL",11X,"NUMBER OF",12X,"FRACTION",11X,"CUMULATIVE"/
     *16X,"(AC-FT)",10X,"OBSERVATIONS",30X,"FRACTION")
      SUM=0.
      EINT=0.0
      DO 250J=1,(NHIST-1)
      BINT=EINT
      EINT=VALUE(2,J)
      FRAC=HI(2,J)/XN
      SUM=SUM+FRAC
      WRITE(IW,170)BINT,EINT,HI(2,J),FRAC,SUM
  250 CONTINUE
      WRITE(IW,260)HI(2,NHIST),VALUE(2,NHIST)
  260 FORMAT(/" # OF OBSERVATIONS=",F10.1,
     *" WHICH EXCEED DEFICIT=",F10.1," AC-FT")
      RETURN
      END
      FUNCTION XNORM(JSEED)
C
C THIS ROUTINE GENERATES NORMAL RANDOM DEVIATES FOR
C THE HARRIS 500 COMPUTER. STANDARD ROUTINES ARE
C AVAILABLE FOR OTHER COMPUTER SYSTEMS
C
      INTEGER *6 JSEED
      XNORM=0
      DO 100 I=1,12
      XNORM=XNORM+RAND(JSEED)
  100 CONTINUE
      XNORM=XNORM-6.0
      RETURN
      END
      FUNCTION RAND(JSEED)
C
C THIS ROUTINE GENERATES UNIFORM RANDOM DEVIATES FOR
C THE HARRIS 500 COMPUTER. STANDARD ROUTINES ARE
C AVAILABLE FOR OTHER COMPUTER SYSTEMS
C
      INTEGER *6 ICON1,ICON2,ICON3,JSEED
      COMMON /GEN/ICON1,ICON2,ICON3,CON3
      JSEED=JSEED*ICON1
      IF(JSEED.LT.0)JSEED=JSEED+ICON3
      RAND=FLOAT(JSEED)*CON3
      RETURN
      END
```

Table A-1 (continued)

*** SAMPLE OUTPUT ***

```
MEAN ANNUAL FLOW (AC-FT)=  372.6000
FLOW STANDARD DEVIATION (AC-FT)=   74.8000
LAG-ONE SERIAL CORRELATION=    0.1700
```

Sample statistics for observed streamflow

```
HISTOGRAM INTERVALS
DURATION INTERVAL (YRS)=      1.00
SEVERITY INTERVAL (AC-FT)=      50.
NUMBER OF CLASS INTERVALS=       20

SIMULATION DATA
# OF SIMULATIONS=       200
PRINTING INTERVALS=       100
```

```
STATISICS AT SIMULATION #=       100
SAMPLE MEAN (AC-FT)=  357.8061
SAMPLE STANDARD DEVIATION (AC-FT)=   75.3450
SAMPLE LAG-ONE CORRELATION COEFFICIENT=    0.0393
```

Sample statistics for first 100 years of synthetic streamflow

HISTOGRAMS FOR DROUGHT DURATION AND SEVERITY

for 100 Years of Synthetic Streamflow

Drought duration between given class interval

INTERVAL (YEARS)		NUMBER OF OBSERVATIONS	DROUGHT DURATION FRACTION	CUMULATIVE FRACTION
0.	1.	11.	0.4583	0.4583
1.	2.	6.	0.2500	0.7083
2.	3.	2.	0.0833	0.7917
3.	4.	3.	0.1250	0.9167
4.	5.	0.	0.0000	0.9167
5.	6.	0.	0.0000	0.9167
6.	7.	1.	0.0417	0.9583
7.	8.	1.	0.0417	1.0000
8.	9.	0.	0.0000	1.0000
9.	10.	0.	0.0000	1.0000
10.	11.	0.	0.0000	1.0000
11.	12.	0.	0.0000	1.0000
12.	13.	0.	0.0000	1.0000
13.	14.	0.	0.0000	1.0000
14.	15.	0.	0.0000	1.0000
15.	16.	0.	0.0000	1.0000
16.	17.	0.	0.0000	1.0000
17.	18.	0.	0.0000	1.0000
18.	19.	0.	0.0000	1.0000

OF OBSERVATIONS= 0. WHICH EXCEED DURATION= 19.0 YEARS

Fraction of total number of observed droughts of given duration, used to estimate drought duration probability

Table A-1 (continued)

Drought severity between given class intervals for 100 years of synthetic flow

DROUGHT SEVERITY

INTERVAL (AC-FT)		NUMBER OF OBSERVATIONS	FRACTION	CUMULATIVE FRACTION
0.	50.	5.	0.2083	0.2083
50.	100.	6.	0.2500	0.4583
100.	150.	6.	0.2500	0.7083
150.	200.	1.	0.0417	0.7500
200.	250.	1.	0.0417	0.7917
250.	300.	1.	0.0417	0.8333
300.	350.	2.	0.0833	0.9167
350.	400.	0.	0.0000	0.9167
400.	450.	1.	0.0417	0.9583
450.	500.	0.	0.0000	0.9583
500.	550.	0.	0.0000	0.9583
550.	600.	1.	0.0417	1.0000
600.	650.	0.	0.0000	1.0000
650.	700.	0.	0.0000	1.0000
700.	750.	0.	0.0000	1.0000
750.	800.	0.	0.0000	1.0000
800.	850.	0.	0.0000	1.0000
850.	900.	0.	0.0000	1.0000
900.	950.	0.	0.0000	1.0000

Fraction of total number of observed droughts of given volume deficit, used to estimate drought severity probability

OF OBSERVATIONS= 0.0 WHICH EXCEED DEFICIT= 950.0 AC-FT

STATISICS AT SIMULATION #= 200
SAMPLE MEAN (AC-FT)= 355.7967
SAMPLE STANDARD DEVIATION (AC-FT)= 76.6668
SAMPLE LAG-ONE CORRELATION COEFFICIENT= 0.0932

HISTOGRAMS FOR DROUGHT DURATION AND SEVERITY

for 100 Years of Synthetic Streamflow

DROUGHT DURATION

INTERVAL (YEARS)		NUMBER OF OBSERVATIONS	FRACTION	CUMULATIVE FRACTION
0.	1.	18.	0.4091	0.4091
1.	2.	10.	0.2273	0.6364
2.	3.	4.	0.0909	0.7273
3.	4.	7.	0.1591	0.8864
4.	5.	1.	0.0227	0.9091
5.	6.	1.	0.0227	0.9318
6.	7.	1.	0.0227	0.9545
7.	8.	1.	0.0227	0.9773
8.	9.	1.	0.0227	1.0000
9.	10.	0.	0.0000	1.0000
10.	11.	0.	0.0000	1.0000
11.	12.	0.	0.0000	1.0000
12.	13.	0.	0.0000	1.0000
13.	14.	0.	0.0000	1.0000
14.	15.	0.	0.0000	1.0000
15.	16.	0.	0.0000	1.0000
16.	17.	0.	0.0000	1.0000
17.	18.	0.	0.0000	1.0000
18.	19.	0.	0.0000	1.0000

OF OBSERVATIONS= 0. WHICH EXCEED DURATION= 19.0 YEARS

Table A-1 (continued)

DROUGHT SEVERITY

INTERVAL (AC-FT)	NUMBER OF OBSERVATIONS	FRACTION	CUMULATIVE FRACTION
0. 50.	6.	0.1364	0.1364
50. 100.	12.	0.2727	0.4091
100. 150.	10.	0.2273	0.6364
150. 200.	4.	0.0909	0.7273
200. 250.	2.	0.0455	0.7727
250. 300.	2.	0.0455	0.8182
300. 350.	3.	0.0682	0.8864
350. 400.	2.	0.0455	0.9318
400. 450.	1.	0.0227	0.9545
450. 500.	0.	0.0000	0.9545
500. 550.	0.	0.0000	0.9545
550. 600.	1.	0.0227	0.9773
600. 650.	0.	0.0000	0.9773
650. 700.	0.	0.0000	0.9773
700. 750.	0.	0.0000	0.9773
750. 800.	1.	0.0227	1.0000
800. 850.	0.	0.0000	1.0000
850. 900.	0.	0.0000	1.0000
900. 950.	0.	0.0000	1.0000

OF OBSERVATIONS= 0.0 WHICH EXCEED DEFICIT= 950.0 AC-FT

SUMMARY OF ANNUAL FLOW STATISTICS

ITERATIONS	SAMPLE MEAN (AC-FT)	SAMPLE STD DEV (AC-FT)	R LAG-ONE
100	357.81	75.34	0.03934
200	355.80	76.67	0.09321

comparison of sample 100 and 200 years of synthetic flows