

CECW-EH-Y

Engineer Manual  
No. 1110-2-1415

5 March 1993

Engineering and Design  
HYDROLOGIC FREQUENCY ANALYSIS

1. **Purpose.** This manual provides guidance and procedures for frequency analysis of: flood flows, low flows, precipitation, water surface elevation, and flood damage.
2. **Applicability.** This manual applies to major subordinate commands, districts, and laboratories having responsibility for the design of civil works projects.
3. **General.** Frequency estimates of hydrologic, climatic and economic data are required for the planning, design and evaluation of flood control and navigation projects. The text illustrates many of the statistical techniques appropriate for hydrologic problems by example. The basic theory is usually not provided, but references are provided for those who wish to research the techniques in more detail.

FOR THE COMMANDER:



WILLIAM D. BROWN  
Colonel, Corps of Engineers  
Chief of Staff

CECW-EH-Y

Engineer Manual  
No. 1110-2-1415

5 March 1993

Engineering and Design  
HYDROLOGIC FREQUENCY ANALYSIS

Table of Contents

	Subject	Paragraph	Page
CHAPTER 1	INTRODUCTION		
	Purpose and Scope .....	1-1	1-1
	References .....	1-2	1-1
	Definitions .....	1-3	1-1
	Need for Hydrologic Frequency Estimates ...	1-4	1-1
	Need for Professional Judgement .....	1-5	1-2
CHAPTER 2	FREQUENCY ANALYSIS		
	Definition .....	2-1	2-1
	Duration Curves .....	2-2	2-3
	Selection of Data for Frequency Analysis ...	2-3	2-6
	Graphical Frequency Analysis .....	2-4	2-10
	Analytical Frequency Analysis .....	2-5	2-12
CHAPTER 3	FLOOD FREQUENCY ANALYSIS		
	Introduction .....	3-1	3-1
	Log-Pearson Type III Distribution .....	3-2	3-1
	Weighted Skew Coefficient .....	3-3	3-6
	Expected Probability .....	3-4	3-7
	Risk .....	3-5	3-10
	Conditional Probability Adjustment .....	3-6	3-11
	Two-Station Comparison .....	3-7	3-13
	Flood Volumes .....	3-8	3-18
	Effects of Flood Control Works on Flood Frequencies .....	3-9	3-22
	Effects of Urbanization .....	3-10	3-28
CHAPTER 4	LOW-FLOW FREQUENCY ANALYSIS		
	Uses .....	4-1	4-1
	Interpretation .....	4-2	4-1
	Application Problems .....	4-3	4-1

	Subject	Paragraph	Page
CHAPTER 5	<b>PRECIPITATION FREQUENCY ANALYSIS</b>		
	General Procedures . . . . .	5-1	5-1
	Available Regional Information . . . . .	5-2	5-2
	Derivation of Flood-Frequency Relations from Precipitation . . . . .	5-3	5-2
CHAPTER 6	<b>STAGE(ELEVATION)-FREQUENCY ANALYSIS</b>		
	Uses . . . . .	6-1	6-1
	Stage Data . . . . .	6-2	6-1
	Frequency Distribution . . . . .	6-3	6-2
	Expected Probability . . . . .	6-4	6-2
CHAPTER 7	<b>DAMAGE-FREQUENCY RELATIONSHIPS</b>		
	Introduction . . . . .	7-1	7-1
	Computation of Expected Annual Damage . .	7-2	7-1
	Equivalent Annual Damage . . . . .	7-3	7-3
CHAPTER 8	<b>STATISTICAL RELIABILITY CRITERIA</b>		
	Objective . . . . .	8-1	8-1
	Reliability of Frequency Statistics . . . . .	8-2	8-1
	Reliability of Frequency Curves . . . . .	8-3	8-1
CHAPTER 9	<b>REGRESSION ANALYSIS AND APPLICATION TO REGIONAL STUDIES</b>		
	Nature and Application . . . . .	9-1	9-1
	Calculation of Regression Equations . . . . .	9-2	9-1
	The Correlation Coefficient and Standard Error . . . . .	9-3	9-3
	Simple Linear Regression Example . . . . .	9-4	9-4
	Factors Responsible of Nondetermination . . .	9-5	9-7
	Multiple Linear Regression Example . . . . .	9-6	9-8
	Partial Correlation . . . . .	9-7	9-8
	Verification of Regression Results . . . . .	9-8	9-10
	Regression by Graphical Techniques . . . . .	9-9	9-10
	Practical Guidelines . . . . .	9-10	9-10
	Regional Frequency Analysis . . . . .	9-11	9-11
CHAPTER 10	<b>ANALYSIS OF MIXED POPULATIONS</b>		
	Definition . . . . .	10-1	10-1
	Procedure . . . . .	10-2	10-1
	Cautions . . . . .	10-3	10-2
CHAPTER 11	<b>FREQUENCY OF COINCIDENT FLOWS</b>		
	Introduction . . . . .	11-1	11-1
	A Procedure for Coincident Frequency Analysis	11-2	11-1

Subject	Paragraph	Page
<b>CHAPTER 12 STOCHASTIC HYDROLOGY</b>		
Introduction .....	12-1	12-1
Applications .....	12-2	12-1
Basic Procedure .....	12-3	12-1
Monthly Streamflow Model .....	12-4	12-3
Data Fill In .....	12-5	12-6
Application In Areas of Limited Data .....	12-6	12-6
Daily Streamflow Model .....	12-7	12-6
Reliability .....	12-8	12-7
<b>APPENDIX A SELECTED BIBLIOGRAPHY</b>		
References and Textbooks .....	A	A-1
Computer Programs .....	B	A-4
<b>APPENDIX B GLOSSARY .....</b>		B-1
<b>APPENDIX C COMPUTATION PROCEDURE FOR EXTREME VALUE (GUMBEL) DISTRIBUTION .....</b>		C-1
<b>APPENDIX D HISTORIC DATA .....</b>		D-1
<b>APPENDIX E EXAMPLES OF RELIABILITY TESTS FOR THE MEAN AND STANDARD DEVIATION ...</b>		E-1
<b>APPENDIX F STATISTICAL TABLES .....</b>		F-1
Median Plotting Positions .....		F-2
Deviates for Pearson Type III Distribution ..		F-4
Normal Distribution .....		F-6
Percentage Points of the One-Tailed t-Distribution .....		F-7
Values of Chi-Square Distribution .....		F-8
Values of the F Distribution .....		F-9
Deviates for the Expected Probability Adjustment .....		F-10
Percentages for the Expected Probability Adjustment .....		F-11
Confidence Limit Deviates for Normal Distribution .....		F-12
Mean-Square Error of Station Skew Coefficient .....		F-16
Outlier Test K Values (10 Percent Significance Level) .....		F-17
Binomial Risk Tables .....		F-18

List of Figures

Title	Number	Page
Histogram and Probability Density Function . . . . .	2-1a	2-2
Sample and Theoretical Cumulative Distribution Functions . . .	2-1b	2-2
Daily Flow-Duration Curve . . . . .	2-2	2-3
Daily Flow-Duration Curves for Each Month . . . . .	2-3	2-5
Illustration of Chronologic Sequence and Arrayed Flood Peaks .	2-4	2-9
Example of Graphical Frequency Analysis . . . . .	2-5	2-12
Partial Duration Frequency Curve, Log-Log Paper . . . . .	2-6a	2-14
Partial Duration Frequency Curve, Probability Paper . . . . .	2-6b	2-14
Annual Frequency Curve . . . . .	3-1	3-4
Confidence Limit Curves based on the Non-central t-Distribution	3-2	3-8
Cumulative Probability Distribution of Exceedances per 100 Years	3-3	3-9
Two-Station Comparison Computations . . . . .	3-4	3-16
Observed and Two-Station Comparison Frequency Curves . . . .	3-5	3-17
Coordination of Flood-Volume Statistics . . . . .	3-6	3-20
Flood-Volume Frequency Curves . . . . .	3-7	3-21
Flood-Volume Frequency Relations . . . . .	3-8a	3-23
Storage Requirement Determination . . . . .	3-8b	3-23
Daily Reservoir Elevation-Duration Curve . . . . .	3-9	3-25
Seasonal Variation of Elevation-Duration Relations . . . . .	3-10	3-25
Example With-Project versus Without-Project Peak Flow Relations	3-11	3-27
Example Without-Project and With-Project Frequency Curves .	3-12	3-27
Typical Effect of Urbanization on Flood Frequency Curves . . .	3-13	3-29
Low-Flow Frequency Curves . . . . .	4-1	4-3
Frequency Curve, Annual Precipitation . . . . .	5-1	5-1
Flow-Frequency Curve, Unregulated and Regulated Conditions	6-1	6-1
Rating Curve for Present Conditions . . . . .	6-2	6-3
Derived Stage-Frequency Curves, Unregulated and Regulated Conditions . . . . .	6-3	6-3
Maximum Reservoir Elevation-Frequency Curve . . . . .	6-4	6-4
Schematic for Computation of Expected Annual Damage . . . . .	7-1	7-2
Frequency Curve with Confidence Limit Curves . . . . .	8-1	8-2
Computation of Simple Linear Regression Coefficients . . . . .	9-1	9-5
Illustration of Simple Regression . . . . .	9-2	9-6
Example Multiple Linear Regression Analysis . . . . .	9-3	9-9
Regression Analysis for Regional Frequency Computations . . .	9-4	9-14
Regional Analysis Computations for Mapping Errors . . . . .	9-5	9-15
Regional Map of Regression Errors . . . . .	9-6	9-15
Annual Peaks and Sequential Computed Skew by Year . . . . .	9-7	9-18
Nonhurricane, Hurricane and Combined Flood Frequency Curves	10-1	10-3
Illustration of Water-Surface Profiles in Coincident Frequency Analysis . . . . .	11-1	11-2
Data Estimation from Regression Line . . . . .	12-1a	12-2
Data Estimation with Addition of Random Errors . . . . .	12-1b	12-2

List of Tables

Title	Number	Page
Daily Flow-Duration Data and Interpolated Values . . . . .	2-1	2-4
Annual Peaks, Sequential and Arrayed with Plotting Positions .	2-2	2-11
Partial Duration Peaks, Arrayed with Plotting Positions . . . . .	2-3	2-13
Computed Frequency Curve and Statistics . . . . .	3-1	3-3
High-Flow Volume-Duration Data . . . . .	3-2	3-18
Low-Flow Volume-Duration Data . . . . .	4-1	4-2

ENGINEERING AND DESIGN  
HYDROLOGIC FREQUENCY ANALYSIS

CHAPTER 1

INTRODUCTION

1-1. **Purpose and Scope.** This manual provides guidance in applying statistical principles to the analysis of hydrologic data for Corps of Engineers Civil Works activities. The text illustrates, by example, many of the statistical techniques appropriate for hydrologic problems. The basic theory is usually not provided, but references are provided for those who wish to research the techniques in more detail.

1-2. **References.** The techniques described herein are taken principally from "Guidelines for Determining Flood Flow Frequency" (46)<sup>1</sup>, "Statistical Methods in Hydrology" (1), and "Hydrologic Frequency Analysis" (41). References cited in the text and a selected bibliography of literature pertaining to frequency analysis techniques are contained in Appendix A.

1-3. **Definitions.** Appendix B contains a list of definitions of terms common to hydrologic frequency analysis and symbols used in this manual.

1-4. **Need for Hydrologic Frequency Estimates.**

a. **Applications.** Frequency estimates of hydrologic, climatic and economic data are required for the planning, design and evaluation of water management plans. These plans may consist of combinations of structural measures such as reservoirs, levees, channels, pumping plants, hydroelectric power plants, etc., and nonstructural measures such as flood proofing, zoning, insurance programs, water use priorities, etc. The data to be analyzed could be streamflows, precipitation amounts, sediment loads, river stages, lake stages, storm surge levels, flood damage, water demands, etc. The probability estimates from these data are used in evaluating the economic, social and environmental effects of the proposed management action.

b. **Objective.** The objective of frequency analysis in a hydrologic context is to infer the probability that various size events will be exceeded or not exceeded from a given sample of recorded events. Two basic problems exist for most hydrologic applications. First the sample is usually small, by statistical standards, resulting in uncertainty as to the true probability. And secondly, a single theoretical frequency distribution does not always fit a particular data-type equally well in all applications. This manual provides guidance in fitting frequency distributions and construction of confidence limits. Techniques are presented which can possibly reduce the errors caused by small sample sizes. Also, some types of data are noted which usually do not fit any theoretical distributions.

c. **General Guidance.** Frequency analysis should not be done without consideration of the primary application of the results. The application will have a bearing on the type of analysis (annual series or partial duration series), number of stations to be included,

---

<sup>1</sup> Numbered references refer to Appendix A, Selected Bibliography.

whether regulated frequency curves will be needed, etc. A frequency study should be well coordinated with the hydrologist, the planner and the economist.

1-5. Need for Professional Judgment. It is not possible to define a set of procedures that can be rigidly applied to each frequency determination. There may be applications where more complex joint or conditional frequency methods, that were considered beyond the scope of this guidance, will be required. Statistical analyses alone will not resolve all frequency problems. The user of these techniques must insure proper application and interpretation has been made. The judgment of a professional experienced in hydrologic analysis should always be used in concert with these techniques.

## CHAPTER 2 FREQUENCY ANALYSIS

### 2-1. Definition.

a. Frequency. Many of the statistical techniques that are applied to hydrologic data (to enable inferences to be made about particular attributes of the data) can be labeled with the term "frequency analysis" techniques. The term "frequency" usually connotes a count (number) of events of a certain magnitude. To have a perspective of the importance of the count, the total number of events (sample size) must also be known. Sometimes the number of events within a specified time is used to give meaning to the count, e.g., two daily flows were this low in 43 years. The probability of a certain magnitude event recurring again in the future, if the variable describing the events is continuous, (as are most hydrologic variables), is near zero. Therefore, it is necessary to establish class intervals (arbitrary subdivisions of the range) and define the frequency as the number of events that occur within a class interval. A pictorial display of the frequencies within each class interval is called a histogram (also known as a frequency polygon).

b. Relative Frequency. Another means of representing the frequency is to compute the relative frequency. The relative frequency is simply the number of events in the class interval divided by the total number of events:

$$f_i = n_i/N \quad (2-1)$$

where:

$f_i$  = relative frequency of events in class interval  $i$

$n_i$  = number of events in interval  $i$

$N$  = total number of events

A graph of the relative frequency values is called a frequency distribution or histogram, Figure 2-1a. As the number of observations approaches infinity and the class interval size approaches zero, the enveloping line of the frequency distribution will approach a smooth curve. This curve is termed the probability density function (Figure 2-1a).

c. Cumulative Frequency. In hydrologic studies, the probability of some magnitude being exceeded (or not exceeded) is usually the primary interest. Presentation of the data in this form is accomplished by accumulating the probability (area) under the probability density function. This curve is termed the cumulative distribution function. In most statistical texts, the area is accumulated from the smallest event to the largest. The accumulated area then represents non-exceedance probability or percentage (Figure 2-1b). It is more common in hydrologic studies to accumulate the area from the largest event to the smallest. Area accumulated in this manner represents exceedance probability or percentage.

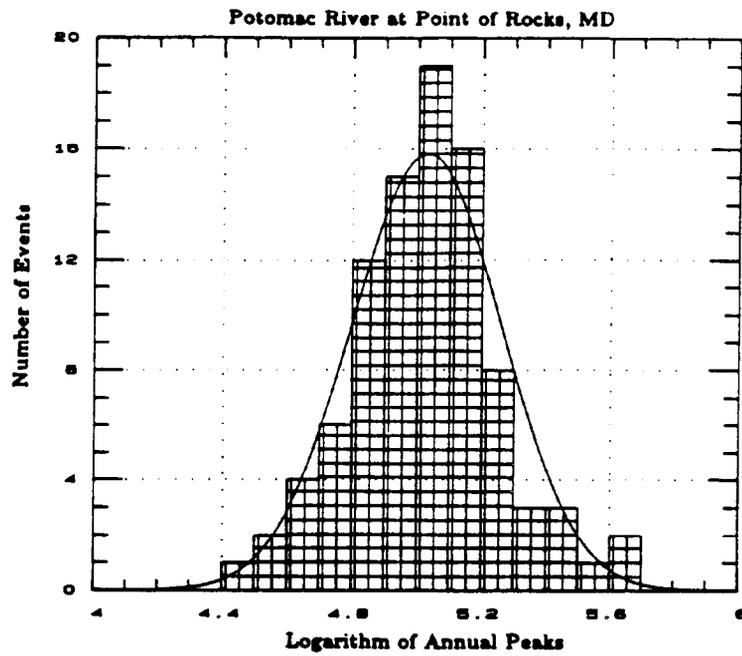


Figure 2-1a. Histogram and Probability Density Function.

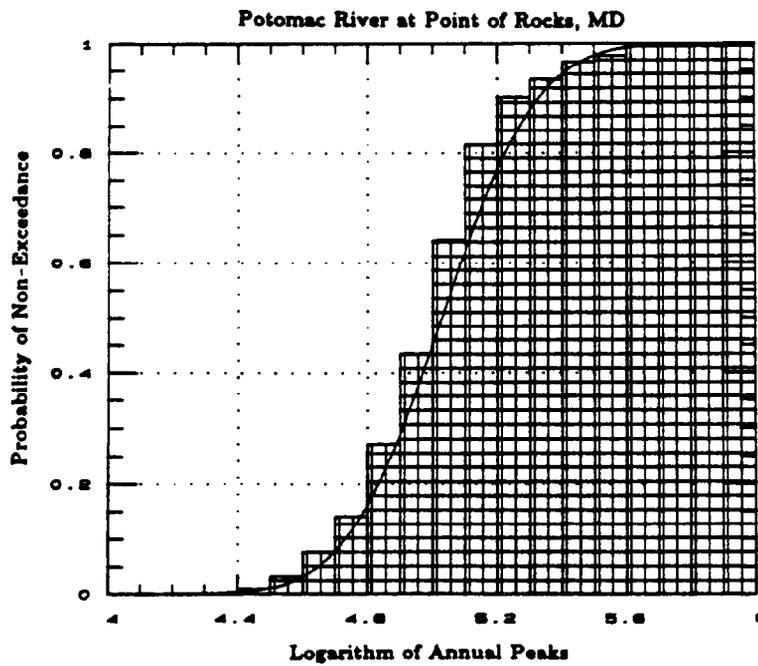


Figure 2-1b. Sample and Theoretical Cumulative Distribution Functions.

2-2. Duration Curves.

a. Computation. The computation of flow-duration curves was probably the first attempt to analyze hydrologic data by statistical techniques. The events for flow-duration curves are usually mean daily flow values. One of the first steps in preparation of a duration curve is dividing the range of the data into class intervals. Table 2-1 shows the class intervals of daily flows input into the computer program STATS (58) for a duration analysis of Fishkill Creek at Beacon, New York. The flows tend to be grouped near the low end with very few large flows. Therefore, the relative frequency curve is skewed to the right. It has been found that making the logarithmic transform reduces the skewness of the curve. The class intervals in Table 2-1 are based on a logarithmic distribution of the flows. Plotting the data in Table 2-1 on log-probability paper, Figure 2-2, provides a plot that is easily read at the extremities of the data. The daily flow-duration curve cannot be considered a frequency curve in the true sense, because the daily flow on a particular day is highly correlated with the flow on the preceding day. For this reason, the abscissa is labeled "percent of time exceeded."

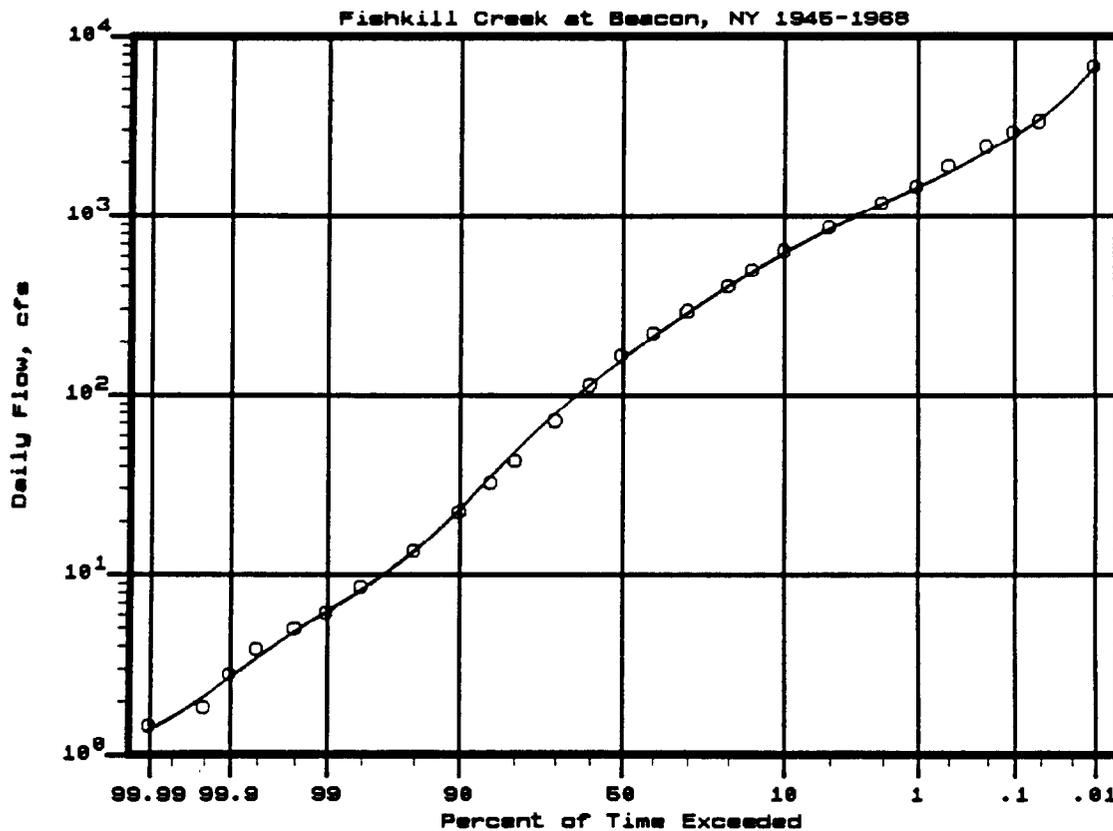


Figure 2-2. Daily Flow-Duration Curve.

Table 2-1. Daily Flow-Duration Data and Interpolated Values.

-DURATION DATA- FISHKILL CR AT BEACON, NY - DAILY FLOWS

* CLASS NUMBER	LOWER CLASS LIMIT FLOW, CFS	NUMBER IN CLASS	ACCUM NUMBER	PERCENT EQUAL OR EXCEED	* CLASS NUMBER	LOWER CLASS LIMIT FLOW, CFS	NUMBER IN CLASS	ACCUM NUMBER	PERCENT EQUAL OR EXCEED
* 1	1.00	5	8766	100.00	* 16	100.00	888	5606	63.95
* 2	2.00	4	8761	99.94	* 17	150.00	783	4718	53.82
* 3	3.00	8	8757	99.90	* 18	200.00	1246	3935	44.89
* 4	4.00	22	8749	99.81	* 19	300.00	822	2689	30.68
* 5	5.00	37	8727	99.56	* 20	400.00	484	1867	21.30
* 6	6.00	66	8690	99.13	* 21	500.00	340	1383	15.78
* 7	8.00	95	8624	98.38	* 22	600.00	465	1043	11.90
* 8	10.00	254	8529	97.30	* 23	800.00	239	578	6.59
* 9	15.00	261	8275	94.40	* 24	1000.00	251	339	3.87
* 10	20.00	423	8014	91.42	* 25	1500.00	47	88	1.00
* 11	30.00	405	7591	86.60	* 26	2000.00	32	41	0.47
* 12	40.00	359	7186	81.98	* 27	3000.00	6	9	0.10
* 13	50.00	332	6827	77.88	* 28	4000.00	0	3	0.03
* 14	60.00	480	6495	74.09	* 29	6000.00	3	3	0.03
* 15	80.00	409	6015	68.62	* 30	7000.00	0	0	0.00

-INTERPOLATED DURATION CURVE- FISHKILL CR AT BEACON, NY - DAILY FLOWS

* PERCENT EQUAL OR EXCEED	INTERPOLATED MAGNITUDE FLOW, CFS	* PERCENT EQUAL OR EXCEED	INTERPOLATED MAGNITUDE FLOW, CFS
* 0.01	6970.0	* 60.00	118.0
* 0.05	3480.0	* 70.00	74.5
* 0.10	3020.0	* 80.00	44.7
* 0.20	2530.0	* 85.00	33.4
* 0.50	1960.0	* 90.00	22.7
* 1.00	1500.0	* 95.00	14.0
* 2.00	1230.0	* 98.00	8.8
* 5.00	903.0	* 99.00	6.3
* 10.00	658.0	* 99.50	5.2
* 15.00	518.0	* 99.80	4.0
* 20.00	420.0	* 99.90	2.9
* 30.00	306.0	* 99.95	1.9
* 40.00	230.0	* 99.99	1.5
* 50.00	171.0	* 100.00	1.1

Output from HEC computer program STATS.

b. Uses. Duration curves are useful in assessing the general low flow characteristics of a stream. If the lower end drops rapidly to the probability scale, the stream has a low ground-water storage and, therefore, a low or no sustained flow. The overall slope of the flow-duration curve is an indication of the flow variability in the stream. Specific uses that have been made of duration curves are: 1) assessing the hydropower potential of run-of-river plants; 2) determining minimum flow release; 3) water quality studies; 4) sediment yield studies; and 5) comparing yield potential of basins. It must be remembered that the chronology of the flows is lost in the assembly of data for duration curves. For some studies, the low-flow sequence, or persistence, may be more important (see Chapter 4).

c. Monthly Curves. Occasionally the distribution of the flows during particular seasons of the year is of interest. Figure 2-3 illustrates a way of presenting daily-flow-duration curves that were computed from the daily flows during each month.

d. Stage-Duration. Stage-duration curves are often used to establish vertical navigation clearances for bridges. If there have been no changes in the discharge versus stage relationship (rating curve), then the stages may be used instead of flows to compute a stage-duration curve. But, if there have been significant changes to the rating curve (because of major levee construction, for instance) then the stage-duration curve should be derived from the flow-duration curve and the latest rating curve. The log transformation is not recommended for stages.

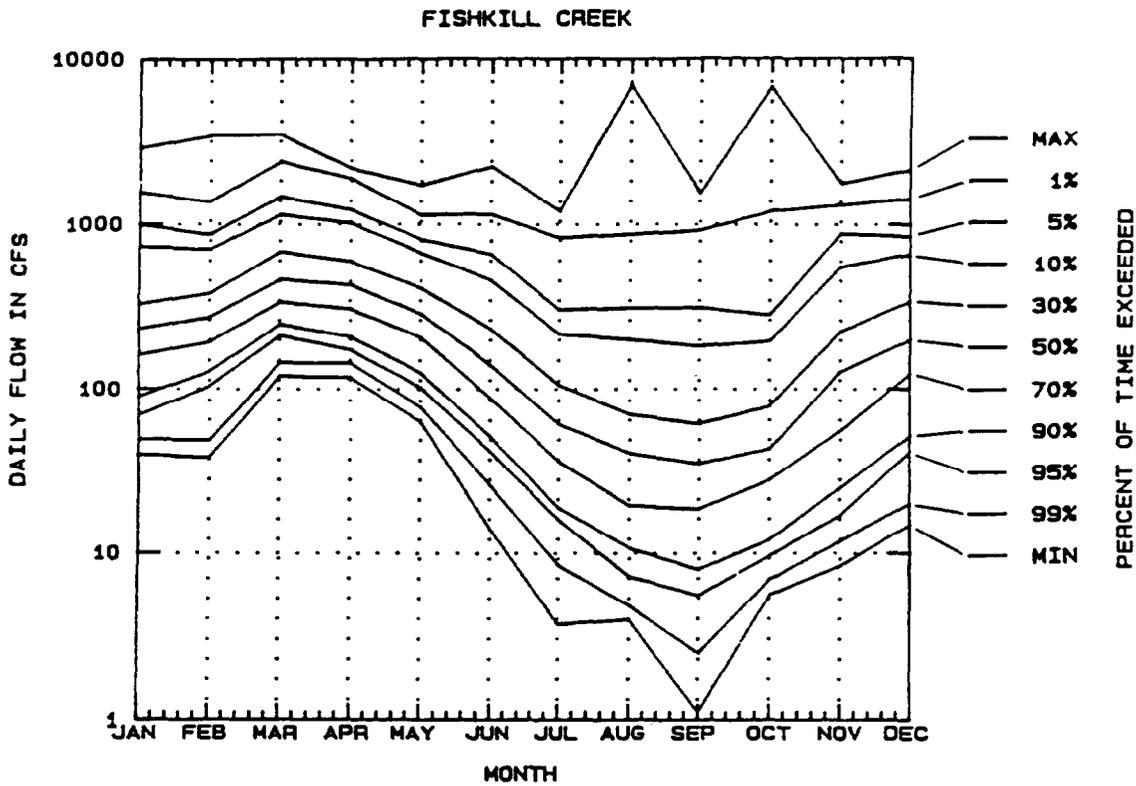


Figure 2-3. Daily Flow-Duration Curves for Each Month.

e. Future Probabilities. A duration curve is usually based on a fairly large sample size. For instance, Figure 2-2 is based on 8766 daily values (Table 2-1). Even though the observed data can be used to make inferences about future probabilities, conclusions drawn from information at the extremities can be misleading. The data indicate there is a zero percent chance of exceeding 6970 cfs, however, it is known that there is a finite probability of experiencing a larger flow. And similarly, there is some chance of experiencing a lower flow than the recorded 1.1 cfs. Therefore, some other means is needed for computing the probabilities of infrequent future events. Section 2-4 describes the procedure for assigning probabilities to independent events.

### 2-3. Selection of Data for Frequency Analysis.

a. Selection Based on Application of Results. The primary question to be asked before selection of data for a frequency study is: how will the frequency estimates be used? If the frequency curve is to be used for estimating damage that is related to the peak flow in a stream, maximum peak flows should be selected from the record. If the damage is best related to a longer duration of flow, the mean flow for several days' duration may be appropriate. For example, a reservoir's behavior may be related to the 3-day or 10-day rain flood volume or to the seasonal snowmelt volume. Occasionally, it is necessary to select a related variable in lieu of the one desired. For example, where mean-daily flow records are more complete than the records of peak flows, it may be more desirable to derive a frequency curve of mean-daily flows and then, from the computed curve, derive a peak-flow curve by means of an empirical relation between mean daily flows and peak flows. All reasonably independent values should be selected, but only the annual maximum events should be selected when the application of analytical procedures discussed in Chapter 3 is contemplated.

#### b. Uniformity of Data.

(1) General Considerations. Data selected for a frequency study must measure the same aspect of each event (such as peak flow, mean-daily flow, or flood volume for a specified duration), and each event must result from a uniform set of hydrologic and operational factors. For example, it would be improper to combine items from old records that are reported as peak flows but are in fact only daily readings, with newer records where the instantaneous peak was actually measured. Similarly, care should be exercised when there has been significant change in upstream storage regulation during the period of record to avoid combining unlike events into a single series. In such a case, the entire record should be adjusted to a uniform condition (see Sections 2-3f and 3-9). Data should always be screened for errors. Errors have been noted in published reports of annual flood peaks. And, errors have been found in the computer files of annual flood peaks. The transfer of data to either paper or a computer file always increases the probability that errors have been accidentally introduced.

(2) Mixed Populations. Hydrologic factors and relationships during a general winter rain flood are usually quite different from those during a spring snowmelt flood or during a local summer cloudburst flood. Where two or more types of floods are distinct and do not occur predominately in mutual combinations, they should not be combined into a single series for frequency analysis. It is usually more reliable in such cases to segregate the data in accordance with type and to combine only the final curves, if necessary. In the Sierra Nevada region of California and Nevada, frequency studies are made separately for rain floods which occur principally during the months of November through March,

and for snowmelt floods, which occur during the months of April through July. Flows for each of these two seasons are segregated strictly by cause - those predominantly caused by snowmelt and those predominantly caused by rain. In desert regions, summer thunderstorms should be excluded from frequency studies of winter rain flood or spring snowmelt floods and should be considered separately. Along the Atlantic and Gulf Coasts, it is often desirable to segregate hurricane floods from nonhurricane events. Chapter 10 describes how to combine the separate frequency curves into one relation.

c. Location Differences. Where data recorded at two different locations are to be combined for construction of a single frequency curve, the data should be adjusted as necessary to a single location, usually the location of the longer record. The differences in drainage area, precipitation and, where appropriate, channel characteristics between the two locations must be taken into account. When the stream-gage location is different from the project location, the frequency curve can be constructed for the stream-gage location and subsequently adjusted to the project location.

d. Estimating Missing Events. Occasionally a runoff record may be interrupted by a period of one or more years. If the interruption is caused by destruction of the gaging station by a large flood, failure to fill in the record for that flood would result in a biased data set and should be avoided. However, if the cause of the interruption is known to be independent of flow magnitude, the record should be treated as a broken record as discussed in Section 3-2b. In cases where no runoff records are available on the stream concerned, it is usually best to estimate the frequency curve as a whole using regional generalizations, discussed in Chapter 9, instead of attempting to estimate a complete series of individual events. Where a longer or more complete record at a nearby station exists, it can be used to extend the effective length of record at a location by adjusting frequency statistics (Section 3-7) or estimating missing events through correlation (Chapter 12).

e. Climatic Cycles. Some hydrologic records suggest regular cyclic variations in precipitation and runoff potential, and many attempts have been made to demonstrate that precipitation or streamflows evidence variations that are in phase with various cycles, particularly the well established 11-year sunspot cycle. There is no doubt that long-duration cycles or irregular climatic changes are associated with general changes of land masses and seas and with local changes in lakes and swamps. Also, large areas that have been known to be fertile in the past are now arid deserts, and large temperate regions have been covered with glaciers one or more times. Although the existence of climatic changes is not questioned, their effect is ordinarily neglected, because the long-term climatic changes have generally insignificant effect during the period concerned in water development projections, and short-term climatic changes tend to be self-compensating. For these reasons, and because of the difficulty in differentiating between stochastic (random) and systematic changes, the effect of natural cycles or trends during the analysis period is usually neglected in hydrologic frequency studies.

f. Effect of Basin Development on Frequency Relations.

(1) Hydrologic frequency estimates are often used for some purpose relating to planning, design or operation of water resources control measures (structural and nonstructural). The anticipated effects of these measures in changing the rate and volume of flow is assessed by comparing the without project frequency curve with the corresponding with project frequency curve. Also, projects that have existed in the past have affected the rates and volumes of flows, and the recorded values must be adjusted to reflect uniform conditions in order that the frequency analysis will conform to the basic

assumption of homogeneity. In order to meet the assumptions associated with analytical frequency analysis techniques, the flows must be essentially unregulated by manmade storage or diversion structures. Consequently, wherever practicable, recorded runoff values should be adjusted to natural (unimpaired) conditions before an analytical frequency analysis is made. In cases where the impairment results from a multitude of relatively small improvements that have not changed appreciably during the period of record, it is possible that analytical frequency analysis techniques can be applied. The adjustment to natural conditions may be unnecessary and, because of the amount of work involved, not cost effective.

(2) One approach to determining a frequency curve of regulated or modified runoff consists of routing all of the observed flood events under conditions of proposed or anticipated development. Then a relationship is developed between the modified and the natural flows, deriving an average or dependable relationship. A frequency curve of modified flows is derived from this relationship and the frequency curve of natural flows. In order to determine frequencies of runoff for extreme floods, routings of multiples of the largest floods of record or multiples of a large hypothetical flood can be used. Techniques of estimating project effects are outlined in Chapter 3-09d.

g. Annual Series Versus Partial Duration Series. There are two basic types of frequency curves used to estimate flood damage. A curve of annual maximum events is ordinarily used when the primary interest lies in the larger events or when the second largest event in any year is of little concern in the analysis. The partial-duration curve represents the frequency of all independent events of interest, regardless of whether two or more occurred in the same year. This type of curve is sometimes used in economic analysis, where there is considerable damage associated with the second largest and third largest floods that occurred in some of the years. Caution must be exercised in selecting events because they must be both hydrologically and economically independent. The selected series type should be established early in the study in coordination with the planner and/or economist. The time interval between flood events must be sufficient for recovery from the earlier flood. Otherwise damage from the later flood would not be as large as computed. When both the frequency curve of annual floods and the partial-duration curve are used, care must be exercised to assure that the two are consistent. A graphic demonstration of the relation between a chronologic record, an annual-event curve and a partial-duration curve is shown on Figure 2-4.

h. Presentation of Data and Results of Frequency Analysis. When frequency curves are presented for technical review, adequate information should be included to permit an independent review of the data, assumptions and analysis procedures. The text should indicate clearly the scope of the studies and include a brief description of the procedure used, including appropriate references. A summary of the basic data consisting of a chronological tabulation of values used and indicating sources of data and any adjustments should be included. The frequency data should also be presented in graphical form, ordinarily on probability paper, along with the adopted frequency curves. Confidence limit curves should also be included for analytically-derived frequency curves to illustrate the relative value of the frequency relationships. A map of the gage locations and tables of the adopted statistics should also be included.

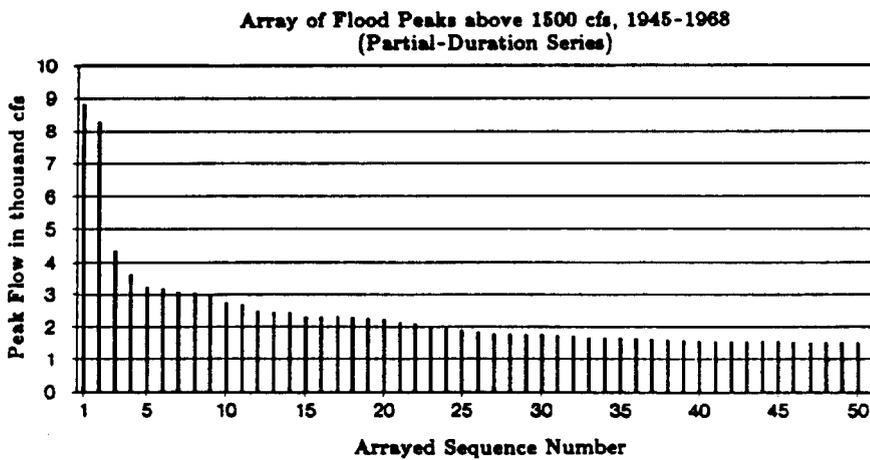
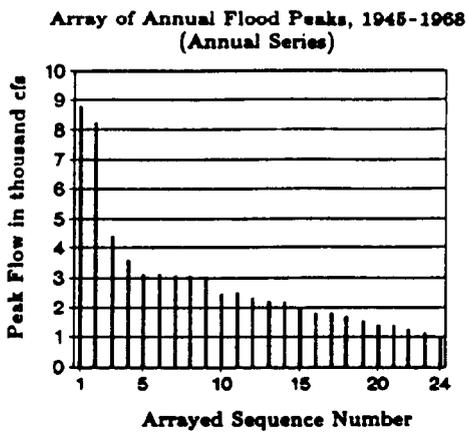
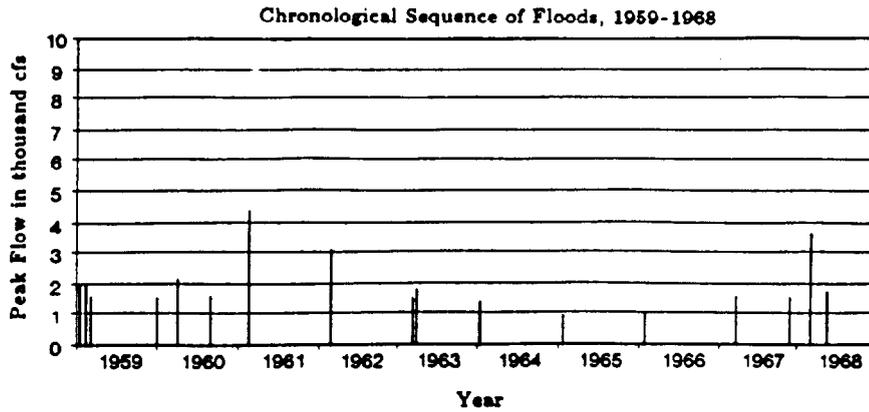


Figure 2-4. Illustration of Chronologic Sequence and Arrayed Flood Peaks.

## 2-4. Graphical Frequency Analysis.

a. Advantages and Limitations. Every set of frequency data should be plotted graphically, even though the frequency curves are obtained analytically. It is important to visually compare the observed data with the derived curve. The graphical method of frequency-curve determination can be used for any type of frequency study, but analytical methods have certain advantages when they are applicable. The principal advantages of graphical methods are that they are generally applicable, that the derived curve can be easily visualized, and that the observed data can be readily compared with the computed results. However, graphical methods of frequency analysis are generally less consistent than analytical methods as different individuals would draw different curves. Also, graphical procedures do not provide means for evaluating the reliability of the estimates. Comparison of the adopted curve with plotted points is not an index of reliability, but it is often erroneously assumed to be, thus implying a much greater reliability than is actually attained. For these reasons, graphical methods should be limited to those data types where analytical methods are known not to be generally applicable. That is, where frequency curves are too irregular to compute analytically, for example, stream or reservoir stages and regulated flows. Graphical procedures should always be to visually check the analytical computations.

b. Selection and Arrangement of Peak Flow Data. General principles in the selection of frequency data are discussed in Section 2-3. Data used in the construction of frequency curves of peak flow consist of the maximum instantaneous flow for each year of record (for annual-event curve) or all of the independent events that exceed a selected base value (for partial-duration curve). This base value must be smaller than any flood flow that is of importance in the analysis, and should also be low enough so that the total number of floods in excess of the base equals or exceeds the number of years of record. Table 2-2 is a tabulation of the annual peak flow data with dates of occurrence, the data arrayed in the order of magnitude, and the corresponding plotting positions.

c. Plotting Formulas. Median plotting positions are tabulated in Table F-1. In ordinary hydrologic frequency work,  $N$  is taken as the number of years of record rather than the number of events, so that percent chance exceedance can be thought of as the number of events per hundred years. For arrays larger than 100, the plotting position,  $P_1$ , of the largest event is obtained by use of the following equation:

$$P_1 = 100 (1 - (.5)^{1/N}) \quad (2-2a)$$

The plotting position for the smallest event ( $P_N$ ) is the complement ( $1 - P_1$ ) of this value, and all the other plotting positions are interpolated linearly between these two. The median plotting positions can be approximated by

$$P_m = 100(m - .3)/(N + .4) \quad (2-2b)$$

where  $m$  is the order number of the event.

For partial-duration curves, particularly where there are more events than years ( $N$ ), plotting positions that indicate more than one event per year can also be obtained using

Table 2-2. Annual Peaks, Sequential and Arrayed with Plotting Positions.

---

-PLOTTING POSITIONS-FISHKILL CREEK AT BEACON, N.Y.  
\*\*\*\*\*

*.....EVENTS ANALYZED.....*				*.....ORDERED EVENTS.....*			
				WATER			
* MON	* DAY	* YEAR	* FLOW,CFS	* RANK	* YEAR	* FLOW,CFS	* MEDIAN PLOT POS *
*	3	5	1945	*	1	1955	8800. 2.87 *
*	12	27	1945	*	2	1956	8280. 6.97 *
*	3	15	1947	*	3	1961	4340. 11.07 *
*	3	18	1948	*	4	1968	3630. 15.16 *
*	1	1	1949	*	5	1953	3220. 19.26 *
*	3	9	1950	*	6	1952	3170. 23.36 *
*	4	1	1951	*	7	1962	3060. 27.46 *
*	3	12	1952	*	8	1949	3020. 31.56 *
*	1	25	1953	*	9	1948	2970. 35.66 *
*	9	13	1954	*	10	1958	2500. 39.75 *
*	8	20	1955	*	11	1951	2490. 43.85 *
*	10	16	1955	*	12	1945	2290. 47.95 *
*	4	10	1957	*	13	1947	2220. 52.05 *
*	12	21	1957	*	14	1860	2140. 56.15 *
*	2	11	1959	*	15	1959	1960. 60.25 *
*	4	6	1960	*	16	1963	1780. 64.34 *
*	2	26	1961	*	17	1954	1760. 68.44 *
*	3	13	1962	*	18	1967	1580. 72.54 *
*	3	28	1963	*	19	1946	1470. 76.64 *
*	1	26	1964	*	20	1864	1380. 80.74 *
*	2	9	1965	*	21	1957	1310. 84.84 *
*	2	15	1966	*	22	1950	1210. 88.93 *
*	3	30	1967	*	23	1866	1040. 93.03 *
*	3	19	1968	*	24	1965	980. 97.13 *

\*\*\*\*\*

Output from HEC computer program HECWRC.

---

Equation 2-2b. This is simply an approximate method used in the absence of knowledge of the total number of events in the complete set of which the partial-duration data constitute a subset.

d. Plotting Grids. The plotting grid recommended for annual flood flow events is the logarithmic normal grid developed by Allen Hazen (ref 13) and designed such that a logarithmic normal frequency distribution will be represented by a straight line, Figure 2-5. The plotting grid used for stage frequencies is often the arithmetic normal grid. The plotting grids may contain a horizontal scale of exceedance probability, exceedance frequency, or percent chance exceedance. Percent chance exceedance (or nonexceedance) is the recommended terminology.

e. Example Plotting of Annual-Event Frequency Curve. Figure 2-5 shows the plotting of a frequency curve of the annual peak flows tabulated in Table 2-2. A smooth curve should be drawn through the plotted points. Unless computed by analytical frequency procedures, the frequency curve should be drawn as close to a straight line as possible on the chosen probability graph paper. The data plotted on Figure 2-5 shows a tendency to curve upward, therefore, a slightly curved line was drawn as a best fit line.

f. Example Plotting of Partial-Duration Curve. The partial-duration curve corresponding to the partial-duration data in Table 2-3 is shown of Figure 2-6a. This curve has been drawn through the plotted points, except that it was made to conform with the annual-event curve in the upper portion of the curve. The annual-event curve was

developed in accordance with the procedures described in Chapter 3. When partial-duration data must include more events than there are years of record (see Subparagraph b) it will be necessary to use logarithmic paper for plotting purposes, as on Figure 2-6a, in order to plot exceedance frequencies greater than 100 percent. Otherwise, the curve can be plotted on probability grid, as illustrated on Figure 2-6b.

2-5. Analytical Frequency Analysis.

a. General Procedures and Common Distributions. The fitting of data by an analytical procedure consists of selecting a theoretical frequency distribution, estimating the parameters of the distribution from the data by some fitting technique, and then evaluating the distribution function at various points of interest. Some theoretical distributions that have been used in hydrologic frequency analysis are the normal (Gaussian), log normal, exponential, two-parameter gamma, three-parameter gamma, Pearson type III, log-Pearson type III, extreme value (Gumbel) and log Gumbel. Chapter 3 describes the fitting of the log-Pearson type III to annual flood peaks and Appendix C describes fitting the extreme value (Gumbel) distribution.

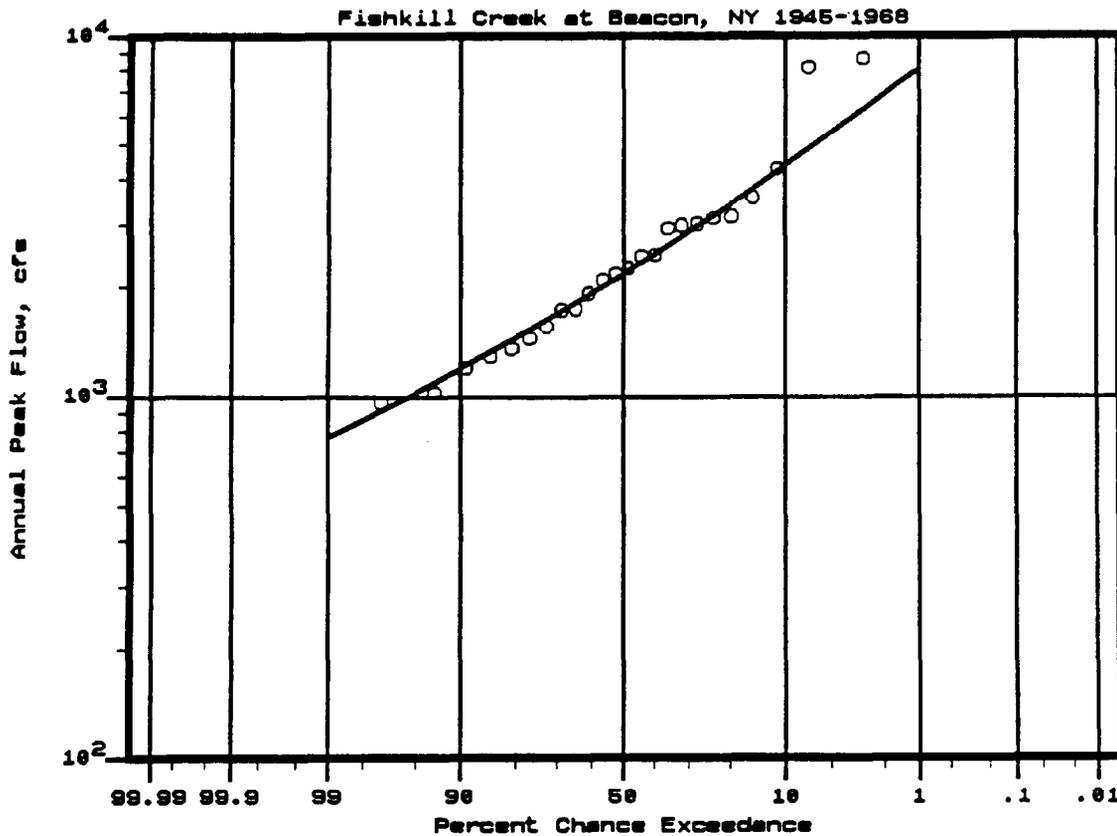


Figure 2-5. Example of Graphical Frequency Analysis

Table 2-3. Partial Duration Peaks, Arrayed with Plotting Positions.

FISHKILL CREEK AT BEACON, NY -- PEAKS ABOVE 1500 CFS				FISHKILL CREEK AT BEACON, NY -- PEAKS ABOVE 1500 CFS			
***** ORDERED EVENTS *****				***** ORDERED EVENTS *****			
RANK	WATER YEAR	FLOW,CFS	MEDIAN PLOT POS	RANK	WATER YEAR	FLOW,CFS	MEDIAN PLOT POS
* 1	1955	8800.	2.87	* 26	1952	1820.	105.33
* 2	1956	8280.	6.97	* 27	1945	1780.	109.43
* 3	1961	4340.	11.07	* 28	1963	1780.	113.52
* 4	1968	3630.	15.16	* 29	1956	1770.	117.62
* 5	1953	3220.	19.26	* 30	1954	1760.	121.72
* 6	1952	3170.	23.36	* 31	1952	1730.	125.82
* 7	1962	3060.	27.46	* 32	1968	1720.	129.92
* 8	1949	3020.	31.56	* 33	1955	1660.	134.02
* 9	1948	2970.	35.66	* 34	1958	1650.	138.11
* 10	1948	2750.	39.75	* 35	1958	1650.	142.21
* 11	1949	2700.	43.85	* 36	1953	1630.	146.31
* 12	1958	2500.	47.95	* 37	1960	1610.	150.41
* 13	1951	2490.	52.05	* 38	1956	1600.	154.51
* 14	1952	2460.	56.15	* 39	1958	1590.	158.61
* 15	1945	2290.	60.25	* 40	1958	1580.	162.70
* 16	1953	2290.	64.34	* 41	1967	1580.	166.80
* 17	1958	2290.	68.44	* 42	1951	1560.	170.90
* 18	1953	2280.	72.54	* 43	1959	1560.	175.00
* 19	1948	2220.	76.64	* 44	1955	1550.	179.10
* 20	1951	2210.	80.74	* 45	1951	1540.	183.20
* 21	1960	2140.	84.84	* 46	1968	1530.	187.30
* 22	1953	2080.	88.93	* 47	1960	1520.	191.39
* 23	1959	1960.	93.03	* 48	1958	1520.	195.49
* 24	1959	1920.	97.13	* 49	1952	1520.	199.59
* 25	1958	1900.	101.23	* 50	1948	1510.	203.69
				* 51	1963	1510.	207.79

b. **Advantages.** Determining the frequency distribution of data by the use of analytical techniques has several advantages. The use of an established procedure for fitting a selected distribution would result in consistent frequency estimates from the same data set by different persons. Error distributions have been developed for some of the theoretical distributions that enable computing the degree of reliability of the frequency estimates (see Chapter 8). Another advantage is that it is possible to regionalize the parameter estimates which allows making frequency estimates at ungaged locations (see Chapter 9).

c. **Disadvantages.** The theoretical fitting of some data can result in very poor frequency estimates. For example, stage-frequency curves of annual maximum stages are shaped by the channel and valley characteristics, backwater conditions, etc. Another example is the flow-frequency curve below a reservoir. The shape of this frequency curve would depend not only on the inflow but the capacities, operation criteria, etc. Therefore, graphical techniques must be used where analytical techniques provide poor frequency estimates.

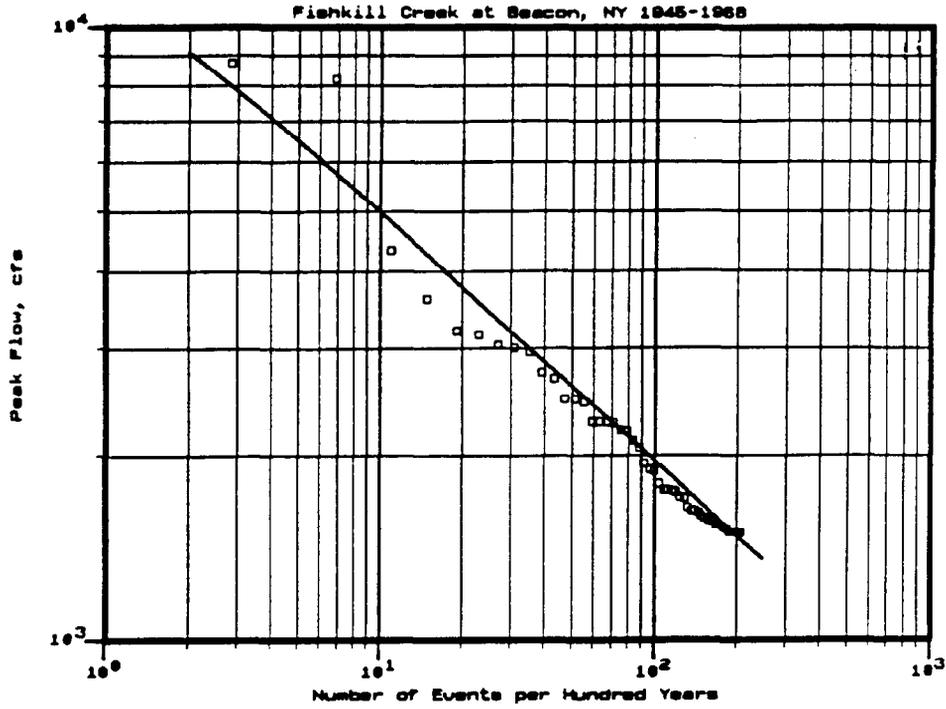


Figure 2-6a. Partial Duration Frequency Curve, Log-Log Paper

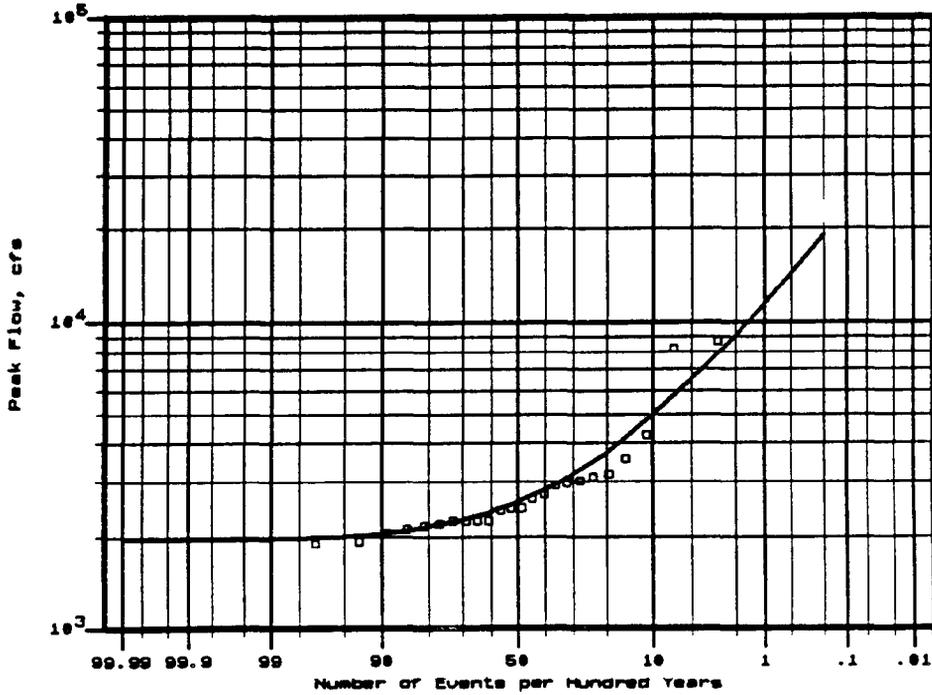


Figure 2-6b. Partial Duration Frequency Curve, Probability Paper.

## CHAPTER 3 FLOOD FREQUENCY ANALYSIS

### 3-1. Introduction.

The procedures that federal agencies are to follow when computing a frequency curve of annual flood peaks have been published in Guidelines for Determining Flood Flow Frequency, Bulletin 17B (46). As stated in Bulletin 17B, "Flood events ... do not fit any one specific known statistical distribution." Therefore, it must be recognized that occasionally, the recommended techniques may not provide a reasonable fit to the data. When it is necessary to use a procedure that departs from Bulletin 17B, the procedure should be fundamentally sound and the steps of the procedure documented in the report along with the frequency curves.

This report contains most aspects of Bulletin 17B, but in an abbreviated form. Various aspects of the procedures are described in an attempt to clarify the computational steps. The intent herein is to provide guidance for use with Bulletin 17B. The step by step procedures to compute a flood peak frequency curve are contained in Appendix 12 of Bulletin 17B and are not repeated herein.

### 3-2. Log-Pearson Type III Distribution.

a. General. The analytical frequency procedure recommended for annual maximum streamflows is the logarithmic Pearson type III distribution. This distribution requires three parameters for complete mathematical specification. The parameters are: the mean, or first moment, (estimated by the sample mean,  $\bar{X}$ ); the variance, or second moment, (estimated by the sample variance,  $S^2$ ); the skew, or third moment, (estimated by the sample skew,  $G$ ). Since the distribution is a logarithmic distribution, all parameters are estimated from logarithms of the observations, rather than from the observations themselves. The Pearson type III distribution is particularly useful for hydrologic investigations because the third parameter, the skew, permits the fitting of non-normal samples to the distribution. When the skew is zero the log-Pearson type III distribution becomes a two-parameter distribution that is identical to the logarithmic normal (often called log-normal) distribution.

#### b. Fitting the Distribution.

(1) The log-Pearson type III distribution is fitted to a data set by calculating the sample mean, variance, and skew from the following equations:

$$\bar{X} = \frac{\sum X}{N} \quad (3-1)$$

$$S^2 = \frac{\sum x^2}{N-1} = \frac{\sum (X-\bar{X})^2}{N-1} \quad (3-2a)$$

$$= \frac{\sum X^2 - (\sum X)^2/N}{N-1} \quad (3-2b)$$

$$G = \frac{N(\sum x^3)}{(N-1)(N-2)S^3} = \frac{N(\sum (X-\bar{X})^3)}{(N-1)(N-2)S^3} \quad (3-3a)$$

$$= \frac{N^2(\sum X^3) - 3N(\sum X)(\sum X^2) + 2(\sum X)^3}{N(N-1)(N-2)S^3} \quad (3-3b)$$

in which:

$\bar{X}$  = mean logarithm

$X$  = logarithm of the magnitude of the annual event

$N$  = number of events in the data set

$S^2$  = unbiased estimate of the variance of logarithms

$x$  =  $X - \bar{X}$ , the deviation of the logarithm of a single event from the mean logarithm

$G$  = unbiased estimate of the skew coefficient of logarithms

The precision of the computed values is more sensitive to the number of significant digits when Equations 3-2b and 3-3b are used.

(2) In terms of the frequency curve itself, the mean represents the general magnitude or average ordinate of the curve, the square root of the variance (the standard deviation,  $S$ ) represents the slope of the curve, and the skew represents the degree of curvature. Computation of the unadjusted frequency curve is accomplished by computing values for the logarithms of the streamflow corresponding to selected values of percent chance exceedance. A reasonable set of values and the results are shown in Table 3-1. The number of values needed to define the curve depends on the degree of curvature (i.e., the skew). For a skew value of zero, only two points would be needed, while for larger skew values all of the values in the table would ordinarily be needed.

(3) The logarithms of the event magnitudes corresponding to each of the selected percent chance exceedance values are computed by the following equation:

$$\log Q = \bar{X} + KS \quad (3-4)$$

where  $\bar{X}$  and  $S$  are defined as in Equations 3-1 and 3-2 and where

log Q = logarithm of the flow (or other variable) corresponding to a specified value of percent chance exceedance

K = Pearson type III deviate that is a function of the percent chance exceedance and the skew coefficient.

c. Example Computation.

(1) As shown in the following example, Equation 3-4 is solved by using the computed values of  $\bar{X}$  and S and obtaining from Appendix V-3 the value of K corresponding to the adopted skew, G, and the selected percent chance exceedance (P). An example computation for P=1.0, where  $\bar{X}$ , S and G are taken from Table 3-1, is:

$$\begin{aligned} \log Q &= 3.3684 + 2.8236 (.2456) \\ &= 4.0619 \\ Q &= 11500 \text{ cfs} \end{aligned}$$

(2) It has been shown (36) that a frequency curve computed in this manner is biased in relation to average future expectation because of uncertainty as to the true mean and standard deviation. The effect of this bias for the normal distribution can be eliminated by an adjustment termed the expected probability adjustment that accounts for the actual sample size. This adjustment is discussed in more detail in Section 3-4. Table 3-1 and Figure 3-1 shows the derived frequency curve along with the expected probability adjusted curves and the 5 and 95 percent confidence limit curves.

Table 3-1. Computed Frequency Curve and Statistics.

-FREQUENCY CURVE- 01-3735 FISHKILL CREEK AT BEACON, NEW YORK						
*****						
.....FLOW,CFS.....*	PERCENT	*...CONFIDENCE LIMITS....*				
* EXPECTED	* CHANCE	*				
* COMPUTED	* PROBABILITY	* EXCEEDANCE	* 0.05 LIMIT	* 0.95 LIMIT	* *	
-----*						
* 19200.	28300.	* 0.2	* 39100.	12300.	* *	
* 14500.	19000.	* 0.5	* 26900.	9740.	* *	
* 11500.	14100.	* 1.0	* 20100.	8080.	* *	
* 9110.	10500.	* 2.0	* 14800.	6640.	* *	
* 7100.	7820.	* 4.0	* 10800.	5380.	* *	
* 4960.	5210.	* 10.0	* 6850.	3950.	* *	
* 3650.	3740.	* 20.0	* 4710.	2990.	* *	
* 2190.	2190.	* 50.0	* 2650.	1790.	* *	
* 1440.	1420.	* 80.0	* 1760.	1110.	* *	
* 1200.	1170.	* 90.0	* 1490.	884.	* *	
* 1040.	1010.	* 95.0	* 1320.	746.	* *	
* 841.	791.	* 99.0	* 1100.	568.	* *	
-----*						
* FREQUENCY CURVE STATISTICS *			* STATISTICS BASED ON *			
-----*						
* MEAN LOGARITHM	3.3684	* HISTORIC EVENTS	0	0	* *	
* STANDARD DEVIATION	0.2456	* HIGH OUTLIERS	0		* *	
* COMPUTED SKEW	0.7300	* LOW OUTLIERS	0		* *	
* GENERALIZED SKEW	0.6000	* ZERO OR MISSING	0		* *	
* ADOPTED SKEW	0.7000	* SYSTEMATIC EVENTS	24		* *	
*****						

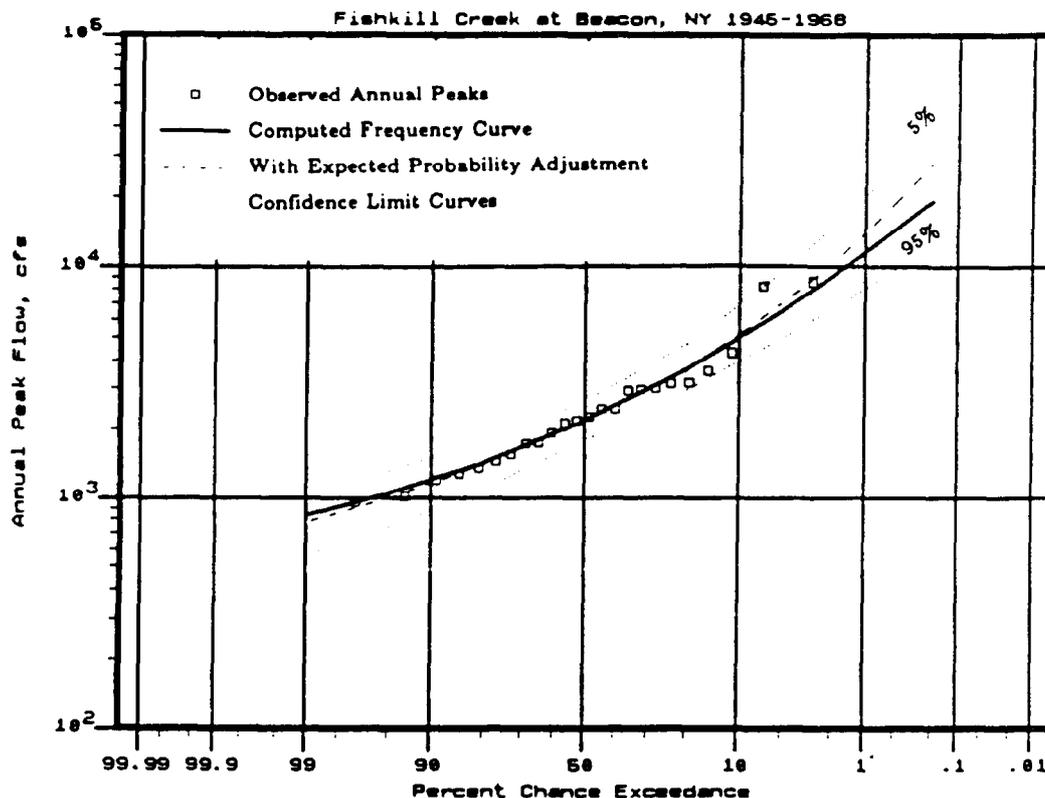


Figure 3-1. Annual Frequency Curve.

d. **Broken Record.** A broken record results when one or more years of annual peaks are missing for any reason not related to the flood magnitude. In other words, the missing events were caused by a random occurrence. The gage may have been temporarily discontinued for budgetary or other reasons. The different segments of the record are added together and analyzed as one record, unless the different parts of the record are considered non-homogeneous. If a portion of the record is missing because the gage was destroyed by a flood or the flood was too low to record, then the observed events should be analyzed as an incomplete record.

e. **Incomplete Record.** An incomplete record can result when some of the peak flow events were either too high or too low. Different analysis procedures are recommended for missing high events and for missing low events. Missing high events may result from the gage being out of operation or the stage exceeding the rating table. In these cases, every effort should be made to obtain an estimate of the missing events. Missing low floods usually result when the flood height is below the minimum reporting level or the bottom of a crest stage gage. In these cases, the record should be analyzed using the conditional probability adjustment described in Appendix 5 of Bulletin 17B and Section 3-6 of this report.

f. Zero-flood years. Some of the gaging stations in arid regions record no flow for the entire year. A zero flood peak precludes the normal statistical analysis because the logarithm of zero is minus infinity. In this case the record should be analyzed using the conditional probability adjustment described in Appendix 5 of Bulletin 17B and Section 3-6 of this report.

g. Outliers.

(1) Guidance. The Bulletin 17B (46) defines outliers as "data points which depart significantly from the trend of the remaining data." The sequence of steps for testing for high and low outliers is dependent upon the skew coefficient and the treatment of high outliers differs from that of low outliers. When the computed (station) skew coefficient is greater than +0.4, the high-outlier test is applied first and the adjustment for any high outliers and/or historic information is made before testing for low outliers. When the skew coefficient is less than -0.4, the low-outlier test is applied first and the adjustment for any low outlier(s) is made before testing for high outliers and adjusting for any historic information. When the skew coefficient is between -0.4 and +0.4, both the high- and low-outlier tests are made to the systematic record (minus any zero flood events) before any adjustments are made.

(2) Equation. The following equation is used to screen for outliers:

$$\bar{X}_o = X \pm K_N S \quad (3-5)$$

where:

$X_o$  = outlier threshold in log units

$\bar{X}$  = mean logarithm (may have been adjusted for high or low outliers, and/or historical information depending on skew coefficient)

$S$  = standard deviation (may be adjusted value)

$K_N$  = K value from Appendix 4 of Bulletin 17B or Appendix F, Table 11 of this report. Use plus value for high-outlier threshold and minus value for low-outlier threshold

$N$  = Sample size (may be historic period (H) if historically adjusted statistics are used)

(3) High Outliers. Flood peaks that are above the upper threshold are treated as high outliers. The one or more values that are determined to be high outliers are weighted by the historical adjustment equations. Therefore, for any flood peak(s) to be weighted as high outlier(s), either historical information must be available or the probable occurrence of the event(s) estimated based on flood information at nearby sites. If it is not possible to obtain any information that weights the high outlier(s) over a longer period than that of the systematic record, then the outlier(s) should be retained as part of the systematic record.

(4) Low Outliers. Flood peaks that are below the low threshold value are treated as low outliers. Low outliers are deleted from the record and the frequency curve computed by the conditional probability adjustment (Section 3-6). If there are one or more values very near, but above the threshold value, it may be desirable to test the sensitivity of the results by considering the value(s) as low outlier(s).

h. Historic Events and Historical Information.

(1) Definitions. Historic events are large flood peaks that occurred outside of the systematic record. Historical information is knowledge that some flood peak, either systematic or historic, was the largest event over a period longer than that of the systematic record. It is historical information that allows a high outlier to be weighted over a longer period than that of the systematic record.

(2) Equations. The adjustment equations are applied to historic events and high outliers at the same time. It is important that the lowest historic peak be a fairly large peak, because every peak in the systematic record that is equal to or larger than the lowest historic peak must be treated as a high outlier. Also a basic assumption in the adjusting equations is that no peaks higher than the lowest historic event or high outlier occurred during the unobserved part of the historical period. Appendix D in this manual is a reprint of Appendix 6 from Bulletin 17B and contains the equations for adjusting for historic events and/or historical information.

3-3. Weighted Skew Coefficient.

a. General. It can be demonstrated, either through the theory of sampling distributions or by sampling experiments, that the skew coefficient computed from a small sample is highly unreliable. That is, the skew coefficient computed from a small sample may depart significantly from the true skew coefficient of the population from which the sample was drawn. Consequently, the skew coefficient must be compared with other representative data. A more reliable estimate of the skew coefficient of annual flood peaks can be obtained by studying the skew characteristics of all available streamflow records in a fairly large region and weighting the computed skew coefficient with a generalized skew coefficient. (Chapter 9 provides guidelines for determining generalized skew coefficients.)

b. Weighting Equation. Bulletin 17B recommends the following weighting equation:

$$G_w = \frac{MSE_{\bar{G}}(G) + MSE_G(\bar{G})}{MSE_{\bar{G}} + MSE_G} \quad (3-6)$$

where:

- $G_w$  = weighted skew coefficient
- $G$  = computed (station) skew
- $\bar{G}$  = generalized skew
- $MSE_{\bar{G}}$  = mean-square error of generalized skew
- $MSE_G$  = mean-square error of computed (station) skew

c. Mean Square Error.

(1) The mean-square error of the computed skew coefficient for log-Pearson type III random variables has been obtained by sampling experiments. Equation 6 in Bulletin 17B provides an approximate value for the mean-square error of the computed (station) skew coefficient:

$$MSE_G \approx 10^{(A-B[\log_{10}(N/10)])} \quad (3-7a)$$

$$\approx 10^{A+B}/N^B \quad (3-7b)$$

$$A = -0.33 + 0.08 |G| \text{ if } |G| \leq 0.90$$

$$= -0.52 + 0.30 |G| \text{ if } |G| > 0.90$$

$$B = 0.94 - 0.26 |G| \text{ if } |G| \leq .50$$

$$= 0.55 \quad \text{if } |G| > 1.50$$

where:

|G| = absolute value of the computed skew

N = record length in years

Appendix F-10 provides a table of mean-square error for several record lengths and skew coefficients based on Equation 3-7a.

(2) The mean-square error (MSE) for the generalized skew will be dependent on the accuracy of the method used to develop generalized skew relations. For an isoline map, the MSE would be the average of the squared differences between the computed (station) skew coefficients and the isoline values. For a prediction equation, the square of the standard error of estimate would approximate the MSE. And, if an arithmetic mean of the stations in a region were adopted, the square of the standard deviation (variance) would approximate the MSE.

3-4. Expected Probability.

a. The computation of a frequency curve by the use of the sample statistics, as an estimate of the distribution parameters, provides an estimate of the true frequency curve. (Chapter 8 discusses the reliability and the distribution of the computed statistics.) The fact of not knowing the location of the true frequency curve is termed uncertainty. For the normal distribution, the sampling errors for the mean are defined by the t distribution and the sampling errors for the variance are defined by the chi-squared distribution. These two error distributions are combined in the formation of the non-central t distribution. The non-central t-distribution can be used to construct curves that, with a specified confidence (probability), encompass the true frequency curve. Figure 3-2 shows

the confidence limit curves around a frequency curve that has the following assumed statistics:  $N=10$ ,  $\bar{X}=0.$ ,  $S=1.0$ .

b. If one wished to design a flood protection work that would be exceeded, on the average, only one time every 100 years (one percent chance exceedance), the usual design would be based on the normal standard deviate of 2.326. Notice that there is a 0.5 percent chance that this design level may come from a "true" curve that would average 22 exceedances per 100 years. On the other side of the curve, instead of the expected one exceedance, there is a 99.5 percent chance that the "true" curve would indicate 0.004 exceedances. Note the large number of exceedances possible on the left side of the curve. This relationship is highly skewed towards the large exceedances because the bound on the right side is zero exceedance. A graph of the number of possible "true" exceedances versus the probability that the true curve exceeds this value, Figure 3-3, provides a curve with an area equal to the average (expected) number of exceedances.

c. The design of many projects with a target of 1 exceedance per 100 years at each project and assuming  $N=10$  for each project, would actually result in an average of 2.69 exceedances (see Appendix F-8).

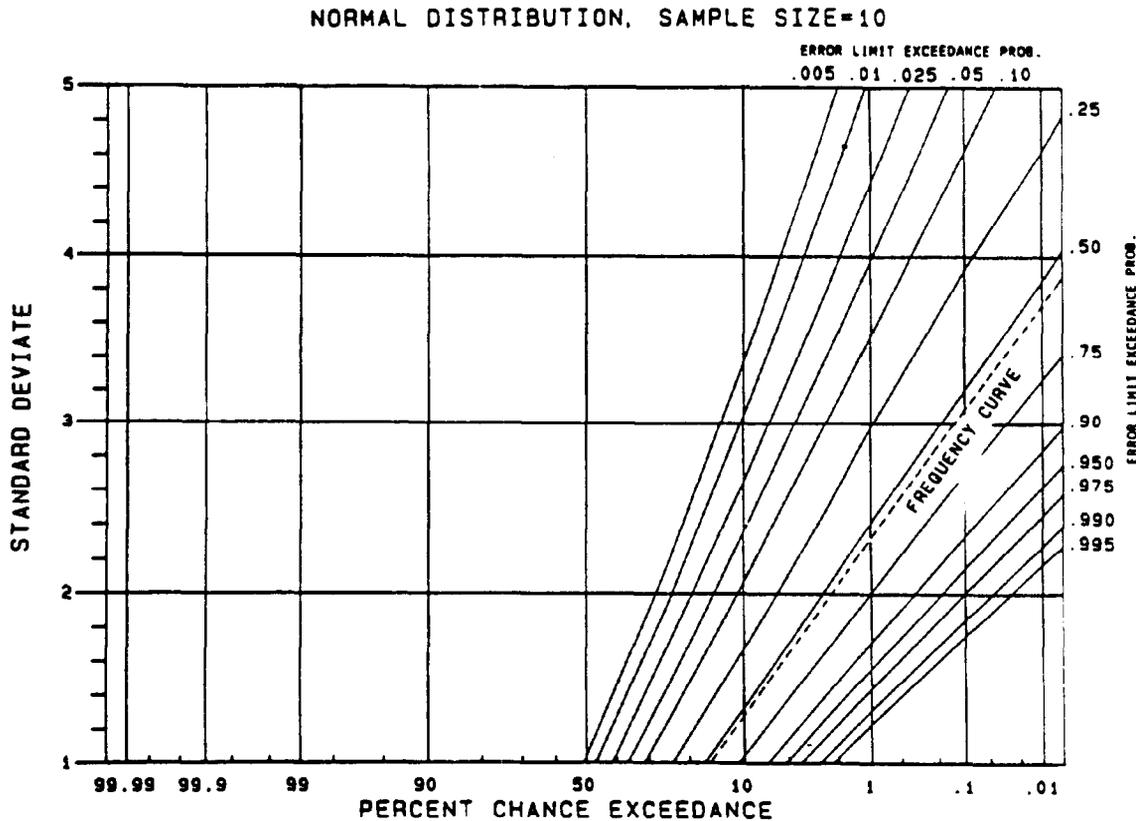


Figure 3-2. Confidence Limit Curves based on the Non-central t Distribution.

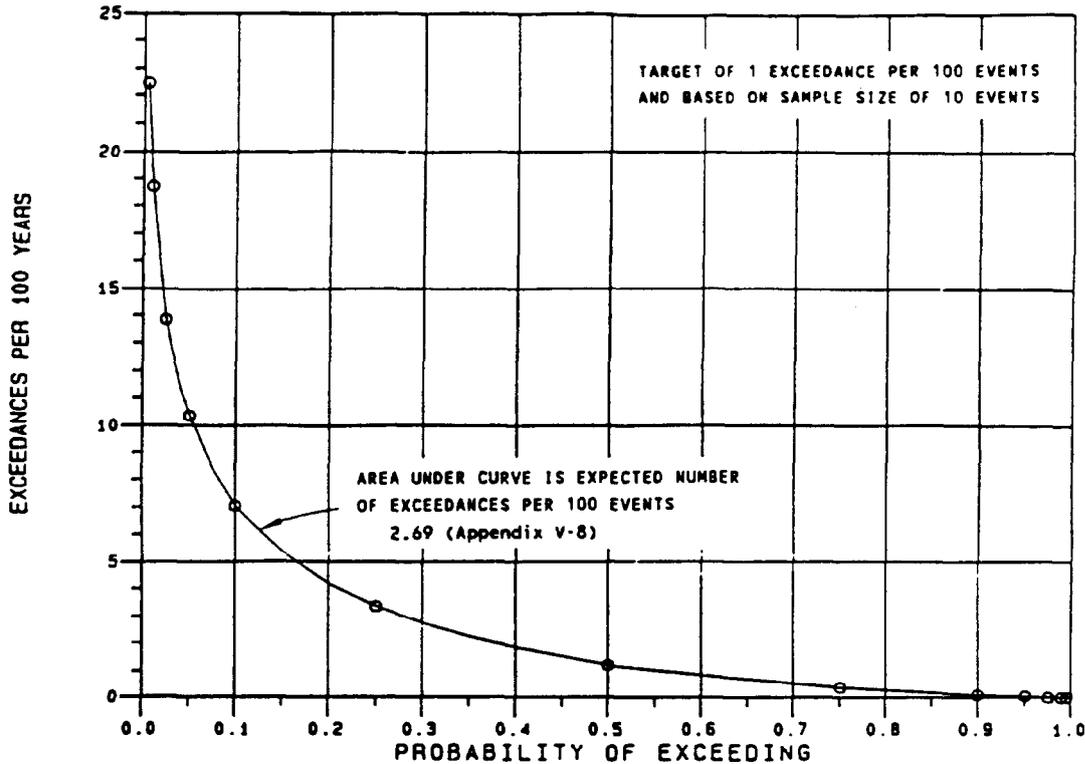


Figure 3-3. Cumulative Probability Distribution of Exceedances per 100 Years.

d. There are two methods that can be used to correct (expected probability adjustment) for this bias. The first method, as described above, entails plotting the curve at the "expected" number of exceedances rather than at the target value, drawing the new curve and then reading the adjusted design level. Appendix F-8 provides the percentages for the expected probability adjustment.

e. The second method is more direct because an adjusted deviate (K value) is used in Equation 3-4 that makes the expected probability adjustment for a given percent chance exceedance. Appendix F-7 contains the deviates for the expected probability adjustment. These values may be derived from the t-distribution by the following equation:

$$K_{p,N} = t_{p, N-1} [(N+1)/N]^{1/2} \quad (3-8)$$

where:

P = exceedance probability (percent chance exceedance divided by 100)

N = sample size

K = expected probability adjusted deviate

t = Student's t-statistic from one-tailed distribution

f. For a sample size of 10 and a 1% percent chance exceedance, the expected probability adjusted deviate is 2.959 as compared to the value of 2.326 used to derive the computed frequency curve.

g. As mentioned in the first paragraph, the non-central t distribution, and consequently the expected probability adjustment, is based on the normal distribution. The expected probability adjustment values in Appendices F-7 and F-8 are considered applicable to Pearson type III distributions with small skew coefficients. The phrase "small skew coefficients" is usually interpreted as being between -0.5 to +0.5. Note also that the uncertainty in the skew coefficient is not considered. In other words, the skew coefficient is treated as if it were the population skew coefficient.

h. The expected probability adjustment can be applied to frequency curves derived by other than analytical procedures if the equivalent worth (in years) of the procedure can be computed or estimated.

### 3-5. Risk.

a. Definition. The term risk is usually defined as the possibility of suffering loss or injury. In a hydrologic context, risk is defined as "the probability that one or more events will exceed a given flood magnitude within a specified period of years" (46). Note that this narrower definition includes a time specification and assumes that the annual exceedance frequency is exactly known. Uncertainty is not taken into account in this definition of risk. Risk then enables a probabilistic statement to be made about the chances of a particular location being flooded when it is occupied for a specified number of consecutive years. The percent chance of the location being flooded in any given year is assumed to be known.

b. Binomial Distribution. The computation of risk is accomplished by the equation for the binomial distribution:

$$R_I = \frac{N!}{I!(N-I)!} P^I(1-P)^{N-I} \quad (3-9)$$

where:

$R_I$  = risk (probability) of experiencing exactly I flood events

N = number of years (trials)

I = number of flood events (successes)

P = exceedance probability, percent chance exceedance divided by 100, of the annual event (probability of success)

(The terms in parentheses are those usually used in statistical texts)

When I equals zero (no floods), Equation 3-9 reduces to:

$$R_0 = (1-P)^N \quad (3-10a)$$

and the probability of experiencing one or more floods is easily computed by taking the complement of the probability of no floods:

$$R_{(1 \text{ or more})} = 1-(1-P)^N \quad (3-10b)$$

c. Application.

(1) Risk is an important concept to convey to those who are or will be protected by flood control works. The knowledge of risk alerts those occupying the flood plain to the fact that even with the protection works, there could be a significant probability of being flooded during their lifetime. As an example, if one were to build a new house with the ground floor at the 1% chance flood level, there is a fair (about one in four) chance that the house will be flooded before the payments are completed, over the 30-year mortgage life. Using Equation 3-10b:

$$\begin{aligned} R_{(1 \text{ or more})} &= 1-(1-.01)^{30} \\ &= 1-.99^{30} \\ &= 1-.74 \\ &= .26 \text{ or } 26\% \text{ chance} \end{aligned}$$

(2) Appendix F-12 provides a table for risk as a function of percent chance exceedance, period length and number of exceedances. This table could also be used to check the validity of a derived frequency curve. As an example, if a frequency curve is determined such that 3 observed events have exceeded the derived 1% chance exceedance level during the 50 years of record, then there would be reason to question the derived frequency curve. From Appendix F-12, the probability of this occurring is 0.0122 or about 1%. It is possible for the situation to occur, but the probability of occurring is very low. This computation just raises questions about the validity of the derived curve and indicates that other validation checks may be warranted before adopting the derived curve.

3-6. Conditional Probability Adjustment. The conditional probability adjustment is made when flood peaks have either been deleted or are not available below a specified truncation level. This adjustment will be applied when there are zero flood years, an incomplete record or low outliers. As stated in Appendix 5 of Bulletin 17B, this procedure is not appropriate when 25 percent or more of the events are truncated. The computation steps in the conditional probability adjustment are as follows:

1. Compute the estimated probability ( $\bar{P}$ ) that an annual peak will exceed the truncation level:

$$\bar{P} = N/n \quad (3-11a)$$

where N is the number of peaks above the truncation level and n is the total number of years of record. If the statistics reflect the adjustments for historic information, then the appropriate equation is

$$\tilde{P} = \frac{H - WL}{H} \quad (3-11b)$$

where H is the length of historic period, W is the systematic record weight and L is the number of peaks truncated.

2. The computed frequency curve is actually a conditional frequency curve. Given that the flow exceeds the truncation level, the exceedance frequency for that flow can be estimated. The conditional exceedance frequencies are converted to annual frequencies by the probability computed in Step 1:

$$P = \tilde{P} P_d \quad (3-12)$$

where P is the annual percent chance exceedance and  $P_d$  is the conditional percent chance exceedance.

3. Interpolate either graphically or mathematically to obtain the discharge values ( $Q_p$ ) for 1, 10 and 50 percent chance exceedances.

4. Estimate log-Pearson type III statistics that will fit the upper portion of the adjusted curve with the following equations:

$$G_s = -2.50 + 3.12 \frac{\log(Q_1/Q_{10})}{\log(Q_{10}/Q_{50})} \quad (3-13)$$

$$S_s = \frac{\log(Q_1/Q_{50})}{K_1 - K_{50}} \quad (3-14)$$

$$X_s = \log(Q_{50}) - K_{50} S_s \quad (3-15)$$

where  $G_s$ ,  $S_s$  and  $\bar{X}_s$  are the synthetic skew coefficient, standard deviation and mean, respectively;  $Q_1$ ,  $Q_{10}$  and  $Q_{50}$  the discharges determined in Step 3; and  $K_1$  and  $K_{50}$  are the Pearson Type III deviates for percent change exceedances of 1 and 50 and skew coefficient  $G_s$ .

5. Combine the synthetic skew coefficient with the generalized skew by use of Equation 3-6 to obtain the weighted skew.
6. Develop the computed frequency curve with the synthetic statistics and compare it with the plotted observed flood peaks.

### 3-7. Two-Station Comparison.

#### a. Purpose.

(1) In most cases of frequency studies of runoff or precipitation there are locations in the region where records have been obtained over a long period. The additional period of record at such a nearby station is useful for extending the record at a short record station provided there is reasonable correlation between recorded values at the two locations.

(2) It is possible, by regression or other techniques, to estimate from concurrent records at nearby locations the magnitude of individual missing events at a station. However, the use of regression analysis produces estimates with a smaller variance than that exhibited by recorded data. While this may not be a serious problem if only one or two events must be estimated to "fill in" or complete an otherwise unbroken record of several years, it can be a significant problem if it becomes necessary to estimate more than a few events. Consequently, in frequency studies, missing events should not be freely estimated by regression analysis.

(3) The procedure for adjusting the statistics at a short-record station involves three steps: (1) computing the degree of correlation between the two stations, (2) using the computed degree of correlation and the statistics of the longer record station to compute an adjusted set of statistics for the shorter-record station, and (3) computing an equivalent "length of record" that approximately reflects the "worth" of the adjusted statistics of the short-record station. The longer record station selected for the adjustment procedure should be in a hydrologically similar area and, if possible, have a drainage area size similar to that of the short-record station.

b. Computation of Correlation. The degree of correlation is reflected in the correlation coefficient  $R^2$  as computed through use of the following equation:

$$R^2 = \frac{[\sum XY - (\sum X \sum Y)/N]^2}{[\sum X^2 - (\sum X)^2/N][\sum Y^2 - (\sum Y)^2/N]} \quad (3-16)$$

where:

- $R^2$  = the determination coefficient
- Y = the value at the short-record station
- X = the concurrent value at the long-record station
- N = the number of years of concurrent record

For most studies involving streamflow values, it is appropriate to use the logarithms of the values in the equations in this section.

c. Adjustment of Mean. The following equation is used to adjust the mean of a short-record station on the basis of a nearby longer-record station:

$$\bar{Y} = \bar{Y}_1 + (\bar{X}_3 - \bar{X}_1) R (S_{Y_1}/S_{X_1}) \quad (3-17)$$

where:

$\bar{Y}$  = the adjusted mean at the short-record station

$\bar{Y}_1$  = the mean for the concurrent record at the short-record station

$\bar{X}_3$  = the mean for the complete record at the longer-record station

$\bar{X}_1$  = the mean for the concurrent record at the longer-record station

R = the correlation coefficient

$S_{Y_1}$  = the standard deviation for the concurrent record at the short-record station

$S_{X_1}$  = the standard deviation for the concurrent record at the longer-record station

All of the above parameters may be derived from the logarithms of the data where appropriate, e.g., for annual flood peaks. The criterion for determining if the variance of the adjusted mean will likely be less than the variance of the concurrent record is:

$$R^2 > 1/(N_1 - 2) \quad (3-18)$$

where  $N_1$  equals the number of years of concurrent record. If  $R^2$  is less than the criterion, Equation 3-17 should not be applied. In this case just use the computed mean at the short-record station or check another nearby long-record station. See Appendix 7 of Bulletin 17B for procedures to compare the variance of the adjusted mean against the variance of the entire short-record period.

d. Adjustment of Standard Deviation. The following equation can be used to adjust the standard deviation:

$$S_Y^2 = S_{Y_1}^2 + (S_X^2 - S_{X_1}^2) R^2 (S_{Y_1}^2/S_{X_1}^2) \quad (3-19)$$

(approximate)

where:

- $S_y$  = the adjusted standard deviation at the short-record station
- $S_{y_1}$  = the standard deviation for the period of concurrent record at the short-record station
- $S_x$  = the standard deviation for the complete record at the base station
- $S_{x_1}$  = the standard deviation for the period of concurrent record at the base station
- $R^2$  = the determination coefficient

All of the above parameters may be derived from the logarithms of the data where appropriate, e.g., for annual flood peaks. This equation provides approximate results compared to Equation 3-19 in Appendix 7 of Bulletin 17B, but in most cases the difference in the results does not justify the additional computations.

e. Adjustment of Skew coefficient. There is no equation to adjust the skew coefficient that is comparable to the above equations. When adjusting the statistics of annual flood peaks either a weighted or a generalized skew coefficient may be used depending on the record length.

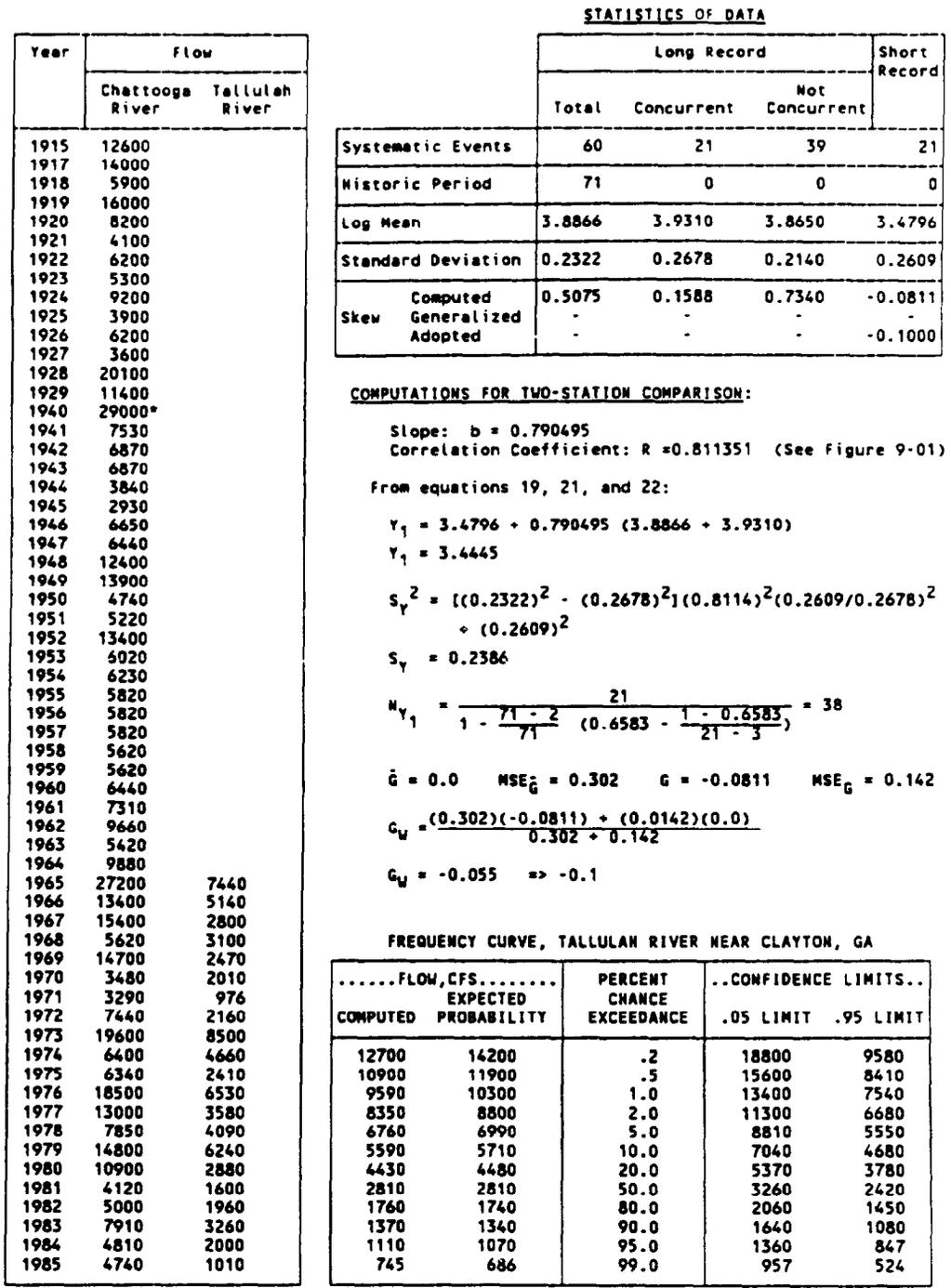
f. Equivalent Record Length. The final step in adjusting the statistics is the computation of the "equivalent record length" which is defined as the period of time which would be required to establish unadjusted statistics that are as reliable (in a statistical sense) as the adjusted values. Thus, the equivalent length of record is an indirect indication of the reliability of the adjusted values of  $Y$  and  $S_y$ . The equivalent record length for the adjusted mean is computed from the following equation:

$$N_y = \frac{N_{y_1}}{1 - [(N_x - N_{y_1})/N_x] [R^2 - (1 - R^2)/(N_{y_1} - 3)]} \quad (3-20)$$

where:

- $N_y$  = the equivalent length of record of the mean at the short-record station
- $N_{y_1}$  = the number of years of concurrent record at the two stations
- $N_x$  = the number of years of record at the longer-record station
- $R$  = the adjusted correlation coefficient

Figure 3-4 shows the data and computations for a two-station comparison for a short record station with 21 events and a long record station with 60 systematic events. It can be seen that the adjustment of the frequency statistics provides an increased reliability in the mean equivalent to having an additional 17 years of record at the short-record station.



\* Historic information, peak largest since 1915.

Figure 3-4. Two-Station Comparison Computations.

5 Mar 93

(Figure 9-1 shows the computations for  $b$  and  $R$  and Figure 9-2 shows the Tallulah River annual peaks plotted against the Chattooga River peaks.) Figure 3-5 shows the resulting unadjusted and adjusted frequency curves based on the computed and adjusted statistics in Figure 3-4. Although  $N_y$  is actually the equivalent years of record for the mean, the value is used as an estimate equivalent record length in the computation of confidence limits and the expected probability adjustment.

g. Summary of Steps. The procedure for computing and adjusting frequency statistics using a longer-record station can be summarized as follows:

- (1) Arrange the streamflow data by pairs in order of chronological sequence.
- (2) Compute  $\bar{Y}_1$  and  $S_{Y_1}$  for the entire record at the short-record station.
- (3) Compute  $\bar{X}$  and  $S_X$  for the entire record at the longer-record station.
- (4) Compute  $\bar{X}_1$  and  $S_{X_1}$  for the portion of the longer-record station which is concurrent with the short-record station.
- (5) Compute the correlation coefficient using Equation 3-16.
- (6) Compute  $\bar{Y}$  and  $S_Y$  for the short-record station using Equations 3-17 and 3-18.

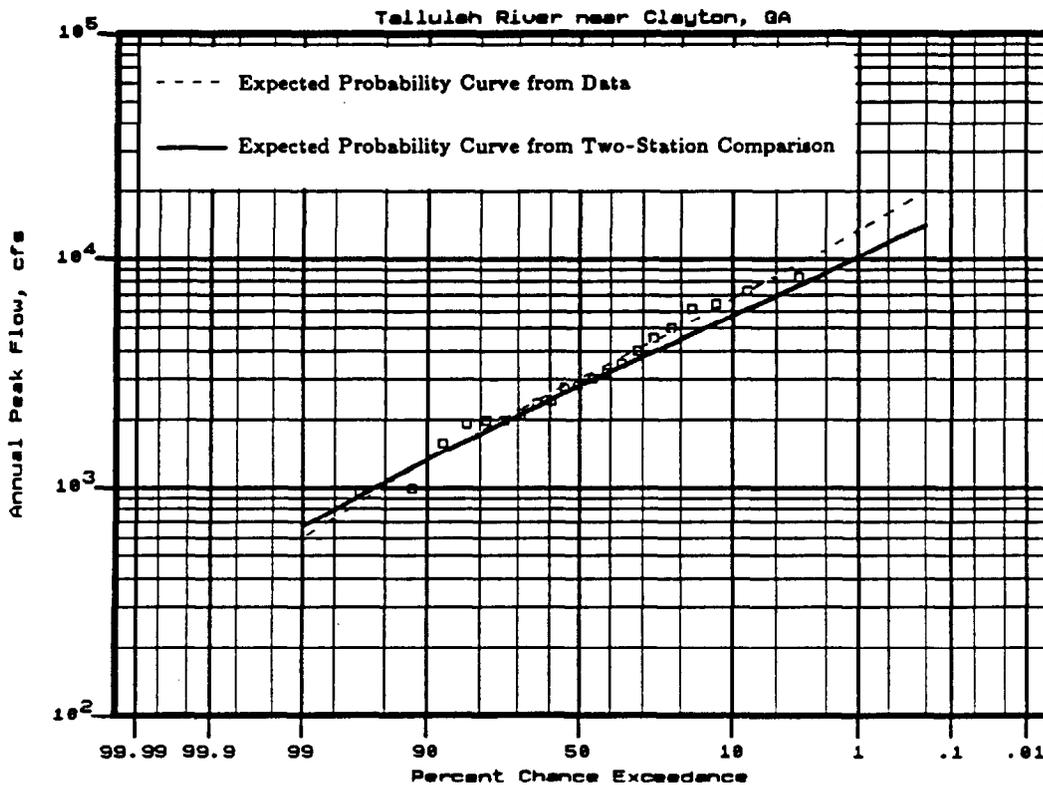


Figure 3-5. Observed and Two-Station Comparison Frequency Curves.

- (7) Calculate the equivalent length of record of the mean for the short-record station using Equation 3-20.
- (8) Compute the frequency curve using adjusted values of  $\bar{Y}$  and S in Equation 3-4 and K values from Appendix F-2 corresponding to the adopted skew coefficient.
- (9) Compute the expected probability adjustment and the confidence limits.

3-8. Flood Volumes.

a. Nature and Purpose. Flood volume frequency studies involve frequency analysis of maximum runoff within each of a set of specified durations. Flood volume-duration data normally obtained from the USGS WATSTORE files consists of data for 1, 3, 7, 15, 30, 60, 90, 120, and 183 days. These same values are the default values in the HEC computer program STATS (Table 3-2). Runoff volumes are expressed as average flows in order that peak flows and volumes can be readily compared and coordinated. Whenever it is necessary to consider flows separately for a portion of the water year such as the rain season or snowmelt season, the same durations (up to the 30-day or 90-day values) are selected from flows during that season only. Flood volume-duration curves are used primarily for reservoir design and operation studies, and should generally be developed in the design of reservoirs having flood control as a major function.

Table 3-2. High Flow Volume-Duration Data

- VOLUME-DURATION DATA - FISHKILL CR AT BEACON, NY - DAILY FLOWS

YEAR	HIGHEST MEAN VALUE FOR DURATION, FLOW, CFS								
	1	3	7	15	30	60	90	120	183
1945	2080.0	1936.7	1714.3	1398.7	1106.8	752.3	742.2	649.4	559.2
1946	1360.0	1180.3	923.0	837.3	657.8	605.3	476.2	451.5	379.9
1947	1800.0	1616.7	1159.1	820.5	687.1	611.9	558.5	485.8	396.4
1948	2660.0	2430.0	2322.9	1641.7	1145.1	862.0	706.2	638.1	512.7
1949	2900.0	2346.7	1715.7	1358.9	888.9	680.7	586.8	522.4	422.4
1950	1050.0	909.7	746.9	639.7	588.1	455.9	423.0	387.2	335.1
1951	2160.0	1886.7	1744.3	1248.1	872.9	832.1	781.2	689.8	568.9
1952	2870.0	2266.7	1557.6	1186.5	1032.8	925.1	854.1	732.6	692.9
1953	2850.0	2233.3	1644.3	1317.2	1145.5	994.6	831.1	794.4	654.5
1954	1520.0	1086.7	811.7	620.4	482.9	397.0	405.7	372.7	348.1
1955	6970.0	4536.7	2546.1	1360.0	758.2	608.0	494.0	463.1	478.7
1956	6760.0	5456.7	3354.3	1959.7	1572.8	1080.9	767.7	635.8	641.7
1957	1230.0	1117.3	1037.7	758.9	524.2	408.8	363.3	373.4	324.4
1958	2130.0	1916.7	1587.1	1354.5	1128.1	872.0	848.2	777.8	654.1
1959	1670.0	986.7	782.1	586.6	517.6	466.7	437.5	398.8	346.2
1960	2080.0	1770.0	1374.3	1046.9	712.3	605.5	530.5	515.1	468.4
1961	3440.0	2966.7	2155.7	1590.2	1152.3	845.2	759.5	656.2	491.4
1962	2570.0	2070.0	1547.7	1105.0	857.7	600.9	461.3	429.4	325.0
1963	1730.0	1616.7	1309.0	1216.0	900.8	569.1	438.0	370.8	305.9
1964	1300.0	1106.7	945.3	737.8	541.2	514.8	486.6	450.1	368.3
1965	900.0	826.3	652.6	455.7	375.8	303.3	275.7	235.0	175.0
1966	930.0	774.7	693.3	546.5	445.7	352.5	296.2	272.5	209.0
1967	1520.0	1416.7	1247.1	1023.5	906.8	701.3	581.4	521.1	436.8
1968	3500.0	2810.0	1934.3	1328.5	878.7	611.7	609.5	567.3	460.3

Note - Data based on water year of October 1 of preceding year through September 30 of given year.

b. Data for Comprehensive Series. Data to be used for a comprehensive flood volume-duration frequency study should be selected from nearly complete water year records. Unless overriding reasons exist, the durations in Table 3-2 should be used in order to assure consistency among various studies for comparison purposes. Maximum flood events should be selected only for those years when recorder gages existed or when the maximum events can be estimated by other means. Where a minor portion of a water year's record is missing, the longer-duration flood volumes for that year can often be estimated adequately. If upstream regulation or diversion is known to have an effect, care should be exercised to assure that the period selected is the one when flows would have been maximum under the specified (usually natural) conditions.

c. Statistics for Comprehensive Series.

(1) The probability distribution recommended for flood volume-duration frequency computations is the log-Pearson type III distribution; the same as that used for annual flood peaks. In practice, only the first two moments, mean and standard deviations are based on station data. As discussed in Section 3-3, the skew coefficient should not be based solely on the station record, but should be weighted with information from regional studies. To insure that the frequency curves for each duration are consistent, and especially to prevent the curves from crossing, it is desirable to coordinate the variation in standard deviation and skew with that of the mean. This can be done graphically as shown in Figure 3-6. For a given skew coefficient, there is a maximum and minimum allowable slope for the standard deviation-versus-mean relation which prevents the curves from crossing within the established limits. For instance, to keep the curves from crossing within 99.99 and .01 percent chance exceedances with a skew of 0., the slope must not exceed .269, nor be less than -.269, respectively. The value of this slope constraint is found by stating that the value of one curve ( $X_A$  for curve A) must equal or exceed the value for a second curve ( $X_B$  for curve B) at the desired exceedance frequency. Each of these values can be found by substitution into Equation 3-4 (the K for zero skew and 99.99 percent chance exceedance is -3.719):

$$\begin{aligned} X_A &\geq X_B \\ \bar{X}_A + (-3.719) S_A &\geq \bar{X}_B + (-3.719) S_B \\ 3.719 (S_B - S_A) &\geq \bar{X}_B - \bar{X}_A \\ \frac{(S_B - S_A)}{(\bar{X}_B - \bar{X}_A)} &\geq 0.269 \end{aligned}$$

where:

- $X_A$  = Value of frequency curve A at 99.99 percent chance exceedance
- $X_B$  = Value of frequency curve B at 99.99 percent chance exceedance
- $\bar{X}_A$  = Mean of frequency curve A
- $\bar{X}_B$  = Mean of frequency curve B
- $S_A$  = Standard deviation of frequency curve A
- $S_B$  = Standard deviation of frequency curve B

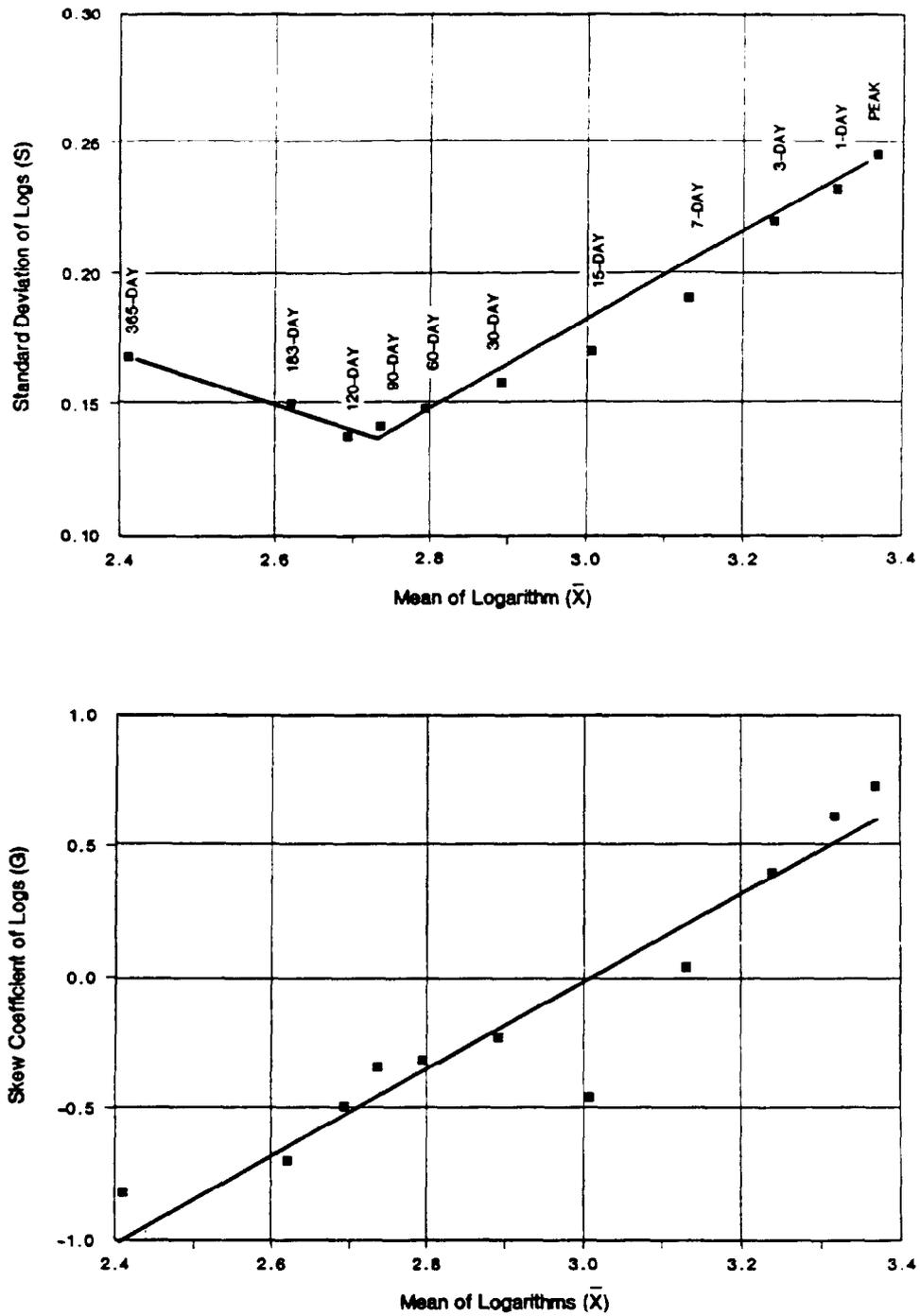


Figure 3-6. Coordination of Flood-Volume Statistics.

(2) When the skew changes between durations, it is probably easiest to adopt smoothed relations for the standard deviation and skew and input the statistics into a computer program that computes the ordinates. The curves can then be inspected for consistency.

(3) If the statistics for the peak flows have been computed according to the procedures in Bulletin 17B, the smoothing relations should be forced through those points. The procedure for computing a least-squares line through a given intersection can be found in texts describing regression analyses.

d. Frequency Curves for Comprehensive Series.

(1) General Procedure. Frequency curves of flood volumes are computed analytically using general principles and methods of Chapters 2 and 3. They should also be shown graphically and compared with the data on which they are based. This is a general check on the analytic work and will ordinarily reveal any inconsistency in data and methodology. The computed frequency curves and the observed data should be plotted on a single sheet for comparison purposes, Figure 3-7.

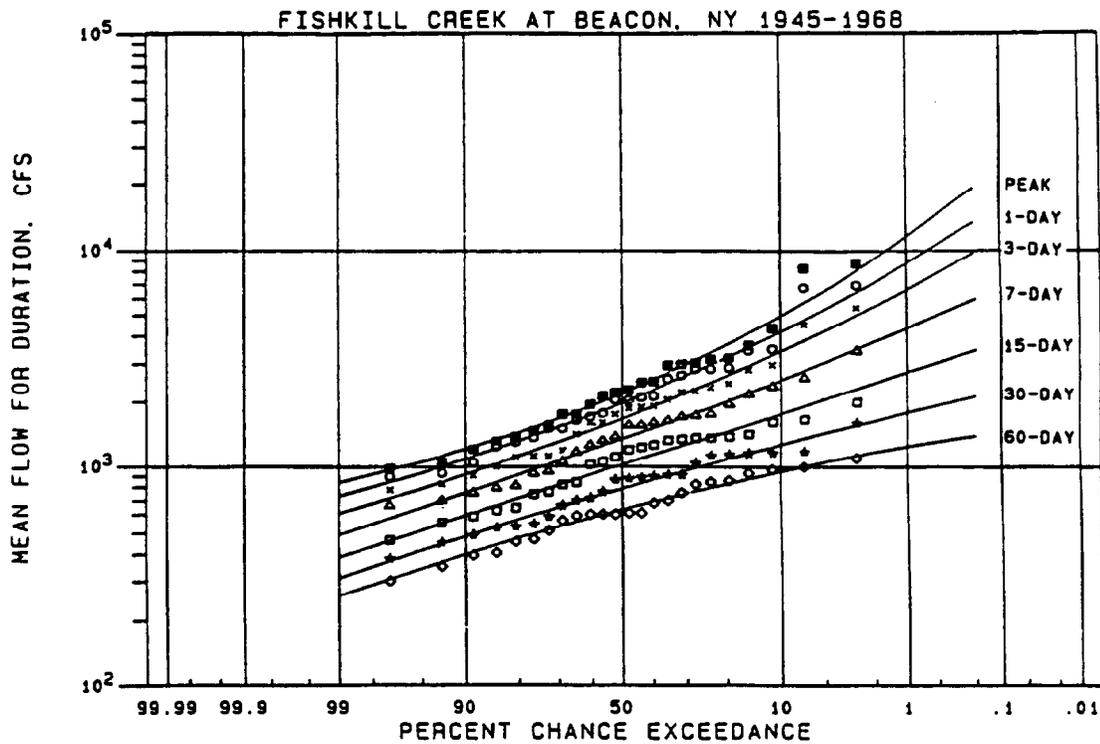


Figure 3-7. Flood-Volume Frequency Curves.

(2) Interpolation Between Fixed Durations. The runoff volume for any specified frequency can be determined for any duration between 1-day and 365-days by drawing a curve on logarithmic paper relating mean discharge (or volume) to duration for that specified frequency (see Figure 3-8a). When runoff volumes for durations shorter than 24 hours are important, special frequency studies should be made. These could be done in the same manner as for the longer durations, using skew coefficients interpolated in some reasonable manner between those used for peak and 1-day flows.

e. Applications of Flood Volume-Duration Frequencies.

(1) Volume-duration Curves. The use of flood volume-duration frequencies in solving reservoir planning, design, and operation problems usually involves the construction of volume-duration curves for specified frequencies. These are drawn first on logarithmic paper for interpolation purposes, as illustrated on Figure 3-8a. The mean discharge values are multiplied by appropriate durations to obtain volumes and are then replotted on an arithmetic grid as shown on the Figure 3-8b. A straight line on this grid represents a constant rate of flow. The straight line represents a uniform flow of 1,500 cfs, and the maximum departure from the 2% chance exceedance curve demonstrates that a reservoir capacity of 16,000 cfs-days (31,700 acre-feet) is required to control the indicated runoff volumes by a constant release of 1,500 cfs. The curve also indicates that a duration of about 8 days is critical for this project release rate and associated flood-control storage space.

(2) Application to a Single Reservoir. In the case of a single flood-control reservoir located immediately upstream of a single damage center, the volume frequency problems are relatively simple. A series of volume-duration curves, similar to that shown on Figure 3-8, corresponding to selected exceedance frequencies should first be drawn. The project release rate should be determined, giving due consideration to possible channel deterioration, encroachment into the flood plain, and operational contingencies. This procedure can be used not only as an approximate aid in selecting a reservoir capacity, but also as an aid in drawing filling-frequency curves.

(3) Application to a Reservoir System. In solving complex reservoir problems, representative hydrographs at all locations can be patterned after one or more past floods. The ordinates of these hydrographs can be adjusted so that their volumes for the critical durations will equal corresponding magnitudes at each location for the selected frequency. A design or operation scheme based on regulation of such a set of hydrographs would be reasonably well balanced. Some aspects of this problem are described in Section 3-9g.

3-9. Effects of Flood Control Works on Flood Frequencies.

a. Nature of the Problem. Flood control reservoirs are designed to substantially affect the frequency of flood flows (or flood stages) at various downstream locations. Many land use changes such as urbanization, forest clearing, etc. can also have significant effects on downstream flood flows (see Section 3-10). Channel improvements (intended to reduce stages) and levee improvements (intended to confine flows) at specified locations can substantially affect downstream flows by eliminating some of the natural storage effects. Levees can also create backwater conditions that affect river stages for a considerable distance upstream. The degree to which flows and stages are modified by various flood control works or land use changes can depend on the timing, areal distribution and magnitude of rainfall (and snowmelt, if pertinent) causing the flood. Accordingly, the studies should include evaluations of the effects on representative flood

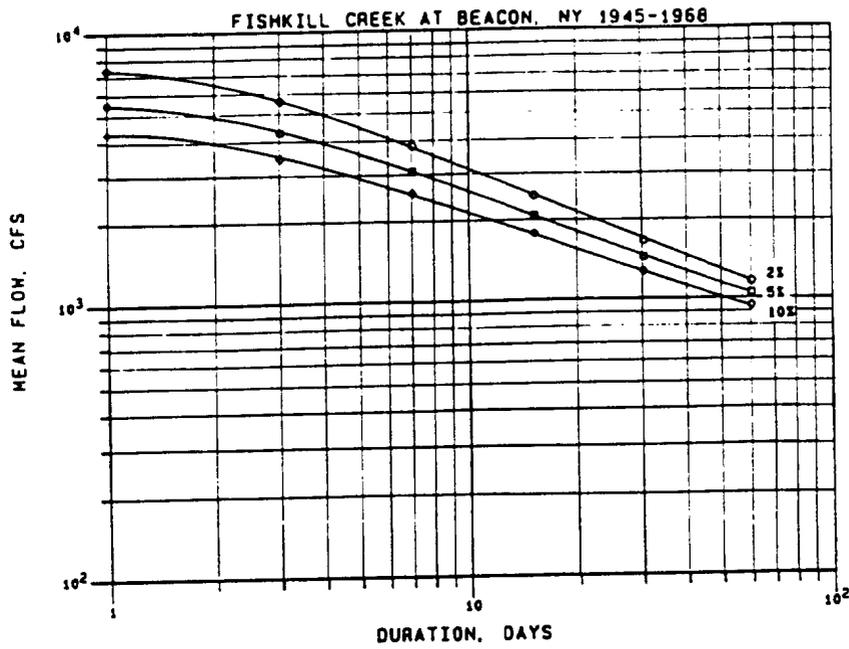


Figure 3-8a. Flood-Volume Frequency Relations.

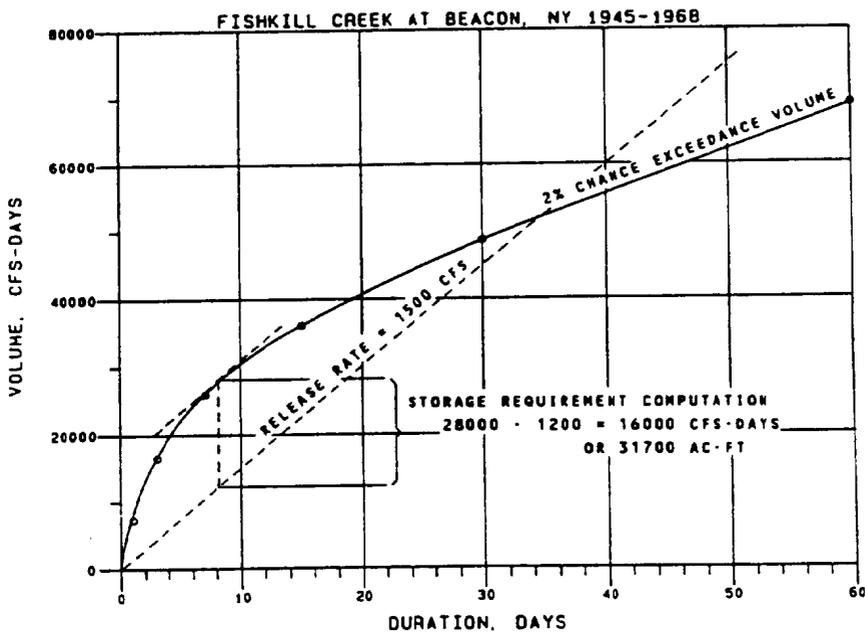


Figure 3-8b. Storage Requirement Determination.

events, with careful consideration given to the effects of different temporal and areal distributions.

b. Terminology.

(1) Natural Conditions. Natural conditions in the drainage basin are defined as hydrologic conditions that would prevail if no regulatory works or other works of man were constructed. Natural conditions, however, include the effects of natural lakes, swamp areas, etc.

(2) Present Conditions. Present or base conditions are defined as the conditions that exist as of the date of the study or some specified time.

(3) Without-Project Conditions. Without-project conditions are defined as the conditions that would exist without the projects under consideration, but with all existing projects and may include future projects whose construction is imminent.

(4) With-Project Conditions. With-project conditions are defined as the conditions that will exist after the projects under consideration are completed.

c. Reservoir-Level Frequency Computation.

(1) Factors to be Considered. Factors affecting the frequency of reservoir levels include historical inflow rates and anticipated future inflow rates estimated by volume-frequency studies, the storage-elevation curves, and the plan of reservoir regulation including location and size of reservoir outlets and spillway. A true frequency curve of annual maxima or minima can only be computed when the reservoir completely fills every year. Otherwise, the events would not be independent. If there is dependence between annual events, the ordinate should be labeled "percent of years exceeded" for maximum events and "percent of years not exceeded" for minimum events.

(2) Computation and Presentation of Results. A frequency curve of annual maximum reservoir elevations (or stages) is ordinarily constructed graphically, using procedures outlined in Section 2-4. Observed elevations (or stages) are used to the extent that these are available, if the reservoir operation will remain the same in the future. Historical and/or large hypothetical floods may also be routed through the reservoir using future operating plans. A typical frequency curve is illustrated on Figure 6-4. Elevation-duration curves are constructed from historical operation data or from routings of historical runoff in accordance with procedures discussed in Section 2-2, Figure 3-9. Such curves may be constructed for the entire period of record or for a selected wet period or dry period. For many purposes, particularly recreation uses, the seasonal variation of reservoir elevation (stages) is important. In this case a set of frequency or duration curves for each month of the year may be valuable. One format for presenting this information is illustrated on Figure 3-10.

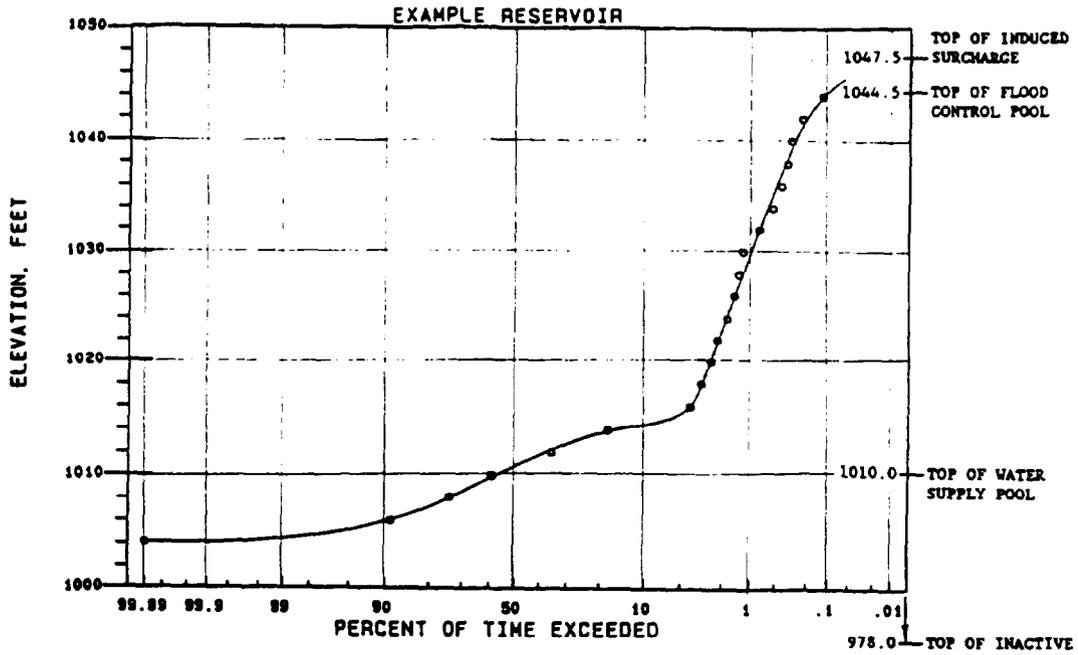


Figure 3-9. Daily Reservoir Elevation-Duration Curve.

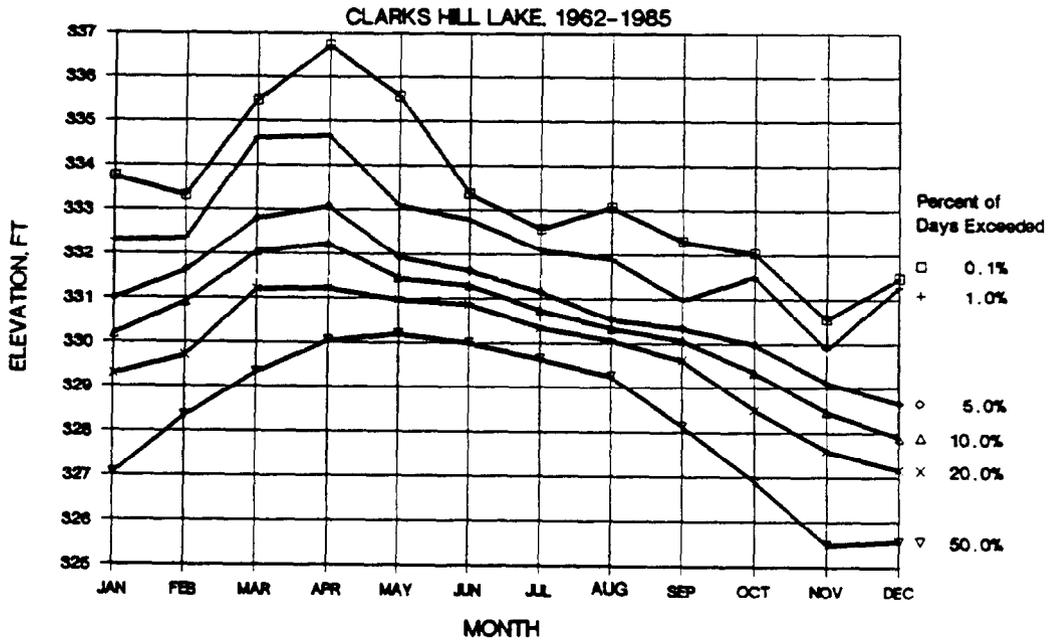


Figure 3-10. Seasonal Variation of Elevation-Duration Relations.

d. Effects of Reservoirs on Flows at Downstream Points.

(1) Routing for Period of Record. The frequency of reservoir outflows or of flows at a downstream location can be obtained from routings of the period-of-record runoff by the following methods:

(a) Determine the annual maximum flow at each location of interest and construct a frequency curve of the regulated flows by graphical techniques (Section 2-4).

(b) Construct a graph of with-project versus without-project flows at the location of interest and draw a curve relating the two quantities as illustrated on Figure 3-11. The points should be balanced in the direction transverse to the curve, but factors such as flood volume of the events and reliability of regulation must be considered in drawing the curve. This curve can be used in conjunction with a frequency curve of without-project flows to construct a frequency curve of with-project flows as illustrated on Figure 3-12. This latter procedure assures consistency in the analysis and gives a graphical presentation of the variability of the regulated events for a given unregulated flow.

(2) Use of Hypothetical-Flood Routings. Usually recorded values of flows are not large enough to define the upper end of the regulated frequency curve. In such cases, it is usually possible to use one or more large hypothetical floods (whose frequency can be estimated from the frequency curve of unregulated flows) to establish the corresponding magnitude of regulated flows. These floods can be multiples of the largest observed floods or of floods computed from rainfall; but it is best not to multiply any one flood by a factor greater than two or three. The floods are best selected or adjusted to represent about equal severity in terms of runoff frequency of peak and volumes for various durations. The routings should be made under reasonably conservative assumptions as to initial reservoir stages.

(3) Incidental Control by Water Supply Space. In constructing frequency curves of regulated flows, it must be recognized that reservoir operation for purposes other than flood control will frequently provide incidental regulation of floods. However, the availability of such space cannot usually be depended upon, and its value is considerably diminished for this reason. Consequently, the effects of such space on the reduction of floods should be estimated very conservatively.

(4) Allowance for Operational Contingencies. In constructing frequency curves of regulated flows, it should be recognized that actual operation is rarely perfect and that releases will frequently be curtailed or diminished because of unforeseen operation contingencies. Also, where flood forecasts are involved in the reservoir operation, it must be recognized that these are subject to considerable uncertainty and that some allowance for uncertainty will be made during operation. In accounting for these factors, it will be found that the actual control of floods is somewhat less than could be expected if full release capacities and downstream channel capacities were utilized efficiently and if all forecasts were exact.

(5) Runoff from Unregulated Areas. In estimating the frequency of runoff at a location that is a considerable distance downstream from one or more reservoir projects, it must be recognized that none of the runoff from the intermediate areas between the reservoir(s) and the damage center will be regulated. This factor can be accounted for by constructing a frequency curve of the runoff from the intermediate area, and using this curve as an indicator of the lower limit for the curve of regulated flows. Streamflow

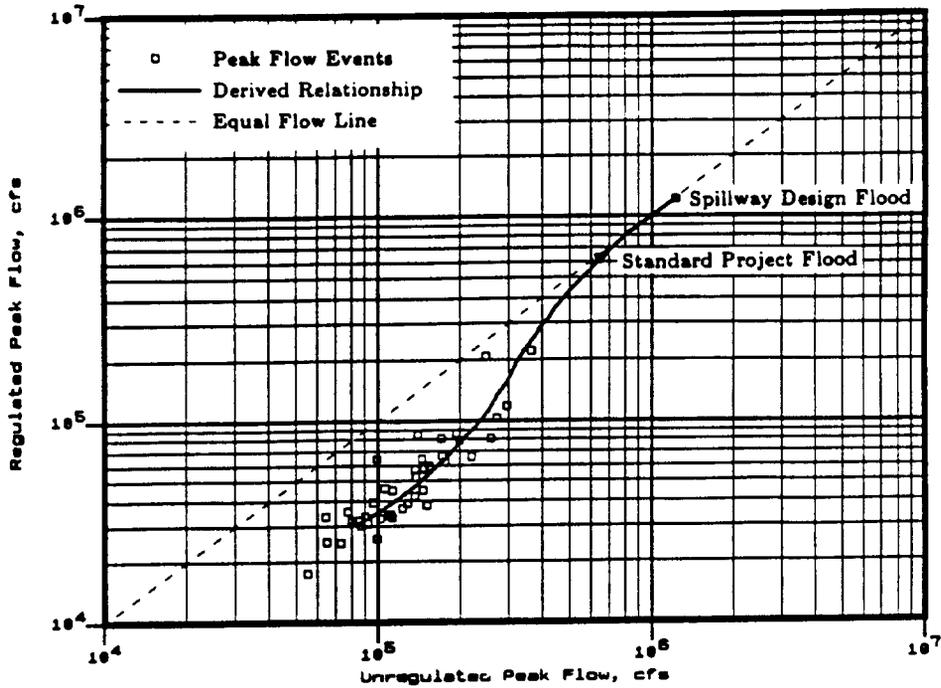


Figure 3-11. Example With-Project versus Without-Project Peak Flow Relations.

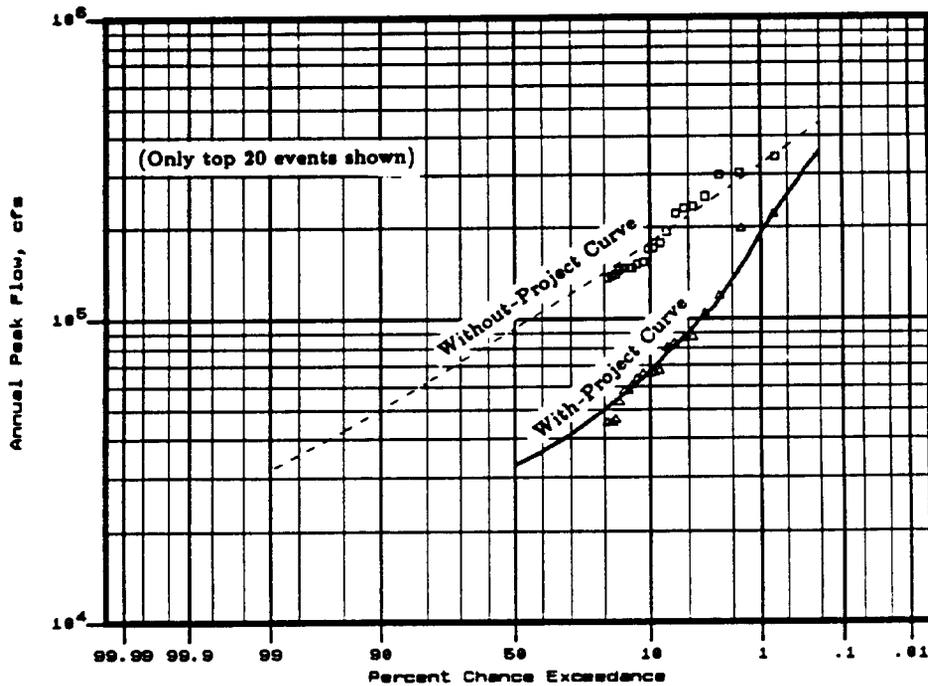


Figure 3-12. Example Without-Project and With-Project Frequency Curves.

routing and combining of both the flows from the unregulated area and those from the regulated area is the best procedure for deriving the regulated frequency curve.

e. Effects of Channel, Levee and Floodway Improvements. The effect of channel, levee and floodway improvements on river stages at the project location and on river discharges downstream from the project location can generally be evaluated by routing several typical floods through the reaches of the improvement and the upstream reaches affected by backwater. The stages or discharges thus derived can be plotted against corresponding without- project values, and a smooth curve drawn. This curve could be used in conjunction with a frequency curve of without-project values to construct a frequency curve of with-project values as discussed in Paragraph 3-09d(1)b. Corresponding stages upstream from the selected control point can be estimated from water-surface profile computations.

f. Changes in Stage-Discharge Relationships. Changes in stage-discharge relations due to channel improvements, levee construction or flow obstructions can best be evaluated by computing theoretical water surface profiles for each of a number of discharges. The resulting relationships for modified conditions can be used to modify routing criteria to enable evaluation of the downstream effects of these changes.

g. Effects of Multiple Reservoir Systems.

(1) Representative events. When more than one reservoir exists above a damage center, the problem of evaluating reservoir stages and downstream flows under project conditions becomes increasingly complex. Whenever practicable, it is best to make complete routings of five to ten historic flood events and a large event that has been developed from a hypothetical rainfall pattern. If necessary, it is possible to supplement these events by using multiples of the flow values. Care must be exercised in selecting events that have representative flood volumes, timings, and areal distributions. Also, there should be a balance of events caused by particular climatic factors, i.e. snowmelt, tropical storm, thunderstorm, etc. Furthermore, the flood-volume-duration characteristics of the hypothetical events should be similar to the recorded events (see Section 3-8). Hypothetical events must be used with caution, however, because certain characteristics of atypical floods may be responsible for critical flooding conditions. Accordingly, such studies should be supplemented by a critical examination of the potential effects of atypical floods.

(2) Computer Program. It is generally impossible to make all of the flood routings necessary to evaluate the effect of a reservoir system by hand computations. Computer programs have been developed to route floods through a reservoir system with complex operational criteria (55).

3-10. Effects of Urbanization.

a. General Effects. Urbanization has two major effects on the watershed which influence the runoff characteristics. First, there is a substantial increase in the impervious area, which results in more water entering the stream system as direct runoff. Second, the drainage system collecting the runoff is generally more efficient and tends to concentrate the water faster in the downstream portion of the channel system. It is important to keep these two effects in mind when considering the changes in the flood peak frequency curve caused by increasing urbanization.

b. Effect on Frequency Relations. A general statement can be made about the effects of urbanization on flood-peak frequency relations. The usual effect on the frequency relation is to cause a significant increase in the magnitude of the more frequent events, but a lesser increase in the less frequent events. This results in an increase in the mean of the annual flood peaks, a decrease in the standard deviation and an unpredictable effect on the skew coefficient (see Figure 3-13). The resulting frequency relation may not fit any of the standard theoretical distributions. Graphical techniques should be applied if a good fit is not possible by an analytical distribution.

c. Other Considerations. The actual effect of urbanization at a specific location is dependent on many factors. Some of the factors that must be considered are basin slope, basin shape, previous land use and ground cover, number of depressional areas drained, magnitude and nature of urban development and the typical flood source (snowmelt, thunderstorm, hurricane, or frontal storm). It is possible for urbanization to cause a decrease in the flood peaks at a particular site. For instance, consider an area downstream of two tributary areas of such size and shape that the large floods are caused by the addition of the nearly coincident peaks from the two tributaries. Urbanization in one of the tributary areas will likely cause the contribution from this area to arrive downstream earlier. This change in the timing of the peaks would result in lower downstream peaks. Of course, when both areas have become equally urbanized, the flood peaks may coincide again. The construction of bridges or other encroachments can reduce the flood peak downstream, but causes backwater flooding upstream.

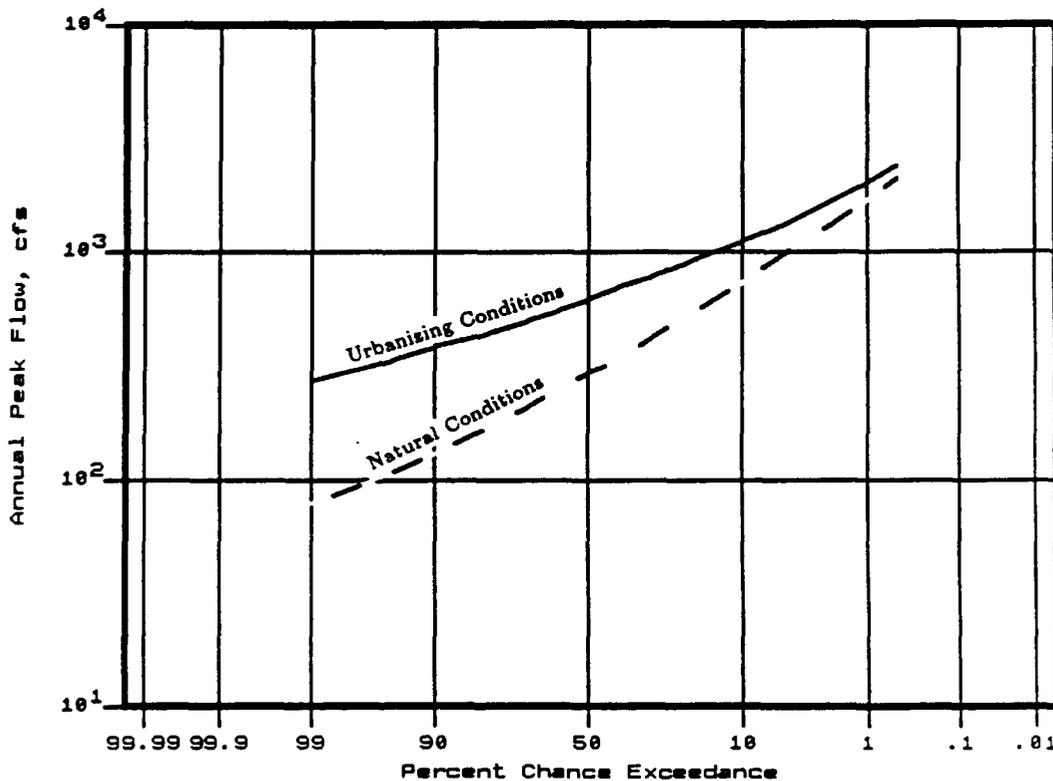


Figure 3-13. Typical Effect of Urbanization on Flood Frequency Curves.

d. Adjustment of a Series of Nonstationary Peak Discharges. When the annual peak discharges have been recorded at the outlet of a basin which has been undergoing progressive urbanization during the period of record, the peak discharges are nonstationary because of the varying basin condition. It is generally necessary to adjust the discharges to a stationary series representative of existing conditions. One approach to adjusting the peaks to a stationary series is as follows:

(1) Develop and calibrate a rainfall-runoff model for existing basin conditions and for conditions at several other points in time during the period of record.

(2) Develop a hypothetical storm for the basin using generalized rainfall criteria, such as that contained in Weather Bureau Technical Paper 40 (14). Select the magnitude of the storm, e.g., a 25-year recurrence interval, to be used. The recurrence interval is arbitrary as it is not assumed in this approach that runoff frequency is equal to rainfall frequency. The purpose of adopting a specific magnitude is to establish a base storm to which ratios can be applied for subsequent steps in the analysis.

(3) Apply several ratios (say 5 to 8) to the hypothetical storm developed in the previous step such that the resulting calculated peak discharges at the gage will cover the range desired for frequency analysis. Input the balanced storms to the rainfall-runoff model for each of the basin conditions selected in step (1), and determine peak discharges at the gaged location.

(4) From the results of step (3), plot curves representing peak discharge versus storm ratio for each basin condition (or point in time).

(5) Use the curves developed in step (4) to adjust the observed annual peak discharges. For example, an observed annual peak discharge that occurred in 1975 is adjusted by entering the "1975" curve (or interpolating) with that discharge, locating the frequency of that event, and reading the magnitude of the adjusted peak from the base-condition curve for the same frequency. The adjusted peak thus obtained is assumed to be the peak discharge that would have occurred for the catchment area and development at the base condition. It is not necessary to adjust to natural conditions. A stationary series could be developed for one or more points in time.

(6) A conventional frequency analysis can be performed on the adjusted peak discharges determined in the preceding step. If the data represent natural conditions, Bulletin 17B procedures would be applicable. If the basin conditions represent significant urbanization, graphical analysis may be appropriate.

e. Development of Frequency Curves at Ungaged Sites. There are several approaches that can be taken to develop frequency curves at ungaged sites that have been subject to urbanization. In order of increasing difficulty, they are: 1) application of simple transfer procedures (e.g.,  $Q = CIA$ ); 2) application of available region-specific criteria, e.g., USGS regression equations; 3) application of rainfall-runoff models to hypothetical storm events; 4) application of simple and detailed rainfall-runoff models with observed storm events and 5) complete period-of-record simulation. As approaches (3) and (4) are often applied, the computational steps are presented in some detail.

(1) Hypothetical Storm Approach

- (a) Develop peak-discharge frequency curve for specific land use conditions from available gaged data and/or regional relationships.
- (b) Develop balanced storms of various frequencies using data from generalized criteria, a nearby gage or the equivalent.
- (c) Develop rainfall-runoff model for the specific watershed with the adopted land-use conditions. Calibrate runoff and routing parameters by reproducing observed hydrographs occurring under natural conditions.
- (d) Input balanced storms (from b) to rainfall-runoff model (from c). Determine exceedance probabilities to associate with balanced storms from adopted specific land-use conditions peak discharge frequency curve (from a) with computed peak discharges.
- (e) Modify parameters of rainfall-runoff model to reflect future urban runoff characteristics. Input balanced storms to the urban- conditions model.
- (f) Plot results assuming frequency of each event is the same for both the adopted land use and the future urban conditions.

(2) "Simple" and Detailed Simulation of Historic Events

- (a) Simulate all major historic events with a relatively simple model to establish the ranking of events and an approximate peak discharge for each. The approximate peaks could be developed by using a multiple linear regression approach, by using a very simple rainfall-runoff model, or by any other approach that will capture the hydrologic response of the basin.
- (b) Perform a conventional frequency analysis of the approximate peaks obtained in step a.
- (c) Make detailed simulations of selected events and correlate the more precise peaks with the approximate peaks.
- (d) Use the relationship developed in step c to determine the desired frequency curve. The same approach can be followed for both existing and future conditions.

## CHAPTER 4

### LOW-FLOW FREQUENCY ANALYSIS

4-1. Uses. Low-flow frequency analyses are used to evaluate the ability of a stream to meet specified flow requirements at a particular location. The analysis can provide an indication of the adequacy of the natural flow to meet a given demand with a stated probability of experiencing a shortage. Additional analyses can indicate the amount of storage that would be required to meet a given demand, again with a stated probability of being deficient. The design of hydroelectric power plants, determination of minimum flow requirements for water quality and/or fish and wildlife, and design of water storage projects can benefit from low-flow frequency analysis.

#### 4-2. Interpretation.

a. Analytical frequency techniques are usually not applicable to low-flow data because most theoretical frequency distributions cannot satisfactorily fit the recorded data. It is recommended that graphical techniques be used and that known geologic and hydrologic conditions be kept in mind when developing the relationships. As the low values are the major interest, the data are arranged with the smallest value first. The probability scale is usually labeled "percent chance nonexceedance."

b. Annual low flows are usually computed for several durations (in days) with the flow rate expressed as the mean flow for the period. For example, the USGS WATSTORE output provides the mean flow values for daily durations of 1, 3, 7, 14, 30, 60, 90, 120 and 183 days. The default values for the HEC program STATS are the same with the exception of using a 15-day duration instead of 14 (Table 4-1). Often a climatic year from April 1 to March 31 is specified to provide a definite separation of the seasonal low-flow periods. Figure 4-1 is a plot of the data in Table 4-1.

#### 4-3. Application Problems.

a. Basin Development. The effects of any basin developments on low flows are usually quite significant. For example, a relatively moderate diversion can be neglected when evaluating flood flow relations, but it would reduce, or even eliminate, low flows. Accordingly, one of the most important aspects of low flows concerns the evaluation of past and future effects of basin developments.

b. Multi-Year Events. In regions of water scarcity and where a high degree of development has been attained, projects that entail carryover of water for several years are often planned. In such projects it is desirable to analyze low-flow volume frequencies for periods ranging from 1-1/2 to 8-1/2 years or more. Because the number of independent low-flow periods of these lengths, in even the longest historical records, is very small and because the concept of multi-annual periods is somewhat inconsistent with the basic concept of an "annual event;" there is no truly satisfactory way for computing the percent chance nonexceedance for low-flow periods that are more than 1 year in length. One procedure described in reference (37) has been used with long sequences of synthetically generated streamflows to derive estimates of drought frequency. Although

Table 4-1. Low-Flow Volume-Duration Data.

- VOLUME-DURATION DATA - FISHKILL CR AT BEACON, NY - DAILY FLOWS

YEAR	LOWEST MEAN VALUE FOR DURATION, FLOW, CFS								
	1	3	7	15	30	60	90	120	183
1945	92.0	104.0	115.1	127.7	143.0	179.5	220.7	254.1	305.3
1946	9.4	12.8	17.6	21.3	28.5	49.8	62.1	58.7	75.4
1947	9.4	12.8	17.3	19.0	21.2	32.1	41.0	62.0	137.0
1948	8.3	10.2	15.7	15.7	18.9	21.6	27.4	33.5	78.1
1949	7.1	8.2	9.0	9.1	10.0	11.3	12.3	14.3	21.2
1950	22.0	22.0	23.9	27.0	32.6	37.0	43.1	51.1	119.0
1951	20.0	33.3	40.9	45.5	58.4	73.2	84.0	88.7	116.4
1952	34.0	39.7	43.0	44.0	46.2	64.6	100.1	97.9	135.3
1953	4.4	4.8	4.9	7.3	10.4	11.0	15.2	25.3	49.9
1954	8.4	9.5	12.3	14.6	16.7	22.9	39.8	99.8	160.8
1955	6.1	6.3	7.0	11.2	14.7	37.2	67.6	148.4	247.9
1956	19.0	21.3	23.0	26.8	29.5	53.5	59.5	71.3	98.2
1957	3.7	5.1	5.8	6.4	8.9	9.8	12.6	15.5	25.6
1958	12.0	13.3	17.4	19.2	23.8	29.6	41.5	50.9	118.5
1959	17.0	17.3	20.6	25.1	39.0	49.8	53.2	60.4	111.0
1960	48.0	48.3	53.3	63.9	77.5	122.1	136.3	149.1	213.9
1961	17.0	17.0	19.7	22.9	27.9	32.1	31.9	37.2	57.2
1962	5.9	6.6	7.0	7.8	9.9	15.0	16.7	20.6	41.9
1963	19.0	19.0	19.6	21.6	27.4	32.1	38.8	58.9	70.5
1964	1.1	1.4	1.8	2.5	4.2	7.1	9.0	11.5	19.1
1965	5.7	6.7	9.9	11.7	12.1	13.7	15.8	20.6	26.1
1966	4.0	4.5	4.7	4.8	5.0	6.4	13.2	21.8	64.8
1967	43.0	44.0	49.3	58.1	62.0	92.4	122.4	147.2	188.6

Note - Data based on Climatic Year of April 1 of given year through March 31 of next year.

the results obtained through the use of this procedure seem reasonable, it is impossible to verify the accuracy of the frequency estimates.

c. Regionalization. Regionalization of low-flow events is usually not very successful. The variations in geologic conditions such as depth to ground water, size of ground water basin, permeability of the aquifer, etc., are not easily quantifiable to enable translation into probable low-flow rates. It may be possible to estimate low-flow rates on a per unit area basis for a given exceedance frequency if the study area is relatively homogeneous with respect to geology, topography, and climate. If information is needed at several unengaged sites, the procedures described by Riggs (28) should be reviewed for applicability.

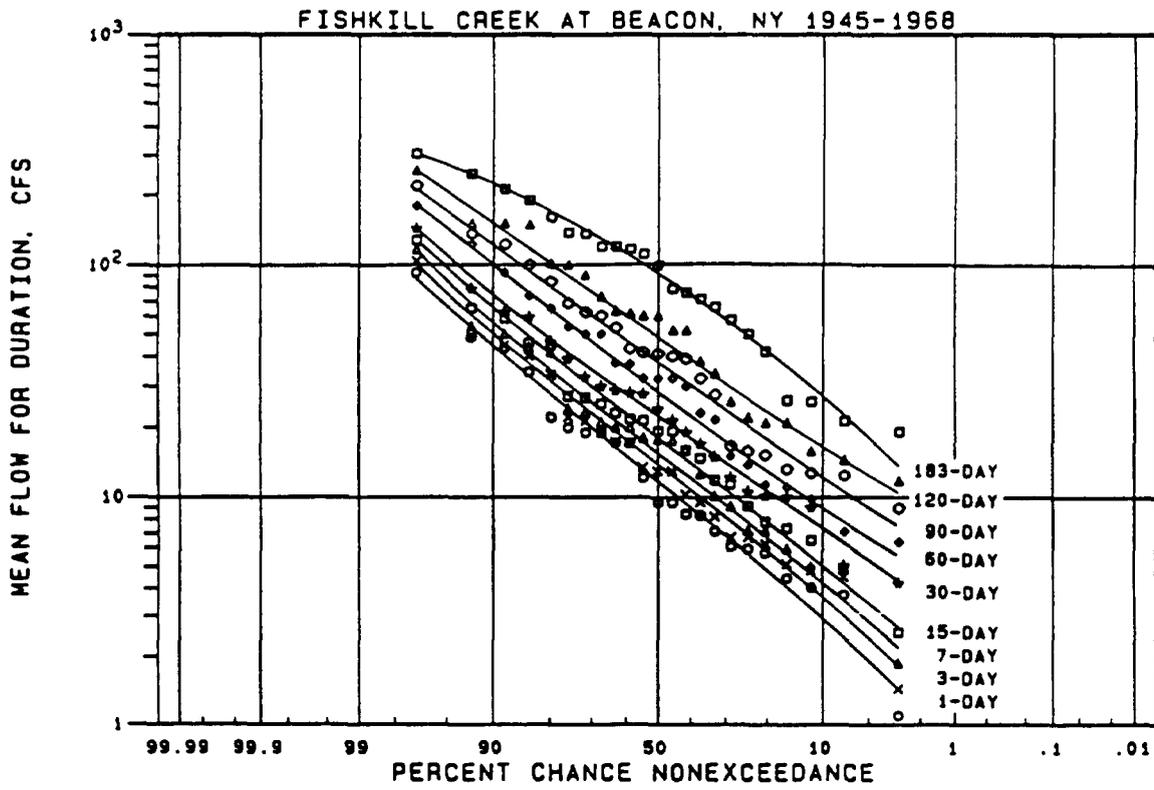


Figure 4-1. Low-Flow Frequency Curves.

## CHAPTER 5 PRECIPITATION FREQUENCY ANALYSIS

5-1. **General Procedures.** The computation of frequency curves of station precipitation can be done by procedures similar to those for streamflow analysis described in the preceding sections. Both graphical and analytical methods may be used. In precipitation studies, however, instantaneous peak intensities are ordinarily not analyzed since they are virtually impossible to measure and are of little practical value. Cumulative precipitation amounts for specified durations are commonly analyzed, mostly for durations of less than 3 or 4 days. The National Weather Service has traditionally used the Fisher-Tippett Type I frequency distribution with Gumbel's fitting procedure. The logarithmic normal, Pearson Type III and log-Pearson Type III (Figure 5-1) distributions, have also been used with success. Station precipitation alone is not adequate for most hydrologic studies, and some method of evaluating the frequency of simultaneous or near-simultaneous precipitation over an area is necessary. Procedures for obtaining depth-area frequency curves are usually available from National Weather Service publications (references are given in subsequent paragraphs).

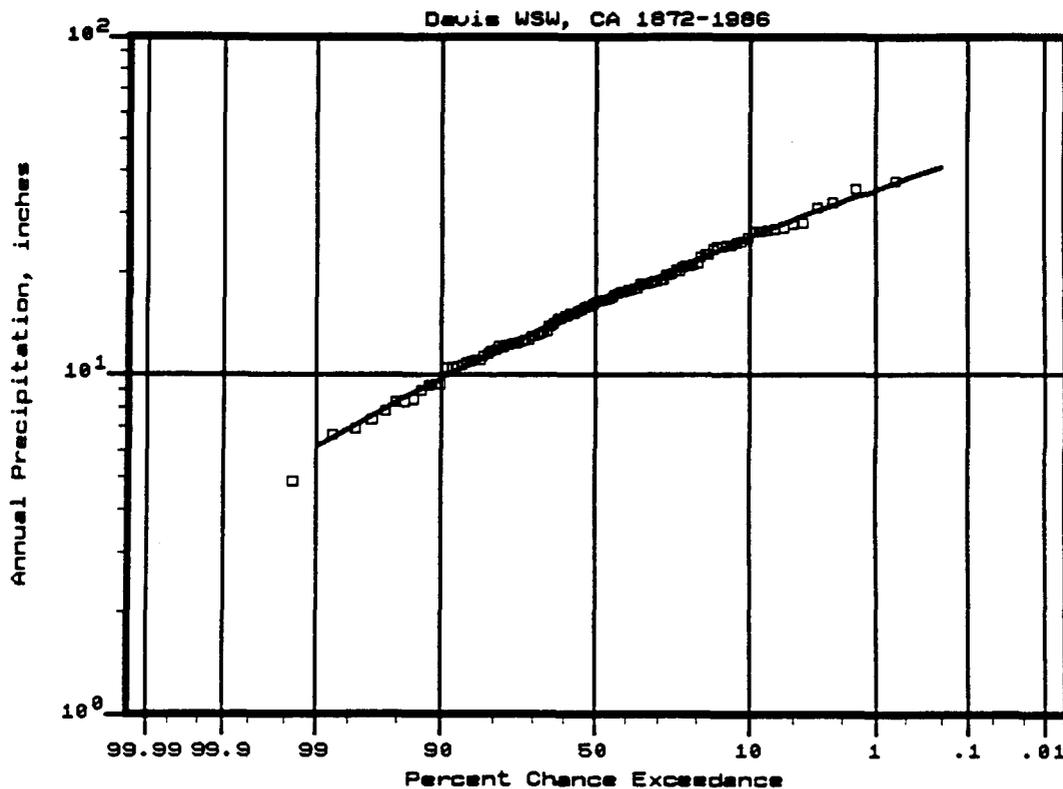


Figure 5-1. Frequency Curve, Annual Precipitation.

5-2. Available Regional Information. Where practical, use should be made of previous precipitation-frequency-duration studies that have incorporated regional information. For durations of 5 to 60 minutes in an area generally east of 105th meridian, see Hydro-35 (9). For durations of 2 to 24 hours in the same area see Technical Paper 40 (10). Because of the orographic effect, individual reports have been prepared for each of the 11 western states (24). These reports have maps for 6- and 24-hour durations with extrapolation procedures to obtain durations less than 6 hours. Longer duration events (2- to 10-days) are presented in references (21), (22) and (23).

5-3. Derivation of Flood-Frequency Relations from Precipitation.

a. Application. Precipitation-frequency relations are often used to derive flood-frequency relations where inadequate flow data are available or where existing (or proposed) watershed changes have modified (or will modify) the rainfall-runoff relationships. Guidelines for developing runoff frequencies from precipitation frequencies are presented in references (10) and (44). Flood-frequency curves developed by rainfall-runoff procedures often have less variance (lower standard deviation) than those developed from annual flood peaks. This results because not all the possible loss rates for a given magnitude of precipitation are modeled. If extensive use will be made of frequency curves derived by rainfall-runoff modeling, an appropriate ratio adjustment for the standard deviation should be developed for the region.

b. Calibration. Reference (44) describes the procedures involved in calibrating a HEC-1 model to a flow-frequency curve based either on gaged data from a portion of the basin or on regional flood-frequency relations. The coefficients from the calibrated model must be consistent with those from nearby basins that have also been modeled. It must be remembered that a frequency curve computed from observed flood peaks is based on a relatively small sample. It is possible that the flow-frequency curve derived from precipitation-frequency data is more representative of the population flow-frequency curve than the one computed from the statistics of the observed flood peaks. But, there are also errors in calibrating the model and establishing loss rates approximate with the different frequency events. Therefore, the derivation of frequency relations by rainfall-runoff modeling requires careful checking for consistency at every step.

c. Partial Duration. The precipitation-frequency relations presented in the National Weather Service publications represent all the events above a given magnitude; therefore, these relations are from a partial-duration series. The resulting flood frequency relations must be adjusted if an annual peak flood frequency relationship is desired. Or, more typically, the partial-duration series precipitation estimates are adjusted to represent annual series estimates prior to use.

## CHAPTER 6

### STAGE (ELEVATION) - FREQUENCY ANALYSIS

6-1. Uses. Maximum stage-frequency relations are often required to evaluate inundation damage. Inundation can result from a flooding river, storm surges along a lake or ocean, wind driven waves (runup), a filling reservoir, or combinations of any of these. Minimum elevation-frequency curves are used to evaluate the recreation benefits at a lake or reservoir, to locate a water supply intake, to evaluate minimum depths available for navigation purposes, etc. (Stages are referenced to an arbitrary datum; whereas, elevations are generally referenced to mean sea level.)

#### 6-2. Stage Data.

a. The USGS WATSTORE Peak Flow File has, in addition to annual peak flows, maximum annual stages at most sites. Also, some sites located near estuaries have only stage information because the flow is affected by varying backwater conditions.

b. River stages can be very sensitive to changes in the river channel and floodway. Therefore, the construction of levees, bridges, or channel modifications can result in stage data that is non-homogeneous with respect to time. For riverine situations, it is usually recommended that the flow-frequency curve (Figure 6-1) and a rating curve (stage versus

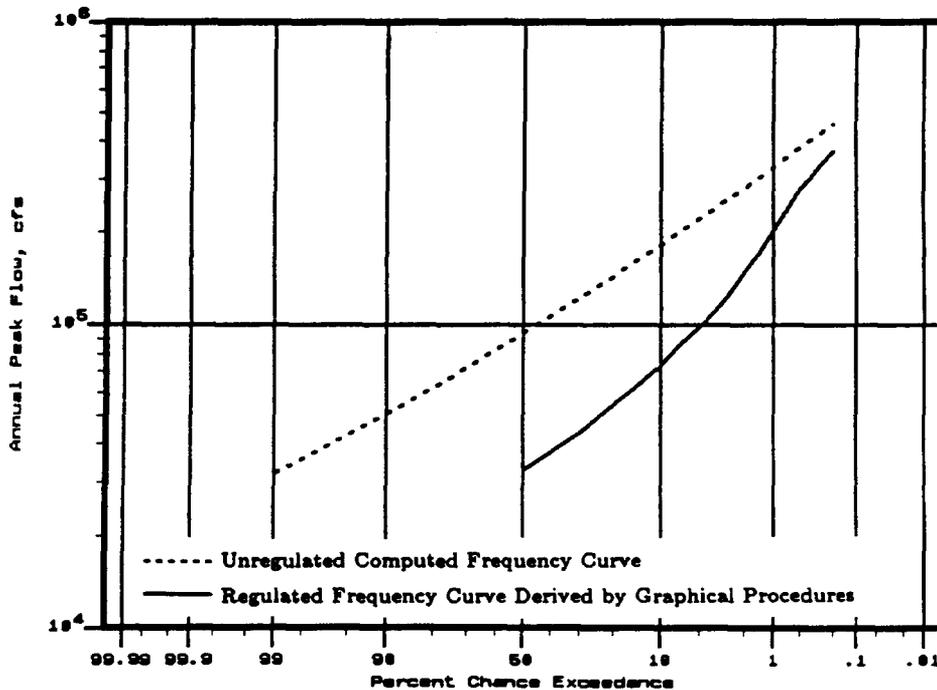


Figure 6-1. Flow Frequency Curve, Unregulated and Regulated Conditions.

5 Mar 93

flow, Figure 6-2) be used to derive a stage-frequency curve (Figure 6-3). A stage-frequency curve derived by indirect methods may not always represent the true relation if the site is subject to occasional backwater situations. Backwater conditions can be caused by an ice jam, a debris flow, a downstream reservoir, a high tide, a storm surge, or a downstream river. A coincident frequency analysis may be necessary to obtain an accurate estimate of the stage-frequency relationship (see Chapter 11).

c. Usually the annual extreme value is used to develop an annual series, but a seasonal series or a partial-duration series could be developed if needed. Caution must be used in selecting independent events. Independent events are not easily determined if the events are elevations of a large lake or reservoir; in fact even the annual events may be significantly correlated.

6-3. Frequency Distribution. Stage (elevation) data are usually not normally distributed (not a straight line on probability paper). Therefore, an analytical analysis should not be made without observing the fit to the plotted points (see Chapter 2). Usually, an arithmetic-probability plot is appropriate for stage or elevation data, but there may be situations where a logarithmic or some other appropriate transformation will make the plot more nearly linear. When drawing the curve, known constraints must be kept in mind. As an example, the bottom elevation, bankfull stage, levee heights, etc., would be important for a riverine site. The minimum pool, top of conservation pool, top of flood control pool, spillway elevation, operation criteria, etc., all influence the elevation-frequency relation for a reservoir, Figure 6-4. These constraints usually make these frequency relations very non-linear. Extrapolation of stage (elevation) frequency relations must be done very cautiously. Again, any constraints acting on the relations must be used as a guide in drawing the curves. Historical information can be incorporated into a graphical analysis of stage (elevation) data by use of the procedures in Appendix 6 of Bulletin 17B (ref 46). The statistical tests (Appendix 4, ref 46) to screen for outliers should not be applied unless the stage (elevation) data can be shown to nearly fit a normal distribution.

6-4. Expected Probability. The expected probability adjustment should not be made to frequency relations derived by graphical methods. The median plotting position formula corrects for the bias caused by small sample sizes. The expected probability adjustment should be made when an analytical method is used directly to derive the stage (elevation) frequency relation. The expected probability adjustment should be made to the flow-frequency curve when the stage (elevation) frequency relation is derived indirectly.

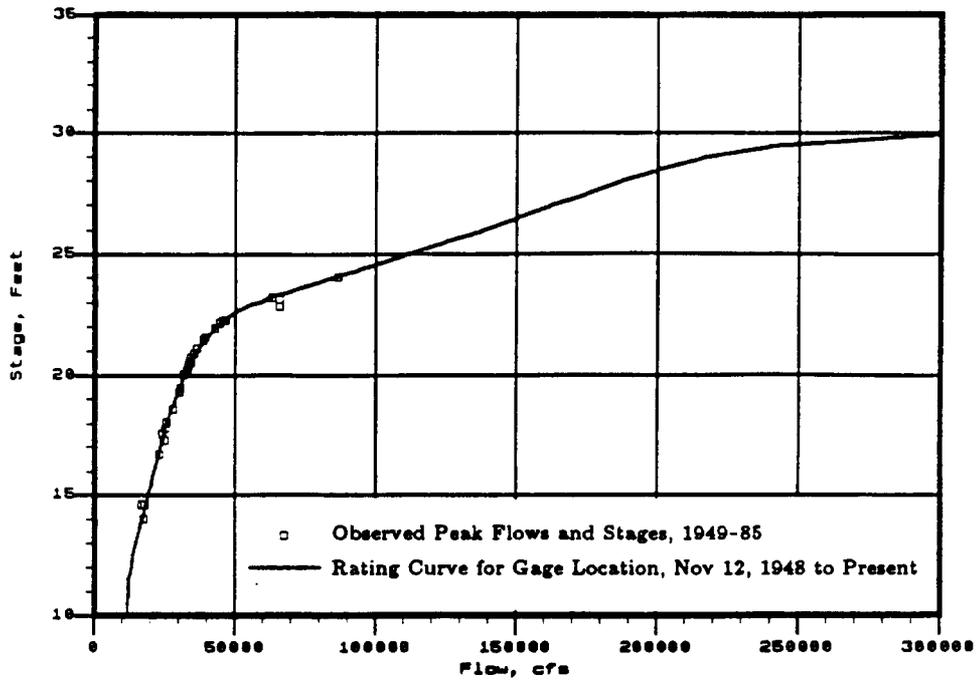


Figure 6-2. Rating Curve for Present Conditions.

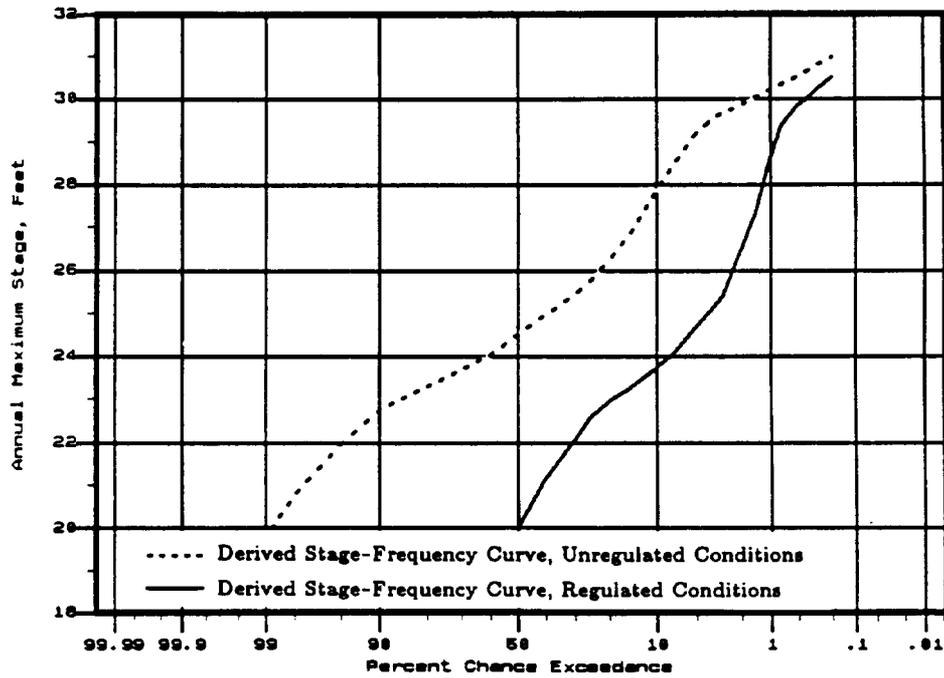


Figure 6-3. Derived Stage-Frequency Curves, Unregulated and Regulated Conditions

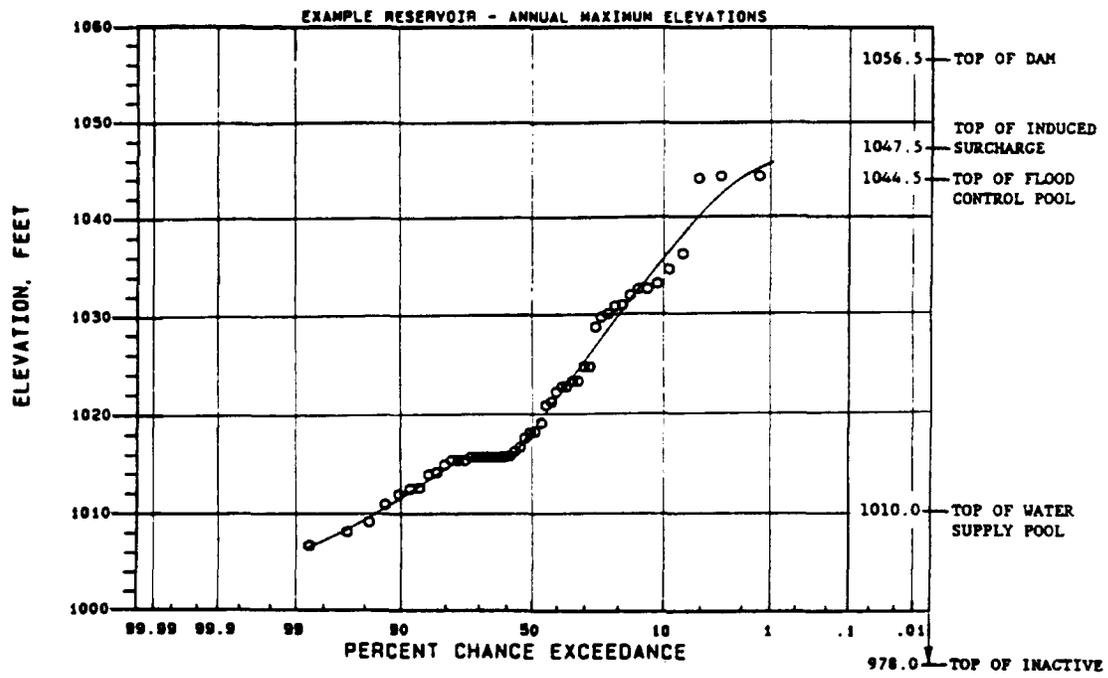


Figure 6-4. Maximum Reservoir Elevation-Frequency Curve.

## CHAPTER 7

### DAMAGE-FREQUENCY RELATIONSHIPS

#### 7-1. Introduction.

a. There are three methods that may be used to compute average annual damage and are herein termed the historic method, the simulation method and the frequency method. If 50 years of damage information were available for an area that has remained in essentially the same land use with a reasonably constant level of economic activity, historic damage could be scaled to the present to account for price differences (inflation) and the average simply computed. This approach is termed the historic method and is the most direct but is seldom used because sufficient data usually do not exist and the land use and economic activity of an area are usually changing.

b. A hydrologic simulation model could be developed, or the historic record used, along with damage functions to generate a time trace of simulated damage. The average of the time trace of damage would be the average annual damage. This would be termed the "simulation" method. The simulation method has the advantage of permitting the use of complex damage functions that can consider more than a single parameter and thus enable a more accurate computation of damage. The disadvantage of this method is that the future floods are assumed to exactly duplicate the historic floods and no consideration given to the possibility of larger floods.

c. The most widely used approach within the Corps of Engineers is the frequency technique. This technique is described in detail in Section 7-2. This technique addresses the disadvantages of the previous two methods, and yet is fairly easily applied. Experience in the development and application of damage functions is essential to computation of reasonable estimates. Care should be taken to assure the rating curve is not looped so that discharge is a unique function of stage. Otherwise more complex functions that correctly relate stage and discharge should be developed and applied. Damage functions in agricultural areas are often a function of the season and the duration of flooding. Sensitivity analysis may be useful in determining the reliability of the computed expected annual damage considering the uncertainties involved.

#### 7-2. Computation of Expected Annual Damage.

a. Figure 7-1 shows a schematic of the application of the three basic damage evaluation functions used to compute the expected value of the annual damage. The term "expected" is used rather than "average" because a frequency curve is used to represent the distribution of future flood events and the expected value of damage is computed by the summation of probability weighted estimates of damage.

b. The steps involved in determining the reduction in annual damage due to project measures are:

- (1) Develop the basic relationships (stage-damage, stage-discharge, and discharge-exceedance frequency functions) for each index location for existing conditions.

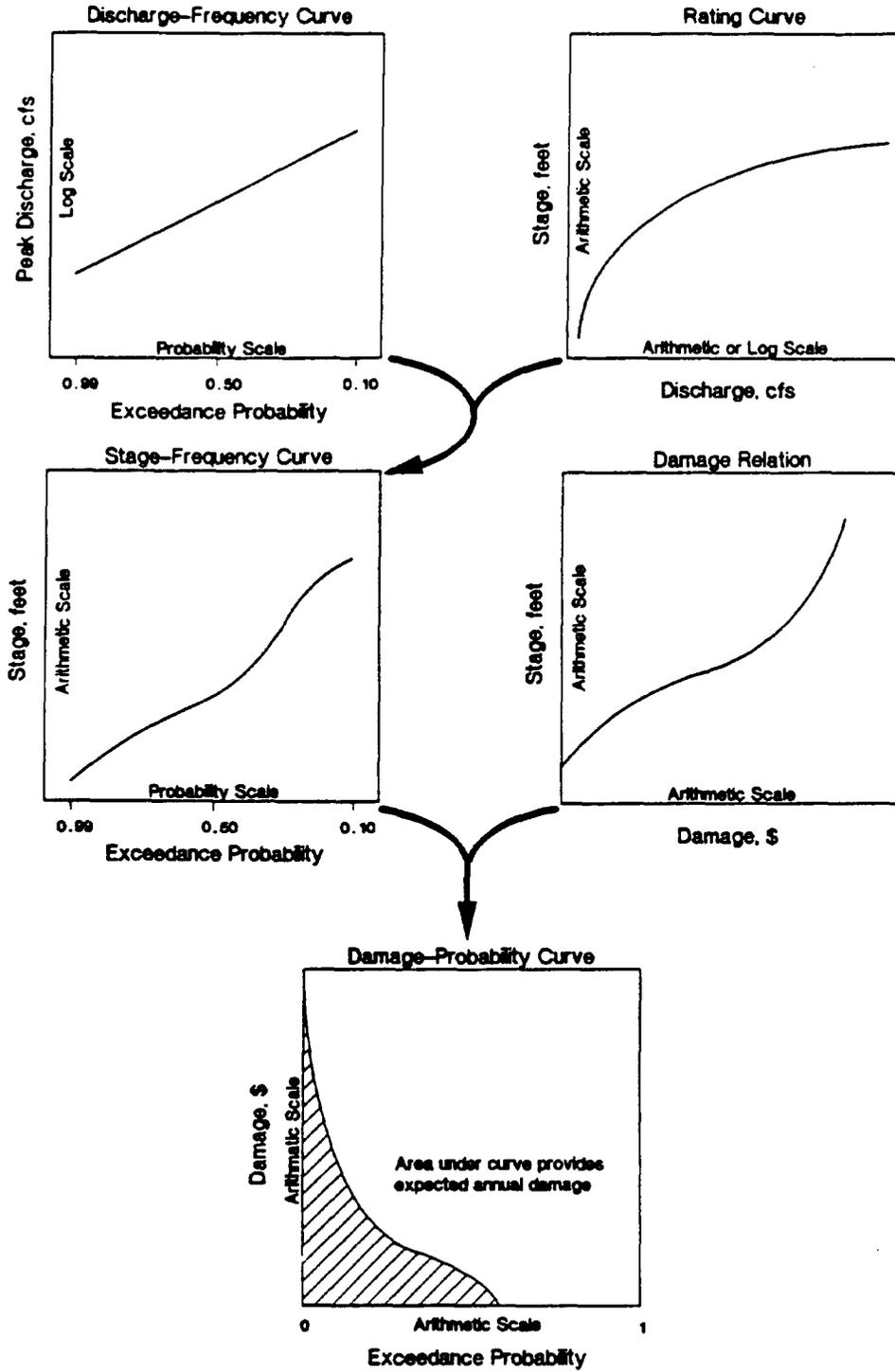


Figure 7-1. Schematic for Computation of Expected Annual Damage.

- (2) Combine the stage-damage and the stage-discharge relations into an intermediate discharge-damage function. Make certain that the stage datum for the stage-damage and stage-discharge functions is consistent for the index location.
- (3) Combine the discharge-exceedance frequency (in events per year) and discharge-damage function into a damage-exceedance frequency relationship.
- (4) Compute the area beneath the damage-exceedance frequency relation (expected annual damage) for each index location and sum to obtain the total expected annual flood damage.
- (5) Repeat step (1) for each alternative flood plain management plan under investigation, i.e., revise the three basic evaluation functions as necessary.
- (6) Repeat steps (2)-(4).
- (7) Subtract results of step (4) (with project) for each plan from results of step (4) for without-project measures. The differences will be expected annual damage reduction (raw damage reduction benefits) for each plan.

### 7-3. Equivalent Annual Damage.

a. To determine the expected annual benefit it is necessary to account for the changes in expected annual damage that might occur over the life of the project. This adjustment can be of substantial significance. Watershed runoff characteristics may be changing with time due to changes in land use, there may be long-term adjustments in alluvial channel flow regimes that would cause the rating curve to change with time, and the damage potential of structures and facilities will certainly change with time resulting in changed stage-damage functions.

b. To develop a single measure of the damage potential, the expected annual damage must be evaluated over time, at say 10 year intervals with revised evaluation functions at each interval. The revised expected annual damage is discounted to the base period and then the raw damage value is amortized over the life of the project to obtain equivalent annual damage. The computer program "Expected Annual Flood Damage Computation" (54) has the capability to make these computations, and describes in detail the basic concepts presented in this chapter.

## CHAPTER 8

### STATISTICAL RELIABILITY CRITERIA

8-1. Objective. One principal advantage of analytical frequency analysis is that there are means for evaluating the reliability of the parameter estimates. This permits a more complete understanding of the frequency estimates and provides criteria for decision-making. For instance, a common statistical index of reliability is the standard error of estimate, which is defined as the root-mean-square error. In general, it is considered that the standard error is exceeded on the positive side one time out of six estimates, and equally frequently on the negative side, for a total of one time in three estimates. An error twice as large as the standard error of estimate is considered to be exceeded one time in 40 in either direction, for a total of one time in 20. These statements are based on an assumed normal distribution of the errors; thus, they are only approximate for other distributions of errors. Exact statements as to error probability must be based on examination of the frequency curve of errors or the distribution of the errors. Both the standard error of estimate and the confidence limits are discussed in this chapter.

8-2. Reliability of Frequency Statistics. The standard errors of estimate of the mean, standard deviation, and skew coefficient, which are the principal statistics used in frequency analysis, are given by the following equations:

$$S_{\bar{x}} = S/(N)^{1/2} \quad (8-1)$$

$$S_s = S/(2N)^{1/2} \quad (8-2)$$

$$S_g = (6N(N-1)/[(N-2)(N+1)(N+3)])^{1/2} \quad (8-3)$$

where:

$S_{\bar{x}}$  = the standard error of estimate for the mean

$S_s$  = the standard error of estimate for the standard deviation

$S_g$  = the standard error for estimate for the skew coefficient, and S and N are defined in Section 3-2.

These have been used to considerable advantage, as discussed in Chapter 9, in drawing maps of mean, standard deviation and skew coefficient for regional frequency studies.

8-3. Reliability of Frequency Curves. The reliability of analytical frequency determinations can best be illustrated by establishing confidence limits. The error of the estimated value at a given frequency based on a sample from a normal distribution is a function of the errors in estimating the mean and standard deviation. (Note that in practical application there are errors introduced by not knowing the true theoretical distribution of the data, often termed model error.) Criteria for construction of confidence limits are based on the non-central t distribution. Selected values are given in Table F-9. Using that appendix, the confidence limit curves shown on Figure 8-1

were calculated. While the expected frequency is that shown by the middle curve, there is one chance in 20 that the true value for any given frequency is greater than that indicated by the .05 curve and one chance in 20 that it is smaller than the value indicated by the .95 curve. There are, therefore, nine chances in 10 that the true value lies between the .05 and .95 curves. Appendix E and Example 1 in Appendix 12 of Bulletin 17B (40) provide additional information and example computations.

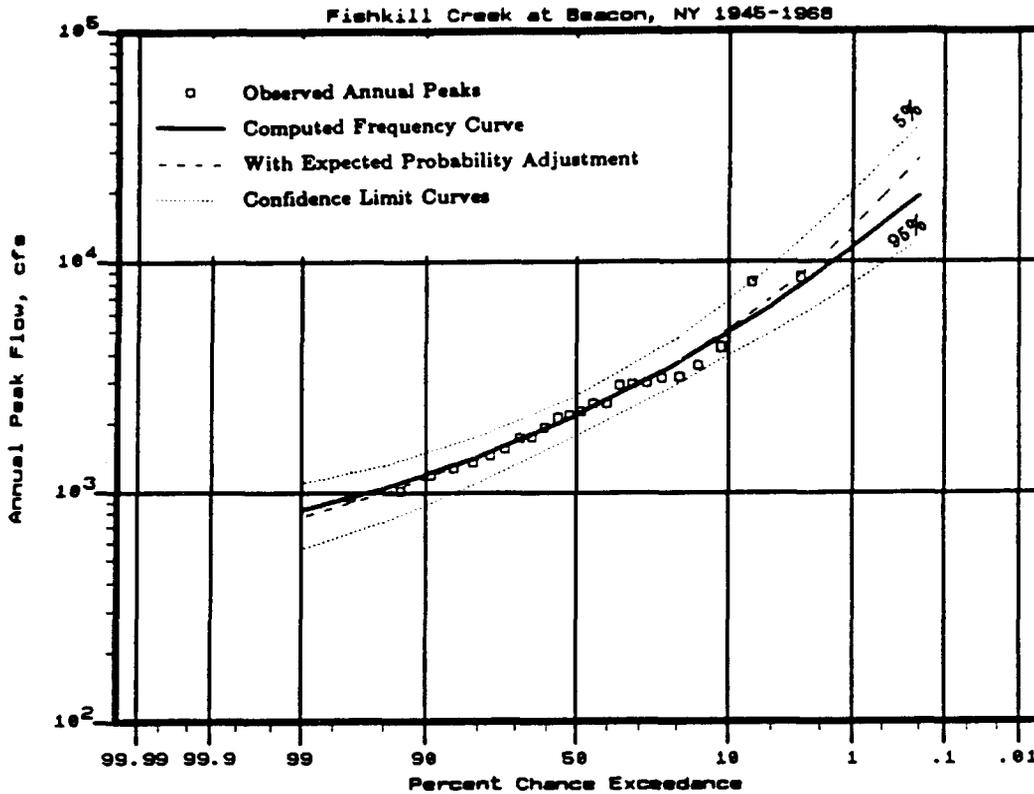


Figure 8-1. Frequency Curve with Confidence Limit Curves.

## CHAPTER 9

### REGRESSION ANALYSIS AND APPLICATION TO REGIONAL STUDIES

#### 9-1. Nature and Application.

a. General. Regression analysis is the term applied to the analytical procedure for deriving prediction equations for a variable (dependent) based on given values of one or more other variables (independent). The dependent variable is the value sought and is to be related to various explanatory variables which will be known in advance, and which will be physically related to the dependent variable. For example, the volume of spring-season runoff from a river basin (dependent variable) might be correlated with the depth of snow cover in the watershed (explanatory variable). Recorded values of such variables over a period of years might be graphed and the apparent relation sketched in by eye. However, regression analysis will generally permit a more reliable determination of the relation and has the additional advantage of providing a means for evaluating the reliability of the relation or of estimates based on the relation.

b. Definitions. The function relating the variables is termed the "regression equation," and the proportion of the variance of the dependent variable that is explained by the regression equation is termed the "coefficient of determination," which is the square of the "correlation coefficient." Correlation is a measure of the association between two or more variables. Regression equations can be linear or curvilinear, but linear regression suffices for most applications, and curvilinear regression is therefore not discussed herein. Often a curvilinear relation can be linearized by using a logarithmic or other transform of one or more of the variables.

#### 9-2. Calculation of Regression Equations.

a. Simple Regression. In a simple regression (one in which there is only one independent, or explanatory, variable), the linear regression equation is written:

$$Y = a + bX \quad (9-1)$$

in which Y is the dependent variable, X is the independent variable, "a" is the regression constant, and "b" is the regression coefficient. The coefficient "b" is evaluated from the tabulated data by use of the following equations:

$$b = \frac{\sum(yx)}{\sum(x)^2} \quad (9-2a)$$

or

$$b = RS_y/S_x \quad (9-2b)$$

in which y is the deviation of a single value  $y_i$  from the mean ( $\bar{Y}$ ) of its series, x is similarly defined,  $S_y$  and  $S_x$  are the respective standard deviations and R is computed by

Equation 9-11. The regression constant is obtained from the tabulated data by use of the following equation:

$$a = \bar{Y} - b\bar{X} \quad (9-3)$$

All summations required for a simple linear regression can be obtained using Equations 9-8 and 9-9a.

b. Multiple Regression. In a multiple regression (one in which there is more than one explanatory variable) the linear regression equation is written:

$$Y = a + b_1X_1 + b_2X_2 \dots + b_NX_N \quad (9-4)$$

In the case of two explanatory variables, the regression coefficients are evaluated from the tabulated data by solution of the following simultaneous equations:

$$\begin{aligned} \sum(x_1)^2b_1 + \sum(x_1x_2)b_2 &= \sum(yx_1) \\ \sum(x_1x_2)b_1 + \sum(x_2)^2b_2 &= \sum(yx_2) \end{aligned} \quad (9-5)$$

In the case of three explanatory variables, the b coefficients can be evaluated from the tabulated data by solution of the following simultaneous equations:

$$\begin{aligned} \sum(x_1)^2b_1 + \sum(x_1x_2)b_2 + \sum(x_1x_3)b_3 &= \sum(yx_1) \\ \sum(x_1x_2)b_1 + \sum(x_2)^2b_2 + \sum(x_2x_3)b_3 &= \sum(yx_2) \\ \sum(x_1x_3)b_1 + \sum(x_2x_3)b_2 + \sum(x_3)^2b_3 &= \sum(yx_3) \end{aligned} \quad (9-6)$$

For cases of more than three explanatory variables, the appropriate set of simultaneous equations can be easily constructed after studying the patterns of the above two sets of equations. In such cases, solution of the equations becomes tedious, and considerable time can be saved by use of the Crout method outlined in reference (51) or (52). Also, programs are available for solution of simple or multiple linear regression problems on practically any type of electronic computer. For multiple regression equations, the regression constant is determined as follows:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 \dots - b_N\bar{X}_N \quad (9-7)$$

In Equations 9-2, 9-5 and 9-6, the quantities  $\sum(x)^2$ ,  $\sum(yx)$  and  $\sum(x_1x_2)$  can be determined by use of the following equations:

$$\sum(x)^2 = \sum(X)^2 - (\sum X)^2/N \quad (9-8)$$

$$\sum(yx) = \sum(XY) - \sum X \sum Y/N \quad (9-9a)$$

$$\sum(x_1x_2) = \sum(X_1X_2) - \sum X_1 \sum X_2/N \quad (9-9b)$$

9-3. The Correlation Coefficient and Standard Error.

a. General. The correlation coefficient is the square root of the coefficient of determination, which is the proportion of the variance of the dependent variable that is explained by the regression equation. A correlation coefficient of 1.0 would correspond to a coefficient of determination of 1.0, which is the highest theoretically possible and indicates that whenever the values of the explanatory variables are known exactly, the corresponding value of the dependent variable can be calculated exactly. A correlation coefficient of 0.5 would correspond to a coefficient of determination of 0.25, which would indicate that 25 percent of the variance is accounted for and 75 percent unaccounted for by the regression equation. The remaining variance (error variance) would be 75 percent of the original variance and the remaining standard error would be the square root of 0.75 (or 87 percent) multiplied by the original standard deviation of the dependent variable. Thus, with a correlation coefficient of 0.5, the average error of estimate would be 87 percent of the average errors of estimate based simply on the mean observed value of the dependent variable without a regression analysis.

b. Determination Coefficient. The sample coefficient of multiple determination ( $R^2$ ) can be computed by use of the following equation:

$$R^2 = \frac{b_1 \sum(yx_1) + b_2 \sum(yx_2) \dots + b_N \sum(yx_n)}{\sum(y)^2} \quad (9-10)$$

In the case of simple correlation, Equation 9-10 resolves to:

$$R^2 = \sum(yx)^2 / \sum(y)^2 \sum(x)^2 \quad (9-11)$$

An unbiased estimate of the coefficient of determination is recommended for most applications, and is computed by the following equation:

$$\bar{R}^2 = 1 - (1 - R^2)(N - 1) / df \quad (9-12)$$

The number of degrees of freedom (df), is obtained by subtracting the number of variables (dependent and explanatory) from the number of events tabulated for each variable.

c. Standard Error. The adjusted standard error ( $S_e$ ) of a set of estimates is the root-mean-square error of those estimates corrected for the degrees of freedom. On the

average, about one out of three estimates will have errors greater than the standard error and about one out of 20 will have errors greater than twice the standard error. The adjusted error variance is the square of the adjusted standard error. The adjusted standard error or error variance of estimates based on a regression equation is calculated from the data used to derive the equation by use of one of the following equations:

$$S_e^2 = \frac{\sum(y)^2 - b_1 \sum(yx_1) - b_2 \sum(yx_2) \dots - b_n \sum(yx_n)}{df} \quad (9-13a)$$

$$= (1 - \bar{R}^2) \sum(y)^2 / (N-1) \quad (9-13b)$$

$$= (1 - \bar{R}^2) S_y^2 \quad (9-13c)$$

Inasmuch as there is some degree of error involved in estimating the regression coefficients, the actual standard error of an estimate based on one or more extreme values of the explanatory variables is somewhat larger than is indicated by the above equations, but this fact is usually neglected.

d. **Reliability.** In addition to considering the amount of variance that is explained by the regression equation, as indicated by the determination coefficient or the standard error, it is important to consider the reliability of these indications. There is some chance that any correlation is accidental, but the higher the correlation and the larger the sample upon which it is based, the less is the chance that it would occur by accident. Also, the reliability of a regression equation decreases as the number of independent variables increases. Ezekiel (8) gives a set of charts illustrating the reliability of correlation coefficients. It shows, for example, that an unadjusted correlation coefficient (R) of 0.8 based on a simple linear correlation with 12 degrees of freedom could come from a relationship that has a true value as low as 0.53 in one case out of 20. On the other hand, the same unadjusted correlation coefficient based on a multiple linear correlation with the same number of degrees of freedom but with seven independent variables, could come from a relationship that has a true value as low as zero in one case out of 20. With only 4 degrees of freedom, an unadjusted correlation coefficient of 0.97 would one time in 20 correspond to a true value of 0.8 or lower, in the case of simple correlation, and as low as zero in a seven-variable multiple correlation. Accordingly, extreme care must be exercised in the use of multiple correlation in cases based on small samples.

#### 9-4. Simple Linear Regression Example.

a. **General.** An example of a simple linear regression analysis is illustrated on Figures 9-1 and 9-2. The data for this example are the concurrent flows at two stations in Georgia for which a two-station comparison is desired (see Section 3-7). The long record station is the Chattooga, so the flows for this station are selected as X; therefore the flows for the short record station (Tallulah) are assigned to Y.

b. **Physical Relationship.** The values in the table are the annual peak flows for the water years 1965-1985 (21 values). These two stations are less than 20 miles apart and are likely to be subject to the same storm events; therefore, the first requirement of a

Year	Chattooga River		Tallulah River	
	Flow X'	Log X	Flow Y'	Log Y
1965	27200	4.434568	7440	3.871572
1966	13400	4.127104	5140	3.710963
1967	15400	4.187520	2800	3.447158
1968	5620	3.749736	3100	3.491361
1969	14700	4.167317	2470	3.392696
1970	3480	3.541579	2010	3.303196
1971	3290	3.517195	976	2.989449
1972	7440	3.871572	2160	3.334453
1973	19600	4.292256	8500	3.929418
1974	6400	3.806179	4660	3.668385
1975	6340	3.802089	2410	3.382017
1976	18500	4.267171	6530	3.814913
1977	13000	4.113943	3580	3.553883
1978	7850	3.894869	4090	3.611723
1979	14800	4.170261	6240	3.795184
1980	10900	4.037426	2880	3.459392
1981	4120	3.614897	1600	3.204119
1982	5000	3.698970	1960	3.292256
1983	7910	3.898176	3260	3.513217
1984	4810	3.682145	2000	3.301029
1985	4740	3.675778	1010	3.004321

$$\begin{aligned} \Sigma X &= 82.55075 & \Sigma Y &= 73.07071 \\ \Sigma X^2 &= 325.93995 & \Sigma Y^2 &= 255.61486 \\ \bar{X} &= 3.93099 & \bar{Y} &= 3.47956 \\ (\Sigma XY)^2 &= 288.37484 \\ \frac{(\Sigma X)^2}{N} &= 324.50602 \\ \frac{(\Sigma Y)^2}{N} &= 254.25371 \\ \frac{\Sigma X \Sigma Y}{N} &= 287.24008 \\ x^2 &= 1.43393 & & \text{(by equation 9-8)} \\ xy &= 1.13351 & & \text{(" " 9-9a)} \\ y^2 &= 1.26115 & & \text{(" " 9-8)} \end{aligned}$$

Computations for a, b, R<sup>2</sup>, and R̄<sup>2</sup>:

$$b = 1.13351/1.433922 \quad \text{(by equation 9-2a)} \\ = 0.79049$$

$$a = 3.47956 - (0.79049)(3.93099) \quad \text{(by equation 9-3)} \\ = 0.37213$$

$$R^2 = (1.13351)^2 / (1.43393)(1.36115) \quad \text{(by equation 9-11)} \\ = 0.658290$$

$$\bar{R}^2 = 1 - (1 - 0.65892)(21 - 1) / (21 - 2) \quad \text{(by equation 9-12)} \\ = 0.64031$$

Computations for standard error:

$$s_e^2 = (1 - 0.64031^2)(1.36115) / (21 - 1) \quad \text{(by equation 9-13b)} \\ = 0.02448$$

$$s_e = 0.15646$$

Regression equation:  $Y = 0.37213 + 0.79049X$  (by equation 9-1)  
 $Y' = 2.356X^{0.79}$  (without logarithms)

Figure 9-1. Computation of Simple Linear Regression Coefficients.

regression analysis (logical physical relationship) is satisfied. Because runoff is a multiplicative factor of precipitation and drainage area, the logarithmic transformation is likely to be appropriate when comparing two stations with different drainage areas. A linear correlation analysis was made, as illustrated on Figure 9-1, using equations given in Section 9-2. The annual peaks for the each station are plotted against each other on Figure 9-2.

c. Regression Equation. The regression equation is plotted as Curve A on Figure 9-2. This curve represents the best estimate of what the annual peak Tallulah River would be given the observed annual peak on the Chattooga River. Although not computed in Figure 9-1, Curve B represents the regression line for estimating the annual peak flow for the Chattooga River given an observed annual peak on the Tallulah River.

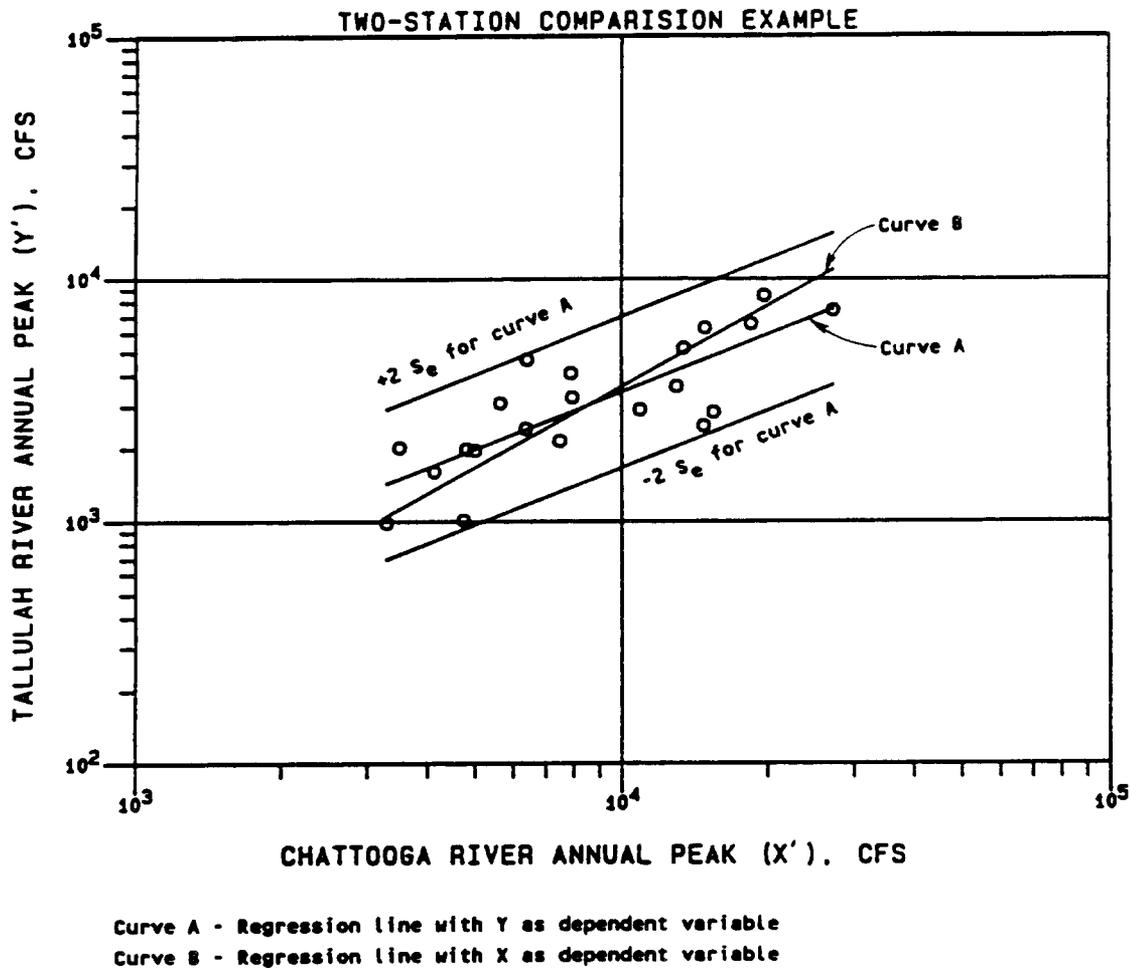


Figure 9-2. Illustration of Simple Regression.

d. Reliability. In addition to the curve of best fit, an approximate confidence interval can be established at a distance of plus and minus 2 standard errors from Curve A. Because logarithms are used in the regression analysis, the effect of adding (or subtracting) twice the standard error to the estimate is equivalent to multiplying (or dividing) the annual peak values by the antilogarithm of twice the standard error. In this case, the standard error is 0.156, and the antilogarithm of twice this quantity is 2.05. Hence, values of annual peak flow represented by the confidence interval curves are those of Curve A multiplied and divided respectively by 2.05. There is a 95 percent chance that the true value of the dependent variable (Y) for a single observed independent value (X) will lie between these limits. The confidence interval is not correct for repeated predictions using the same sample (6).

#### 9-5. Factors Responsible for Nondetermination.

a. General. Factors responsible for correlations being less than 1.0 (perfect correlation) consist of pertinent factors not considered in the analysis and of errors in the measurement of those factors considered. If the effect of measurement errors is appreciable, it is possible in some cases to evaluate the standard error of measurement of each variable (see Paragraph 9-3c) and to adjust the correlation results from such effects.

b. Measurement Errors. If an appreciable portion of the variance of Y (dependent variable) is attributable to measurement errors and these errors are random, then the regression equation would be more reliable than is indicated by the standard error of estimate computed from Equation 9-13. This is because the departure of some of the points from the regression line on Figure 9-2 is artificially increased by measurement errors and therefore exaggerates the unreliability of the regression function. In such a case, the curve is generally closer to the true values than to the erroneous observed values. Where there is large measurement error of the dependent variable, the standard error of estimate should be obtained by taking the square root of the difference of the error variance obtained from Equation 9-13 and the measurement error variance. If well over half of the variance of the points from the best-fit line is attributable to measurement error in the dependent variable, then the regression line would actually yield a better estimate of a value than the original measurement. If appreciable errors exist in the values of an explanatory variable, the regression coefficient and constant will be affected, and erroneous estimates will result. Hence, it is important that values of the explanatory variables be accurately determined, if possible.

c. Other Factors. In the example used in Section 9-4 there may well be factors responsible for brief periods of high intensities that do not contribute appreciably to annual precipitation. Consequently, some locations with extremely high mean annual precipitation may have maximum short-time intensities that are not correspondingly high, and vice versa. Therefore, the station having the highest mean annual precipitation would not automatically have the highest short-time intensity, but would in general have something less than this. On the other hand, if mean annual precipitation were made the dependent variable, the station having the highest short-time intensity would be expected to have something less than the highest value of mean annual precipitation. Thus, by interchanging the variables, a change in the regression line is effected. Curve B of Figure 9-2 is the regression curve obtained by interchanging the variables Y and X. As there is a considerable difference in the two regression curves, it is important to use the variable whose value is to be calculated from the regression equation as the dependent variable in those cases where important factors have not been considered in the analysis.

d. Average Slope. If it is obvious that all of the pertinent variables are included in the analysis, then the variance of the points about the regression line is due entirely to measurement errors, and the resulting difference in slope of the regression lines is entirely artificial. In cases where all pertinent variables are considered and most of the measurement error is in one variable, that variable should be used as the dependent variable. Its errors will then not affect the slope of the regression line. In other cases where all pertinent variables are considered, an average slope should be used. An average slope can be obtained by use of the following equation:

$$b = S_y/S_x \quad (9-14)$$

#### 9-6. Multiple Linear Regression Example.

a. General. An example of a multiple linear regression analysis is illustrated on Figure 9-3. In this case, the volume of spring runoff is correlated with the water equivalent of the snow cover measured on April 1, the winter low-water flow (index of ground water) and the precipitation falling on the area during April. Here again, it was determined that logarithms of the values would be used in the regression equation. Although loss of 4 degrees of freedom of 12 available, as in this case, is not ordinarily desirable, the adjusted correlation coefficient attained (0.94) is particularly high, and the equation is consequently fairly reliable. The computations in Figure 9-3 were made with the HEC computer program MLRP (reference 50).

b. Logarithmic Transformation. In determining whether logarithms should be used for the dependent variable as above, questions such as the following should be considered: "Would an increase in snow cover contribute a greater increment to runoff under conditions of high ground water (wet ground conditions) than under conditions of low ground water?" If the answer is yes, then a logarithmic dependent variable (by which the effects are multiplied together) would be superior to an arithmetic dependent variable (by which the effects are added together). Logarithms should be used for the explanatory variables when they would increase the linearity of the relationship. Usually logarithms should be taken of values that have a natural lower limit of zero and a natural upper limit that is large compared to the values used in the study.

c. Function of Multiple Regression. It should be recognized that multiple regression performs a function that is difficult to perform graphically. Reliability of the results, however, is highly dependent on the availability of a large sampling of all important factors that influence the dependent variable. In this case, the standard error of an estimate as shown on Figure 9-3 is approximately 0.038, which, when added to a logarithm of a value, is equivalent to multiplying that value by 1.09. Thus, the standard error is about 9 percent, and the 1-in-20 error is roughly 18 percent. As discussed in Paragraph 9-3d, however, the calculated correlation coefficient may be accidentally high.

9-7. Partial Correlation. The value gained by using any single variable (such as April precipitation) in a regression equation can be measured by making a second correlation study using all of the variables of the regression equation except that one. The loss in correlation by omitting that variable is expressed in terms of the partial correlation coefficient. The square of the partial correlation coefficient is obtained as follows:

INPUT DATA

OBS NO	OBS ID	LOG Q	LOG SNO	LOG GW	LOG PRCP
1	1936	.939	.399	.325	.710
2	1937	.945	.343	.385	.634
3	1938	1.052	.369	.408	.886
4	1939	.744	.246	.428	.581
5	1940	.666	.181	.316	1.027
6	1941	1.081	.297	.460	1.315
7	1942	1.060	.299	.511	1.097
8	1943	.892	.354	.379	.707
9	1944	1.021	.295	.395	1.240
10	1945	.920	.321	.376	1.091
11	1946	.755	.168	.413	1.038
12	1947	.960	.280	.410	.979

STATISTICS OF DATA

VARIABLE	AVERAGE	VARIANCE	STANDARD DEVIATION	
LOG SNO	.2960	.0050	.0704	
LOG GW	.4005	.0028	.0531	
LOG PRCP	.9421	.0572	.2392	
LOG Q	.9196	.0181	.1346	DEPENDENT VARIABLE

UNBIASED CORRELATION COEFFICIENTS (R)

VARIABLE	LOG SNO	LOG GW	LOG PRCP	LOG Q
LOG SNO	1.0000	.0000	-.0459	.6308
LOG GW	.0000	1.0000	.1275	.4170
LOG PRCP	-.0459	.1275	1.0000	.2011
LOG Q	.6308	.4170	.2011	1.0000

REGRESSION RESULTS

INDEPENDENT VARIABLE	REGRESSION COEFFICIENT	PARTIAL DETERMINATION COEFFICIENT	
LOG SNO	1.621806	.9106	
LOG GW	1.012912	.6814	
LOG PRCP	.273390	.7451	

REGRESSION CONSTANT	R SQUARE	UNBIASED R SQUARE	STANDARD ERROR OF ESTIMATE
-.223698	.9437	.9226	.0375

Figure 9-3. Example Multiple Linear Regression Analysis.

$$r_{Y3.12}^2 = 1 - (1 - R_{Y.123}^2) / (1 - R_{Y.12}^2) \quad (9-15)$$

in which the subscript to the left of the decimal indicates the variable whose partial correlation coefficient is being computed, and the subscripts on the right of the decimal indicate the independent variables. An approximation of the partial correlation can sometimes be made by use of beta coefficients. After the regression equation has been calculated, beta coefficients are very easy to obtain by use of the following equation:

$$\beta_n = b_n S_n / S_Y \quad (9-16)$$

The beta coefficients of the variables are proportional to the influence of each variable on the result. While the partial correlation coefficient measures the increase in correlation that is obtained by addition of one more explanatory variable to the correlation study, the beta coefficient is a measure of the proportional influence of a given explanatory variable on the dependent variable. These two coefficients are related closely only when there is no interdependence among the various explanatory variables. However, some explanatory variables naturally correlate with each other, and when one is removed from the equation, the other will take over some of its weight in the equation. For this reason, it must be kept in mind that beta coefficients indicate partial correlation only approximately.

**9-8. Verification of Regression Results.** Acquisition of basic data after a regression analysis has been completed will provide an opportunity for making a check of the results. This is done simply by comparing the values of the dependent variable observed, with corresponding values calculated from the regression equation. The differences are the errors of estimate, and their root-mean-square is an estimate of the standard error of the regression-equation estimates (Paragraph 9-3). This standard error can be compared to that already established in Equation 9-13. If the difference is not significant, there is no reason to suspect the regression equation of being invalid, but if the difference is large, the regression equation and standard error should be recalculated using the additional data acquired.

**9-9. Regression by Graphical Techniques.** Where the relationships among variables used in a regression analysis are expected to be curvilinear and a simple transformation cannot be employed to make these relationships linear, graphical regression methods may prove useful. A satisfactory graphical analysis, however, requires a relatively large number of observations and tedious computations. The general theory employed is similar to that discussed above for linear regression. Methods used will not be discussed herein, but can be found in references 8 and 27.

**9-10. Practical Guidelines.** The most important thing to remember in making correlation studies is that accidental correlations occur frequently, particularly when the number of observations is small. For this reason, variables should be correlated only when there is reason to believe that there is a physical relationship. It is helpful to make preliminary examination of relationships between two or more variables by graphical plotting. This is particularly helpful for determining whether a relationship is linear and in selecting a transformation for converting curvilinear relationships to linear relationships. It should also be remembered that the chance of accidentally high

correlation increases with the number of correlations tried. If a variable being studied is tested against a dozen other variables at random, there is a chance that one of these will produce a good correlation, even though there may be no physical relation between the two. In general, the results of correlation analyses should be examined to assure that the derived relationship is reasonable. For example, if streamflow is correlated with precipitation and drainage area size, and the regression equation relates streamflow to some power of the drainage area greater than one, a maximum exponent value of one should be used, because the flow per square mile usually does not increase with drainage area when other factors remain constant.

#### 9-11. Regional Frequency Analysis.

a. General. In order to improve flood frequency estimates and to obtain estimates for locations where runoff records are not available, regional frequency studies may be utilized. Procedures described herein consist of correlating the mean and standard deviation of annual maximum flow values with pertinent drainage basin characteristics by use of multiple linear regression procedures. The same principles can be followed using graphical frequency and correlation techniques where these are more appropriate.

b. Frequency Statistics. A regional frequency correlation study is based on the two principal frequency statistics: the mean and standard deviation of annual maximum flow logarithms. Prior to relating these frequency statistics to drainage basin characteristics, it is essential that the best possible estimate of each frequency statistic be made. This is done by adjusting short-record values by the use of longer records at nearby locations. When many stations are involved, it is best to select long-record base stations for each portion of the region. It might be desirable to adjust the base station statistics by use of the one or two longest-record stations in the region, and then adjust the short-record station values by use of the nearest or most appropriate base station. Methods of adjusting statistics are discussed in Section 3-7.

c. Drainage-Basin Characteristics. A regional analysis involves the determination of the main factors responsible for differences in precipitation or runoff regimes between different locations. This would be done by correlating important factors with the long-record mean and with the long-record standard deviation of the frequency curve for each station (the long-record values are those based on extension of the records as discussed in Section 3-7). Statistics based on precipitation measurements in mountainous terrain might be correlated with the following factors:

- Elevation of station
- General slope of surrounding terrain
- Orientation of that slope
- Elevation of windward barrier
- Exposure of gage
- Distance of leeward controlling ridge

Statistics based on runoff measurements might be correlated with the following factors:

- Drainage area (contributing)
- Stream length
- Slope of drainage area or of main channel
- Surface storage (lakes and swamps)
- Mean annual rainfall
- Number of rainy days per year
- Infiltration characteristics
- Urbanized Area

d. Linear Relationships. In order to obtain satisfactory results using multiple linear regression techniques, all variables must be expressed so that the relation between the independent and any dependent variable can be expected to be linear, and so that the interaction between two independent variables is reasonable. An illustration of the first condition is the relation between rainfall and runoff. If the runoff coefficient is sensibly constant, as in the case of urban or airport drainage, then runoff can be expected to bear a linear relation to rainfall. However, in many cases initial losses and infiltration losses cause a marked curvature in the relationship. Ordinarily, it will be found that the logarithm of runoff is very nearly a linear function of rainfall, regardless of loss rates, and in such cases, linear correlation of logarithms would be most suitable. An illustration of the second condition is the relation between rainfall,  $D$ , drainage area,  $A$ , and runoff,  $Q$ . If the relation used for correlation is as follows:

$$Q = aD + bA + c \quad (9-17)$$

then it can be seen that one inch change in precipitation would add the same amount of flow, regardless of the size of drainage area. This is not reasonable, but again a transformation to logarithms would yield a reasonable relation:

$$\log Q = d \log D + e \log A + \log f \quad (9-18)$$

or transformed:

$$Q = fD^d A^e \quad (9-19)$$

Thus, if logarithms of certain variables are used, doubling one independent quantity will multiply the dependent variable by a fixed ratio, regardless of what fixed values the other independent variables have. This particular relationship is reasonable and can be easily visualized after a little study. There is no simple rule for deciding when to use

logarithmic transformation. It is usually appropriate, however, when the variable has a fixed lower limit of zero. The transformation should provide for near-uniform variance throughout the range of data.

e. Example of Regional Correlation. An illustrative example of a regional correlation analysis for the mean log of annual flood peaks (Y) with several basin characteristics is shown on Figure 9-4. In this example, the dependent variable is primarily related to the drainage area size, but precipitation and slope added a small amount to the adjusted determination coefficient. The regression equation selected for the regional analysis included only drainage area as an independent variable.

f. Selection of Useful Variables. In the regression equations shown on Figure 9-4, the adjusted determination coefficient increases as variables are deleted according to their lack of ability to contribute to the determination. This increase is because there is a significant increase in the degrees of freedom as each variable is deleted for this small sample of 20 observations. Both the adjusted determination coefficient and standard error of estimate should be reviewed to determine how many variables are included in the adopted regression equation. Even in the case of a slight increase in correlation obtained by adding a variable, consideration of the increased unreliability of R as discussed in Section 9-3 might indicate that the factor should be eliminated in cases of small samples. The simplest equation that provides an adequate predictive capability should be selected. In this example, there is some loss in determination in only using drainage area, but this simple equation is adopted to illustrate regional analysis. The adopted equation is:

$$\log Y = 1.586 + 0.962 \log (\text{AREA}) \quad (9-20)$$

or

$$Y = 38.5 \text{ AREA}^{.962} \quad (9-21)$$

The  $\bar{R}^2$  for this equation is 0.839.

g. Use of Map. Many hydrologic variables cannot be expressed numerically. Examples are soil characteristics, vegetal cover, and geology. For this reason, numerical regional analysis will explain only a portion of the regional variation of runoff frequencies. The remaining unexplained variance is contained in the regression errors, which varies from station to station. These regression errors are computed by subtracting the predicted values from the observed values for each station. These errors can then be plotted on a regional map at the centroid of each station's area, and lines of equal values drawn (perhaps using soils, vegetation, or topographic maps as a guide). Combining this regional error with the regression equation should be much better than using the single constant for the entire region. In smoothing lines on such a map, consideration should be given to the reliability of computed statistics. Equations 8-1 and 8-2 can be used to compute the standard errors of estimating means and standard deviations. In Figure 9-5 for example, Station 5340 (observation 11) had 66 years of record and the standard error for the mean was 0.028. There is about one chance in three that the mean is in error by more than 0.029 or about one chance in twenty that the mean is in error by more than 0.056 (twice the standard error). Figure 9-6 shows a map of the errors and Figure 9-5 shows the regional map values for each station and evaluates the worth of the map. The map has a mean square error of 0.0112 compared to that of 0.0356 for the regression equation alone.

INPUT DATA

OBS NO	OBS ID	AREA	SLOPE	LENGTH	LAKES	ELEV	FOREST	PRECIP	SOILS	MEAN
1	5090	292.0	6.3	31.5	1.5	1.230	35.0	40.0	3.5	3.783
2	5140	185.0	14.3	30.3	1.0	1.024	52.0	38.2	3.2	3.783
3	5180	282.0	44.0	29.8	1.0	1.740	64.0	35.0	3.3	4.030
4	5200	298.0	20.1	37.3	1.0	1.600	30.0	36.5	3.3	4.044
5	5205	771.0	24.4	44.8	1.0	1.447	33.0	36.0	3.2	4.333
6	5260	114.0	35.8	17.5	1.0	1.383	30.0	34.0	3.0	3.751
7	5270	52.2	29.9	17.4	1.0	1.489	26.0	33.0	3.0	2.637
8	5280	66.8	12.6	20.1	1.0	1.305	41.0	33.6	3.0	3.186
9	5305	77.5	45.2	18.9	1.0	1.123	27.0	34.6	3.0	3.348
10	5320	215.0	17.7	27.5	1.1	.966	57.0	35.5	3.0	3.995
11	5340	383.0	21.3	36.7	3.5	1.370	54.0	42.0	3.3	4.122
12	5375	15.7	291.0	5.6	1.0	1.350	81.0	40.0	3.5	2.722
13	5380	43.8	52.2	22.1	1.0	1.300	85.0	43.0	3.5	3.078
14	5390	274.0	39.6	32.3	1.6	1.200	70.0	43.0	2.8	3.930
15	5445	136.0	37.4	22.7	1.0	1.800	94.0	43.0	4.9	3.590
16	5485	604.0	22.8	44.5	1.0	1.900	83.0	37.0	3.2	4.092
17	5495	37.7	54.8	15.2	1.0	1.350	67.0	37.8	3.2	3.284
18	5500	173.0	45.7	24.2	1.0	1.700	65.0	39.0	3.9	3.816
19	5520	443.0	24.4	56.0	1.3	1.600	89.0	44.0	3.2	4.275
20	5525	23.8	115.7	10.0	1.0	1.800	80.0	47.0	4.3	3.249

STATISTICS OF DATA

VARIABLE	AVERAGE	VARIANCE	STANDARD DEVIATION	
AREA	2.1488	.2249	.4743	
SLOPE	1.5105	.1255	.3542	
LENGTH	1.3832	.0550	.2345	
STORAGE	.0540	.0171	.1308	
ELEV	1.4339	.0707	.2659	
FOREST	1.7293	.0353	.1879	
PRECIP	1.5845	.0020	.0444	
SOILS	.5230	.0034	.0581	
MEAN	3.6524	.2455	.4955	DEPENDENT VARIABLE

UNBIASED CORRELATION COEFFICIENTS

VARIABLE	AREA	SLOPE	LENGTH	LAKES	ELEV	FOREST	PRECIP	SOILS	MEAN
AREA	1.0000	-.6327	.9304	-.2749	.0000	.0000	.0000	.0000	.9159
SLOPE	-.6327	1.0000	-.7144	-.1318	.1187	.3635	.1867	.2053	-.4521
LENGTH	.9304	-.7144	1.0000	.2345	.0000	.0000	.0000	-.1096	.8263
STORAGE	.2749	-.1318	.2345	1.0000	.0000	.0000	.2812	.0000	.2596
ELEV	.0000	.1187	.0000	.0000	1.0000	.2791	.0000	.5297	.0000
FOREST	.0000	.3635	.0000	.0000	.2791	1.0000	.6877	.4304	.0000
PRECIP	.0000	.1867	.0000	.2812	.0000	.6877	1.0000	.5412	.0000
SOILS	.0000	.2053	-.1096	.0000	.5297	.4304	.5412	1.0000	.0000
MEAN	.9159	-.4521	.8263	.2596	.0000	.0000	.0000	.0000	1.0000

SUMMARY OF REGRESSION ANALYSES FOR MEAN LOG OF ANNUAL PEAKS

REGRESSION CONSTANT	AREA (LOG)	SLOPE (LOG)	LENGTH (LOG)	LAKES (LOG)	ELEV (NONE)	FOREST (LOG)	PRECIP (LOG)	SOILS (LOG)	ADJUSTED DETERMINATION COEFFICIENT	STANDARD ERROR OF ESTIMATE	MEAN SQUARE ERROR
-1.522	1.261	0.182	-0.328	-0.272	-0.184	0.055	1.739	0.140	0.8097	0.2162	0.0257
-1.633	1.269	0.169	-0.364	-0.289	-0.169	0.054	1.874	-----	0.8254	0.2070	0.0257
-1.808	1.267	0.179	-0.350	-0.301	-0.165	-----	2.023	-----	0.8386	0.1990	0.0258
-1.668	1.130	0.243	-----	-0.251	-0.167	-----	1.753	-----	0.8469	0.1939	0.0263
-1.130	1.104	0.250	-----	-----	-0.129	-----	1.399	-----	0.8537	0.1896	0.0269
-1.034	1.069	0.198	-----	-----	-----	-----	1.319	-----	0.8584	0.1865	0.0278
-1.134	0.975	-----	-----	-----	-----	-----	1.699	-----	0.8553	0.1885	0.0302
1.586	0.962	-----	-----	-----	-----	-----	-----	-----	0.8390	0.1988	0.0356

Figure 9-4. Regression Analysis for Regional Frequency Computations.

REGIONAL ANALYSIS WITH REGRESSION ON DRAINAGE AREA ONLY

OBS NO	STATION	OBSERVED	COMPUTED	ERROR	MAP VALUE	DIFF	DIFF <sup>2</sup>	YEARS OF RECORD	STANDARD DEVIATION	S <sub>y</sub>
1	5090	3.783	3.957	-0.174	-0.18	0.006	0.000036	43	0.190	0.029
2	5140	3.783	3.766	0.017	0.01	0.007	0.000049	52	0.195	0.027
3	5180	4.030	3.942	0.088	0.09	-0.002	0.000004	39	0.289	0.046
4	5200	4.044	3.965	0.079	0.07	0.009	0.000081	27	0.256	0.049
5	5205	4.333	4.362	-0.029	0.08	-0.109	0.011881	50	0.251	0.035
6	5260	3.751	3.564	0.187	0.11	0.077	0.005929	35	0.293	0.050
7	5270	2.637	3.238	-0.601	-0.22	-0.381	0.145161	31	0.206	0.037
8	5280	3.186	3.341	-0.155	-0.17	0.015	0.000225	45	0.186	0.028
9	5305	3.348	3.403	-0.055	-0.04	-0.015	0.000225	44	0.128	0.019
10	5320	3.995	3.829	0.166	0.16	0.006	0.000036	66	0.288	0.035
11	5340	4.122	4.070	0.052	0.02	0.032	0.001024	66	0.227	0.028
12	5375	2.722	2.736	-0.014	-0.05	0.036	0.001296	40	0.323	0.051
13	5380	3.078	3.164	-0.086	-0.08	-0.006	0.000036	60	0.226	0.029
14	5390	3.930	3.930	0.000	0.00	0.000	0.000000	41	0.261	0.041
15	5445	3.590	3.638	-0.048	-0.15	0.102	0.010404	39	0.278	0.045
16	5485	4.092	4.260	-0.168	-0.08	-0.088	0.007744	61	0.242	0.031
17	5495	3.284	3.102	0.182	0.04	0.142	0.020164	39	0.242	0.039
18	5500	3.816	3.738	0.078	0.08	-0.002	0.000004	66	0.237	0.029
19	5520	4.275	4.131	0.144	0.17	-0.026	0.000676	54	0.277	0.038
20	5525	3.249	2.910	0.339	0.20	0.139	0.019321	39	0.291	0.047
Sum					0.06	-0.058	0.224296			
Average					0.003	-0.003	0.0112			

Figure 9-5. Regional Analysis Computations for Mapping Errors.

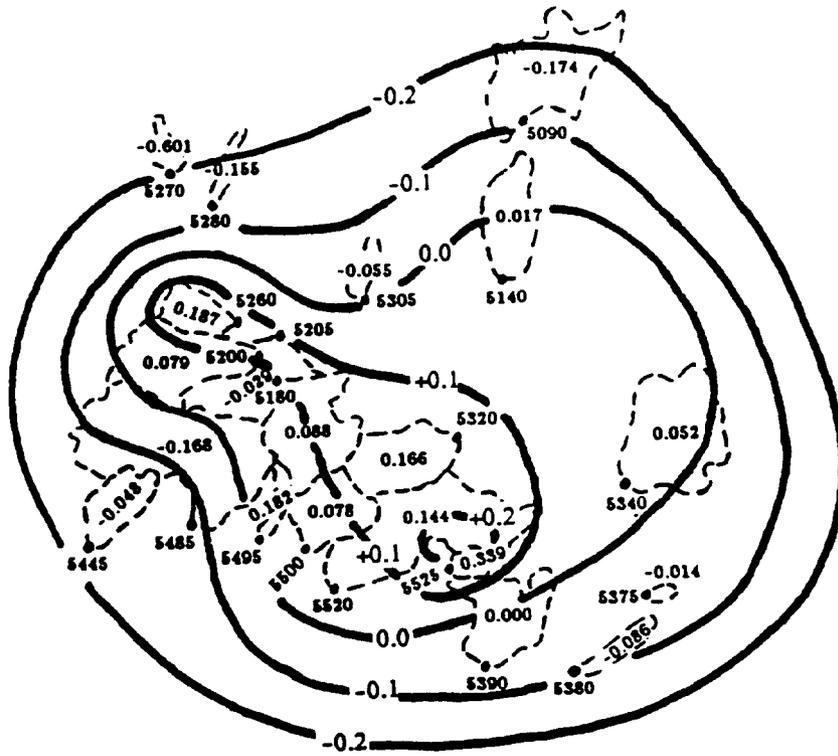


Figure 9-6. Regional Map of Regression Errors.

h. Summary of Procedure. A regional analysis of precipitation or flood-flow frequencies is generally accomplished by performing the following steps:

- (1) Select long-record base stations within the region as required for extension of records at each of the short-record stations.
- (2) Tabulate the maximum events of each station.
- (3) Transform the data to logarithms and calculate  $\bar{X}$ , S and, if appropriate, G (Equations 3-1, 3-2 and 3-3) for each base station.
- (4) Calculate  $\bar{X}$  and S for each other station and for the corresponding values of the base station, and calculate the correlation coefficient (Equation 3-16).
- (5) Adjust all values of  $\bar{X}$  and S by use of the base station, (Equations 3-17 and 3-19). (If any base station is first adjusted by use of a longer-record base station, the longer-record statistics should be used for all subsequent adjustments.)
- (6) Select meteorological and drainage basin parameters that are expected to correlate with  $\bar{X}$  and S, and tabulate the values for each drainage basin or representative area.
- (7) Calculate the regression equations relating  $\bar{X}$  and S to the basin characteristics, using procedures explained in Section 9-2, and compute the corresponding determination coefficients.
- (8) Eliminate variables in turn that contribute the least to the determination coefficient, recomputing the determination coefficient each time, and select the regression equation having the highest adjusted determination coefficient, or one with fewer variables if the adjusted determination coefficient is nearly the same.
- (9) Compute the regression errors for each station, plot on a suitable map, and draw isopleths of the regression errors for the regression equations of  $\bar{X}$  (see Figures 9-5 and 9-6 for an example) and S considering the standard error for each computed, or adjusted,  $\bar{X}$  and S. Note that an alternate procedure is to add the regression constant to each error value and develop a map of this combined value. This procedure eliminates the need to keep the regression constant in the regression equation as the mapped value now includes the regression constant.
- (10) A frequency curve can be computed for any ungaged basin in the area covered within the mapped region by using the adopted regression equations and appropriate map values to obtain  $\bar{X}$  and S, and then using the procedures discussed in Section 3-2 to compute several points to define the frequency curve. (It may also be necessary to develop regional (generalized) values of the skew coefficient if the Pearson type III distribution is considered appropriate. The next section describes the necessary steps to compute a generalized skew coefficient.)

i. Generalized Skew Determinations. Skew coefficients for use in hydrologic studies should be based on regional studies. Values based on individual records are highly unreliable. Figure 9-7 is a plot of skew coefficients sequentially recomputed after adding the annual peak for the given year. Note that, after 1950, the skew coefficient was at a minimum of about 0.5 in 1954 and maximum of about 1.9 in 1955, only one year

apart. The procedures for developing generalized skew values are generally set forth in Bulletin 17B (pages 10-15).

In summary, it is recommended that:

- (1) the stations used in the study have 25 or more years of data,
- (2) at least 40 stations be used in the analysis, or at least all stations surrounding the area within 100 miles should be included,
- (3) the skew values should be plotted at the centroid of the basins to determine if any geographic or topographic trends are present,
- (4) a prediction equation should be developed to relate the computed skew coefficients to watershed and climate variables,
- (5) the arithmetic mean of at least 20 stations, if possible, in an area of reasonably homogeneous hydrology should be computed, and
- (6) then select the method that provides the most accurate estimation of the skew coefficient (smallest mean-square error).

In addition to the above guidelines, care should be taken to select stations without significant man-made changes such as reservoirs, urbanization, etc.

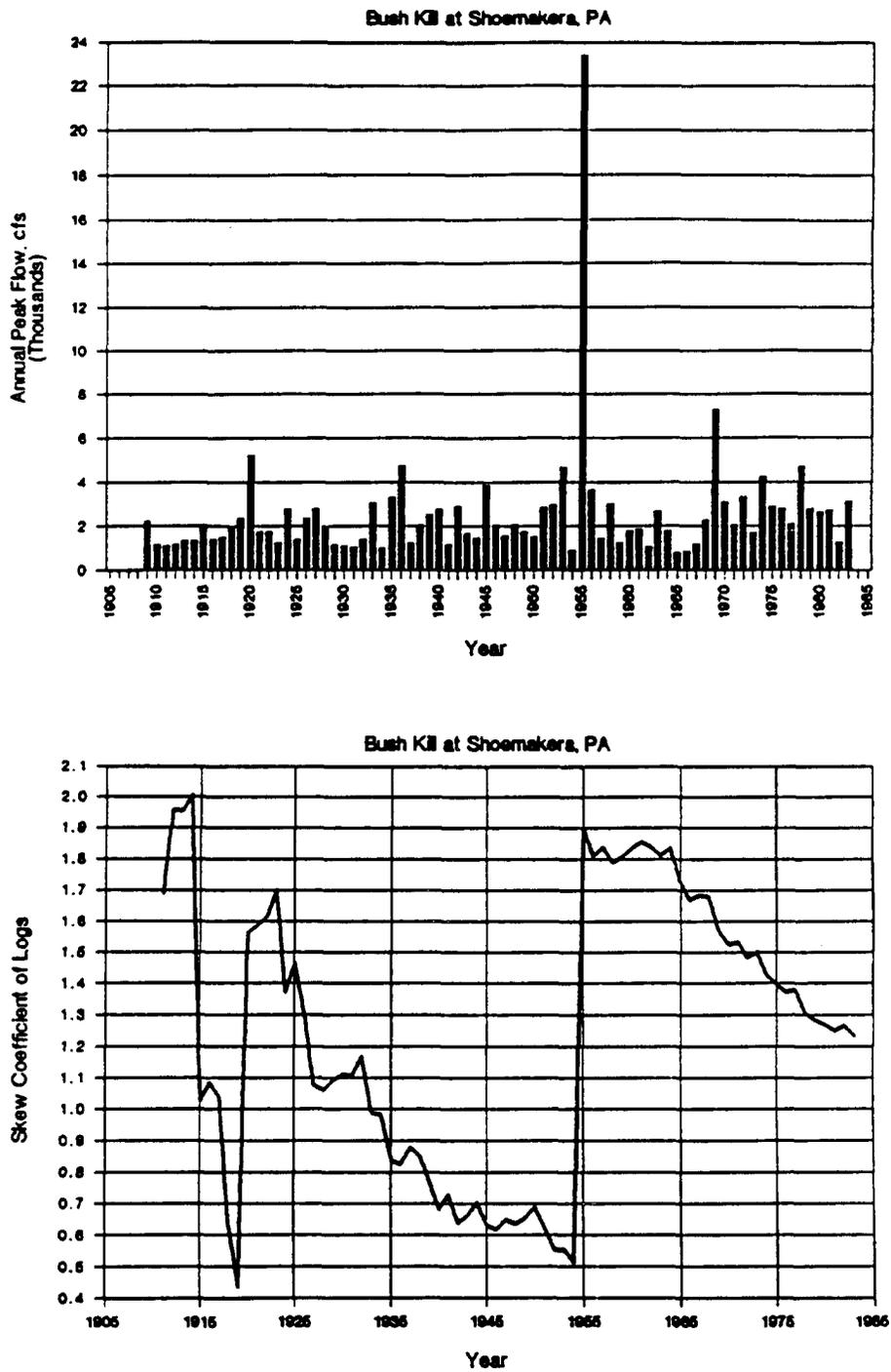


Figure 9-7. Annual Peaks and Sequential Computed Skew by Year.

CHAPTER 10  
ANALYSIS OF MIXED POPULATIONS

10-1. Definition. The term mixed population, in a hydrologic context, is applied to data that results from two or more different, but independent, causative conditions. For example, floods originating in a mountainous area or the northern part of the United States at a given site could be caused by melting snow or by rain storms. Along the Gulf and Atlantic coasts, floods can be caused by general cyclonic storms or by intense tropical storms. A frequency curve representing the events caused by one of the climatic conditions may have a significantly different slope (standard deviation) than for the other condition. A frequency plot of the annual events, irrespective of cause, may show a rather sudden change in slope and the computed skew coefficient may be comparatively high. In these situations, a frequency curve derived by combining the frequency curves of each population can result in a computed frequency relation more representative of the observed events.

10-2. Procedure.

a. The largest annual event is selected for each causative condition. As Bulletin 17B (46) cautions, "If the flood events that are believed to comprise two or more populations cannot be identified and separated by an objective and hydrologically meaningful criterion, the record shall be treated as coming from one population." Also, Bulletin 17B states, "Separation by calendar periods in lieu of separation by events is not considered hydrologically reasonable unless the events in the separated periods are clearly caused by different hydrometeorologic conditions."

b. The frequency relations for each separate population can be derived by the graphical or analytical techniques described in Chapter 2 and then combined to yield the mixed population frequency curve. The individual annual frequency curves are combined by "probability of union." For two curves, the equation is:

$$P_c = P_1 + P_2 - P_1P_2 \quad (10-1)$$

where:

$P_c$  = Annual exceedance probability of combined populations for a selected magnitude.

$P_1$  = Annual exceedance probability of same selected magnitude for population series 1.

$P_2$  = Annual exceedance probability of same magnitude selected above for population series 2.

c. Figure 10-1 illustrates a combined annual-event frequency curve derived by combining a hurricane event frequency curve with a nonhurricane event curve for the Susquehanna River at Harrisburg, PA. For more than two population series,  $n$  curves may be more easily combined by the following form:

$$P_c = 1 - (1-P_1)(1-P_2) \dots (1-P_n) \quad (10-2)$$

NOTE: The exceedance probability (percent chance exceedance divided by 100) must be used in the above equations.

d. If partial duration curves are to be added, the equation is simply  $P_c = P_1 + P_2$ . This assumes that the events in both series are hydrologically independent. When the combined curve is used in an economic analysis, the events in both series must also be economically independent.

### 10-3. Cautions.

a. If annual flood peaks have been separated by causative factors, a generalized skew must be derived for each separate series to apply the log-Pearson Type III distribution as recommended by Bulletin 17B. Plate 1 of Bulletin 17B or any other generalized skew map based on the maximum annual event, irrespective of cause, will not be applicable to any of the separated series. Derivation of generalized skew relations for each series can involve much effort.

b. Some series may not have an event each year. For example, tropical storms do not occur every year over most drainage areas in the United States, and quite often there are only a few flood events for the series. Extensive regionalization may be necessary to reduce the probable error in the frequency relations which results from small sample sizes.

c. Sometimes frequency relations of particular seasons are of interest, i.e., quarterly or monthly, and the curves are combined to verify the annual series curve. The combined curve will very likely fit the annual curve only in the middle parts of the curve. The lower end of the curve will have a partial duration shape as many small events have been included in the analysis. Also, it is possible that the slope of the frequency relation will be higher at the upper end of the curve as the one season or month with the maximum event included in its series will likely have a higher slope than that of the annual series.

d. A basic assumption of this procedure is that each series is independent of the other. Coincidental frequency analysis techniques must be used where dependance is a factor. For instance, the frequency curves of two or more tributary stations cannot be combined by the above equation to derive the frequency curve of a downstream site. This is because the downstream flow is a function of the summation of the coincident flows on each of the tributaries.

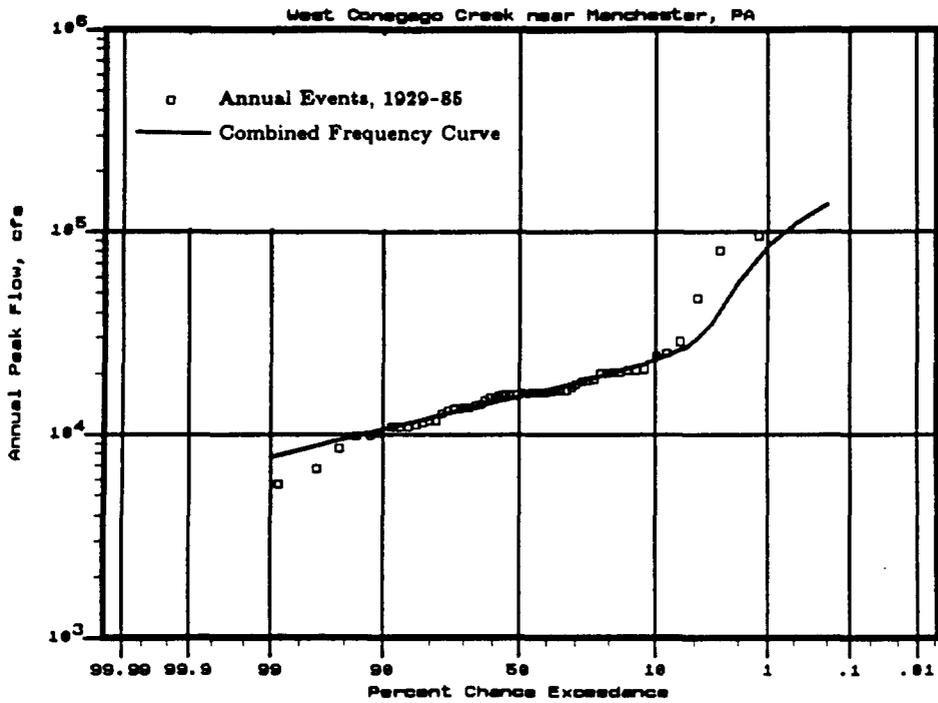
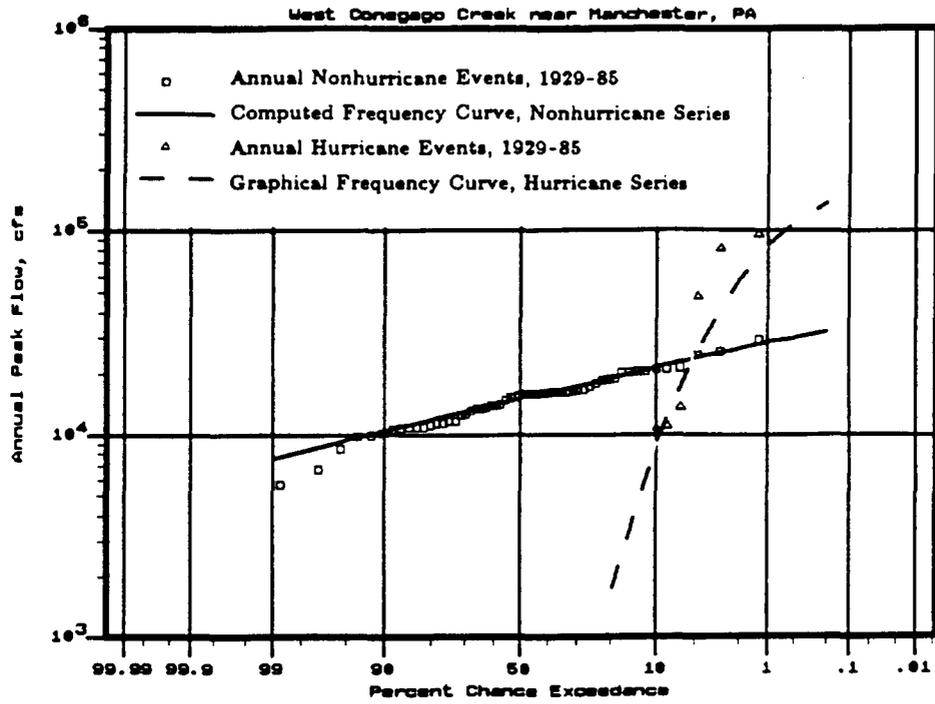


Figure 10-1. Nonhurricane, Hurricane, and Combined Flood Frequency Curves.

CHAPTER 11  
FREQUENCY OF COINCIDENT FLOWS

11-1. Introduction. In many cases of hydrologic design, it is necessary to consider only those events which occur coincidentally with other events. For example, a pump station is usually required to pump water only when interior runoff occurs at a time that the main river stage is above interior ponding levels. In constructing a frequency curve of interior runoff that occurs only at such times, data selected for direct use should be limited to that recorded during high river stages. In some cases, such data might not be adequate, but it is possible in these cases where the two types of events are not highly correlated to make indirect use of noncoincident data in order to establish a more reliable frequency curve of coincident events.

11-2. A Procedure for Coincident Frequency Analysis.

a. Objective. Determine an exceedance-frequency relationship for a variable C. Variable C is a function of two variables, A and B.

b. Selection of Dominant Variable. The variable that has the largest influence on variable C is designated as variable A; the less influential variable is designated as variable B. The significance of "influential" will be indicated by means of an example. Figure 11-1 shows water surface profiles along a tributary near the junction with a main river. Stage on the tributary (variable C) is a function of main river stage and tributary discharge. In Region I, main river stage, will tend to have the dominant influence on tributary stage, whereas in Region II, tributary discharge will tend to dominate. The boundary between Regions I and II cannot be precisely defined and will vary with exceedance frequency. Stage-frequency determinations will be least accurate in the vicinity of the boundary where both variables have a substantial impact on the combined result.

c. Procedure.

(1) Construct a duration curve for variable B. Discretize the duration curve with a set of "index" values of B. Index values should represent approximately equal ranges of magnitude of variable B. The area under the resulting discretized duration curve should equal the area under the original duration curve. The number of index values of B required for discretization depends on the range of variation of B and the sensitivity of variable C to B. Therefore, the number of points selected should adequately define the relationships.

(2) For each of the index values of variable B, develop a relationship between variable A and the combined result C. In the illustration (Figure 11-1) the relationship linking variables A, B and C would be obtained with a set of water surface profile calculations for various combinations of main river stage and tributary discharge.

(3) If variables A and B are independent of each other, construct an exceedance-frequency curve of variable A. If the variables are not independent,

construct a conditional exceedance-frequency curve of variable A for each index value of variable B.

(4) Using the relationship developed in step (2) and frequency curve(s) developed in step (3), construct a conditional exceedance-frequency curve of variable C for each index value of variable B.

(5) For a selected magnitude of variable C, multiply the exceedance-frequencies from each curve developed in step (4) by the corresponding proportions of time represented, and sum these products to obtain the exceedance-frequency of variable C. Repeat this step for other selected magnitudes of C until a complete exceedance-frequency curve for variable C is defined. This step is an application of the total probability theorem.

d. Seasonal Effects. The duration of frequency curves from steps (1) and (3) are assumed to represent stationary processes. That is, it is assumed that probabilities and exceedance frequencies obtained from the curves do not vary with time. In order for this assumption to be reasonably valid, it is generally necessary to follow the above procedure on a seasonal basis. Once seasonal exceedance-frequency curves have been obtained (step e), they may be combined to obtain an all-season exceedance-frequency curve.

e. Assumption of Independence. Although step (3) enables application of the procedure to situations where variables A and B are not independent, data is generally not available to establish the conditional exceedance frequency curves required by that step. Consequently, application of the procedure presented here is generally limited to situations where it is reasonable to assume that variables A and B are independent.

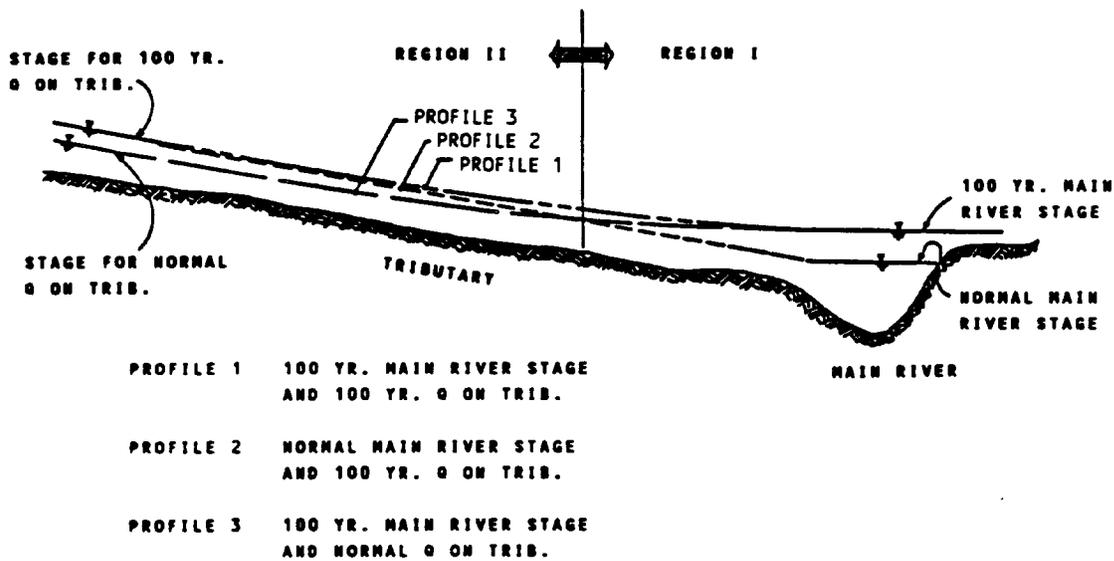


Figure 11-1. Illustration of Water Surface Profiles in Coincident Frequency Analysis.

## CHAPTER 12

### STOCHASTIC HYDROLOGY

#### 12-1. Introduction.

a. A stochastic process is one in which there is a chance component in each successive event and ordinarily some degree of correlation between successive events. Modeling of a stochastic process involves the use of the "Monte Carlo" method of adding a random (chance) component to a correlated component in order to construct each new event. The correlated component can be related, not only to preceding events of the same series, but also to concurrent and preceding events of series of related phenomena.

b. Work in stochastic hydrology has related primarily to annual and monthly streamflows, but the results often apply to other hydrologic quantities such as precipitation and temperatures. Some work on daily streamflow simulation has been done.

#### 12-2. Applications.

a. Hydrologic records are usually shorter than 100 years in length, and most of them are shorter than 25 years. Even in the case of the longest records, the most extreme drought or flood event can be far different from the next most extreme event. There is often serious question as to whether the extreme event is representative of the period of record. The severity of a long drought can be changed drastically by adding or subtracting 1 year of its duration. In order that some estimate of the likelihood of more severe sequences can be made, the stochastic process can be simulated, and long sequences of events can be generated. If the generation is done correctly, the hypothetical sequence would have as equal likelihood of occurrence in the future as did the observed record.

b. The design of water resource projects is commonly based on assumed recurrence of past hydrologic events. By generating a number of hydrologic sequences, each of a specified desired length, it is possible to create a much broader base for hydrologic design. While it is not possible to create information that is not already in the record, it is possible to use the information more systematically and more effectively. In selecting the number and length of hydrologic sequences to be generated, it is usually considered that 10 to 20 sequences would be adequate and that their length should correspond to the period of project amortization.

c. It must be recognized that the more hydrologic events that are generated, the more chance there is that an extreme event or combination of events will be exceeded. Consequently, it is not logical that a design be based on the most extreme generated event, but rather on some consideration of the total consequences that would prevail for a given design if all generated events should occur. The more events that are generated, the less proportional weight each event is given. If a design is tested on 10 sequences of hydrologic events, for example, the benefits and costs associated with each sequence would be divided by 10 and added in order to obtain the "expected" net benefits.

12-3. Basic Procedure. Successful simulation of stochastic processes in hydrology has been based generally on the concept of multiple linear regression, where the regression

equation determines the correlated component, and the standard error of estimate determines the random component. Figure 12-1 illustrates the general nature of the process. In this case, a low degree of correlation is illustrated, in order to emphasize important aspects of the process. It can be seen that, if every estimate of the dependent variable is determined by the regression line (Figure 12-1a), the estimated points would be perfectly correlated with the independent variable and would have a much smaller range of magnitude than the actual observed values of the dependent variable. In order to avoid such unreasonable results, it is necessary to add a random component to each estimate (Figure 12-1b), and this random component should conform to the scatter of the observed data about the regression line.

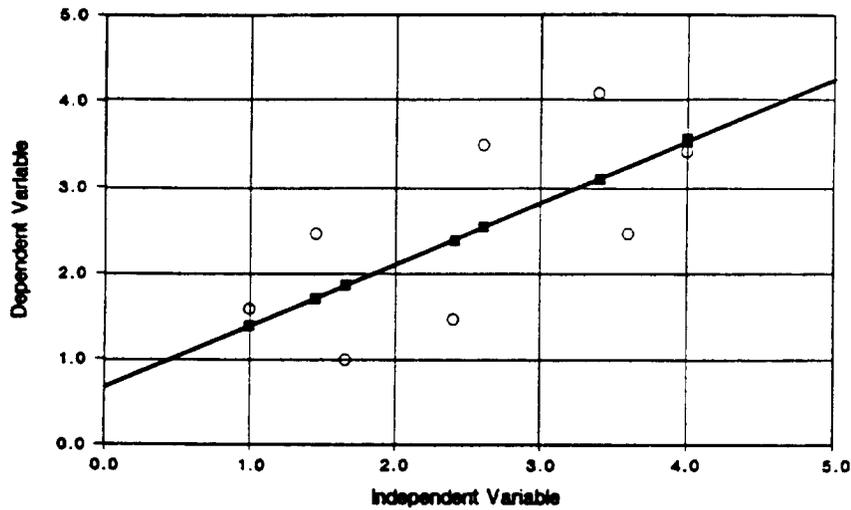


Figure 12-1a. Data Estimation from Regression Line.

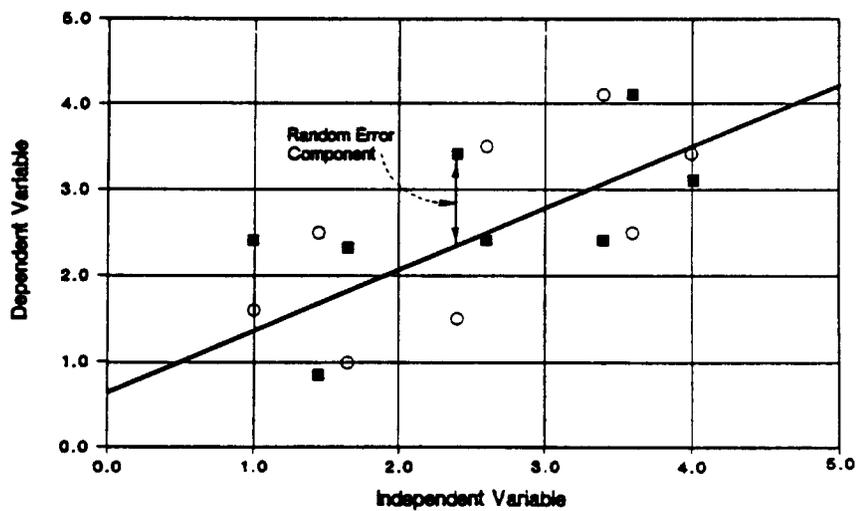


Figure 12-1b. Data Estimation with Addition of Random Errors.

12-4. Monthly Streamflow Model.

a. In accordance with the above basic procedure, a simulation model for generating values of a variable which can be defined only partially by a deterministic relation is:

$$Y = a + b_1X_1 + b_2X_2 + ZS_Y(1-R^2)^{1/2} \quad (12-1)$$

where:

- Y = dependent variable
- a = regression constant
- $b_1, b_2$  = regression coefficients
- $X_1, X_2$  = independent variables
- Z = random number from normal standard population with zero mean and unit variance
- $S_Y$  = standard deviation of dependent variable
- R = multiple correlation coefficient

b. This type of simulation model can be used to generate related monthly streamflow values at one or more stations. Multiple linear regression theory is based on the assumed distribution of all variables in accordance with the Gaussian normal distribution. Therefore, mathematical integrity requires that each variable be transformed to a normal distribution, if it is not already normal. It has been found that the logarithms of streamflows are approximately normally distributed in most cases. For computational efficiency it is convenient to work with deviations from the mean which have been normalized by dividing by the standard deviation. This deviate is sometimes called the Pearson Type III deviate and can be computed as follows:

$$t_i = (X_{i,j} - \bar{X}_i)/S_i \quad (12-2)$$

where:

- t = Pearson Type III deviate
- i = month number
- j = year number
- X = logarithm of flow
- $\bar{X}$  = mean of flow logarithms
- S = standard deviation of flow logarithms

c. If these deviates exhibit a skewness, they can be further transformed, if necessary, to a distribution very close to normal by use of the following approximate Pearson Type III transform equation:

$$K_i = (6/G_i) \{[(G_i t_i / 2) + 1]^{1/3} + 1\} + G_i / 6 \quad (12-3)$$

where:

**K** = normal standard deviate

**i** = month number

**G** = skew coefficient

**t** = Pearson Type III deviate as defined in Equation 12-2

An equation for generating monthly streamflow is:

$$K'_{i,k} = \beta_1 K'_{i,1} + \beta_2 K'_{i,2} + \dots + \beta_{k-1} K'_{i,k-1} + \beta_k K'_{i-1,k} \\ + \beta_{k+1} K'_{i-1,k+1} + \beta_n K'_{i-1,n} + Z_{i,k} (1-R_{i,k}^2)^{1/2} \quad (12-4)$$

where:

**K'** = monthly flow logarithm, expressed as a normal standard deviate

**$\beta$**  = beta coefficient, defined as  $b_{i,m} S_{i,m} / S_{i,k}$  where *m* is a station not equal to *k* and *b* is the regression coefficient.

**i** = month number for value being generated

**k** = station number for value being generated

**n** = number of interrelated stations

**R** = multiple correlation coefficient

**Z** = random number from normal standard population

For the case of a single station, this resolves to:

$$K_i' = R_{i,i-1} K_{i-1}' + Z_i (1 - R_{i,i-1}^2)^{1/2} \quad (12-5)$$

d. Note that Equation 12-5 is very similar to Equation 12-1. The differences result from using normal standard deviates. When this is done, the regression constant,  $a$ , equals zero, the regression coefficients,  $b$ , become beta coefficients,  $\beta$ , and the standard deviation,  $S$ , does not appear in the random component since it equals 1. Note also that one of the independent variables is the flow for the preceding month in order to preserve the inherent serial correlation. The flow value in the original units is computed by reversing the transformation process, i.e., from normal standard deviate to Pearson Type III deviate, to logarithm of flow and finally flow value.

e. A step-by-step procedure for generating monthly streamflows for a number of interrelated locations having simultaneous records is as follows:

- (1) Compute the logarithm of each streamflow quantity. If a value of zero streamflow is possible, it is necessary to add a small increment, such as 0.1 percent of the mean annual flow, to each monthly quantity before taking the logarithm.
- (2) Compute the mean, standard deviation and skew coefficient of the values for each location and each month, using equations given in Chapter 2.
- (3) For each month and location, subtract the mean from each event and divide by the standard deviation (Equation 12-2).
- (4) Transform these "standardized" quantities to a normal distribution by use of Equation 12-3.
- (5) Arrange the locations in any sequence, and compute a regression equation for each location in turn for each month. In each case, the independent variables will consist of concurrent monthly values at preceding stations and preceding monthly values at the current and subsequent stations.
- (6) Generate standardized variates for each location in turn for each month, starting with the earliest month of generated data. This is accomplished by computing a regression value and adding a random component. The random component, according to Equation 12-5, is a random selection from a normal distribution with zero mean and unit standard deviation, multiplied by the alienation coefficient which is  $(1 - R^2)^{1/2}$ .
- (7) Transform each generated value by reversing the transform of Equation 12-3 with the appropriate skew coefficient, multiplying by the standard deviation and adding to mean in order to obtain the logarithm of streamflow.
- (8) Find the antilogarithm of the value determined in step (7) and subtract the small increment added in step (1). If a negative value results, set it to zero.

f. It is obviously not feasible to accomplish the above computations without the use of an electronic computer. A computer program, HEC-4 Monthly Streamflow Simulation (51) can be used for this purpose.

12-5. Data Fill In. Ordinarily, periods of recorded data at different locations do not cover the same time span, and therefore, it is necessary to estimate missing values in order to obtain a complete set of data for analysis as described above. In estimating the missing values, it is important to preserve all statistical characteristics of the data, including frequency and correlation characteristics. To preserve these characteristics, it is necessary to estimate each individual value on the basis of multiple correlation with the preceding value at that location and with the concurrent or preceding values in all other locations. A random component is also required, as indicated in Equation 12-1.

12-6. Application In Areas of Limited Data. The streamflow generation models discussed so far have assumed that sufficient records were available to derive the appropriate statistics. For instance, the monthly streamflow model requires four frequency and correlation coefficients for each of the 12 months, or 48 values for one station simulation. A model has been developed (51) that combines the coefficients into a few generalized coefficients for the purpose of generating monthly streamflow at ungaged locations. (Procedures for determining generalized statistics for use in generating daily flows have not yet been developed.) The generalized model considers the following:

- season of maximum runoff
- lag to season of minimum runoff
- average runoff
- variation between maximum and minimum runoff
- standard deviation of flows
- interstation and serial correlations of flows

12-7. Daily Streamflow Model.

a. Generation of daily streamflows can be accomplished in a manner very similar to the generation of monthly streamflow quantities. Although a computer program has been prepared for this purpose, it is capable only of generating flows at a single location and does not provide a totally satisfactory hydrograph. Since it is desired in many reservoir operation studies to use a monthly interval most of the time, and to perform daily operation computations for only a few critical periods, the program has been designed to generate daily flows after the monthly total runoff has been generated by another program. Flows for any particular day are correlated with flows for the preceding day and for the second antecedent day.

b. A procedure that will give a reasonable shaped hydrograph, as well as coordinated hydrographs at many locations in a basin, would consist of (1) stochastic generation of

precipitation over the basin, and (2) using a precipitation-runoff model to derive the resulting streamflow.

12-8. Reliability. While the simulation of stochastic processes can add reliability in hydrologic design, the techniques have not yet developed to the stage that they are completely dependable. All mathematical models are simplified representations of the physical phenomena. In most applications, simplifying assumptions do not cause serious discrepancies. It is important at this "state of the art," however, to examine carefully the results of hydrologic simulation to assure that they are reasonable in each case.