Environment Canada
Environnement Canada

Flood Frequency and Risk

G. W. Kite

INLAND WATERS DIRECTORATE,
WATER RESOURCES BRANCH,
APPLIED HYDROLOGY DIVISION,
NETWORK PLANNING AND FORECASTING SECTION,
OTTAWA, CANADA, 1974.

FLOOD FREQUENCY AND RISK

ABSTRACT

Every year floods cause loss of life and millions of dollars worth of damage. The use of hydrologic forecasts can reduce this toll. While warning if impending high water is the most obvious form of forecast, the use of frequency and risk analyses at the planning stage can aid in the efficient and safe design and location of all the various structures and institutions making up today's society. In these ways the risk of flooding and its consequent losses can be minimized.

This paper discusses the background and justification for probability and risk analyses and recommends the use of specific procedures to determine plotting positions, choose a probability distribution and compute design event magnitudes and their confidence limits. Other topics discussed include the advantages of regional analysis and design procedures based on a specified project risk or hydrologic risk.

It is concluded that the mean frequency should be used as a plotting position, that the lognormal distribution is of more practical use than many other distributions and that some form of risk analysis should be used for all hydrologic designs.

Chapter 6, Conclusions and Recommendations, has been written so that it may be read as a complete text, independent of the main report.

FRÉQUENCE ET RISQUE D'INONDATIONS

RÉSUMÉ

Chaque année, les crues causent des pertes de vie et des millions de dollars de dégâts. Il est possible de réduire ce fleau par les prévisions hydrologiques. Bien que la forme la plus sûre de prévision soit de donner un avertissement de crue imminente, l'emploi des analyses de fréquence et de risque au stade de la planification peut aider à la conception et à l'aménagement efficases et sûrs de toutes les diverses structures et institutions qui constituent la société d'aujourd'hui. Les risques d'inondations et leurs pertes conséquentes peuvent ainsi être minimises.

Le present document étudie la base et la justification des analyses de probabilité et de risque et recommande l'emploi de procédes pour déterminer la position de traçage, choisir la distribution des probabilités et programmer l'importance prévue des crues ainsi que leurs limites de diabilité. Les autres sujets traités sont les avantages de l'analyse régional et les procédés de prévision bases sur un risque éventuel ou hydrologique.

L'auteur conclut que la fréquence moyenne devrait servir de position de traçage,que la distribution lognormal est beaucoup plus pratique que bon nombre d'autres distributions et que certains types d'analyses de risque devraient être employés pour toutes les prévisions hydrologiques.

FLOOD FREQUENCY AND RISK

TABLE OF CONTENTS

Table of Contents (Cont'd)

List of Tables

List of Tables (Cont'd) Page

List of Tables (Cont'd) Page

List of Figures

CHAPTER 1

Introduction

1.1 Introduction

Every year floods threaten life and property in locations across Canada and even more so in other areas of the world. It has been estimated (23) that over the period 1948 to 1970 floods in Canada have caused an average damage of $5 million per year with the average cost to the Federal Government of $2 million per year. On top of these direct costs are the loss of life, injury, inconvenience and other indirect losses caused by floods.

Forecasts of flood events can generally reduce the damages caused by floods. There are two ways in which this beneficial effect may be achieved. Firstly, and most obviously, warning of an event enables people to evacuate a danger area. If sufficient lead-time is provided vulnerable possessions may be removed from the danger zone and preparations can be made, such as sandbagging, to minimize property damage.

The second method of achieving benefits through forecasting is to use flood frequency analysis in the design of structures within the flood plain and for flood plain zoning. As examples, a knowledge of magnitude-frequency relationships should be used in the design of dams, highway bridges, railway bridges, culverts, water-supply systems, and flood control structures. The American Water Works Association has reported (1) that out of 293 dam failures in the U.S. and other countries since 1799, about 20% of the failures were due to faulty spillway design. Flood plain zoning ensures that housing and industrial developments are not located in high-risk zones. Or at least, if they are located in such areas there can be no justification for compensation in the inevitable event of flood damage. High-risk zones along the banks of rivers susceptible to floods should be used for activities compatible with this risk such as parkland, recreation areas, some types of crop production or grazing.

Frequency analysis is not only of use as an aid in averting disaster but is also a means of introducing efficient designs. When a hydraulic structure, through inadequate or inaccurate data or methods, is underdesigned, the results are regretably obvious; the dam may fail, the highway may flood or the bridge collapse. This does not happen very often and so the hydrologist, equating non-failure with success, is satisfied with his design techniques. Non-failure however, does not necessarily mean an efficient design. Frequently, structures are over-designed, and hence very safe, but also very expensive (16). A truly efficient design will be achieved only as the result of studies relating cost to risk and frequency analysis.

The Water Resources Council of the U.S. government (23) has recently noted that because of the range of uncertainty in design flood anlaysis there is a need for continued research and development to solve the many unresolved problems. Current methods of providing design floods for hydraulic structures include the deterministic use of meteorologic data in techniques such as dynamic flow equations and the so-called Probable Maximum Flood method (PMF), and the stochastic use of frequency analysis techniques. The PMF and similar methods suffer from the major disadvantages of being entirely subjective and of having no associated probability level. This latter is particularly important since to non-technical people it implies that no risk is involved, that the maximum flood cannot exceed this certain limit. This, of course, is untrue and can sometimes have disasterous consequences. Yevjevich (28) has characterised the difference between the PMF method and the frequency analysis method as being between "expediency" and "truth".

Neglecting the PMF approach, this report discusses only the techniques to be used in flood frequency analysis and, to a lesser degree, drought analysis, together with an analysis of the associated risk.

1.2 Arrangement of Text

The final section of this introductory chapter discusses the basic data requirements for frequency analysis and the underlying assumptions of the technique.

Once the necessary data have been abstracted then the frequency analysis proper can begin. To plot extreme data it is necessary to associate each event with an appropriate recurrence interval or return period. This subject is discussed in detail in Chapter 2. An assumption is then made of a theoretical frequency distribution for the population of events and the statistical parameters of the distribution are computed from the sample data. Chapter 3 describes, for some of the distributions commonly used in hydrology, the form of the distribution, estimation of parameters, estimation of events at given return periods and estimation of confidence limits. The objective of Chapter 3 is to provide sufficient background information to enable an hydrologist to intelligently select a distribution to use in frequency analysis.

A flood frequency relation over a region is usually preferable to one developed for a specific site for two reasons:

(a) Because of the sample variation possible at a single station any single station analysis is subject to large error. This error can be reduced by combining data from many sites.

(b) There are many more sites where hydrologic data are needed, than there are sites at which data are collected. This means that some form of analysis is required which can transfer data from gauged sites to ungauged sites.

Chapter 4 discusses the various methods of regional analysis presently available.

Both single station and regional frequency analyses involve risks. In determining the design flood for a project, the length of data available, the project life and the allowable probability of failure are all factors to be considered. Chapter 5 reviews the statistical methods available to analyse these risks.

Chapter 6 presents conclusions and recommends procedures to be followed. Chapter 6 has been written so that it may stand apart from the rest of the report. That is, for those readers with insufficient time to read the complete text, Chapter 6 can be read as an abridged version of the complete report.

1.3 Data and Assumptions

Starting with the original recorded hydrograph, or with the tabulated data abstracted from the hydrograph, there are two ways in which this data may be used in a frequency analysis. The first method, direct frequency analysis, is to select from the total data only that information which is required in the design process, e.g. maximum instantaneous flow in each year (annual series) all instantaneous flows above a certain base flow (partial-duration), minimum 7-day flow in each year (annual series), etc.

This represents a considerable loss of information, however, since such a few data points are utilised. Dhyr-Nielson (9) has discussed the effects of this loss of information on the estimation of probabilities of extremes.

The second method of using the basic hydrologic data, simulation, is to design a mathematical model which will describe the observed hydrograph. This model can then be used to generate many sets of data from which the required events can be abstracted. As an example, if the object of the frequency analysis is to define the flow having a magnitude such that it will occur on the average once every 50 years, then 50 years of hydrologic data can be generated and the maximum flow during that period can be found. In addition, by generating many different sets of 50-year records, a distribution of 50-year flows can be found and confidence limits set on the mean (10). The disadvantage of the data generation method is that it is very unwieldy. It is practicable when dealing with monthly data, less so for daily means and becomes difficult when the need is to generate flows that can be considered instantaneous. In addition, most models cannot successfully generate the extreme peaks and lows of a variable; they only operate well for average conditions.

This report describes only the first method of analysis. Within the first method, direct frequency analysis, there are two ways in which the required data may be abstracted from the original recorded or tabulated data. As indicated earlier, these are known as annual series and partial-duration series.

An annual series takes one event, and only one event, from each year of the record. A disadvantage of this abstraction technique is that the second or third, etc., highest events in a particular year may be higher than the maximum event in another year and yet they are totally disregarded. This disadvantage is remedied in the partial-duration series method in which all events above a certain base magnitude are included in the analysis. The base is generally selected low enough that at least one event in each year is included. Each event, to be included in the partial-duration series, must be separate and distinct, i.e. including two consecutive daily flows caused by the same meteorologic event is not valid. If the total number of events which occurred during the entire period of record are ranked without regard to the year in which they occurred and then the n top ranking events are selected, where n is the number of years of record, the events are termed annual exceedences (20).

The recurrence interval of an event of given magnitude is defined as the average length of time between occurrences of that event. This is purely a statistical term and contains no inference of periodicity. The recurrence intervals of annual series and partial duration series have different meanings. In the first case the recurrence interval means the average number of years between the occurrence of an event of a given magnitude as an annual maximum. In the second case the recurrence interval carries no implication of annual maximum.

Chow (5) investigated the theoretical relationship between these two recurrence intervals and their corresponding probabilities. If P_E is the probability of an event in a partial-duration series being equal to or greater than x and if the number of events in the partial-duration series is Nm, where N is the number of years and m is the average number of events per year, then P_E/m is the annual probability of an event being equal to or greater than x. The probability of an event x being the largest of the m events in a year must then be

$$P_M = (1 - P_E/m)^m \qquad 1.1$$

But P_M is then the probability of an annual event of magnitude x (say, P_E) and corresponds to the annual series. Substituting the approximation $(1 - P_E/m)^m$ equal to e^{-P_E} and letting $T_E = 1/P_E$ and $T_M = 1/P_M$ where T_E and T_M are the recurrence intervals of the partial-duration and annual series respectively, then

$$T_E = \frac{1}{\ln T_M - \ln (T_M - 1)} \qquad 1.2$$

The following table from Dalrymple (7) compares the recurrence intervals of the two types of series.

Table 1.1

Comparison of Recurrence Intervals for Annual and Partial Duration Series

Recurrence intervals (years)	
Partial-duration	Annual Series
0.5	1.16
1	1.58
1.44	2
2	2.54
5	5.52
10	10.5
20	20.5
50	50.5
100	100.5

The difference amounts to about 10% when P_E is 5 years and about 5% when P_E is 10 years. The distinction is only of importance at low recurrence intervals.

To some extent the decision to use an annual series or a partial-duration series depends on the use to which the frequency analysis will be put. Some types of structure, for example erosion protection works, are susceptible not particularly to one peak flow but more to closely repeated high flows and so a partial-duration series analysis might be more suitable. If the design flood is likely to have a low recurrence interval then again the partial duration series may be more suitable since the smallest recurrence interval for the annual maximum series is one year.

If flood flows are being investigated, they should preferably be maximum instantaneous flows derived from a continuous hydrograph. Usually, however, instantaneous flows are only available for a comparatively few years, but older data may be available as maximum mean daily flows. In this case, by correlation between the two series, it may be possible to extend the instantaneous maxima back in time (13).

In some regions it may be necessary to carry out more than one frequency analysis on the data. Stoddart and Watt (21) have described how watersheds in southern Ontario have two distinct types of flood; those due to precipitation only, generally occurring in the summer, and those due to snowmelt (sometimes combined with precipitation) generally occurring in the winter or spring. The hydrographs of these two types of flood are quite different and cannot be considered to be from the same population.

Suppose that a flood of a given magnitude x would have a recurrence interval of T_p if due to rainfall and T_S if due to snowmelt. The probabilities of not equalling or exceeding x are given as

$$q_p = 1 - 1/T_p \tag{1.3}$$

and

$$q_S = 1 - 1/T_S \tag{1.4}$$

So that the probability of not equalling or exceeding x in any year, q_A, is given by

$$q_A = q_p . q_S \tag{1.5}$$

and the recurrence interval for the annual flood equalling or exceeding x is

$$T_A = 1/(1 - q_A) \tag{1.6}$$

$$T_A = \frac{T_p . T_S}{(T_p + T_S - 1)} \tag{1.7}$$

The frequency curve of the annual series will be asymptotic to both the rainfall and snowmelt frequency curves. Stoddart and Watt (21) list four possible combinations of these frequency curves.

Most data series are incomplete. For various reasons such as mechanical failure, inaccessibility, flood damage to the recorder, etc., some flood peaks are usually missing. This may have a large effect on the recurrence intervals assigned to the floods on record.

Dalrymple (7) has described a method of overcoming this problem. By regression with a gauged stream in the same hydrologic region it is possible to estimate the magnitudes of the missing floods at the stream being analysed. As an example of a technique, Langbein (14) has correlated the logarithms of flows standardised on a monthly basis. Dalrymple (7) used the computed flows as an aid in sorting the observed flows by magnitude and assigning recurrence intervals. The computed flows were then discarded and not used in the further analysis.

On the other hand, properly authenticated historic events, antedating periods of consecutive records, can be used in frequency analysis to increase the accuracy of the analysis. Discharge estimated on the basis of authenticated stages of historic floods that occurred prior to the modern continuous stage records may be used in conjunction with the modern records to obtain a more accurate probability curve (22).

Benson (2) studying the Susquehanna River at Harrisburg, Pa., found 7 historic floods with stages greater than 18.0 feet in the period 1786 to 1873 and a record of continuous annual maxima ranging from 14.3 to 30.3 feet for the period 1874 to 1947. Since the period of historical floods may have contained an unknown number of events of less than 18.0 feet stage the problem was to combine the two types of record in the proper proportions to obtain an array of events properly representative of the total period. Benson (2) arrayed all events, historic and recent, in order of descending magnitude using the plotting position m/(n+1). The order number of those flood peaks which were lower than the lowest historical event (the base event, 18.0 feet on the Susquehanna) were then adjusted as:

$$m_c = t_b + \frac{n - t_b}{t - t_b} (m - t_b) \qquad 1.8$$

where m_c is the corrected order number, m is the original order number, n is the period in years from the first historical event to the most recent recorded event, t_b is the number of events equalling or exceeding the base event, and t is the total number of events.

If a single historic event is known and there is not much difference in magnitude between that event and the highest event in the recent record, then the Equation 1.8 can be used with confidence (2). If there is a large gap in magnitude then there should be some reasonable certainty that there were few or no events during the ungauged interval which exceeded the highest in the recent period. With a large gap in magnitude and no intimation of what may have happened in between, then Equation 1.8 is no longer applicable. Dalrymple (6) has included further examples of the application of Benson's (2) method.

Glos and Krause (10) have described the augmentation of recorded data with historical data for the Rivers Dnepr in the USSR, Elbe in Czechoslovakia, Main in West Germany and the Spree and Werra in East Germany. As an example, the Elbe River at Decin, Czechoslovakia, has a record of annual flood peaks for 111 years (1851 to 1962). For purposes of flood frequency analysis historical data can be used to extend this record to 530 years (back to 1432). In this case only the relative magnitudes of the historical floods were used since obviously the absolute magnitudes were not measured.

Glos and Krause (10) used the historical data to increase the accuracy of estimation of the sample mean, $\bar{x}$, and variance; S^2, as;

$$\bar{x} = \Sigma x_i p_i \qquad 1.9$$

$$S^2 = \Sigma \ (x_i - \bar{x})^2 \ p_i \qquad 1.10$$

where p_i is a weight given to each event, $\Sigma \ p_i = 1$. For historical floods the weights were assigned as:

$$p = 1/N \qquad 1.11$$

where N is the length of the historical period. The events of the annual record were weighted as:

$$p = \frac{N - k}{N.n} \qquad 1.12$$

where k is the number of historical floods and n is the number of years in the annual record. The computed mean and standard deviation were then used in the standard frequency equation:

$$x(K) = \bar{x} + K.S \qquad 1.13$$

where K, the frequency factor, depends upon the return period required and the distribution characteristics.

Leese (15) has shown how historical flood marks may be used to achieve greater precision in the estimates of flood events and hence greater efficiency in design. The Type I extremal distribution (see Chapter 3 for description) was used on the 29-year record of annual maximum discharges of the River Avon at Bath to estimate floods with return periods of 10, 25, 50, 100 and 1000 years. The procedure was then repeated adding to the data 13 historic floods determined from old flood marks. It was found that the sampling error of the flood estimates was reduced in the second case by between 8% for the 10-year flood and 18% for the 1000-year flood. By applying expressions for the benefits and costs of an imaginary structure Leese found that if the structure were designed on the basis of a 50-year flood then the incorporation of the historic data would reduce the cost of the structure by approximately 1%. More importantly it was found that to obtain the same cost reduction from a continuous record would require a further 20 years of streamflow data.

Two assumptions implicit in any frequency analysis are (19):

(a) that the data to be analysed describe random events, and

(b) that the natural processes involved are stationary with respect to time.

This chapter has already discussed the implications of assumption (a) on data selection, emphasising the fact that all events used must be independent. Assumption (b) is more difficult to guarantee. The earth is in a constant state of flux with innumerable processes affecting the hydrologic cycle and its various components. Statistical tests are available (26) to check for stationarity of time series. Non-stationarity in hydrologic time series is generally due to one of two basic causes:

(a) a slow change in hydrologic parameters such as might be caused by the gradual urbanisation of watersheds or (on a different time scale) long-term changes in temperature or precipitation distribution.

(b) rapid change in parameters caused by, for example, earthquakes, landslides, building of dams.

In a recent paper, Yevjevich (27) gave a plot of annual maximum flows of the Danube River at Orshova in Romania which shows a pronounced upward trend. This trend in discharge is mainly due to the construction of flood protection levels along the Danube and its major tributaries. Non-homogeneities such as this shift the mean value of the distribution and increase the variance.

Another source of error is basic inconsistency in the data due to systematic measurement and computational errors. This subject has been discussed in detail by Dickinson (8), Robertson (17), and Herschy (11), and the particular errors involved in determination of winter flows in Canada have been discussed by Rosenberg and Pentland

(18). It might be noted at this point that the major emphasis in this report is on flood flows and that these are the very events subject to the maximum measurement error. In fact maximum flows are seldom, if ever, measured because of the difficulties involved firstly in predicting the time at which maximum flow will occur, secondly the difficulty in getting to the gauging site at that time and finally the difficulties of actually carrying out the gauging at the high stage and corresponding high velocities. As a result high flows are normally estimated by extrapolation of the rating curve, estimation of mean velocity from an isolated surface velocity measurement, use of the slope-area method (6) or other similar procedures. The resulting estimates of peak discharge contain a high error component which has been estimated by Blench (4) to be at least ±25%.

Yen and Ang (25) have termed these errors "subjective uncertainty" and have shown that if the individual uncertainties are denoted by σ_1, σ_2. . . then the overall subjective uncertainty σn, can be written as:

$$\sigma_n = [\sigma_1{}^2 + \sigma_2{}^2 + \ldots]^{\frac{1}{2}} \qquad 1.14$$

The first assumption made, in any frequency analysis and a very important one, is that the sample data available are good estimates of the population of events. This assumption is necessary so that estimates of population statistics such as mean, variance, skew, etc. may be derived from the sample. Benson (3) used a theoretical frequency curve to obtain 1000 random events. This base data was then divided into shorter records e.g. 40 records of 25 events each, 20 records of 50 events each, etc. and Benson investigated the variability of the frequency curves of these samples compared to the original theoretical curve. Benson's conclusions provide estimates of the number of events needed in a sample before sample estimates of magnitudes at various return periods are comparable to the population values.

References for Chapter 1

1. American Water Works Association, 1966, Spillway Design Practice, AWWA Manual M13, New York.

2. Benson, M.A., 1950, Use of Historical Data in Flood Frequency Analysis, Trans. Am. Geophys. Union, V. 31, pp. 419-424.

3. Benson, M.A., 1960, Characteristics of Frequency Curves Based on a Theoretical 1000 Year Record, USGS Water Supply Paper No. 1543-A, pp. 51-73.

4. Blench, T., 1959, Empirical Methods, Proceedings of Symposium No. 1, Spillway Design Floods, NRC, Ottawa, pp. 36-48.

5. Chow, V.T., 1964, Handbook of Hydrology, McGraw-Hill.

6. Dalrymple, T., 1956, Measuring Floods, Proceedings of IASH Symposia Darcy, Dijon, Vol. 3, pp. 380-404.

7. Dalrymple, T., 1960, Flood Frequency Analyses, USGS Water Supply Paper No. 1543-A, pp. 1-47.

8. Dickinson, W.T., 1967, Accuracy of Discharge Determinations, Hydrology Paper No. 20, Colorado State University, Fort Collins, Colorado.

9. Dyhr-Nielson, M., 1972, Loss of Information by Discretizing Hydrology Series, Hydrology Paper No. 54, Colorado State University, Fort Collins, Colorado.

10. Glos, E., and R. Krause, 1967, Estimating the Accuracy of Statistical Flood Values by Means of Long-term Discharge Records and Historical Data, Proc. Leningrad Symposium on Floods and their Computations, pp. 144-151.

11. Herschy, R.W., 1969, The Evaluation of Errors at Flow Measurement Stations, Technical Note No. 11, Water Resources Board, Reading, England.

12. Kite, G.W., and R.L. Pentland, 1971, Data Generating Methods in Hydrology, Technical Bulletin No. 36, Inland Waters Directorate, Department of the Environment, Ottawa.

13. Kite, G.W., 1973, Flood Frequency for Mackenzie Highway Culverts, unpublished paper, Water Resources Branch, Department of the Environment, Ottawa.

14. Langbein, W.B., 1960, Plotting Positions in Frequency Analysis, USGS Water Supply Paper No. 1543-A, pp. 48-51.

15. Leese, M.N., 1973, The Use of Censored Data in Estimating T-Year Floods, Proceedings of the UNESCO/WMO/IAHS Symposium on the Design of Water Resources Projects with Inadequate Data, Madrid, Vol. 1, pp. 235-247.

16. Maasland, D.E.L., 1966, Discussion of "Extending Hydrologic Data", by R.H. Clark and J.P. Bruce, Proceedings of Hydrology Symposium No. 5, Statistical Methods in Hydrology, McGill University, Montreal, pp. 143-144.

17. Robertson, A.I.G.S., 1966, The Magnitude of Probable Errors in Water Level Determination at a Gauging Station, Technical Note, No. 7, Water Resources Board, Reading, England.

18. Rosenberg, H.B., and R.L. Pentland, 1966, Accuracy of Winter Streamflow Records, Proc. Eastern Snow Conference, Hartford, Connecticut.

19. Spence, E.S., 1973, Theoretical Frequency Distributions for the Analysis of Plains Streamflow, Can. J. Earth Sci, V. 10, pp. 130-139.

20. Stall, J.B., and J.C. Neill, 1961, A Partial Duration Series for Low-Flow Analyses, Journal of Geophysical Research, Vol. 66 No. 12, pp. 4219-4225.

21 Stoddart, R.B.L., and W.E. Watt, 1970, Flood Frequency Prediction for Intermediate Drainage Basins in Southern Ontario, C.E. Research Report No. 66, Queen's University at Kingston.

22. Subcommitte of the Joint Division Committee on Floods, 1953, Review of Flood Frequency Methods, Trans. ASCE, Vol. 118, pp. 1220-1231.

23. United States Water Resources Council, 1967, A Uniform Technique for Determining Flood Flow Frequencies, Hydrol. Comm. Bull. No. 15.

24. Water Management Service, 1973, Scope and Plan of Work for Study of Floods and Flood Damages in Canada, unpublished proposal, Canada Department of the Environment.

25. Yen, B.C., and A.H-S. Ang, 1971, Risk Analysis in Design of Hydraulic Projects, Proceedings of Symposium on Stochastic Hydraulics, Univ. of Pittsburg, pp. 694-709.

26. Yevjevich, V., 1972, Probability and Statistics in Hydrology, Water Resources Publications, Fort Collins, Colorado.

27. Yevjevich, V., 1972, New Vistas for Flood Investigations, Academia Nazionale Dei Lincei, Roma, Quaderno N. 169, pp. 515-546.

28. Yevjevich, V., 1973, Discussion of Session I, Precipitation and Precipitation Probability, Proceedings of the Second International Symposium in Hydrology, Fort Collins, Colorado, Water Resources Publications, p. 106.

CHAPTER 2

Plotting Position

2.1 Introduction

Consider a series of 10 annual maximum discharges. Suppose that it is desired to plot these annual maxima on a graph in order to better interpret the data, perhaps detect errors, or to get an idea of which probability distribution to use to describe the data. The ordinate of such a graph conventionally contains the event magnitudes either on a linear or logarithmic scale while the abscissa will be some measure of the probability of occurrence of each event or the average time interval between occurrences (since this is simply the inverse of the probability of occurrence). Frequently the scale of the abscissa will be arranged so that events distributed according to a given distribution will plot as a straight line.

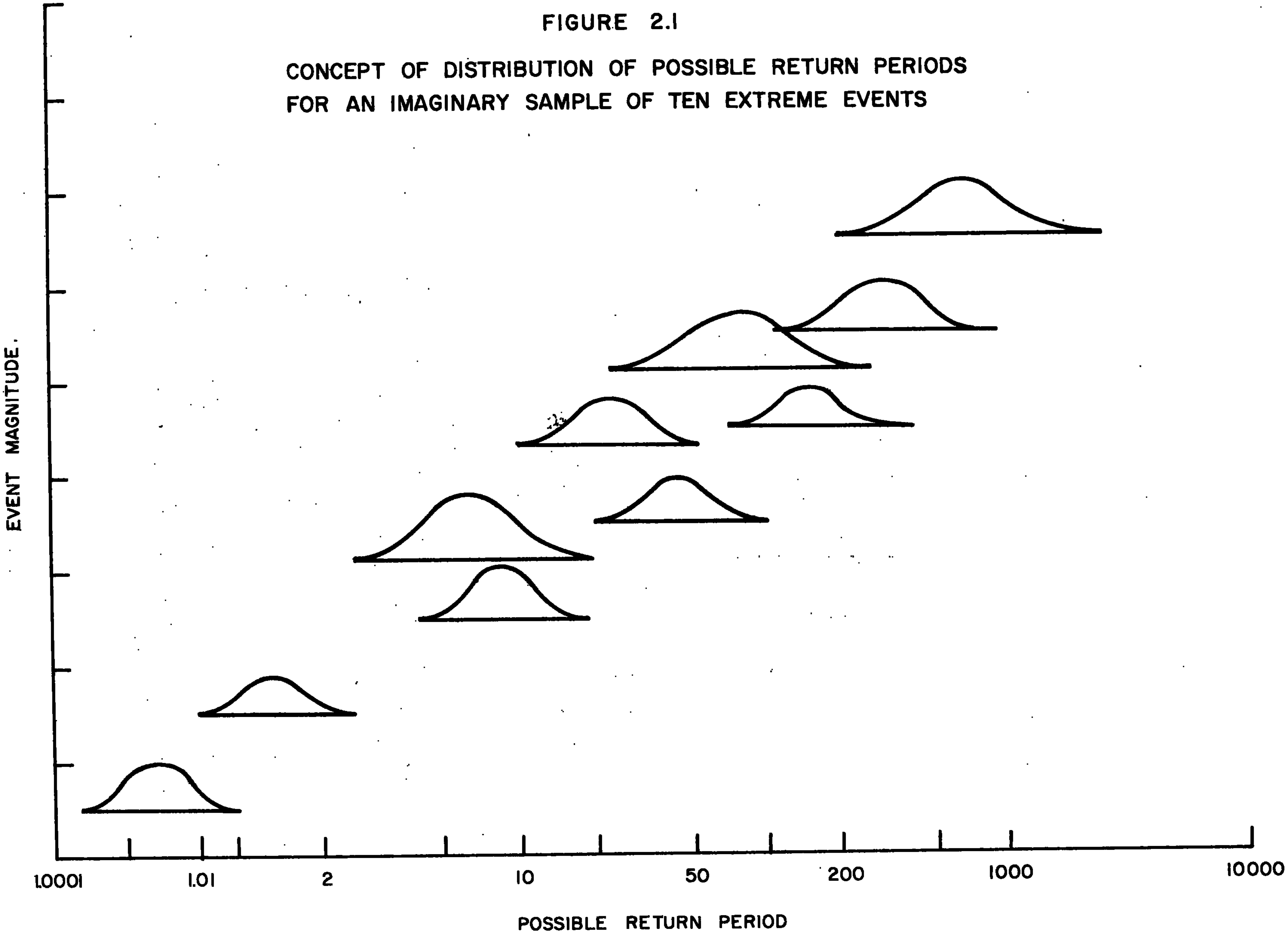
The question then arises as to how to derive the probability of occurrence or average return period of each of the set of annual maximum floods. Sorting the annual events in order of magnitude it is apparent that the largest flood occurs once in the 10 years. But is 10 years the true average return period (i.e. the average interval between occurrences) of this flood? We do not know. In the particular sample a flood of this magnitude occurred only once but in other equal length samples of annual maxima the same magnitude of flood might occur several times or not at all. That is, the maximum flood in the 10-year sample may have a true return period of 5 years, 50 years, 500 years or any other number of years. And yet, for practical reasons, some probability of occurrence must be assigned to this flood.

For the maximum of a series of 10 annual maximum values it can be shown (11) that the true return period, T, has a 10% probability of being as low as 5 years and a 10% probability of being as high as 100 years. As a matter of theory (1), if the basic data are truly representative, a flood having a return period of 10 years will be equalled or exceeded in a great length of time, on an average of once in 10 years. In 1000 years of record there would be 1000/10, or 100 such floods. However, if these 1000 years were divided into 100 periods of 10 years each, about 37% of such periods would not experience a flood of that magnitude; about 37% of the periods would experience 1 such flood; about 18% would experience 2; about 6% would experience 3; about 1.5% would experience 4; and about 0.5% would experience 5 or more such floods.

As an illustration of the problem consider a large rain storm occurring simultaneously over several adjacent streamflow basins. Imagine that one stream has been gauged for 5 years, another for 10 years and another for 15 years. If the resulting flood runoff is the maximum in the 15 years it will have an apparent return period of 5 years at one gauge, 10 years at another and 15 years at the third. Which is correct? Obviously the estimate based on 15 years is more correct than those based on 10 or 5 years. In the absence of infinite records some method of correcting for this

FIGURE 2.1

CONCEPT OF DISTRIBUTION OF POSSIBLE RETURN PERIODS FOR AN IMAGINARY SAMPLE OF TEN EXTREME EVENTS



variation in apparent return period with available record must be derived.

Figure 2.1 illustrates the concept of all 10 annual maxima having distributions of possible return periods. The scales used here, for purposes of illustration, are linear and normal probability and the distributions of the possible return periods are shown as arbitrary bell-shaped distributions. The problem of plotting position is then to locate the "correct" average return period for each event so that the distributions can be replaced by point positions.

The large variation possible in T can be shown more formally by considering T as an independent random variable which can take the values 1, 2, ... t-1, t, t+1...∞ where t is the number of years from the occurrence of one event until the occurrence of the next event. The distribution of T is then of the form:

$$P(T = t) = p(1 - p)^{t-1} \qquad 2.1$$

where p is the probability of the event occuring at any one time. The average value of T or return period is then given by:

$$E(T) = \sum_{t=1}^{\infty} t.P(T = t) = 1/p \qquad 2.2$$

The variance of T can similarly be expressed as:

$$\text{var } T = E(T - E(T))^2 \qquad 2.3$$

Following Lloyd (10) this expression reduces to:

$$\text{var } T = (1 - p)/p^2 \qquad 2.4$$

For small p the variance of the return period can be approximated as 1/p i.e. the same as the expected value. This explains the large observed variations in values of T. Lloyd (10) has also shown that for non-independent sequences of events the variance of the return period is even larger.

The problem of apparent return period varying with sample length is approached in practice by defining a plotting position for the frequency of occurrence, p. The plotting position is generally based on some assumption of the position of the sample estimate of the frequency within a population distribution of frequencies. A few general requirements of any plotting position can be stated (7) as:

(a) The plotting position should be such that all observations can be plotted.

(b) The plotting position should lie between the observed frequencies (m - 1)/n and m/n and should be distribution free (m is the order of the particular event in the series of n maximum events. For the largest of the n events, m = 1).

(c) The return period of a value equal to or larger than the largest observation should approach n, the number of observations.

(d) The observations should be equally spaced on the frequency scale, i.e. the difference between the plotting positions of the (m + 1)th and the mth observations should be a function of n only and be independent of m.

(e) The plotting position should have an intuitive meaning and should be analytically simple.

2.2 Plotting Position as an Extreme

The simplest assumption regarding the sample frequency in its population is that the events correspond directly to their observed frequency i.e. the maximum event in a series of 10 independent maxima would have a return period of 10 and a frequency of occurrence of 0.1. This is known as the California method (10) and is given by the general equation

$$p = \frac{m}{n} \qquad 2.5$$

That is, considering the frequency interval 0.1 - 0, the California method uses a plotting position at the upper extreme of this interval. By first considering the observed maxima arranged in a decreasing order of magnitude the observed frequency increasing is equally legitimately given by

$$p = \frac{m-1}{n} \qquad 2.6$$

which corresponds to the lower extreme of the frequency interval noted above. Since the frequencies zero and unity do not exist for an unlimited variate the largest observation of the series cannot be plotted using the function (m-1)/n and the smallest observation cannot be plotted using the function m/n. These two functions therefore fail Gumbel's conditions (7) and are not acceptable as plotting positions unless an upper or lower limit to the population can be envisaged.

2.3 Plotting Position as the Mean

Foster (6) has pointed out that there is no reasonable basis for assuming that the maximum event in a sample represents exactly the simple (California) frequency. Instead he contended that this maximum event is representative of the whole class of possible events occurring with frequencies less than that of the maximum in the sample, i.e. for a sample of 10 the maximum event represents the interval 0.10 to 0, and therefore should be plotted at the mean of this class interval (i.e. at 0.05 in our example). The Foster (or sometimes called Hazen) plotting position is given by the general equation

$$p = \frac{2m - 1}{2n} \qquad 2.7$$

Equation 2.7 is a compromise between Equations 2.5 and 2.6. When m = 1, the largest event of the sample, this plotting position claims that an event which has already happened once in n years will occur, in the mean, once in 2n years.

Similarly if the observed frequency of an event is assumed to be the mean of the population of frequencies for that event, $\bar{p}$, then

$$\bar{p} = \frac{\int_0^1 (1-Z^{1/n})\ dZ}{\int_0^1 dZ} \qquad 2.8$$

where Z is the frequency of occurrence of the event in the n year period. For the maximum value, $\bar{p} = \frac{1}{n+1}$ so that for a 10-year sample the maximum value would have an assigned average return period of 11 years. Use of the mean frequency leads to the general equation

$$p = \frac{m}{n+1} \qquad 2.9$$

where m is the order of the flood, m being 1 for the largest and n for the smallest event in the n years of record. The probability p is thus the average of the probabilities of all events with rank m in a series of periods each of n years.

2.4 Plotting Position as the Mode

Another assumption possible is that the observed frequency is the mode (by definition the event which occurs most frequently) of the population of frequencies. Equation 2.1 has no mode and so the type of probability distribution must be incorporated. As an example Equation 2.1 can be used with Gumbel or Type I extremal distribution (7):

$$Z = (e^{-e^{-y}})^n \qquad 2.10$$

where y is a linear function of discharge. For the mode of any distribution the probability density is a maximum i.e. $dZ/dy = 0$ and $d^2Z/dy^2 < 0$, which for Equation 2.10 yields

$$p = 1 - (\frac{1}{e})^{1/n} \qquad 2.11$$

For the maximum annual event in a 10-year record this equation indicates an average return period of $10\frac{1}{2}$ years. The plotting positions for the maximum events in samples of different size are given in the following table:

Table 2.1

Plotting Positions for Maximum Event Under Modal Assumption

Sample Size(n)	Plotting Position(p)
2	0.393
5	0.181
10	0.0951
20	0.0488
50	0.0198
100	0.00995
200	0.00499
500	0.00199
1000	0.000999

It is clear that as the sample size gets larger, p is not significantly different from the simple plotting position $p = m/n$.

Blythe (4) computed the mode of a distribution of extreme values as

$$\text{mode} = \mu - \sigma \frac{\beta_1^{\frac{1}{2}} (\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \qquad 2.12$$

where μ, σ, β_1 and β_2 are respectively the mean, standard deviation, coefficient of skew and coefficient of kurtosis. From the normal distribution the probability of occurrence of this mode can be determined. For the maximum values of samples of different sizes Blythe (4) provided the following table:

Table 2.2

Plotting Positions for Maximum Event of Extremal Distribution

Sample Size (n)	Plotting Position (p)
2	0.306
5	0.143
10	0.077
20	0.041
60	0.014
100	0.0087
200	0.0046
500	0.0018
1000	0.00092

No plotting positions are available by this method for events other than the maximum of the sample, and Blythe recommended that for practical applications the Foster (6) plotting position (Equation 2.7) should be used.

2.5 Plotting Position as the Median

A further possible assumption is that at any given order of magnitude within a sample set of events, the return period used should be the median of all possible return periods obtained from a population of equally sized samples. The return period to be used at any given order of magnitude, j, within a sample of size n events is then given by

$$T = 1/p_j \qquad 2.13$$

where p_j is the solution of the binomial equation

$$\sum_{i=0}^{j-1} \binom{n}{1} p_j^i (1 - p_j)^{n-i} = 0.5 \qquad 2.14$$

that is, p_j is a solution of a polynomial of order equal to the sample size n and having a number of terms equal to the numerical value of j.

In effect,this method results in a return period of approximately 1.44 n for the largest flood in a period of n years. This occurs because, for example, the event with a return period of 144 years has an equal chance of being exceeded or not exceeded in any period of 100 years.

As an example of the use of the median assumption in assigning plotting positions, the maximum event in a sample of 10 maxima would have computed return period, T, of 14.92 years. Beard (3) gave the plotting positions for the sequence of 10 maxima as in Table 2.3

Table 2.3

Plotting Positions Under Median Assumption

magnitude order m	plotting position p
1	.067
2	.164
3	.258
4	.355
5	.452
6	.548
7	.645
8	.742
9	.836
10	.933

The theoretical values of the plotting positions are tedious to compute requiring the solution of a polynomial of degree equal to the sample size and Beard (3) has recommended that the Foster equation (Equation 2.7) which can be transformed (2) to

$$p_j = p_1 + (j - 1)\ (1 - 2p_1)/n - 1 \qquad 2.15$$

where $p_1 = 1/2n$ should be used as an approximate solution.

Banerji and Gupta (2) proposed an alternate solution to Equation 2.15 as

$$p_1 = 1 - 0.5^{1/n} \qquad 2.16$$

In practice the procedure advocated by Banerji and Gupta (2) is to compute p_1 from Equation 2.16, compute the increment $(1-2p_1)/n-1$ and then compute p_j for $j = 2, 3....n/2$ from Equation 2.15. The remaining plotting positions p_j, $j = n/2, (n/2+1)...n$ can be calculated as

$$p_{(n+1)/2} = 0.5 \qquad 2.17$$

$$p_{n-j+1} = 1.0 - p_j \qquad 2.18$$

Beard (3) and Hardison and Jennings (8) have demonstrated that for a normal distribution the average exceedence probability, $\bar{p}$, for an event with a 10 year return period estimated from a sample of 10 events is 0.1261. As the length of the sample increases so the average exceedence probability decreases until as $n \to \infty$, $\bar{p} \to p$ where p is the simple probability exceedence estimated as $p = m/n$.

Chow (5) lists many other plotting positions most of which have no theoretical basis.

References for Chapter 2

1. ASCE, Subcommitte of the Joint Division Committee on Floods, 1953, Review of Flood Frequency Methods, Trans. ASCE, Paper No. 2574, Vol. 118, pp. 1220-1231.

2. Banerji, S., and D.K. Gupta, 1967, On a General Theory of Duration Curve and its Application to Evaluate the Plotting Position of Maximum Probable Precipitation or Discharge, Proc. Symposium on Floods and their Computations, Leningrad, pp. 183-193.

3. Beard, L.R., 1943 Statistical Analysis in Hydrology, Trans., ASCE, Vol. 69, No. 8, Pt. 2, pp 1110-1160.

4. Blythe, R.H., 1943, Discussion of Paper No. 2201, Trans. ASCE, Vol. 69, No. 8, Pt. 2, pp. 1137-1138.

5. Chow, V.T., 1964, Handbook of Applied Hydrology, McGraw-Hill.

6. Foster, H.A., 1936, Methods for Estimating Floods, USGS Water supply Paper 771.

7. Gumbel, E.J., 1958, Statistics of Extremes, Columbia University Press.

8. Hardison, C.H., and M.E. Jennings, 1972, Bias in Computed Flood Risk, Proc. ASCE, Vol. 98, No. HY3 pp. 415-427.

9. Langbein, W.B., 1960, Plotting Positions in Frequency Analysis, USGS Water Supply Paper 1543-A, pp. 48-51.

10. Lloyd, E.H., 1970, Return Periods in the Presence of Persistence, Journal of Hydrology, Vol. 10, No. 3, pp. 291-298.

11. USGS, 1936, Floods in the United States, Water Supply Paper No. 771.

12. Yen, B.C., 1970 Risks in Hydrologic Design of Engineering Projects, Proc. ASCE, Vol. 96, No. HY4, pp. 959-966.

CHAPTER 3

Frequency Distributions

3.1 Introduction

One of the most common problems faced in hydrology is the estimation of a design flood or drought from a fairly short record of streamflows. Plotting the magnitude of the measured events (annual maxima for example) some kind of pattern is generally apparent. The question is how to use this pattern to extend the available data and enable the design event to be derived.

If a large number of observed or measured events are available from a period of record at least as long as the return period of the required design event then the problem is simplified. In the extreme if a large enough sample were available (say one million events) then the design event and its confidence interval could be derived directly from the sample data. This amount of data will not be available, however, and so the sample data is generally used to fit a frequency distribution which is then used to extrapolate from the recorded events to the required design events. The fitted frequency curve can be extrapolated either graphically or by estimating the parameters of some standard frequency distribution which is assumed to describe the recorded events.

Graphical methods have the advantages of simplicity and visual presentation and the fact that no assumption of distribution type is made. These advantages are outweighed, however, by the disadvantage that, given twenty engineers to fit a curve through a set of points, it is highly probable that at least twenty different curves would result. In other words the method is highly subjective and is not compatible with the other phases of engineering design.

Numerous different probability or frequency distributions have been used in hydrology (6). Discrete distributions such as the binomial and Poisson have been used to define the average intervals between events (18) and to evaluate risks (22). Continuous distributions such as the normal and lognormal have been used for both annual series (19) and partial duration series (6) to define the magnitude of an event corresponding to a given probability of occurrence. The two types of distribution have also been combined (14), (37), (42) to give models of frequency of occurrence and frequency of magnitude of extreme events.

There are two sources of error in using a frequency distribution to estimate event magnitudes. The first source of error is that it is not known which of the many distributions available is the "true" distribution, i.e., which distribution, if any, the events naturally follow. This is important because the sample events available are usually for relatively low return periods (i.e. around the centre of the probability distribution) while the events it is required to estimate are generally of large return period (i.e. in the tail of the distribution). Many distributions have similar shape in their centres but differ widely in the tails. It is thus possible to fit several distributions to the sample data and end up with several different estimates of the T-year event. Chi-square and similar tests of goodness of fit can be used to choose the distribution which best describes the sample data but this does not overcome the basic problem.

Once a distribution has been chosen then the second source of error becomes apparent. The statistical parameters of the probability distribution must be estimated from the sample data. Since the sample data is subject to error the method of fitting must minimise these errors and must therefore be as efficient as possible. There are four parameter estimation techniques in current use:

1. method of moments,
2. method of maximum likelihood,
3. least squares, and
4. graphical.

The method of moments (44) defines the rth moment, μ_r', about the origin, $x = 0$, as:

$$\mu_r' = \frac{1}{n} \sum_{i=1}^{m} f_i \, x_i^r \qquad 3.1$$

where n is the number of events, f_i is the frequency of the event x_i and m is the number of distinct frequencies such that:

$$n = \sum_{i=1}^{m} f_i \qquad 3.2$$

The rth moment, μ_r, about the mean, μ_1, is given by:

$$\mu_r = \frac{1}{n} \sum_{i=1}^{m} f_i(x_i - \mu_1')^r \qquad 3.3$$

Clearly then the arithmetic mean is equal to the first moment about the origin and the variance is equal to the second moment about the mean.

The principle of maximum likelihood (15) states that for a distribution with a probability density function $f(x;\alpha,\beta,\ldots)$ where α, β ... are parameters to be estimated (e.g. mean, variance, etc.), then the probability of obtaining a given value of x, x_i, is proportional to $f(x_i;\alpha,\beta,\ldots)$ and the joint probability, L, of obtaining a sample of n values x_1, x_2, ... x_n is proportional to the product

$$L = \prod_{i=1}^{n} f(x_i;\alpha,\beta,\ldots) \qquad 3.4$$

This is called the likelihood; the method of maximum likelihood is to estimate α, β, ..., such that L is maximised. This is obtained by partially differentiating ln L with respect to each of the parameters and equating to zero.

The least squares estimation method (47) consists of fitting a theoretical function to an empirical distribution. The sum of squares of all deviations of observed points from the fitted function is then minimised. Thus to fit a function

$$\hat{y} = f(x;\alpha,\beta\ldots.) \qquad 3.5$$

the sum to be minimised is

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad 3.6$$

$$S = \sum_{i=1}^{n} (y_i - f(x_i;\alpha,\beta\ldots))^2 \qquad 3.7$$

where x_i and y_i are coordinates of observed points, α, β, ... are parameters and n is the sample size. To obtain the minimum sum of squares Equation 3.8 is partially differentiated with respect to the parameter estimates a, b ...

$$\frac{\partial \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\partial a} = o, \quad \frac{\partial \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\partial b} = o \qquad 3.8$$

These partial derivatives give a number of equations equal to the number of parameters to be estimated. In order that the least squares method be an efficient estimator three conditions must be satisfied:

1. the deviations $y_i - \hat{y}_i$ be normally or at least symmetrically distributed.
2. the population variance of the deviations be independent of the magnitude of y_i.
3. the population variance of the deviations along the least squares curve be constant.

The graphical method consists of fitting a function

$$\hat{y} = f(x;\alpha,\beta\ldots) \qquad 3.9$$

visually through the set of coordinate pairs. To estimate m parameters, m points on the curve are selected giving m equations to solve. The process may be simplified by trying various types of graph paper using transformed coordinates until a straight line fit is possible.

In ascending order of efficiency the four methods of estimation may be listed as graphical, least squares, method of moments, maximum likelihood. To offset its great efficiency. however, the method of maximum likelihood is somewhat more difficult to apply.

There are also general criteria with which a distribution should comply before being used in hydrology. As an example, since negative flows are unacceptable a distribution should be bounded on the lower tail. This criterion would eliminate both the normal and the double exponential or Gumbel distributions. In fact, both of these distributions are used in hydrology by ignoring negative flows, replacing them with zeros, or treating the probability of zero flow as a probability mass (47).

Having selected a distribution and estimated its parameters the question is how to use this distribution in the frequency analysis. Chow (6) has proposed a general equation for hydrologic frequency analysis

$$x(K) = \mu + K.\sigma \tag{3.10}$$

where x(K) is the event magnitude at a given frequency of occurrence.

μ and σ are estimates of the population mean and standard deviation and K is a frequency factor which is a function of the recurrence interval and the distribution. For any chosen distribution a relationship can be derived between the recurrence interval and the frequency factor.

A measure of the variability of the resulting event magnitudes is the standard error of estimate. Each method of estimating the parameters of a distribution can also be used to derive the variance of estimates. The standard error of estimates is derived by the method of moments (32) as follows:

$$S = \left\{ \left[\sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \right] / n \right\}^{\frac{1}{2}} \qquad 3.11$$

where $\hat{x}_i$ is the computed estimate of recorded event x_i.

The differences between the recorded and the computed events may have two origins:

1. The choice of a theoretical distribution for the sample may be wrong, and
2. The errors in the parameters of the chosen distribution due to the shortness of the sample data. The standard error of estimate accounts for only the second of these sources.

Cramer (10) has shown that any function of the first two moments of a distribution, such as the general frequency equation, Equation 3.12, tends to normality as n increases with a variance given by:

$$S^2(K) = \text{var } x(K) = \text{var } \mu \left(\frac{\partial x}{\partial \mu}\right)^2 + \text{var } \sigma^2 \left(\frac{\partial}{\partial \sigma^2}\right)^2 +$$

$$2 \text{ cov } (\mu, \sigma^2) \frac{\partial x}{\partial \mu} \frac{\partial x}{\partial \sigma^2} \qquad 3.12$$

where μ and σ^2 are the mean and variance of the original distribution, estimated from the sample as:

$$\mu = \sum_{i=1}^{n} x_i/n \qquad 3.13$$

and

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2/(n-1) \qquad 3.14$$

From Equation 3.12 the derivatives are obtained as:

$$\frac{\partial x}{\partial \mu} = 1 \qquad 3.15$$

and

$$\frac{\partial x}{\partial \sigma^2} = \frac{K}{2\sigma} \qquad 3.16$$

so that Equation 3.14 becomes:

$$S^2(K) = \text{var } \mu + \frac{K^2}{4\sigma^2} \text{ var } \sigma^2 + \frac{K}{\sigma} \text{ cov } (\mu,\sigma^2) \qquad 3.17$$

Alternatively, an expression in terms of var σ instead of var σ^2 may be derived as follows:-

$$S^2(K) = \text{var } x \text{ } (K) = \text{var } [\mu + K\sigma] \qquad 3.18$$

$$S^2(K) = \text{var } \mu + K^2 \text{ var } \sigma + 2K \text{ cov } (\mu,\sigma) \qquad 3.19$$

The expression for var μ is independent of the original distribution and is given by:

$$\text{var } \mu = \frac{\sigma^2}{n} \qquad 3.20$$

Kendall and Stuart (24) have given general expressions for var σ and cov (μ,σ) as follows

$$\text{var } \sigma = \frac{\mu_4 - \mu_2^2}{4n\mu_2} \qquad 3.21$$

and

$$\text{cov } (\mu,\sigma) = \frac{\mu_3}{2n\sqrt{\mu_2}} \qquad 3.22$$

where μ_2, μ_3 and μ_4 are the second, third and fourth central moments of the distribution.

Substituting Equations 3.20, 3.21 and 3.22 in Equation 3.19 and simplifying yields a general expression

$$S(K) = \delta\sqrt{\frac{\mu_2}{n}} \qquad 3.23$$

where

$$\delta = \left[1 + \frac{K\mu_3}{\mu_2{}^{3/2}} + \frac{K^2(\mu_4 - \mu_2^2)}{4\mu_2^2} \right]^{\frac{1}{2}} \qquad 3.24$$

which can be used with any form of distribution. Since the moments are not distribution-free it must be evaluated separately for each distribution.

For distributions such as the lognormal, extremal Type III and log-Pearson Type III, which are logarithmic transformations of simple distributions there are two possible methods of computing event magnitudes and standard errors of event magnitudes. The first method is to develop analytical relationships for the frequency factor K and the parameter δ of the transformed distributions and use these with the mean and standard deviation of the original data, in Equations 3.10 and 3.23. The second method is to logarithmically transform the original data and use the mean and standard deviation of the logarithms in Equations 3.10 and 3.23 together with the K and δ values for the simple untransformed distribution.

If the distribution of the T-year event were known, then confidence limits could be derived for the event. Two methods can be used to find this distribution: analytical and empirical.

The analytical approach utilises the probability distribution fitted to the observed data. Cramer (10) has shown that for a sample of n values fitted with a distribution having a probability density function f(x) and a cumulative probability function F(x) then the density function, g(x), of a random variable x(K) is given by

$$g(x) = \binom{n}{m} (n-m)(F(x))^{m}(1-F(x))^{n-m-1}f(x) \qquad 3.25$$

where m is n.P and P is the cumulative probability associated with event magnitude x(K). Now if G(x) represents the cumulative probability function of the random variable x_T

$$G(x) = \int_{o}^{x_o} g(x)\, dx \qquad 3.26$$

then the upper or lower confidence limits, x_o, for the T-year event, x(K), may be found by solving Equation 3.26 for different levels of significance. For example, for the 95% upper confidence limit x_o must be found for G(x) = 0.95.

Unfortunately the evaluation of Equation 3.26 for x_o involves a lot of approximations and still must be carried out by a numerical iteration process. Because of these computational difficulties the second method of deriving confidence limits, the empirical method, is frequently used. In the empirical method Equations 3.10 and 3.23 are used to compute the mean T-year event, x(K), and the standard error of the T-year event, S(K). The assumption is then made that the distribution of T-year events is normal so that the confidence interval is given by

$$x(K) \pm t.S(K) \qquad 3.27$$

where t is the standard normal deviate corresponding to the required confidence level.

Section 3.4 of this report discusses in detail the circumstances in which this assumption of normality may be applicable.

3.2 Discrete Distributions

3.2.1 Binomial

Tossing a coin or drawing a card from a pack are examples of a Bernoulli trial. Bernoulli trials operate under three conditions:

1. Any trial can have only one of two possible outcomes; success or failure, true or false, rain or no rain, etc.
2. Successive trials are independent.
3. Probabilities are stable.

Under these conditions the probability of x successes in n trials is given by the binomial distribution as

$$p(x) = \binom{n}{x} p^x q^{n-x} \qquad 3.28$$

where $\binom{n}{x}$, sometimes written as nCx (only if $\geq x > 0$) or C^n_x, is the number of combinations of n events taken x at a time,

$$\binom{n}{x} = \frac{n!}{x!\ (n-x)!} \qquad 3.29$$

p is the probability of occurrence of an event, for example the probability of success in tossing a coin, q is the probability of failure,

$$q = 1-p \qquad 3.30$$

and x is the variate or the number of successful trials.

As an example of the use of the binomial distribution suppose that a dam has a projected life of 50 years and we wish to evaluate the probability that a flood with a return period of 100 years will occur once during the life of the dam. Then $p = 1/T = .01$, $q = 1-p = 0.99$, $x = 1$ and $n = 50$, so that

$$p(1) = \binom{50}{1} (.01)^1 \; (.99)^{49} = 0.306 \qquad 3.31$$

i.e., there is about a 31% chance that an event of that magnitude will occur once in the life of the dam.

3.2.2 Poisson

The terms of a binomial expansion are a little inconvenient to compute in any large number. Provided that p is small (say < 0.1) and n is large (say n > 30) and the mean n.p is constant and well defined, it can be shown (44) that

$$(p+q)^n \longrightarrow e^{-\lambda} \cdot e^{\lambda} = e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots\ldots \qquad 3.32$$

as $p \rightarrow 0$, $q \rightarrow 1$ and $n \rightarrow \infty$.

This is known as the Poisson expansion and is generally written

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad 3.33$$

where $\lambda = n.p$ is the mean. The finite binomial distribution can thus be approximated by the infinite Poisson distribution provided that the following four conditions apply:

1. The number of events is discrete.
2. Two events cannot coincide.
3. The mean number of events in unit time is constant, and
4. Events are independent.

Repeating the previous example, the probability that a 100 year return period flood will occur once in a 50 year period is seen to be

$$p(1) = \frac{0.5^1 e^{-0.5}}{1\cdot} \qquad 3.34$$

$$p(1) = 0.303 \qquad 3.35$$

which agrees well with the result obtained from the binomial expansion.

Equation 3.33 will give not only the probability of one event occuring in a given time but also the probability that two events may occur in that time, that three may occur, etc., etc. The probability that one or more events occur will therefore be given as a summation of Equation 3.33.

$$P(1,2...\infty) = \sum_{x=1}^{\infty} p(x) \qquad 3.36$$

but

$$\sum_{x=1}^{\infty} p(x) = \sum_{x=0}^{\infty} p(x) - p(0) \qquad 3.37$$

which from Equation 3.32 gives

$$P(1,2...\infty) = e^{-\lambda}e^{\lambda}-e^{-\lambda} \qquad 3.38$$

or,

$$P(1,2...\infty) = 1 - e^{-\lambda} \qquad 3.39$$

Table 3.1 shows, for $\lambda = 0.5$, the variation of P(x), x = 1, 2, 3, 4 and at $x = \infty$.

Table 3.1

Some Values of Probability of One or More Events for a Poisson Distribution with $\lambda = 0.5$

Variate Value, x	Probability, P(1...x)
1	0.30326
2	0.37908
3	0.39171
4	0.39329
.	.
.	.
.	.
∞	0.39347

Abbreviating the probability of one or more occurrences P(1, 2, ... ∞) to P and replacing λ, the average number of events per time period, by Δt/T where Δt is the time interval being considered (e.g. project life) and T is the event return period, then Equation 3.39 becomes:

$$P = 1 - e^{-\Delta t/T} \qquad 3.40$$

Hall and Howell (18) have prepared a table (Table 3.2) showing values of P for different time intervals and return periods

Table 3.2

Probabilities of One or More Occurrences of Events With Different Return Periods in Different Time Intervals

Time Interval Δt	Return Period T										
	.1	.2	.4	1	2	4	10	20	40	100	200
.1	.632	.393	.221	.095	.049	.025	.010	.005	.003	.001	.0005
.2	.865	.632	.393	.181	.095	.049	.020	.010	.005	.002	.001
.4	.982	.865	.632	.380	.181	.095	.039	.020	.010	.004	.002
1	1.000	.993	.918	.632	.393	.221	.095	.049	.025	.010	.005
2	1.000	1.000	.993	.865	.632	.393	.181	.095	.049	.020	.010
4	1.000	1.000	1.000	.982	.865	.632	.330	.181	.095	.039	.020
10	1.000	1.000	1.000	1.000	.993	.918	.632	.393	.221	.095	.049
20	1.000	1.000	1.000	1.000	1.000	.993	.865	.632	.393	.181	.095
40	1.000	1.000	1.000	1.000	1.000	1.000	.982	.865	.632	.330	.181
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.993	.918	.632	.393
200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.993	.865	.632

Note that, in this table, both Δt and T must be measured in the same time units. Hall and Howell (18) have extended this type of table to time intervals of 5 to 35 days. Since the probabilities in such tables are cumulative, by taking differences it is possible to compute probabilities of one or more occurrences of events with return periods between different values. For example the probability of one or more occurrences within 10 days of flood events with return periods between 10 and 100 years is 0.0024.

Under the assumptions that flood exceedences are independent identically distributed random variables and that the counting process for exceedences is a nonhomogeneous Poisson process, Todorovic and Woolhiser (42) have derived a one-dimensional distribution function for the time of occurence of the largest event in some time interval.

If z(t) represents the number of flood peak exceedances in the time interval (0,t) and $\Lambda(t)$ is the expected value of z(t)

$$\Lambda(t) = E\left\{ z(t) \right\} \qquad 3.41$$

then the probability that the time of occurrence T(t) of the largest momentary flood exceedance in the time interval (0,t) will be less than or equal to U, $P(T(t) \leq U)$, is given by:

$$P(\mathbf{T}(t) \leq \mathbf{U}) = \exp\left\{-\Lambda(t)\right\} + \frac{\Lambda(\mathbf{U})}{\Lambda(t)}\left(1 - \exp\left\{-\Lambda(t)\right\}\right) \qquad 3.42$$

The expression $\Lambda(t)$ was derived by Todorovic and Woolhiser (42) as a finite Fourier series.

Risk is discussed in more detail in a later chapter but it may be pointed out in passing that since from Equation 3.33, using $\lambda = \Delta t/T$,

$$\mathbf{p}(o) = e^{-\Delta t/T} \qquad 3.42$$

then the return period, T, of a design flood with a risk of failure p(0) in a project life Δt is:

$$T = \Delta t/\ln p(o) \qquad 3.44$$

e.g. for a 5% risk of failure (i.e. risk of an event of given magnitude occurring) in a 50 year life the project must be designed for a flood with return period of 975 years!

3.3 Continuous Distributions

3.3.1 Normal or Gaussian

General

A distribution is said to be normal if the variable can take any value from -∞ to +β and the probability density function is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad 3.45$$

where μ and σ are the population mean and standard deviation of the variable. The normal distribution is applicable if:

1. The variable is continuous,
2. Consecutive values are independent, and
3. Probabilities are stable.

The normal distribution can be shown to be (44) a limiting case of the binomial when $p \to q \to \frac{1}{2}$ and $n \to \infty$.

One of the features of the normal distribution is that the mean, mode and median are all the same. The normal distribution is also unusual in that the means and standard deviations derived by the methods of moments, least squares and maximum likelihood are identical. For the higher orders, all odd moments are zero and all even moments can be expressed in terms of σ so that the normal distribution is completely defined by the first two moments.

If the variable, x, is standardised, i.e. forced to a mean of zero and unit variance by subtracting the mean and dividing by the standard deviation, and is denoted by t then Equation 3.45 becomes

$$f(t) = \frac{1}{\sqrt{2\pi}} \quad e^{-t^2/2} \qquad 3.46$$

which is known as the standard normal distribution. Equation 3.46 has been approximated (accuracy > 2.27 x 10^{-3}) by a series of polynomials (1) such as:

$$f(t) = (a_o + a_1t^2 + a_2t^4 + a_3t^6)^{-1} \qquad 3.47$$

where

$a_o = 2.490895$, $a_2 = -0.024393$

$a_1 = 1.466003$, $a_3 = 0.178257$

Tables of the ordinates of the normal curve are also available (44) such as Table 3.3.

The probability corresponding to any interval in the range of the variate is represented by the area under the probability density curve,

$$P(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad dx \qquad 3.48$$

Standardising, the cumulative probability corresponding to Equation 3.46 is:

$$P(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \quad dt \qquad 3.49$$

Similarly, Abramowitz and Stegun (1) list several.approximations for Equations 3.49. A convenient polynomial approximation with an error term less than 1 x 10^{-5} is

$$P(t) = 1-f(t)(a_1q + a_2q^2 + a_3q^3) \qquad 3.50$$

where q is 1.0/(1.0+p.t), t is the positive standard normal deviate and p, a_1, a_2 and a_3 are constants with values

$p = 0.33267$ $\qquad a_2 = -0.12017$

$a_1 = 0.43618$ $\qquad a_3 = 0.93730$

Tables of the area under the standard normal curve are also available such as Table 3.4.

Table 3.3

Ordinates of the Normal Curve

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0395	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Note: $t = (x-\mu)/\sigma$

Estimation of Parameters

The parameters of the normal distribution can be derived by the method of maximum likelihood as previously described. The joint probability function of a sample of n values $x_1 \ldots x_n$ is given by the likelihood

$$L = f\,(x_i;\alpha,\beta\ldots)\;f\,(x_2;\alpha,\beta\ldots)\ldots\ldots f\,(x_n;\alpha,\beta\ldots) \qquad 3.51$$

which, for the normal distribution, is,

$$L = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^n \exp \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} \qquad 3.52$$

Taking logarithms:

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} \qquad 3.53$$

Differentiating with respect to the two parameters μ and σ^2 and equating to zero:

$$\frac{\partial L}{\partial\mu} = \sum_{i=1}^{n}(x_i - \mu)/\sigma^2 = o \qquad 3.54$$

and

$$\frac{\partial L}{\partial\sigma^2} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^3} = o \qquad 3.55$$

Now, from Equation 3.54

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = o \qquad 3.56$$

and since μ is a constant

$$\sum_{i=1}^{n} \mu = n\mu \qquad 3.57$$

So that from Equation 3.56 the maximum likelihood estimate of μ is

$$\hat{\mu} = \sum_{i=1}^{n} x_i/n \qquad 3.58$$

Similarly, from Equation 3.55

$$\frac{n}{\sigma} = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^3} \qquad 3.59$$

so that the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \mu)^2/n \qquad 3.60$$

Frequency Factor

It has been explained in the introduction to this chapter that a standard equation

$$x(K) = \mu + K\sigma \qquad 3.61$$

can be used in the frequency analyses. For the case of the normal distribution the frequency factor, K, is given by the standard normal deviate, t, in Equation 3.49. Thus, from Equation 3.49, knowing the probability of occurrence, the corresponding value of the standard normal deviate can be derived and substituted in Equation 3.61 in place of K to give the required event magnitude.

The easiest way of obtaining the standard normal deviate is from tables of the area under the normal curve such as Table 3.4. As an example consider the determination of t for an event with a mean return period of 100 years. The cumulative probability of non-exceedence, as a percentage, associated with this T value is 99%, (1-1/T). Of this, 50% is constributed by the integral of the standard normal density curve from -∞ to 0 leaving 49% from the integral from 0 to x. Looking up 0.49 in the body of the table it is found that t is given as 2.33.

Table 3.4

Area Under the Standard Normal Curve

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0.1	.0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	.0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	.1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	.1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	.1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	.2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0.7	.2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0.8	.2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	.3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	.3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	.3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	.3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	.4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	.4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	.4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1.6	.4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1.7	.4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	.4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	.4713	4719	4726	4732	4738	4744	4750	4756	4762	4767
2.0	.4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	.4821	4826	4830	4835	4838	4842	4846	4850	4854	4857
2.2	.4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2.3	.4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	.4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	.4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	.4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	.4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	.4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2.9	.4981	4982	4983	4983	4984	4984	4985	4985	4986	4986
3.0	.4986	4987	4987	4988	4988	4989	4989	4989	4990	4990
3.1	.4990	4991	4991	4991	4992	4992	4992	4992	4993	4993

Note: $t = (x-\mu)/\sigma$

The corresponding event magnitude is therefore $\bar{x} + 2.33\sigma$. It should be noted that this is using the standard normal curve in a one-tail manner as compared to the more usual two-tail method commonly employed in statistical tests.

Figures 3.1 and 3.2 compare the concepts of the one and two tail applications of the standard normal curve for an area under the curve of 0.95.

As a convenience, the frequency factors (standard normal deviates) for the normal (and lognormal) distribution are given in Table 3.5 for some commonly used cumulative probabilities. Table 3.5 also shows the standard normal deviates for the two-tail situtation which would be used in many statistical tests.

Table 3.5

Frequency Factor for Use in Normal and Lognormal Distributions

Cumulative Probability, P, %					
50	80	90	95	98	99
Corresponding Return Period, T, Years					
2	5	10	20	50	100
Frequency Factor					
0	0.842	1.282	1.645	2.054	2.326
Two-Tail Standard Normal Deviate					
0.674	1.282	1.645	1.960	2.326	2.576

It is sometimes useful to combine tables of the standard normal deviate with the plotting position m/(n+1) in which m is the rank of the event in descreasing order of magnitude. Table 3.6 is a sample table of this type.

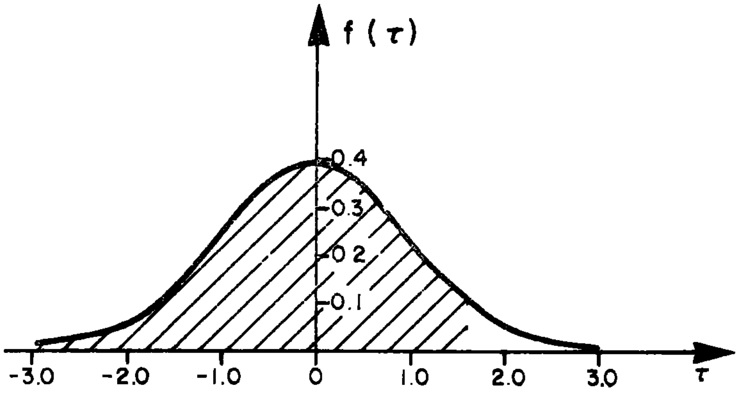


FIGURE 3.1
AREA UNDER THE STANDARD NORMAL CURVE
ONE - TAIL

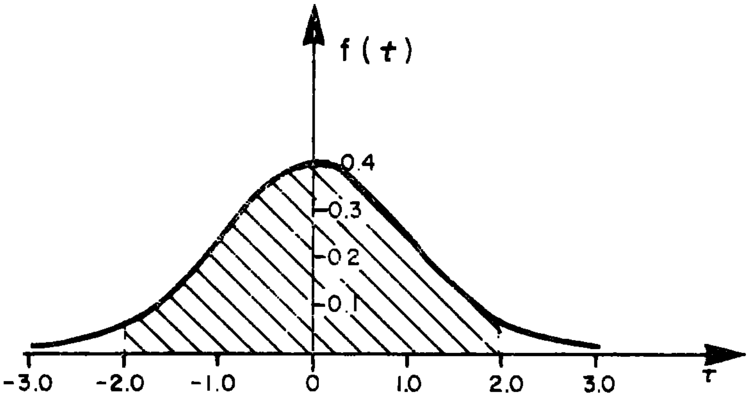


FIGURE 3.2
AREA UNDER THE STANDARD NORMAL CURVE
TWO- TAIL

Polynomial approximations are also available (1) to obtain the standard normal deviate corresponding to a given probability level. As an example, if the cumulative probability is P(t), then Q(t) = 1 - P(t). Now if

$$w = \sqrt{\ln(1/Q(t)^2)} \qquad 3.62$$

then

$$t \simeq w - \frac{c_o + c_1 w + c_2 w^2}{1 + d_1 w + d_2 w^2 + d_3 w^3} \qquad 3.63$$

where

$c_o = 2.515517;$ $d_1 = 1.432788$

$c_1 = 0.802853;$ $d_2 = 0.189269$

$c_2 = 0.010328;$ $d_3 = 0.001308$

This approximation is particularly useful when digital computers are used since it avoids the reverse integration in Equation 3.49. The error term in the approximation is stated (1) to be less than 4.5×10^{-4}. As a sample calculation, substituting P(t) = 0.95 gives w = 2.44775 and t = 1.64521.

Standard Error of Estimate

The general equation for the standard error of estimate from the method of moments is:

$$s^2(K) = \frac{\mu_2}{n} \left[1 + \frac{K\mu_3}{\mu_2^{3/2}} + \frac{K^2(\mu_4 - \mu_2^2}{4\mu_2^2} \right] \qquad 3.64$$

For the normal distribution the central moments are given by:

$$\mu_2 = \sigma^2 \qquad 3.65$$

$$\mu_3 = 0 \qquad 3.66$$

and

$$\mu_4 = 3\sigma^4 \qquad 3.67$$

Substituting Equations 3.65 to 3.67 into Equation 3.64, taking the square root and simplifying, results in

$$s(K) = \delta \sigma / \sqrt{n} \qquad 3.68$$

where

$$\delta = (1 + t^2/2)^{\frac{1}{2}} \qquad 3.69$$

since for the normal distribution the frequency factor, K, in Equation 3.64 is equal to the standard normal deviate, t. Table 3.7 provides values of the parameter δ for some common cumulative probabilities.

Table 3.6

Plotting Positions, P, and Standard Normal Deviate, t, for a Range of Samples of Size n Events

Event Rank	n = 19		n = 20		n = 21		n = 22		n = 23	
No. (m)	P %	t	P %	t	P %	t	P %	t	P %	t
1	5.0	1.64	4.8	1.67	4.5	1.69	4.3	1.71	4.2	1.73
2	10.0	1.28	9.5	1.31	9.1	1.34	8.7	1.36	8.3	1.38
3	15.0	1.04	14.3	1.07	13.6	1.10	13.0	1.12	12.5	1.15
4	20.0	.84	19.0	.87	18.2	.91	17.4	.94	16.7	.97
5	25.0	.67	23.8	.71	22.7	.75	21.7	.78	20.8	.81
6	30.0	.52	28.6	.57	27.3	.61	26.1	.64	25.0	.67
7	35.0	.39	33.3	.43	31.8	.47	30.4	.51	29.2	.55
8	40.0	.25	38.1	.30	36.4	.35	34.8	.39	33.3	.43
9	45.0	.13	42.9	.18	40.9	.23	39.1	.27	37.5	.32
10	50.0	.00	47.6	.06	45.5	.12	43.5	.16	41.7	.21
11	55.0	-.13	52.4	-.06	50.0	.00	47.8	.05	45.8	.10
12	60.0	-.25	57.1	-.18	54.5	-.12	52.2	-.05	50.0	.00
13	65.0	-.39	61.9	-.30	59.1	-.23	56.5	-.16	54.2	-.10
14	70.0	-.52	66.7	-.43	63.6	-.35	60.9	-.27	58.3	-.21
15	75.0	-.67	71.4	-.57	68.2	-.47	65.2	-.39	62.5	-.32
16	80.0	-.84	76.2	-.71	72.7	-.61	69.6	-.51	66.7	-.43
17	85.0	-1.04	81.0	-.87	77.3	-.75	73.9	-.64	70.8	-.55
18	90.0	-1.28	85.7	-1.07	81.8	-.91	78.3	-.78	75.0	-.67
19	95.0	-1.64	90.5	-1.31	86.4	-1.10	82.6	-.94	79.2	-.81
20			95.2	-1.67	90.9	-1.34	87.0	-1.12	83.3	-.97
21					95.5	-1.69	91.3	-1.36	87.5	-1.15
22							95.7	-1.71	91.7	-1.38
23									95.8	-1.73

Notes: $P = m/(n+1)$, $t = (x-\mu)/\sigma$

Table 3.7

Parameter δ for Use in Standard Error of Normal and Lognormal Distributions

Cumulative Probability, P, %					
50	80	90	95	98	99
Corresponding Return Period, T, Years					
2	5	10	20	50	100
1.0000	1.1638	1.3497	1.5340	1.7634	1.9249

An alternate method of tabulating Equation 3.68 is as the ratio S(K)/σ. Hardison (19) has provided tables of this type from which Table 3.8 has been derived. Since both δ and the ratio of deviations S(K)/σ are dimensionless, Tables 3.7 and 3.8 can equally well be used with the lognormal distribution.

Table 3.8
Dimensionless Ratio of the Standard Error of the T-Year Event to the Standard Deviation of the Annual Events for Normal and Lognormal Distributions

Return Period	Sample Length, n					
T	2	5	10	20	50	100
2	0.707	0.447	0.316	0.224	0.141	0.100
5	0.782	0.495	0.350	0.247	0.156	0.116
10	0.954	0.604	0.427	0.302	0.191	0.135
20	1.083	0.685	0.484	0.342	0.217	0.153
50	1.208	0.764	0.540	0.382	0.242	0.176
100	1.364	0.863	0.610	0.431	0.273	0.193

If v(K) is the expected coefficient of variation of the estimate x(K) and V(K) is the coefficient of variation of the true event y(K) defined respectively as

$$v(K) = S(K)/y(K) \qquad 3.70$$

and

$$V(K) = \sigma/\mu \qquad 3.71$$

then from Equation 3.69:

$$v(K) = \frac{V(K)}{\sqrt{n}} \cdot \sqrt{\frac{1 + t^2/2}{(1 + t.V(K))}} \qquad 3.72$$

Nash and Amorocho (32) have shown that as $K \to \infty$, $v(K) \to 1/\sqrt{2N}$. Graphs of $v(K)\sqrt{2n}$ versus t for different values of V(K) show that the coefficient of variation has a minimum value between t = 0 and t = 2 and that for values of V(K) > 0.2 the mean annual event (t = 0) is less well defined than the event corresponding to t = 1.

As an example of the computation of a confidence interval for a normal distribution consider the estimate of the 100 year event from a sample record of 50 years. The cumulative probability of non-exceedence of the 100 year event (area under the normal curve) is 99% and so from Table 3.4 the standard normal deviate, t, is 2.33. In Equation 3.68, S(K) is computed as 0.273σ, where σ is the sample estimate of the population standard deviation. Using a 95% confidence level (two-tail) the confidence interval around the 100 year event, x_{100}, is given as $x_{100} \pm 1.96 * 0.273\ \sigma$.

3.3.2 Lognormal

General

If the logarithms, ln x, of a variable x are normally distributed, then the variable x is said to be lognaithmic-normally distributed so that

$$f(x) = \frac{1}{x\sigma_y \sqrt{2\pi}} e^{-\frac{[\ln x - \mu_y]^2}{2\sigma_y^2}} \qquad 3.73$$

where μ_y and σ_y are the mean and standard deviation of the natural logarithms of x.

Chow (6) has provided a theoretical justification for the use of the lognormal distribution. The causative factors for many hydrologic variables act multiplicatively rather than additively and so the logarithms of these factors will satisfy the four basic conditions for normal distributions. The hydrologic variable will then be the product of these causative factors.

The mean, μ_y, and standard deviation, σ_y, of the logarithms of x can be related to the mean of x, μ_x, and the standard deviation of x, σ_x, from the generalised moment function so that the first three moments are (5):

$$\mu_1' = \mu_x = e^{\mu_y + \sigma_y^2/2} \qquad 3.74$$

$$\mu_2 = \sigma_x^2 = (e^{\sigma_y^2} - 1)\ (e^{\mu_y + \sigma_y^2/2})^2 \qquad 3.75$$

$$\mu_3 = (e^{3\sigma_y^2} - 3e^{\sigma_y^2} + 2)\, e^{3\mu_y + 3\sigma_y^2/2} \qquad 3.76$$

The coefficient of variation can be obtained from Equations 3.74 and 3.75 as:

$$z = \frac{\sigma_x}{\mu_x} = (e^{\sigma_y^2} - 1)^{1/2} \qquad 3.77$$

while the coefficient of skew from Equations 3.75 and 3.76 is:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{e^{3\sigma_y^2} - 3e^{\sigma_y^2} + 2}{(e^{\sigma_y^2} - 1)^{3/2}} \qquad 3.78$$

Comparing Equations 3.77 and 3.78 the relationship between the coefficients of variation and skew is given as:

$$\gamma_1 = 3z + z^3 \qquad 3.79$$

Singh and Sinclair (38) described a mixed, or compound, probability distribution made up of two lognormal distributions, as:

$$P(x) = a_1P_1(x) + a_2P_2(x) \qquad 3.80$$

where

$$P_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x - \mu_1)^2/2\sigma_1^2}\, dx \qquad 3.81$$

$$P_2(x) = \frac{1}{\sigma_2\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x - \mu_2)^2/2\sigma_2^2}\, dx \qquad 3.82$$

and

$$a_1 + a_2 = 1 \qquad 3.83$$

The mean, variance and coefficient of skew of the distribution P(x) may be estimated from the sample, or computed from

$$\mu = a_1\mu_1 + a_2\mu_2 \qquad 3.84$$

$$\sigma^2 = a_1\sigma_1^2 + a_2\sigma_2^2 + a_1a_2(\mu_2 - \mu_1)^2 \qquad 3.85$$

$$\gamma_1 = \frac{3a_1a_2(\mu_1 - \mu_2)\ (\sigma_1^2 - \sigma_2^2)}{\sigma^3} + \frac{a_1a_2(a_2 - a_1)(\mu_1 - \mu_2)^3}{\sigma^3} \qquad 3.86$$

The advantage of this method (38) is that it has the versatility of high parameter models without the errors and uncertainties which result from the use of higher order sample moments.

Estimation of Parameters

Following the maximum likelihood procedure the likelihood expression for Equation 3.73 is

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma_y^2 + \ln \sum_{i=1}^{n} 1/x_i - \sum_{i=1}^{n} (\ln x_i - \mu_y)^2/2\sigma_y^2 \qquad 3.87$$

Differentiating Equation 3.87 with respect to μ_y and σ_y^2 and equating to zero yields the maximum likelihood estimates

$$\hat{\mu}_y = \sum_{i=1}^{n} \ln x_i/n \qquad 3.88$$

$$\hat{\sigma}_y^2 = \sum_{i=1}^{n} (\ln x_i - \mu_y)^2/n \qquad 3.89$$

For streamflows, this method presents a problem. Many streamflow records have the occasional zero flow and, when taking logarithms, this becomes $-\infty$ and cannot be processed. Several solutions to this problem have been proposed (24) such as

1. Add 1.0 to all data.

2. Add small positive value (such as 0.1, 0.01, 0.001, etc.) to all data.

3. Substitute 1.0 in place of all zero readings.

4. Substitute small positive value in place of all zero readings.

All of these solutions affect the parameters of the distribution (1. and 2. affect the mean, 3. and 4. affect both the mean and variance) but the least damaging solution is 3., to substitute 1.0 in place of all zero readings. Substitution of small positive values is to be avoided because of the large effect this has in a logarithmic scale.

An alternate solution is to consider the probability distribution as the sum of a probability mass at zero and a probability density distribution over the remainder of the range (46), so that:

$$P[o \leq x \leq x_o] = P[x=o] + P[o < x \leq x_o] \qquad 3.90$$

In this way, if p_o is the probability of occurrence of $x = 0$ in any one year and p_x is the conditional probability of occurrence of x in any one year given that x is not zero, then:

$$P[o < x < x_o] = p_x(1 - p_o) \qquad 3.91$$

Frequency factor

If y = ln x is normally distributed, the general frequency equation (e.g. Equation 3.10) can be written as:

$$y(K) = \mu_y + t\sigma_y \qquad 3.92$$

where, now, t is the standard normal deviate.

To avoid computation of the mean and standard deviation of logarithms in Equation 3.92, the general frequency equation, Equation 3.10, may be modified by substituting for μ and σ from Equations 3.74 and 3.75 and using $e^{y(K)}$ in place of x(K):

$$e^{y(K)} = e^{\mu_y + \sigma_y^2/2} \left[1 + K(e^{\sigma_y^2} - 1)^{\frac{1}{2}}\right] \qquad 3.93$$

From Equation 3.93 the frequency factor for the lognormal distribution, K, is given by

$$K = \frac{e^{y(K) - \mu_y - \sigma_y^2/2} - 1}{(e^{\sigma_y^2} - 1)^{\frac{1}{2}}} \qquad 3.94$$

But, from Equation 3.92

$$y(K) - \mu_y = t\sigma_y \qquad 3.95$$

which, when substituted in Equation 3.94, yields:

$$K = \frac{e^{\sigma_y t - \sigma_y^2/2} - 1}{(e^{\sigma_y^2} - 1)^{\frac{1}{2}}} \qquad 3.96$$

where, as before, t is the standard normal deviate. Equation 3.96 still involves expressions in σ_y. This can be avoided, however, by using Equation 3.77 relating the coefficient of variation of the observed events, z, to the standard deviation of the logarithms, σ_y. Substituting in Equation 3.96 yields

$$K = \frac{e^{[\ln(1+z^2)]^{\frac{1}{2}}t - [\ln(1+z^2)]/2} - 1}{z} \qquad 3.97$$

Table 3.9 shows values of K computed from Equation 3.97 for some commonly used return periods and various values of z, the coefficient of variation. Chow (5), (6) gave more comprehensive tables which, however, require somewhat awkward interpolation to use for values of z. It is worth noting that in Chow's tables the first line, for a zero coefficient of skew, is equivalent to the normal distribution. Since the relationship between the coefficients of variation and skew (Equation 3.79) does not hold for the normal distribution the frequency factors are independent of the coefficient of variation (35) at that particular coefficient of skew.

Table 3.9

Frequency Factor for Lognormal Distribution

Coefficient			Cumulative Probability, P, %			
of	50	80	90	95	98	99
Variation			Corresponding Return Period, T, Years			
z	2	5	10	20	50	100
0.0500	-0.0250	0.8334	1.2965	1.6863	2.1341	2.4370
0.1000	-0.0496	0.8222	1.3078	1.7247	2.2130	2.5489
0.1500	-0.0738	0.8085	1.3156	1.7598	2.2899	2.6607
0.2000	-0.0971	0.7926	1.3200	1.7911	2.3640	2.7716
0.2500	-0.1194	0.7746	1.3209	1.8183	2.4348	2.8805
0.3000	-0.1406	0.7547	1.3183	1.8414	2.5016	2.9866
0.3500	-0.1604	0.7333	1.3126	1.8602	2.5638	3.0890
0.4000	-0.1788	0.7106	1.3037	1.8746	2.6212	3.1870
0.4500	-0.1957	0.6870	1.2920	1.8848	2.6734	3.2799
0.5000	-0.2111	0.6626	1.2778	1.8909	2.7202	3.3673
0.5500	-0.2251	0.6379	1.2613	1.8931	2.7615	3.4488
0.6000	-0.2375	0.6129	1.2428	1.8915	2.7974	3.5241
0.6500	-0.2485	0.5879	1.2226	1.8866	2.8279	3.5930
0.7000	-0.2582	0.5631	1.2011	1.8786	2.8532	3.6556
0.7500	-0.2667	0.5387	1.1784	1.8677	2.8735	3.7118
0.8000	-0.2739	0.5148	1.1548	1.8543	2.8891	3.7617
0.8500	-0.2801	0.4914	1.1306	1.8388	2.9002	3.8056
0.9000	-0.2852	0.4686	1.1060	1.8212	2.9071	3.8437
0.9500	-0.2895	0.4466	1.0810	1.8021	2.9103	3.8762
1.0000	-0.2929	0.4254	1.0560	1.7815	2.9098	3.9035

There are then two possible procedures for using a lognormal distribution to estimate T-year event magnitudes. The standard normal deviate, t, may be used with the mean and standard deviation of the logarithms of the recorded events, or the frequency factor, K, may be used with the mean and standard deviation of the recorded events.

As an example of these two procedures consider the following data from Collier (7). Annual maximum discharges of the Saint John River at Fort Kent, New Brunswick for the 37 years 1927 to 1963 yield a mean discharge of 81,000 cfs, a standard deviation of 22,800 cfs and a coefficient of variation of 0.28. The mean and standard deviation of the logarithms of these 37 events are 11.263 and 0.284.

Using the first method, the magnitude of the 100-year event ($t = 2.326$) is given by

$$x(100) = e^{11.263 \times 2.326 \times 0.286} = 151{,}000 \text{ cfs} \qquad 3.98$$

Using the second method the frequency factor, K, from Table 3.9 is 2.944 so that

$$x(100) = 81{,}000 + 2.944 \times 22{,}800 = 148{,}000 \text{ cfs} \qquad 3.99$$

The two methods thus produce comparable results and so, if the lognormal distribution is to be used, there is nothing to gain from the extra work involved in taking logarithms and computing μ_y and σ_y.

Standard Error of Estimate

In the previous section two frequency factors were derived for the lognormal distribution. The standard normal deviate, t, can be used with the mean and standard deviation of the logarithms of the events (Equation 3.92) and the frequency factor, K, can be used directly with the mean and standard deviation of the sample events (Equation 3.10). Development of these two relationships by the method of moments leads to two expressions for the standard error of estimate, S(K).

Firstly, if the standard normal deviate is used then the standard error S(K), in logarithmic units, is given by Equations 3.68 and 3.69 provided that σ_y is substituted for σ. Values of the parameter δ are given in Table 3.7 for various commonly used return periods. From the standard error in logarithmic units, S(K), the positive and negative standard errors can be derived as

$$PSE = x(T)(e^{S(K)} - 1) \qquad 3.100$$

and

$$NSE = x(T)(-[1 - e^{-S(K)}]) \qquad 3.101$$

Hardison (19) has published graphs showing the variation of S(K) with the standard deviation of the logarithms, σ_y, and has provided tables to convert S(K) from logarithmic units back to the units of the basic data.

Alternatively, to avoid the computation of the standard deviation of the logarithms, use can be made of the equation derived earlier for standard error:

$$S^2(K) = \frac{\mu_2}{n} \; [1 + \frac{K\mu_3}{\mu_2^{3/2}} + \frac{K^2(\mu_4 - \mu_2^2)}{4\mu_2^2}] \qquad 3.102$$

where $S^2(K)$ is in the same units as x^2. Kendall and Stuart (24) define the second, third and fourth central moments of the lognormal distribution as:

$$\mu_2 = w^2\rho^2(w^2-1) \qquad 3.103$$

$$\mu_3 = w^3\rho^3(w^2-1)^2(w^2+2) \qquad 3.104$$

$$\mu_4 = w^4\rho^4(w^2-1)^2(w^8+2w^6+3w^4-3) \qquad 3.105$$

where

$$w = \exp \sigma_y^2/2 \qquad 3.106$$

and

$$\rho = \exp \mu_y \qquad 3.107$$

Substituting Equations 3.103 to 3.105 into Equation 3.102 and making use of the relationship between the coefficient of variation, z, and the standard deviation of the logarithms, σ_y, (Equation 3.80) results in:

$$S^2(K) = \frac{\sigma^2}{n} \; [1+(z^3+3z)K + (z^8+6z^6+15z^4+16z^2+2)K^{2/}4] \qquad 3.108$$

Simplifying this equation to the standard form:

$$S(K) = \delta_y \sigma/\sqrt{n} \qquad 3.109$$

Table 3.10 provides values of δ_y for some commonly used values of return period and coefficient of variation.

Confidence limits on the event magnitude are then computed by this method as:

$$x(K) \pm t\delta_y \sigma/\sqrt{n} \qquad 3.110$$

assuming the normality of the distribution of x(K).

Kaczmarek (21) has used a slightly different approach to derive values of δ_y which differ from those in Table 3.10.

Table 3.10

Parameter δ_y for Use in Standard Error of Lognormal Distribution

Coefficient		Cumulative Probability, P, %				
of	50	80	90	95	98	99
Variation		Corresponding Return Period, T, Years				
Z	2	5	10	20	50	100
0.0500	0.9983	1.2163	1.4325	1.6442	1.9087	2.0964
0.1000	0.9932	1.2700	1.5224	1.7682	2.0767	2.2974
0.1500	0.9848	1.3242	1.6190	1.9055	2.2676	2.5292
0.2000	0.9733	1.3785	1.7214	2.0557	2.4819	2.7932
0.2500	0.9589	1.4325	1.8292	2.2185	2.7202	3.0908
0.3000	0.9420	1.4857	1.9420	2.3937	2.9829	3.4235
0.3500	0.9229	1.5380	2.0596	2.5813	3.2708	3.7929
0.4000	0.9021	1.5892	2.1816	2.7812	3.5845	4.2007
0.4500	0.8801	1.6392	2.3080	2.9937	3.9251	4.6489
0.5000	0.8575	1.6879	2.4389	3.2189	4.2935	5.1395
0.5500	0.8351	1.7354	2.5742	3.4573	4.6910	5.6749
0.6000	0.8138	1.7818	2.7142	3.7093	5.1190	6.2574
0.6500	0.7945	1.8271	2.8592	3.9756	5.5790	6.8899
0.7000	0.7784	1.8714	3.0095	4.2570	6.0729	7.5754
0.7500	0.7669	1.9148	3.1655	4.5542	6.6024	8.3171
0.8000	0.7615	1.9576	3.3276	4.8682	7.1698	9.1185
0.8500	0.7635	1.9997	3.4962	5.2001	7.7773	9.9834
0.9000	0.7746	2.0414	3.6719	5.5509	8.4272	10.9157
0.9500	0.7959	2.0828	3.8552	5.9217	9.1221	11.9196
1.0000	0.8284	2.1239	4.0466	6.3136	9.8646	12.9995

3.3.3 Three Parameter Lognormal

General

Just as the lognormal distribution represents the normal distribution of the logarithms of the variable x, so the 3-parameter lognormal represents the normal distribution of the logarithms of the reduced variable (x-a) where a is a lower boundary. The probability density distribution is then given by:

$$f(x) = \frac{1}{(x-a)\sigma_y\sqrt{2\pi}} e^{-\frac{[\ln(x-a)-\mu_y]^2}{2\sigma_y^2}} \qquad 3.111$$

where μ_y and σ_y are now the mean and standard deviation of the logarithms of (x-a).

The mean and standard deviation of the distribution $y = \ln(x-a)$ are related to the mean, μ_x, and standard deviation, σ_x, of the original distribution x by

$$\mu_x = a + e^{\mu_y + \sigma_y^2/2} \qquad 3.112$$

$$\sigma_x = (e^{\sigma_y^2} - 1)^{\frac{1}{2}} e^{\mu_y + \sigma_y^2/2} \qquad 3.113$$

These equations may be compared with Equations 3.74 and 3.75.

Similarly an expression is available relating the coefficients of variation of the distributions x and ln(x-a). Since the presence of the parameter a does not affect the variance of the distribution the coefficients of variation of the distributions x and (x-a) can be defined as:

$$z_1 = \sigma_x/\mu_x \qquad 3.114$$

and

$$z_2 = \sigma_x/(\mu_x-a) \qquad 3.115$$

so that

$$z_2 = \mu_x z_1/(\mu_x-a) \qquad 3.116$$

The coefficient of skew of the 3-parameter lognormal distribution is related to its coefficient of variation in the same way as the lognormal distribution (35) (see Equation 3.79).

Estimation of Parameters

If the lower boundary, a is known (perhaps some physical reason why x cannot be lower than a) then the reduced variable (x-a) can be determined and the analysis performed as for the lognormal distribution. However, if a is not known then it must be evaulated in terms of the statistical measures of the variable x. The range of possible values for a is between zero and the magnitude of the smallest recorded event.

The maximum likelihood method yields expressions for μ_y and σ_y^2 the mean and variance of the distribution ln(x-a) similar to Equations 3.88 and 3.89 but substituting x-a for a:

$$\hat{\mu}_y = \sum_{i=1}^{n} \ln(x-a)/n \qquad 3.117$$

$$\hat{\sigma}_y^2 = \sum_{i=1}^{n} [\ln(x_i-a)-\hat{\mu}_y]^2/n \qquad 3.118$$

No direct maximum likelihood expression for parameter a is possible, however. Differentiating the likelihood equation with respect to a and equating to zero yields the expression (47)

$$\sum_{i=1}^{n} \frac{1}{x_i - \hat{a}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln^2(x_i-\hat{a}) - \left[\frac{1}{n} \sum_{i=1}^{n} \ln(x_i-\hat{a})\right]^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i-\hat{a}) \right\} + \sum_{i=1}^{n} \frac{\ln(x_i-\hat{a})}{x_i-\hat{a}} = 0 \qquad 3.119$$

This expression can be solved only by an iterative procedure.

The method of moments permits a direct solution for a but necessitates the compution of the sample coefficient of skew, γ_1. From Equation 3.116 the parameter a is defined as

$$a = \mu_x[1-z_1/z_2] \qquad 3.120$$

The value of z_1 can be computed directly from the observed events using Equation 3.114. Since the second and third moments of the distribution are independent of a, z_2 can be obtained (29) from Equation 3.79. Yevjevich (47) has given the solution of Equation 3.79 as:

$$z_2 = \left[\frac{(\gamma_1^2 + 4)^{1/2} + \gamma_1}{2}\right]^{1/3} - \left[\frac{-(\gamma_1^2 + 4)^{1/2} + \gamma_1}{2}\right]^{1/3} \qquad 3.121$$

where γ_1 is the sample coefficient of skew.

Computation of the coefficient of skew from small samples is notoriously subject to error. Sangal and Biswas (35) have derived a method of estimating the parameter a using only the mean, median and standard deviation of the original data. Their solution is:

$$a = \delta - \frac{\sigma_x^2}{2(\mu_x - \delta)} \qquad 3.122$$

where δ is the median of x, determined as the mean of the middle 1/5th of the data.

As pointed out by Condie (8) the determination of a is very sensitive to the difference of the mean and median, $(\mu_x - \delta)$. When this difference becomes small, then a takes a ridiculously large negative value. The assumption then would be that the original data are symetrically but not necessarily normally, distributed. This follows from an empirical relationship for moderately skewed distributions:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median}) \qquad 3.123$$

From 3.123 if Mean = Median, then Mean = Mode; thus the mean, mode and median **coincide**, which only occurs in symetrical distributions.

Condie (8) has described a simple graphical method of determining the parameter a which is applicable provided that:

a) at least one of the graph scales, vertical or horizontal, is logarithmic.

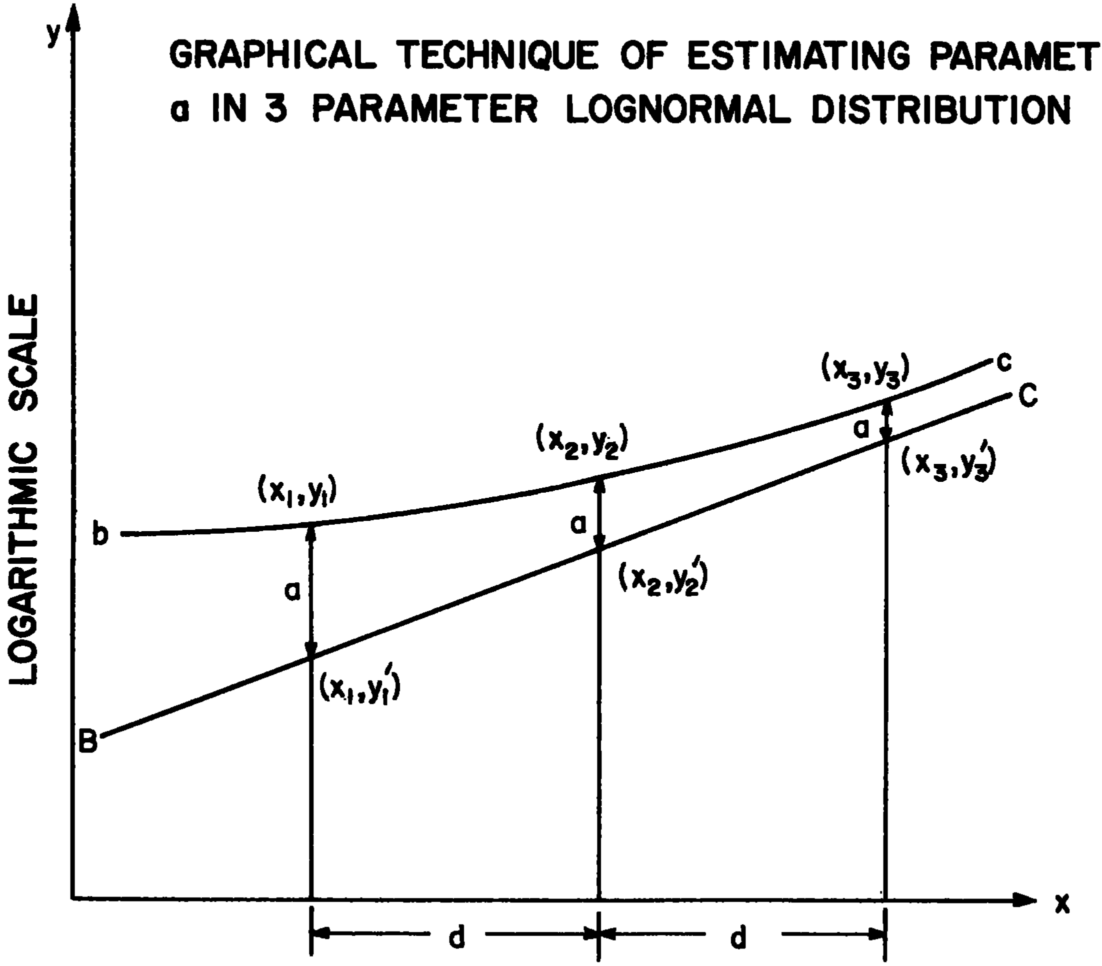
b) the curvature of the best fitting line is gradually decreasing.

The other scale can be linear, logarithmic, normally probabilistic, time scale etc.

Referring to Figure 3.3; (x_1, y_1) (x_2, y_2), (x_3, y_3) are three points on the best-fitting line through the plotted event magnitudes (b-c) such that

FIGURE 3-3

GRAPHICAL TECHNIQUE OF ESTIMATING PARAMETER a IN 3 PARAMETER LOGNORMAL DISTRIBUTION



$$x_3 - x_1 = x_2 - x_3 \qquad 3.124$$

The object is then to find a constant a, which will move points (x_1,y_1), (x_2,y_2), (x_3,y_3) to points (x_1,y_1'), (x_2,y_2'), (x_3,y_3') all on a straight line (B-C). The straight line then is easier to extrapolate than the original arbitrarily curved line.

Considering the slope of the logarithmically straight line B-C:

$$\frac{\ln(y_3-a)- \ln(y_1-a)}{x_3 - x_1} = \frac{\ln(y_2-a) - \ln(y_3-a)}{x_2 - x_1} \qquad 3.125$$

but from Equation 3.124 this reduces to

$$\ln\frac{y_3-a}{y_1-a} = \ln\frac{y_2-a}{y_3-a} \qquad 3.126$$

from which

$$a = \frac{y_1y_2-y_3^2}{y_1+y_2-2y_3} \qquad 3.127$$

The method is simple and can be applied to any distribution with a logarithmic scale such as lognormal, log-Gumbel, log-Pearson Type III.

Frequency Factor

The general frequency equation for the 3-parameter lognormal distribution is given by

$$y = \ln(x-a) = \mu_y + t\sigma_y \qquad 3.128$$

where t, the frequency factor, is again the standard normal deviate.

Since the central moments of the 3-parameter lognormal are the same as for the lognormal distribution, the alternate approach of using a frequency factor with μ_x and σ_x can also be taken.

Standard Error of Estimate

The standard error of the T-year event for the 3-parameter lognormal distribution is the same as for the lognormal, either

$$S(K) = \frac{\sigma_y}{\sqrt{n}} \sqrt{1+t^2/2} \qquad 3.129$$

where S(K) is in log units, σ_y is the standard deviation of the logarithms of the n events, x-a, and t is the standard normal deviate, or,

$$S(K) = \delta_y \sigma / \sqrt{n} \qquad 3.130$$

where S(K) is in the same units as x, δ_y can be taken from Table 3.10 and σ is the standard deviation of x.

3.3.4 Extreme Value Distributions

Suppose that from N samples each containing n events the maximum or minimum event in each sample is selected. As n increases, the distribution of the N maxima or minima approaches a limiting or assymptotic form. The type of the limiting form depends on the type of the initial distribution of the Nn values. The distribution of the maxima or minima is given by the functional equation

$$F^n(x) = F(a_n x+b_n) \qquad 3.131$$

where a_n and b_n are functions of n.

Fisher and Tippett (quoted in (46)) have shown that there are three possible solutions to the functional equation. These are known, logically enough, as Types I, II and III extremal distributions. The Type I distribution is unbounded, the Type II has a lower limit and the Type III has an upper limit.

3.3.4.1 Type 1 Extremal

General

The Type I distribution (or Gumbel (16)) is often used for maximum type events and results from any initial unlimited distribution of exponential type which converges to an exponential function (6). Examples of this type of distribution include the normal and lognormal distributions. The derivation of the Type I distribution for a simple exponential function can be described (41) as follows:

(a) Let ε_1 ε_2 ... ε_n be a series of independent random variables with a cumulative probability distribution given by:

$$F(x) = P(\varepsilon_v \leq x) \qquad 3.132$$

(b) Define x_n as the maximum value of ε in a sample of length n i.e. $\max\limits_{1 \leq v \leq 1} \varepsilon_v$ so that

$$P(x_n \leq y) = P(\varepsilon_v \leq y,\ \varepsilon_2 \leq y.....) \qquad 3.133$$

or

$$P(x_n \leq y) = [F(y)]^n \qquad 3.134$$

(c) Now assume that the tail of the distribution F(y) is exponential such that

$$F(y) = 1 - \alpha e^{-y} \qquad 3.135$$

(d) From Equation 3.134 if ln (αn) is a normalising constant

$$P(\chi_n \leq y+\ln(\alpha n) = [F(y+\ln(\alpha n))]^n \tag{3.136}$$

and from Equation 3.135

$$F(y+\ln(\alpha n)) = 1-\alpha e^{-(y+\ln(\alpha n))} \tag{3.137}$$

so that

$$P(\chi_n \leq y+\ln(\alpha n)) = [1-\alpha e^{-(y+\ln(\alpha n))}]^2 \tag{3.138}$$

or

$$P(\chi_n \leq y+\ln(\alpha n)) = [1-e^{-y}/n]^n \tag{3.139}$$

(e) If n → ∞, then:

$$\lim_{n\to\infty} P(\chi_n \leq y+\ln(\alpha n)) = \lim_{n\to\infty} [1-e^{-y}/n]^n \tag{3.140}$$

or

$$\lim_{n\to\infty} P(\chi_n \leq y+\ln(\alpha n)) = e^{-e^{-y}} \tag{3.141}$$

This is the reduced form of the cumulative probability distribution. Substituting the expression

$$y = \alpha(x-\beta) \tag{3.142}$$

the cumulative probability of the Type I extremal distribution becomes

$$P(x) = e^{-e^{-\alpha(x-\beta)}} \qquad 3.143$$

and the probability density becomes

$$p(x) = \alpha e^{\{-\alpha(x-\beta)-e^{-\alpha(x-\beta)}\}} \qquad 3.144$$

where $P(x)$ is the probability of an event not exceeding x, α is a concentration parameter and β is a measure of central tendency.

Estimation of Parameters

Re-arranging Equation 3.143 the event magnitude, x, corresponding to a return period, T, can be expressed as:

$$x = \beta - \frac{1}{\alpha} \ln(-\ln(1-1/T)) \qquad 3.145$$

where

$$T = 1/(1-P(x)) \qquad 3.146$$

Now, provided α and β are known, the event magnitude for any required return period can be determined directly.

The maximum likelihood method of estimating α and β postulates (33) that α and β should be such that the probability of n individual maximum events $x_1 \ldots x_n$ actually being observed as n annual peaks should be a maximum. The probability that x_1 occurs as an annual peak event is:

$$P(x_1) = \alpha e^{\{-\alpha(x_1-\beta) -e^{-\alpha(x_1-\beta)}} \qquad 3.147$$

and for x_2:

$$p(x_2) = \alpha e^{\{-\alpha(x_2-\beta)-e^{-\alpha(x_2-\beta)}} \qquad 3.148$$

now

$$p(x_1,\ldots.,x_n) = p(x_1)p(x_2)\ldots.p(x_n) \qquad 3.149$$

so that:

$$p(x_1,\ldots,x_n) = \alpha^n e^{\{-\alpha \sum_{i=1}^{n} (x_i-\beta) - \sum_{i=1}^{n} e^{-\alpha(x_i-\beta)}\}} \qquad 3.150$$

The method of maximum likelihood then takes the logarithm of Equation 3.150, partially differentiates with respect to α and β and equates to zero:

$$L = \ln p(x_i,\ldots,x_n) \qquad 3.151$$

$$L = n \ln \alpha - \alpha \sum_{i=1}^{n} (x_i - \beta) - \sum_{i=1}^{n} e^{-\alpha(x_i-\beta)} \qquad 3.152$$

$$\frac{\partial L}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} (x_i-\beta) + \sum_{i=1}^{n} (x_i-\beta)e^{-\alpha(x_i-\beta)} \qquad 3.153$$

$$\frac{\partial L}{\partial \beta} = n\alpha - \alpha \sum_{i=1}^{n} e^{-\alpha(x_i-\beta)} \qquad 3.154$$

Setting Equation 3.154 equal to zero:

$$\sum_{i=1}^{n} e^{-\alpha(x_i-\beta)} = n \qquad 3.155$$

so that:

$$e^{\alpha\beta} = n/\sum_{i=1}^{n} e^{-\alpha x_i} \qquad 3.156$$

or

$$\beta = \frac{1}{\alpha} \ln \left[n/\sum_{i=1}^{n} e^{-\alpha x_i}\right] \qquad 3.157$$

If the arithmetic mean of the series $x_1 \ldots x_n$ is denoted by μ, then Equation 3.153 can be written as:

$$\frac{\partial L}{\partial \alpha} = \frac{n}{\alpha} - n(\mu-\beta) + e^{\alpha\beta} \sum_{i=1}^{n} (x_i-\beta)e^{-\alpha x_i} \qquad 3.158$$

Substituting for $e^{\alpha\beta}$ from Equation 3.156:

$$\frac{\partial L}{\partial \alpha} = \frac{n}{\alpha} - n(\mu-\beta) + \frac{n \sum_{i=1}^{n} (x_i-3)e^{-\alpha x_i}}{\sum_{i=1}^{n} e^{-\alpha x_i}} \qquad 3.159$$

Equating this to zero and simplifying:

$$F(\alpha) = \sum_{i=1}^{n} x_i e^{-\alpha x_i} - (\mu-1/\alpha) \sum_{i=1}^{n} e^{-\alpha x_i} = o \qquad 3.160$$

Equation 3.160 cannot be solved for α analytically and so a Taylor's expansion is often used (33)

$$F(\alpha_{j+1}) = F(\alpha_j+h_j) \qquad 3.161$$

$$F(\alpha_{j+1}) = F(\alpha_j) + h_j F'(\alpha_j) \qquad 3.162$$

where $F'(\alpha_j)$ is the first order derivative of $F(\alpha)$ with respect to α

$$F'(\alpha) = -\sum_{i=1}^{n} x_i^2 e^{-\alpha x_i} + (\mu-1/\alpha) \sum_{i=1}^{n} x_i e^{-\alpha x_i} - \frac{1}{\alpha^2} \sum_{i=1}^{n} e^{-\alpha x_i} \qquad 3.163$$

and α_j and α_{j+1} are successive approximations to α. The procedure adopted by Panchang (33) is to estimate α_1 from the method of moments (47) (see later discussion). By evaluating $F(\alpha_1)$ and $F'(\alpha_1)$ from Equations 3.160 and 3.163 then:

$$h_1 = - F(\alpha_1)/F'(\alpha_1) \qquad 3.164$$

and

$$\alpha_2 = \alpha_1 + h_1 \qquad 3.165$$

This procedure is repeated until a sufficiently small value of $F(\alpha_j)$ is obtained when β can be obtained from Equation 3.157. In most cases only 3 or 4 successive steps will be required.

Samuelsson (34) has described a similar procedure using different first estimates of α and β and suggests that the true parameter values can be estimated to within 1% in only 3 iterations.

Computer programs have been written to carry out the maximum likelihood estimation of α and β and calculate event magnitudes at different return periods. As an example a sample sheet of the output of a program by Cuthbert and Latham (11) is included here as Table 3.11.

Other methods of deriving the maximum likelihood parameter estimates are available (16), (46) but these are generally more complex. Leese (27) has described the modifications to the maximum likelihood equations which are needed to accomodate missing data and historic flood records.

The method of moments computes estimates of the population mean and standard deviation as:

$$\mu_x = \sum_{i=1}^{n} x_i/n \qquad 3.166$$

and

$$\sigma_x = \sum_{i=1}^{n} (x_i-\bar{x})^2/(n-1) \qquad 3.167$$

Now, from Equation 3.142

$$x = \beta + y/\alpha \qquad 3.168$$

and Gumbel (17) has shown that the mean, μ_y and standard deviation, σ_y, of the reduced variable, y, are given by:

$$\mu_y = \gamma \qquad 3.169$$

$$\sigma_y = \frac{\pi}{\sqrt{6}} \qquad 3.170$$

where γ is Euler's constant, approximately 0.5772157.

Expressing Equation 3.168 in terms of means and variances yields:

$$\alpha = \pi/\sigma_x\sqrt{6} \qquad 3.171$$

and

$$\beta = \mu_x - \chi_x\sqrt{6}\ \gamma/\pi \qquad 3.172$$

A simple approximation has been used by Verma and Advani (43) to estimate the parameters α and β. If x_{max} is the largest event in a series of n maxima and x_{min} is the smallest event, then the reduced events y_{max} and y_{min} are defined as

$$y_{max} = \alpha(x_{max} - \beta) \qquad 3.173$$

and

$$y_{min} = \alpha(x_{min} - \beta) \qquad 3.174$$

By taking the probability of exceedence of the largest event, x_{max}, as 1/n and the probability of exceedence of the smallest event, x_{min}, as 1/1.01, then the following expressions apply

$$y_{max} = -\ln(-\ln(1-1/n)) \qquad 3.175$$

and

$$y_{min} = -\ln(-\ln(1-1/1.01)) \qquad 3.176$$

Evaluation of Equations 3.175 and 3.176 and substitution into Equations 3.173 and 3.174 will, by simultaneous solution, produce rough estimates of α and β. By using the expansion of ln and neglecting terms above second order, Verma and Advani (43) have produced expressions for α and β.

Yevjevich (47) has described a process by which estimates of α and β can be determined graphically. If, in Equation 3.143, $x = \beta$, then

$$P(\beta) = e^{-1} = 0.368 \qquad 3.177$$

Table 3.11

Example of Maximum Likelihood Estimation of Parameters of Type I Extremal (Gumbel) Distribution

Example: 01AD002 St. John River at Fort Kent

No. of Trial	j^{th} Estimate of α	$F(\alpha)$
1	.00005213	-8255.38307212
2	.00004692	-2829.58839064
3	.00004222	7157.46947343
4	.00004471	1075.79197549
5	.00004523	38.09131435
6	.00004525	.05290183
7	.00004525	.00000012
8	.00004525	-.00000000

Final value for β is 60146.236146

Plot the event magnitudes, x, versus return period, $T = (n + 1)/m$, on graph paper with a double exponential scale and fit a straight line through the plotted points. By entering the graph at $T = 1/P(\beta) = 2.717$ the value of β is determined. The slope of the best fitting straight line is then equal to $1/\alpha$. This method is very easy to apply but its accuracy is not to be compared with the method of maximum likelihood.

Lowery and Nash (28) have compared various methods of estimating the parameters of the Type I extremal distribution. They recognise the greater efficiency of the maximum likelihood technique but recommend moments because of the methods simplicity and lack of bias.

Frequency Factor

From the cumulative probability distribution (Equation 3.141) the expression relating the reduced variable, y, to return period, T, is

$$y = -\ln(-\ln((T-1)/T)) \qquad 3.178$$

For convenience, Table 3.12 gives values of the reduced variable y for some common return periods.

Table 3.12

Values of the Reduced Variable, y, of the Type I Extremal Distribution for Some Commonly Used Return Periods, T

Return Period T	Reduced Variable y
2	0.3665
5	1.4999
10	2.2504
20	2.9702
50	3.9019
100	4.6001

If the n recorded events are placed in order of magnitude so that m = 1 for the largest event and m = n for the smallest event then T = (n+1)/m and Equation 3.178 can be written as

$$y_m = -\ln[-\ln\{(n+1-m)/(n+1)\}] \qquad 3.179$$

If the mean, μ_y, and the variance, σ_y^2, of the series y_m, m = 1,2...n, are computed from the reduced sample as:

$$\mu_y = \sum_{m=1}^{n} y_m/n \qquad 3.180$$

and

$$\sigma_y^2 = \sum_{m=1}^{n} (y_m-\bar{y})^2/n \qquad 3.181$$

and if μ_x and σ_x^2 are the mean and variance of the recorded maximum events defined in Equations 3.166 and 3.167, then the parameters α and β can be defined (23) as:

$$\alpha = \sigma_y/\sigma_x \qquad 3.182$$

and

$$\beta = \mu_x-\mu_y/\alpha \qquad 3.183$$

Now, introducing the relationship between x and y,

$$y = \alpha(x-\beta) \qquad 3.184$$

substituting Equations 3.182 and 3.183 for α and β and rearranging for x, gives:

$$x = \mu_x + (y-\mu_y)\ \sigma_x/\sigma_y \qquad 3.185$$

Comparing Equation 3.185 with the general frequency equation (e.g. Equation 3.12) it is apparent that for the Type I extremal distribution the frequency factor, K, is defined as:

$$K = \frac{y_m-\mu_y}{\sigma_y} \qquad 3.186$$

Since μ_y and σ_y are functions of the sample size only, they can be tabulated. Table 3.13 is an example of this type of table. Alternatively, for a predetermined set of return periods and sample sizes the frequency factor, K, can be tabulated (23) as in Table 3.14. Coulson (9) includes a more comprehensive table.

As an example, for a sample size of 55, Table 3.13 gives values of μ_y and σ_y of 0.5504 and 1.1681 respectively. For a 100 year return period, Table 3.12 or Equation 3.178 gives the reduced variable, y_m, as 4.6001 so that, from Equation 3.186, the frequency factor, K, is 3.4670. Alternatively this figure can be found directly from Table 3.14. To determine the 100 year event magnitude it is then only necessary to estimate the population mean, μ_x and standard deviation, σ_x, from the 55 recorded events and substitute, μ_x, K and σ_x in the general frequency equation (e.g. Equation 3.12).

Table 3.13

Mean and Standard Deviation of Order Statistics, m/(n+1), for Various Sample Sizes, n

Sample Size n	Mean μ_y	Standard Deviation σ_y
10	0.4952	0.9496
15	0.5128	1.0206
20	0.5236	1.0628
25	0.5309	1.0914
30	0.5362	1.1124
35	0.5403	1.1285
40	0.5436	1.1413
45	0.5463	1.1518
50	0.5485	1.1607
55	0.5504	1.1682
60	0.5521	1.1747
65	0.5535	1.1803
70	0.5548	1.1854
75	0.5559	1.1898
80	0.5569	1.1938
85	0.5578	1.1974
90	0.5586	1.2007
95	0.5593	1.2037
100	0.5600	1.2065

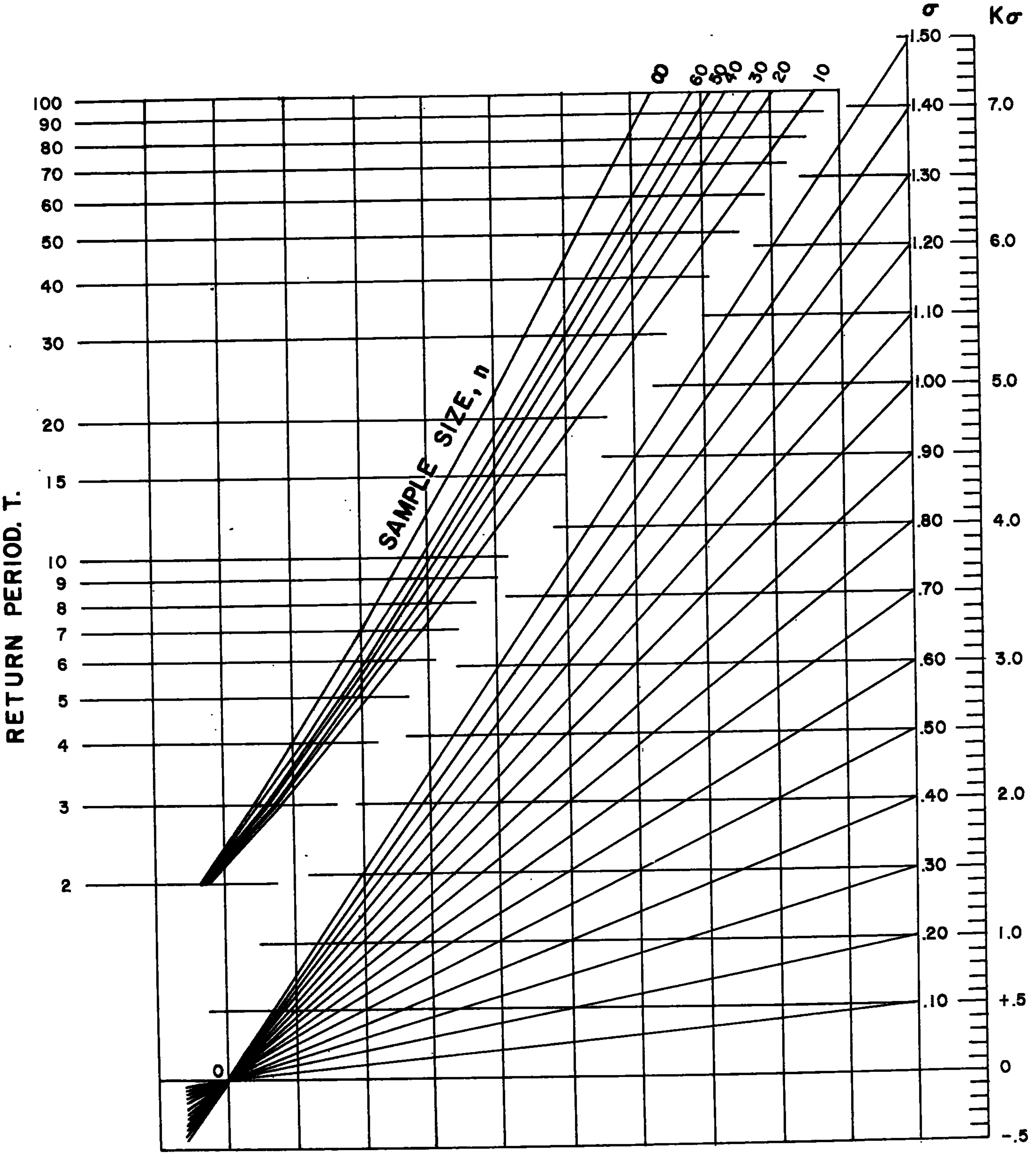
Table 3.14

Frequency Factor for Type I
Extremal Distribution

Sample Size	Cumulative Probability, P, %					
n	50	80	90	95	98	99
	Corresponding Return Period, T, Years					
	2	5	10	20	50	100
10	-0.1355	1.0580	1.8483	2.6063	3.5874	4.3227
15	-0.1434	0.9672	1.7025	2.4078	3.3208	4.0049
20	-0.1478	0.9187	1.6248	2.3020	3.1787	3.8356
25	-0.1506	0.8879	1.5754	2.2350	3.0886	3.7284
30	-0.1526	0.8664	1.5410	2.1881	3.0257	3.6534
35	-0.k540	0.8504	1.5154	2.1532	2.9789	3.5976
40	-0.1552	0.8379	1.4954	2.1261	2.9425	3.5543
45	-0.1561	0.8279	1.4794	2.1044	2.9133	3.5195
50	-0.1568	0.8197	1.4663	2.0865	2.8892	3.4908
55	-0.1574	0.8128	1.4552	2.0714	2.8690	3.4667
60	-0.1580	0.8069	1.4458	2.0586	2.8518	3.4461
65	-0.1584	0.8018	1.4376	2.0475	2.8368	3.4284
70	-0.1588	0.7974	1.4305	2.0377	2.8238	3.4128
75	-0.1592	0.7934	1.4242	2.0291	2.8122	3.3991
80	-0.1595	0.7900	1.4185	2.0215	2.8020	3.3868
85	-0.1597	0.7868	1.4135	2.0146	2.7928	3.3758
90	-0..1600	0.7840	1.4090	2.0084	2.7844	3.3659
95	-0.1602	0.7814	1.4048	2.0028	2.7769	3.3569
100	-0.1604	0.7791	1.4011	1.9977	2.7700	3.3487
Corresponding Frequency Factors from Equation 3.188						
	-0.1643	0.7194	1.3046	1.8658	3.5923	3.1667

Weiss (45) has devised a convenient nomogram for performing graphically the solution of Equation 3.185 given the sample mean and standard deviation. Shown here as Figure 3.4 this nomogram is entered on the left hand side with the required return period, T. From the intersection of the horizontal line through T with the slanting line through the appropriate sample size, n, draw a vertical line to intersect the sloping line corresponding to the sample standard deviation, σ. From this second intersection draw a horizontal line to cut the right hand edge of the diagram at the value of Kσ, the numerical value to be added to the mean, μ_x, to give the required event magnitude, x. Examples of the use of this nomogram are given in Weiss (45) and Kendall (23).

FIGURE 3.4
NOMOGRAM FOR USE WITH TYPE I EXTREMAL DISTRIBUTION[1]



[1] WEISS (45)

Chow (5) has considered the Type I extremal distribution as a special case of the lognormal distribution for which the coefficient of skew, γ_1, is a constant at 1.139. Following this procedure, substitution of α and β from Equations 3.17, and 3.172 into Equation 3.145, relating x to return period, yields:

$$x = \mu - \frac{\sqrt{6}}{\pi}\{\gamma + \ln[-\ln((T-1)/T)]\}\sigma \qquad 3.187$$

Comparison of this equation with the standard frequency equation gives the following expression for the frequency factor of the Type I extremal distribution:

$$K = -\frac{\sqrt{6}}{\pi}\{\gamma + \ln[-\ln((T-1)/T)]\} \qquad 3.188$$

The last line of Table 3.14 gives the values of the frequency factor determined from this equation. These values correspond to the asymptotic results of using Equation 3.186 as n → ∞.

Standard Error of Estimate

The general equation for the standard error of estimate from the method of moments:

$$S^2(K) = \frac{\mu_2}{n}\left[1 + K\frac{\mu_3}{\mu_2^{3/2}} + \frac{K^2(\mu_4 - \mu_2^2)}{4\mu_2^2}\right] \qquad 3.189$$

has been developed earlier in the chapter.

Gumbel (16) gives the central moments of the Type I extremal distribution as:

$$\mu_2 = \pi^2/6 = 1.6449 \qquad 3.190$$

$$\mu_3 = 2.4041 \qquad 3.191$$

$$\mu_4 = 3\pi^2/20 = 14.6114 \qquad 3.192$$

Substituting these constants in Equation 3.189 the following expression results:

$$S^2(K) = \frac{\sigma_x^2}{n} [1+1.1396K+1.1000K^2] \qquad 3.193$$

Taking the square root and simplifying to:

$$S(K) = \delta \ \sigma_x/\sqrt{n} \qquad 3.194$$

values of δ depend only on K and can thus be predetermined and tabulated for typical values of return period, T, and sample size, n. Table 3.15 is a sample of this type of table. Kaczmarek (21), Kendall (23) and Coulson (9) have given similar tables, although as pointed out by Lowery and Nash (28) there is a computational error in the table of Kaczmarek.

Using the same example as before, the 100 year return period event computed from a 55 year sample will have an δ value of 4.265. Knowing the standard deviation of the recorded events, σ_x, the standard error, S(K), is computed from Equation 3.194 and the 95% confidence limits are applied as $x(K) \pm 1.96\ S(K)$

Nash and Amorocho (32) have used the basic standard error equation

$$S^2(K) = \text{var}\ \mu + K^2\text{var}\ \sigma + 2K\text{cov}(\mu,\sigma) \qquad 3.195$$

in an experimental approach. Using the relationship,

$$\text{cov}(\mu,\sigma) = \rho\ (\text{var}\ \mu)^{1/2}\ (\text{var}\ \sigma)^{1/2} \qquad 3.196$$

where ρ is the sample linear correlation coefficient between μ and σ

Equation 3.195 becomes:

$$S^2(K) = \text{var}\ \mu + K^2\text{var}\ \sigma + 2k(\text{var}\ \mu)^{1/2}\ (\text{var}\ \sigma)^{1/2} \qquad 3.197$$

Table 3.15

Parameter φ for Use in Standard Error of Type I Extremal Distribution

Sample Size n	Cumulative Probability, P, %					
	50	80	90	95	98	99
	Corresponding Return Period, T, Years					
	2	5	10	20	50	100
10	0.9305	1.8539	2.6199	3.3826	4.3869	5.1459
15	0.9269	1.7695	2.4756	3.1814	4.1127	4.8174
20	0.9250	1.7249	2.3990	3.0745	3.9670	4.6427
25	0.9238	1.6968	2.3506	3.0069	3.8747	4.5320
30	0.9229	1.6772	2.3169	2.9597	3.8103	4.4548
35	0.9223	1.6627	2.2919	2.9247	3.7624	4.3974
40	0.9218	1.6514	2.2725	2.8975	3.7252	4.3527
45	0.9214	1.6424	2.2569	2.8756	3.6954	4.3169
50	0.9211	1.6350	2.2441	2.8577	3.6708	4.2874
55	0.9208	1.6288	2.2333	2.8426	3.6502	4.2627
60	0.9206	1.6235	2.2241	2.8297	3.6326	4.2415
65	0.9204	1.6189	2.2162	2.8186	3.6173	4.2232
70	0.9202	1.6149	2.2093	2.8089	3.6040	4.2075
75	0.9201	1.6114	2.2032	2.8003	3.5923	4.1951
80	0.9199	1.6083	2.1977	2.7926	3.5818	4.1806
85	0.9198	1.6055	2.1928	2.7858	3.5724	4.1693
90	0.9197	1.6030	2.1884	2.7796	3.5639	4.1591
95	0.9196	1.6007	2.1844	2.7739	3.5562	4.1498
100	0.9195	1.5986	2.1808	2.7688	3.5491	4.1414

Now, var μ is distribution-free at σ^2/n and Nash and Amorocho (32) give var σ as $1.1\sigma^2/n$ so that Equation 3.197 reduces to:

$$S^2(K) = \frac{\sigma^2}{n} + \frac{1.1\ K^2\sigma^2}{n} + \frac{2.1\ K\rho\sigma^2}{n} \qquad 3.198$$

Nash and Amorocho (32) have shown experimentally that ρ in Equation 3.180 is independent of n and has a mean value of 0.56 and standard deviation of approximately 0.02. Substituting the mean value of ρ in Equation 3.198 yields:

$$S^2(K) = \frac{\sigma^2}{n}\ [1+1.18K+1.1K^2] \qquad 3.199$$

which is almost identical to the theoretical relationship given in Equation 3.193.

Dalrymple (12) has given the following equation for the standard error of the reduced variate, y, of the Type I extremal distribution

$$S' = \frac{e^y}{\sqrt{n}} \sqrt{\frac{1}{T - 1}} \qquad 3.200$$

where y, the reduced variable, is given by Equation 3.142. The confidence interval for the reduced variable is then computed as

$$y_T \pm t\ S' \qquad 3.201$$

Dalrymple (12) used this confidence region to determine the homogeneous hydrologic region in his index flood method of regional flood frequency analysis (see Chapter 4.)

Panchang (33) has given the maximum likelihood estimate of the standard error of the Type I extremal distribution as:

$$S'' = \frac{1}{\alpha\sqrt{n}} \left[1+\frac{6}{\pi^2} \left(1-e-\ln\left(\ln\frac{T}{T-1}\right)\right)\right]^{\frac{1}{2}} \qquad 3.202$$

3.3.4.2 Type II Extremal

The Type II extremal distribution is derived as the logarithmic transformation of the Type I extremal distribution (16). In hydrology it is known as the log-Gumbel distribution and is used by assuming that the logarithms of the recorded events follow a Type I extremal distribution (3), (40).

Since the distribution is a logarithmic transformation, the magnitude and standard error of the T-year event can be computed either from the K and δ values for the Type I extremal distribution together with the mean and standard deviation of the logarithms of the recorded events or from the mean and standard deviation of the recorded events together with analytically derived values of K and δ for the Type II extremal distribution. Since this analysis is complex, the first method is to be preferred.

3.3.4 Type III Extremal

General

The Type III, or Weibull, distribution results from an initial distribution in which an upper boundary applies. The distribution is commonly used in hydrology for drought analysis (4).

The cumulative probability distribution is given (6) as:

$$P(x_1 \leq x) = 1-e^{-\{\frac{x-\gamma}{\beta-\gamma}\}^{\alpha}} \qquad 3.203$$

and the probability density is:

$$p(x) = \frac{\alpha}{\beta-\gamma}\{\frac{x-\gamma}{\beta-\gamma}\}^{\alpha-1} e^{-\{\frac{x-\gamma}{\beta-\gamma}\}^{\alpha}} \qquad 3.204$$

where α is a scale parameter equal to the order of the lowest derivative of the probability function that is not zero at $x = \gamma$, β is the location or central value parameter and γ is the lower limit to x.

Commonly, the tranformation

$$y = \{\frac{x-\gamma}{\beta-\gamma}\}^{\alpha} \qquad 3.205$$

is made (17) reducing the cumulative probability and probability density equations to

$$P(x) = 1-e^{-y} \qquad 3.206$$

and

$$p(x) = \frac{\alpha}{\beta-\gamma} y^{(\alpha-1)/\alpha} e^{-y} \qquad 3.207$$

The notation used so far in this chapter for distributions of flood events has been

$$P(x) = 1-m/n+1 = 1-1/T \qquad 3.208$$

where P(x) is the cumulative probability of an event being less than or equal to x and m is the order number of the recorded event, m being 1 for the maximum event and m being n for the minimum event. Following this notation, the larger the return period, T, the larger is the magnitude of the expected event.

In the analysis of droughts, however, it is required that smaller events be associated with larger return periods and so a different notation is commonly used (4). If the recorded events are arranged in order of increasing magnitude with m being 1 for the minimum event and m being n for the maximum event then the cumulative probability of an event being less than or equal to x is given by

$$P(x) = m/n+1 = 1/T \qquad 3.209$$

This convention will be used for the remaining discussion of the Type III extremal distribution.

Estimation of Parameters

For non-negative variables the calculation of the population moment μ_r of order r about the origin may be made with the equation

$$\mu_r = -\int_0^\alpha x^r \, d[1-P(x)] \qquad 3.210$$

Substituting variables for the reduced Type III extremal distribution gives the equation (16):

$$\overline{\left\{\frac{x-\gamma}{\beta-\gamma}\right\}}_r = -\int_\gamma^\alpha \left\{\frac{x-\gamma}{\beta-\gamma}\right\}^r d\, e^{-\left\{\frac{x-\gamma}{\beta\ \gamma}\right\}^\alpha} \qquad 3.211$$

or

$$\overline{\left\{\frac{x-\gamma}{\beta-\gamma}\right\}}_r = -\int_0^\alpha y^{r/\alpha} \, d\, e^{-y} \qquad 3.212$$

Now, introducing the Gamma function,

$$\Gamma(n) = \int_0^\alpha e^{-x} x^{n-1} dx \qquad 3.213$$

Equation 3.212 reduces to

$$\overline{\left\{\frac{x-\gamma}{\beta-\gamma}\right\}}_r = \Gamma\left(1+\frac{r}{\alpha}\right) \qquad 3.214$$

From which the first four moments, μ_1 to μ_4 are derived as

$$\mu_1 = \mu = \gamma+(\beta-\gamma)\Gamma(1+1/\alpha) \qquad 3.215$$

$$\mu_2 = \sigma^2 = (\beta-\gamma)^2 \{\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)\} \qquad 3.216$$

$$\mu_3 = (\beta-\gamma)^3 \{\Gamma(1+3/\alpha) -3\Gamma(1+2/\alpha)\Gamma(1+1/\alpha) + 2\Gamma^3(1+1/\alpha)\} \qquad 3.217$$

and

$$\mu_4 = (\beta-\gamma)^4 \{\Gamma(1+4/\alpha) -4\Gamma(1+3/\alpha)\Gamma(1+1/\alpha) +$$

$$6\Gamma(1+2/\alpha)\Gamma^2(1+1/\alpha) -3\Gamma^4(1+1/\alpha)\} \qquad 3.218$$

If two new variables are defined, A_α and B_α, such that A_α is the standardised difference between the characteristic value and the mean and B_α is the standardised difference between the lower limit and the characteristic value,

$$A_\alpha = \frac{\beta-\mu}{\sigma} \qquad 3.219$$

and

$$B_\alpha = \frac{\beta-\gamma}{\sigma} \qquad 3.220$$

then, by substituting μ and σ from Equations 3.215 and 3.216

$$B_\alpha = \{\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)\}^{-1/2} \qquad 3.221$$

and

$$A_\alpha = \{1-\Gamma(1+1/\alpha)\} B_\alpha \qquad 3.222$$

If the coefficient of skew, γ_1, is defined as usual

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \qquad 3.223$$

then from Equations 3.215, 3.216 and 3.220:

$$\gamma_1 = \{\Gamma(1+3/\alpha) - 3\Gamma(1+2/\alpha)\Gamma(1+1/\alpha) + 2\Gamma^3(1+1/\alpha)\} \ B_\alpha^3 \qquad 3.224$$

an expression involving only functions of α.

Thus if the sample coefficient of skew is computed as:

$$\hat{\gamma}_1 = \frac{n\ \Sigma\ (x-\bar{x})^3}{(n-2)\,[\Sigma(x-\bar{x})^2]^{3/2}} \qquad 3.225$$

then α can be found by the solution of Equation 3.224. Knowing $B\alpha$, the parameter β can be obtained from Equation 3.219 and subsequently γ can be found from Equation 3.220.

To solve Equation 3.224 tables are available, (16), (4), relating α (generally as $1/\alpha$) to γ_1, A_α and B_α. These tables are usually arranged in incremental steps of $1/\alpha$ so that for a computed sample skew a great deal of interpolation is needed to determine the corresponding values of $1/\alpha$, A_α and B_α. In order to avoid this interpolation the following regression equation has been developed to enable α to be calculated directly from γ_1:

$$\alpha = 1/[a_1+a_2\gamma_1+a_3\gamma_1^2+a_4\gamma_1^3+a_5\gamma_1^4] \qquad 3.226$$

$a_1 = 0.2777757913,$ $\qquad a_4 = -0.0013038566$

$a_2 = 0.3132617714,$ $\qquad a_5 = -0.0081523408$

$a_3 = 0.0575670910,$

This polynomial is valid for a range of γ_1 from -1.02 to +2.00, has a multiple correlation coefficient of 0.9999 and a standard error of 0.0006575. Table 3.16 has been derived from this equation. For convenience Table 3.16 also gives values of the parameters A_α and B_α as computed from Equations 3.222 and 3.221.

Table 3.16

Parameter α for Type III Extremal Distribution Tabulated as a Function of the Sample Coefficient of Skewness, γ_1

γ_1	α	A_α	B_α
-1.00	65.63043	0.44760	52.24465
-0.90	26.26360	0.44229	21.47978
-0.80	16.30207	0.43629	13.68443
-0.70	11.73785	0.42952	10.10381
-0.60	9.10978	0.42193	8.03409
-0.50	7.39676	0.41343	6.67757
-0.40	6.18962	0.40397	5.71462
-0.30	5.29236	0.39350	4.99218
-0.20	4.59923	0.38198	4.42770
-0.10	4.04809	0.36938	3.97273
0.00	3.59997	0.35571	3.59692
0.10	3.22914	0.34098	3.28029
0.20	2.91791	0.32523	3.00911
0.30	2.65366	0.30851	2.77366
0.40	2.42717	0.29089	2.56682
0.50	2.23149	0.27246	2.38329
0.60	2.06133	0.25334	2.21910
0.70	1.91253	0.23367	2.07116
0.80	1.78181	0.21360	1.93718
0.90	1.66654	0.19329	1.81524
1.00	1.56457	0.17291	1.70391
1.10	1.47416	0.15265	1.60204
1.20	1.39386	0.13268	1.50873
1.30	1.32247	0.11318	1.42324
1.40	1.25900	0.09432	1.34501
1.50	1.20261	0.07626	1.27360
1.60	1.15260	0.05914	1.20866
1.70	1.10840	0.04311	1.14991
1.80	1.06954	0.02828	1.09714
1.90	1.03562	0.01477	1.05020
2.00	1.00634	0.00268	1.00900

If the parameter γ can be assumed to be zero, then a much simpler, if less accurate, method of estimating the remaining parameters, α and β, exists. If $x = \beta$ then, in Equation 3,203

$$P(\beta) = 1-e^{-1} = 0.632 \qquad 3.227$$

The median of the Type III extremal distribution, M, is obtained by substituting P = 0.50 in Equation 3.203,

$$M = \beta(\ln 2)^{1/\alpha} \qquad 3.228$$

If, therefore, a graph of original variable, x, versus cumulative probability is drawn, where cumulative probability is estimated from the sample using Equation 3.209, then values of β and M can be extracted from the graph at P = 0.632 and P = 0.50. Substitution of β and M in Equation 3.228 will then yield the sample estimate of α.

Other methods of estimating the parameters are available such as the use of an order statistic and the use of the smallest observed event, but none of these methods of estimating parameters is foolproof. The results are only acceptable if the computed value of the parameter γ lies between 0 and x_n, where x_n is the smallest observed event.

Deininger and Westfield (13) have compared the results of several methods, and recommend a conbination of least squares and Fibonacci search. This method uses an initial estimate of γ to compute values of the characteristic event, β, and the scale parameter, α, by least squares. A second estimate of γ, $0 < \gamma \leq x_n$, is then chosen and the procedure repeated. This is continued until the minimum sum of squared deviations between observed and computed events is obtained. The optimal technique for searching the interval 0 to x_n is to use Fibonacci numbers (13) to choose the sample points.

Frequency Factor

From Equations 3.205, 3.206 and 3.208 the following expressions are derived for the Type III extremal distribution:

$$x = \gamma + y^{1/\alpha}(\beta - \gamma) \qquad 3.229$$

and

$$y = -\ln(1-1/T) \qquad 3.230$$

Values of y, the reduced variable, are given in Table 3.17 for some commonly used return periods.

Table 3.17

Values of the Reduced Variable, y, of the Type III Extremal Distribution for some Commonly Used Return Periods, T

Return Period T	Reduced Variable y
2	0.69315
5	0.22314
10	0.10536
20	0.05129
50	0.02020
100	0.01005

Knowing α, β and γ, and obtaining the reduced variable, y, from Table 3.17, the event magnitude, x, corresponding to return period, T, can be computed using Equation 3.229.

If Equations 3.229 and 3.230 are combined, then the following expression results:

$$x = \gamma + [-\ln(1-1/T)]^{1/\alpha} \ (\beta - \gamma) \qquad 3.231$$

But from previous developments,

$$\beta - \gamma = B_\alpha \sigma \qquad 3.232$$

and

$$\gamma = \mu - \sigma\ (B_\alpha - A_\alpha) \qquad 3.233$$

where μ and σ are the mean and standard deviation of the population event magnitudes as estimated from the sample, and A_α and B_α are as defined in Equations 3.222 and 3.221.

Substituting Equations 3.232 and 3.233 into Equation 3.231:

$$x = \mu + \sigma\ \{(A_\alpha - B_\alpha) + B_\alpha[-\ln(1-1/T)]^{1/\alpha}\} \qquad 3.234$$

Comparing Equation 3.234 with the standard frequency equation it is apparent that K, the frequency factor, is given by

$$K = A_\alpha + B_\alpha\{[-\ln(1-1/T)]^{1/\alpha} - 1\} \qquad 3.235$$

This expression is dependent only upon the return period, T, and the coefficient of skew, γ_1, of the recorded events. Table 3.18 provides values of K for some typical values of T and γ_1.

The procedure to be followed is therefore to compute the mean, standard deviation and coefficient of skew from the sample data, look up the value of K for the required return period in Table 3.18 and compute the corresponding event magnitude, x, from the general frequency equation.

Table 3.18

Frequency Factor for Use in Type III Extremal Distribution

Coefficient of skew	Cumulative Probability, P, %					
	50	80	90	95	98	99
γ_1	Corresponding Return Period, T, Years					
	2	5	10	20	50	100
-1.00	0.1567	-0.7329	-1.3134	-1.8641	-2.5680	-3.0889
-0.90	0.1446	-0.7501	-1.3215	-1.8546	-2.5232	-3.0089
-0.80	0.1321	-0.7666	-1.3282	-1.8430	-2.4766	-2.9282
-0.70	0.1189	-0.7825	-1.3332	-1.8294	-2.4280	-2.8465
-0.60	0.1051	-0.7977	-1.3366	-1.8134	-2.3771	-2.7634
-0.50	0.0906	-0.8122	-1.3382	-1.7950	-2.3239	-2.6788
-0.40	0.0754	-0.8258	-1.3379	-1.7741	-2.2683	-2.5928
-0.30	0.0595	-0.8385	-1.3356	-1.7506	-2.2103	-2.5055
-0.20	0.0428	-0.8502	-1.3313	-1.7245	-2.1502	-2.4172
-0.10	0.0255	-0.8607	-1.3248	-1.6960	-2.0881	-2.3282
0.00	0.0075	-0.8699	-1.3161	-1.6650	-2.0244	-2.2390
0.10	-0.0110	-0.8778	-1.3053	-1.6318	-1.9595	-2.1500
0.20	-0.0300	-0.8842	-1.2923	-1.5966	-1.8938	-2.0619
0.30	-0.0493	-0.8891	-1.2773	-1.5595	-1.8277	-1.9752
0.40	-0.0689	-0.8923	-1.2603	-1.5210	-1.7616	-1.8902
0.50	-0.0885	-0.8939	-1.2415	-1.4812	-1.6961	-1.8075
0.60	-0.1081	-0.8938	-1.2209	-1.4405	-1.6315	-1.7275
0.70	-0.1275	-0.8921	-1.1989	-1.3992	-1.5682	-1.6506
0.80	-0.1466	-0.8888	-1.1757	-1.3578	-1.5068	-1.5770
0.90	-0.1651	-0.8840	-1.1515	-1.3165	-1.4473	-1.5071
1.00	-0.1829	-0.8777	-1.1266	-1.2757	-1.3903	-1.4409
1.10	-0.2000	-0.8703	-1.1013	-1.2358	-1.3359	-1.3787
1.20	-0.2162	-0.8617	-1.0758	-1.1969	-1.2842	-1.3204
1.30	-0.2313	-0.8522	-1.0505	-1.1594	-1.2356	-1.2661
1.40	-0.2454	-0.8421	-1.0255	-1.1236	-1.1901	-1.2159
1.50	-0.2583	-0.8314	-1.0013	-1.0896	-1.1477	-1.1696
1.60	-0.2701	-0.8206	-0.9780	-1.0577	-1.1086	-1.1272
1.70	-0.2807	-0.8097	-0.9558	-1.0279	-1.0728	-1.0887
1.80	-0.2900	-0.7990	-0.9351	-1.0006	-1.0403	-1.0540
1.90	-0.2983	-0.7887	-0.9159	-0.9758	-1.0112	-1.0231
2.00	-0.3053	-0.7790	-0.8985	-0.9536	-0.9854	-0.9959

Standard Error of Estimate

The general expression for the standard error of estimate by the method of moments has been given as:

$$S^2(K) = \frac{\mu_2}{n} \left[1 + \frac{K\mu_3}{\mu_2^{3/2}} + \frac{K^2(\mu_4 - \mu_2^2)}{4\mu_2^2} \right] \qquad 3.236$$

Substituting for the second, third and fourth moments from Equations 3.216, 3.217 and 3.218 and using the simplifications

$$x = \{\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)\} \qquad 3.237$$

$$y = \{\Gamma(1+3/\alpha) - 3\Gamma(1+2/\alpha)\Gamma(1+1/\alpha) + 2\Gamma^3(1+1/\alpha)\} \qquad 3.238$$

$$z = \{\Gamma(1+4/\alpha) - 4\Gamma(1+3/\alpha)\Gamma(1+1/\alpha) +$$

$$6\Gamma(1+2/\alpha)\Gamma^2(1+1/\alpha) - 3\Gamma^4(1+1/\alpha)\} \qquad 3.239$$

then the standard error of estimate can be given by:

$$S^2(K) = \frac{\sigma^2}{n} \left[1 + \frac{K.y}{x^{3/2}} + \frac{K^2(z-x^2)}{4x^2}\right] \qquad 3.240$$

Table 3.19 provides values of δ in the equation

$$S(K) = \delta \frac{\sigma}{\sqrt{n}} \qquad 3,241$$

for some commonly used values of T, the return period, and γ_1, the coefficient of skew estimated from the sample. Confidence limits can then be computed assuming the normality of x(K).

Table 3.19

Values of Parameter δ for Use in Standard Error of Type III Extremal Distribution

Coefficient	Cumulative Probability, P, %					
of	50	80	90	95	98	99
Skew	Corresponding Return Period, T, Years					
γ_1	2	5	10	20	50	100
-1.00	0.9184	1.3162	1.5208	1.6921	1.8885	2.0217
-0.90	0.9325	1.2949	1.4807	1.6351	1.8102	1.9275
-0.80	0.9455	1.2709	1.4373	1.5745	1.7284	1.8303
-0.70	0.9574	1.2443	1.3907	1.5106	1.6436	1.7305
-0.60	0.9681	1.2153	1.3414	1.4438	1.5563	1.6289
-0.50	0.9773	1.1842	1.2896	1.3747	1.4671	1.5260
-0.40	0.9851	1.1512	1.2358	1.3036	1.3765	1.4224
-0.30	0.9912	1.1164	1.1801	1.2310	1.2851	1.3187
-0.20	0.9958	1.0798	1.1228	1.1569	1.1929	1.2150
-0.10	0.9988	1.0413	1.0635	1.0811	1.0996	1.1110
0.00	1.0000	1.0007	1.0018	1.0031	1.0046	1.0057
0.10	0.9995	0.9576	0.9373	0.9221	0.9069	0.8982
0.20	0.9971	0.9117	0.8695	0.8374	0.8053	0.7869
0.30	0.9927	0.8631	0.7984	0.7487	0.6991	0.6706
0.40	0.9864	0.8116	0.7231	0.6546	0.5855	0.5454
0.50	0.9781	0.7571	0.6434	0.5534	0.4597	0.4033
0.60	0.9676	0.6995	0.5580	0.4411	0.3078	0.2125
0.70	0.9551	0.6389	0.4653	0.3077	0.0786	0.2338
0.80	0.9406	0.5752	0.3611	0.0845	0.3024	0.3705
0.90	0.9240	0.5092	0.2352	0.2564	0.3990	0.4485
1.00	0.9057	0.4417	0.0828	0.3497	0.4563	0.4953
1.10	0.8858	0.3750	0.2319	0.4005	0.4865	0.5181
1.20	0.8648	0.3143	0.2867	0.4206	0.4918	0.5176
1.30	0.8431	0.2694	0.2980	0.4108	0.4701	0.4910
1.40	0.8215	0.2562	0.2644	0.3639	0.4133	0.4299
1.50	0.8009	0.2859	0.1470	0.2531	0.2963	0.3096
1.60	0.7825	0.3531	0.2447	0.1993	0.1770	0.1706
1.70	0.7676	0.4447	0.4216	0.4265	0.4350	0.4389
1.80	0.7577	0.5507	0.5777	0.6032	0.6221	0.6292
1.90	0.7537	0.6644	0.7259	0.7643	0.7894	0.7981
2.00	0.7564	0.7806	0.8680	0.9154	0.9446	0.9544

3.3.5 Pearson Type III

General

The probability density distribution of the Pearson Type III distribution is of the form

$$p(x) = \frac{1}{\alpha\Gamma(\beta)} \{\frac{x-\gamma}{\alpha}\}^{\beta-1} e^{-\{\frac{x-\gamma}{\alpha}\}} \qquad 3.242$$

where α, β and γ are parameters to be defined and $\Gamma(\beta)$ is, as before, the Gamma function.

If the substitution $y = \frac{x-\gamma}{\alpha}$ is made, then Equation 3.242 simplifies to

$$p(y) = \frac{y^{\beta-1} e^{-y}}{\Gamma(\beta)} \qquad 3.243$$

which is a one parameter Gamma distribution described in many statistics texts (e.g. (47)).

Estimation of Parameters

The likelihood function is set up, as usual, as the logarithm of the sum of the probability density distributions (31):

$$L = -n \ln\Gamma(\beta) - \frac{1}{\alpha} \sum_{i=1}^{n} (x_i - \gamma) + (\beta - 1) \sum_{i=1}^{n} \ln(x_i - \gamma) - n\beta \ln\alpha \qquad 3.244$$

Differentiating with respect to α, β and γ and equating to zero gives the following three equations (29):

$$\frac{\partial L}{\partial\alpha} = \frac{1}{\alpha^2} \sum_{i=1}^{n} (x_i - \gamma) - \frac{n\beta}{\alpha} = o \qquad 3.245$$

$$\frac{\partial L}{\partial \beta} = -n\, \Gamma'(\beta)/\Gamma(\beta) + \sum_{i=1}^{n} \ln(x_i-\gamma) - n\ln\alpha = o \qquad 3.246$$

$$\frac{\partial L}{\partial \gamma} = \frac{n}{\alpha} - (\beta-1) \sum_{i=1}^{n} (1/(x_i-\gamma)) = o \qquad 3.247$$

The maximum likelihood solution then depends on a simultaneous solution of these three equations. The Psi or Digamma function, $\Gamma'(\beta)/\Gamma(\beta)$ is given in many books of statistical tables e.g. Abramowitz and Stegun (1). Yevjevich (47) gives the solution of these equations as follows:

(a) First of all $\hat{\gamma}$ is found by trial and error as the solution of the equation:

$$\frac{1 + [1+4A/3]^{1/2}}{1 + [1+4A/3]^{1/2} - 4A} - \frac{(\bar{x} - \hat{\gamma})}{n} \sum_{i=1}^{n} \frac{1}{(x_i-\hat{\gamma})} = o \qquad 3.248$$

where

$$A = [\ln(\bar{x}-\hat{\gamma}) - \frac{1}{n} \sum_{i=1}^{n} \ln(x_i-\hat{\gamma})] \qquad 3.249$$

(b) Using the determined value of $\hat{\gamma}$, an estimate of β is found from

$$\hat{\beta} = \frac{1 + [1+4A/3]^{1/2}}{4A} \qquad 3.250$$

A correction factor, $\Delta\beta$, tabulated by Yevjevich (47) must then be subtracted from the computed value of β. Condie (8) has shown that the correction factor $\Delta\beta$ may be determined directly from β as:

$$\Delta\beta = 0.0014 + 0.0465 \exp(-1.7731\beta) \qquad 3.251$$

(c) An estimate of α is now found from:

$$\hat{\alpha} = \sum_{i=1}^{n} (x_i - \hat{\gamma})/n\hat{\beta} = (\bar{x} - \hat{\gamma})/\hat{\beta} \qquad 3.252$$

Matalas and Wallis (30) have noted that for samples exhibiting very small absolute values of skewness a solution by maximum likelihood may not be possible because of computer time constraints. Similarly if $\beta < 1$ then a maximum likelihood solution is not possible.

Greenwood and Durand (15) have also noted that for the general Pearson Type III distribution there are no sufficient estimators using maximum likelihood.

If, however, the value of γ can be derived from a source other than the data then maximum likelihood estimates of α and β are available. If the arithmetic mean of the reduced series x-γ is given by

$$A = \frac{1}{n} \sum_{i=1}^{n} (x-\gamma) \qquad 3.253$$

then Equation 3.245 reduces to

$$\alpha = A/\beta \qquad 3.254$$

Now if the geometric mean of the reduced series x-γ is given by

$$B = \left[\prod_{i=1}^{n} (x-\gamma) \right]^{1/n} \qquad 3.255$$

and if

$$C = \ln A - \ln B \qquad 3.256$$

then substitution of Equations 3.254 and 3.256 into Equation 3.246 and simplification yields

$$\ln \beta - \Gamma'(\beta)/\Gamma(\beta) = C \qquad 3.257$$

Solution of Equation 3.257 gives an estimate of β and substitution back into Equation 3.254 will result in an estimate of α. Solution of this equation is made easier by the availability of tables of C versus βC. Thus calculation of C from the reduced data and interpolation from a tables gives βC from which β can be quickly found. Greenwood and Durand (15) have also given the following polynomial approximations to estimate β directly from C:

for $0 \leq C \leq 0.5772$

$$\hat{\beta} = (0.5000876 + 0.1648852C - 0.0544274C^2)/C \qquad 3.258$$

for which the maximum error is 0.0088% and for $0.5772 \leq C \leq 17.0$

$$\hat{\beta} = \frac{8.898919 + 9.059950C + 0.9775373C^2}{C(17.79728 + 11.968477C + C^2)} \qquad 3.259$$

for which the maximum error is 0.0054%.

Similarly a trial and error procedure based on the likelihood equations has been described by Matalas and Wallis (30). Knowing γ, Equations 3.245 and 3.247 can be solved explicitly for α and β. First of all β is obtained from

$$\hat{\beta} = \sum_{i=1}^{n} \left(\frac{1}{x_i - \hat{\gamma}}\right) \Bigg/ \left[\sum_{i=1}^{n}\left(\frac{1}{x_i - \hat{\gamma}}\right) - \frac{n^2}{\sum_{i=1}^{n}(x_i - \hat{\gamma})}\right] \qquad 3.260$$

and than $\hat{\alpha}$ can be obtained from Equation 3.252. Substituting $\hat{\alpha}$, β and $\hat{\gamma}$ back into Equation 3.246 a value of ∂L/∂β is obtained. If $\left|\partial L/\partial\beta\right| > 10^{-8}$ the procedure should be repeated with a new value of $\hat{\gamma}$.

Estimating the parameters of the Pearson Type III distribution by the method of moments (29) utilises the following expressions for the first three central moments:

$$\mu_1 = \gamma + \alpha\beta \qquad 3.261$$

$$\mu_2 = \alpha^2\beta \qquad 3.262$$

$$\mu_3 = 2\alpha^3\beta \qquad 3.263$$

Since the fourth central moment will be used later in this section it is given here for convenience:

$$\mu_4 = 3\beta\alpha^4(\beta+2) \qquad 3.264$$

If the population mean, standard deviation and skewness are estimated from the sample and denoted by μ, σ and γ_1, then by using the relationship:

$$\gamma_1 = \frac{\mu 3}{\mu_2^{3/2}} \qquad 3.265$$

and substituting Equations 3.262 and 3.263 the parameter β can be estimated from:

$$\hat{\beta} = (2/\gamma_1)^2 \qquad 3.266$$

Once β is determined then by solution of Equations 3.261 and 3.262, α and γ can be estimated as

$$\hat{\alpha} = \sigma/\hat{\beta}^{1/2} \qquad 3.267$$

and

$$\hat{\gamma} = \mu - \sigma\hat{\beta}^{1/2} \qquad 3.268$$

Frequency Factor

The cumulative probability distribution of the Pearson Type III can be expressed as:

$$P(x) = \frac{1}{\alpha\Gamma(\beta)} \int_{0}^{x_0} e^{-\{\frac{x-\gamma}{\alpha}\}} \{\frac{x-\gamma}{\alpha}\}^{\beta-1} \quad dx \qquad 3.269$$

If α, β and γ are derived as in the previous section and the probability required for a given return period, T, is $P(x) = 1 - \frac{1}{T}$ then to define the event magnitude, Equation 3.269 must be solved for x_o.

Making the substitution $y = (x-\gamma)/\alpha$ in Equation 3.269 the distribution is given by:

$$P(y) = \frac{1}{\Gamma(\beta)} \int_{0}^{y_0} y^{\beta-1} \; e^{-y} \, dy \qquad 3.270$$

but from (1)

$$P(y) = P(\chi^2|v) = 1-Q(\chi^2|v) \qquad 3.271$$

where $P(x_2|v)$ is the Chi-Square distribution with 2β degrees of freedom and $\chi^2 = 2y$.

So, looking up the tabulated value of χ^2 for probability 1 - 1/T and 2β degrees of freedom the reduced event magnitude, y_o, is obtained as:

$$y_o = \chi^2/2 \qquad 3.272$$

and the expected event magnitude is given by:

$$x = \frac{\chi^2 \alpha}{2} + \gamma \qquad 3.273$$

Tables of Chi-Square distribution are commonly given in statistical texts, but for convenience Table 3.20 provides values of χ^2 for some commonly used probabilities and for various degrees of freedom. Note that in Table 3.20 the probabilities are arranged so that larger event magnitudes correspond to larger cumulative probabilities and smaller probabilities of exceedence (larger return periods). This table would be suitable for an analysis of maxima such as flood events. In the study of minima such as drought events, for which this distribution is sometimes used, (c.f. Chin (4)), these probabilities should be reversed.

It has been found (24) that the expression

$$\left\{\left(\frac{\chi^2}{v}\right)^{1/3} + \frac{2}{9v} - 1\right\}\left\{\frac{9v}{2}\right\}^{1/2} \qquad 3.274$$

is approximately normally distributed with zero mean and unit variance for $v>30$. Thus the value of χ^2 for a particular probability level, P, and number of degrees of freedom, v, can be approximately computed by substituting the corresponding standard normal deviate, t, in the equation:

$$\chi^2 = v\left\{1 - \frac{2}{9v} + t\cdot\sqrt{\frac{2}{9v}}\right\}^3 \qquad 3.275$$

Table 3.20

Percentage Points of the Chi-Square Distribution[1]

Degrees	Cumulative Probability, $P(\chi^2 \mid v)$, %					
of	50	75	90	95	97.5	99
Freedom	Corresponding Return Period, T, Years					
v	2	4	10	20	40	100
1	0.46	1.32	2.71	3.84	5.02	6.63
2	1.39	2.77	4.61	5.99	7.38	9.21
3	2.37	4.11	6.25	7.81	9.35	11.34
4	3.36	5.39	7.78	9.49	11.14	13.28
5	4.35	6.63	9.24	11.07	12.83	15.09
6	5.35	7.84	10.64	12.59	14.45	16.81
7	6.35	9.04	12.02	14.07	16.01	18.48
8	7.34	10.22	13.36	15.51	17.53	20.09
9	8.34	11.39	14.68	16.92	19.02	21.67
10	9.34	12.55	15.99	18.31	20.48	23.21
11	10.34	13.70	17.28	19.68	21.92	24.73
12	11.34	14.84	18.55	21.03	23.34	26.32
13	12.34	15.98	19.81	22.36	24.74	27.69
14	13.34	17.11	21.06	23.68	26.12	29.14
15	14.34	18.25	22.31	25.00	27.49	30.58
16	15.34	19.37	23.54	26.30	28.85	32.00
17	16.34	20.48	24.77	27.59	30.19	33.41
18	17.34	21.61	25.99	29.87	31.53	34.81
19	18.34	22.72	27.20	30.14	32.85	36.19
20	19.34	23.83	28.41	31.41	34.17	37.57
21	20.34	24.93	29.62	32.67	35.48	38.93
22	21.34	26.04	30.81	33.92	36.78	40.29
23	22.34	27.14	32.01	35.17	38.08	41.64
24	23.34	28.24	33.20	36.42	39.36	42.98
25	24.34	29.34	34.38	37.65	40.65	44.31
26	25.34	30.43	35.56	38.89	41.92	45.64
27	26.34	31.53	36.74	40.11	43.19	46.96
28	27.34	32.62	37.92	41.34	44.46	48.28
29	28.34	33.71	39.09	42.56	45.72	49.59
30	29.34	34.80	40.26	43.77	46.98	50.89
40	39.34	45.62	51.81	55.76	59.34	63.69
50	49.33	56.33	63.17	67.50	71.42	76.17
60	59.33	66.98	74.40	79.08	83.30	88.38
70	69.33	77.57	85.53	90.53	95.02	100.42
80	79.33	88.13	96.58	101.88	106.63	112.33
90	89.33	98.65	107.56	113.14	118.14	124.12
100	99.33	109.14	118.49	124.34	129.56	135.81

[1] The function, χ^2_α, tabulated is that value of χ^2 with v degrees of freedom beyond which (1-P)% of the distribution lies.

A refinement to the approximation may be made by substituting $(t-h_v)$ for t in Equation 3.275 where

$$h_v = \frac{60}{v} \cdot h_{60} \qquad 3.276$$

and h_{60} is tabulated in Abramowitz and Stegun (1).

Combining Equations 3.273 and 3.275 and substituting the degrees of freedom, $v = 2\beta$,

$$x = \alpha\beta \left\{ 1 - \frac{1}{9\beta} + t\sqrt{\frac{1}{9\beta}} \right\}^3 + \gamma \qquad 3.277$$

Thus knowing, α, β and γ the value of x corresponding to any given probability level can be computed.

Substituting for α and γ from Equations 3.267 and 3.268 in Equation 3.273 an expression:

$$x = \mu + [\frac{\chi^2\gamma_1}{4} - \frac{2}{\gamma_1}]\sigma \qquad 3.278$$

is obtained, where γ_1 is the coefficient of skew of the sample data.

Comparing Equation 3.271 with the general frequency equation it can be seen that, for the Pearson Type III distribution, the frequency factor, K, is given by:

$$K = \frac{\chi^2 \gamma_1}{4} - \frac{2}{\gamma_1} \qquad 3.279$$

Since K is dependent only on χ^2 and the coefficient of skew and χ^2 is dependent only upon the coefficient of skew (through the degrees of freedom, v) and the probability, then K can be tabulated directly. Table 3.21 is an example. The Soil Conservation Service, U.S. Dept. of Agriculture (39) has prepared very comprehensive tables of frequency factors for the Pearson Type III distribution.

Chin (4) has given a table of frequency factors for probability levels suitable for analysis of minimum events such as droughts.

Table 3.21

Frequency Factor for Use in Pearson Type III Distribution

Coefficient	Cumulative Probability, P, %					
of Skew	50	80	90	95	98	99
γ_1	Corresponding Return Period, T, Years					
	2	5	10	20	50	100
-2.0	0.3068	0.7769	0.8946	0.9487	0.9798	0.9900
-1.8	0.2815	0.7987	0.9450	1.0197	1.0686	1.0870
-1.6	0.2542	0.8172	0.9942	1.0934	1.1658	1.1970
-1.4	0.2254	0.8322	1.0414	1.1683	1.2700	1.3182
-1.2	0.1952	0.8437	1.0861	1.2431	1.3793	1.4494
-1.0	0.1640	0.8516	1.1276	1.3168	1.4919	1.5884
-0.8	0.1320	0.8561	1.1657	1.3886	1.6060	1.7327
-0.6	0.0995	0.8572	1.2003	1.4576	1.7203	1.8803
-0.4	0.0665	0.8551	1.2311	1.5236	1.8336	2.0293
-0.2	0.0333	0.8499	1.2582	1.5861	1.9450	2.1784
0.0	0.0000	0.8416	1.2816	1.6449	2.0538	2.3264
0.2	-0.0333	0.8304	1.3011	1.6997	2.1594	2.4723
0.4	-0.0665	0.8164	1.3167	1.7505	2.2613	2.6154
0.6	-0.0995	0.7995	1.3285	1.7970	2.3593	2.7551
0.8	-0.1320	0.7799	1.3364	1.8392	2.4530	2.8910
1.0	-0.1640	0.7575	1.3404	1.8768	2.5421	3.0226
1.2	-0.1952	0.7326	1.3405	1.9099	2.6263	3.1494
1.4	-0.2254	0.7051	1.3367	1.9384	2.7056	3.2713
1.6	-0.2542	0.6753	1.3290	1.9621	2.7796	3.3880
1.8	-0.2815	0.6434	1.3176	1.9812	2.8485	3.4994
2.0	-0.3069	0.6094	1.3026	1.9957	2.9120	3.6052

Standard Error of Estimate

The central moments of the Pearson Type III distribution are given earlier in the chapter in Equations 3.261 to 3.264 Substituting these values in the general standard error equation

$$S^2(K) = \frac{\mu_2}{n} \; [1 + \frac{K\mu_3}{\mu_2^{3/2}} + \frac{K^2(\mu_4 - \mu_2^2)}{4\mu_2^2}] \qquad 3.280$$

yields:

$$S^2(K) = \frac{\alpha^2\beta}{n} \; [1 + \frac{2K\alpha^3\beta}{(\alpha^2\beta)^{3/2}} + \frac{K^2(3\beta\alpha^2(\beta+2)-\alpha^4\beta^2)}{4\alpha^2\beta^2}] \qquad 3.281$$

which, upon substituting

$$\sigma^2 = \alpha^2\beta \qquad 3.282$$

and

$$\gamma_1 = 2/\sqrt{\beta} \qquad 3.283$$

and simplyifying, gives the expression

$$S^2(K) = \frac{\sigma^2}{n} \; [1 + K.\gamma_1 + \frac{K^2}{2} \; (1+3\gamma_1^2/4)] \qquad 3.284$$

where γ_1 is the coefficient of skew estimated from the sample.

Since the frequency factor, K, is dependent only upon the return period, T, and the coefficient of skew, γ_1, then tables of δ can be prepared where:

$$S(K) = \delta \frac{\sigma}{\sqrt{n}} \qquad 3.285$$

Table 3.22 gives values of δ for some typical values of T and γ_1.

Table 3.22

Parameter δ for Use in Standard Error of Pearson Type III Distribution

Coefficient of Skew	Cumulative Probability, P, %					
	50	80	90	95	98	99
γ_1	Corresponding Return Period, T, Years					
	2	5	10	20	50	100
-2.0	0.7581	0.8083	0.9008	0.9501	0.9800	0.9901
-1.8	0.7932	0.8102	0.9113	0.9735	1.0173	1.0343
-1.6	0.8292	0.8170	0.9233	0.9980	1.0578	1.0848
-1.4	0.8644	0.8308	0.9388	1.0247	1.1018	1.1404
-1.2	0.8974	0.8531	0.9610	1.0561	1.1504	1.2023
-1.0	0.9271	0.8849	0.9924	1.0956	1.2065	1.2725
-0.8	0.9525	0.9260	1.0359	1.1472	1.2743	1.3548
-0.6	0.9729	0.9758	1.0930	1.2143	1.3591	1.4549
-0.4	0.9879	1.0332	1.1646	1.3002	1.4661	1.5794
-0.2	0.9970	1.0964	1.2505	1.4065	1.5998	1.7344
0.0	1.0000	1.1637	1.3495	1.5339	1.7632	1.9251
0.2	0.9970	1.2334	1.4602	1.6816	1.9579	2.1546
0.4	0.9879	1.3038	1.5804	1.8483	2.1836	2.4242
0.6	0.9729	1.3732	1.7082	2.0319	2.4393	2.7337
0.8	0.9525	1.4401	1.8414	2.2304	2.7231	3.0818
1.0	0.9271	1.5032	1.9780	2.4411	3.0326	3.4665
1.2	0.8974	1.5612	2.1160	2.6619	3.3652	3.8852
1.4	0.8644	1.6128	2.2534	2.8904	3.7186	4.3354
1.6	0.8292	1.6572	2.3885	3.1241	4.0899	4.8145
1.8	0.7932	1.6935	2.5197	3.3612	4.4769	5.3198
2.0	0.7580	1.7209	2.6455	3.5996	4.8768	5.8485

Moran (31) has described a procedure for evaluating the standard error of estimate of a two parameter Pearson Type III distribution (2 parameter Gamma). When extended to the three parameter distribution discussed in this report the derivation is as follows:

If x is the event magnitude for a given return period using a Pearson Type III distribution with fitted parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ then

$$dx = \frac{\partial x}{\partial \alpha}\, d\alpha + \frac{\partial x}{\partial \beta}\, d\beta + \frac{\partial x}{\partial \gamma}\, d\gamma \qquad 3.286$$

and the standard error of estimate, $S^2(x)$, is given as

$$S^2(x) = \left(\frac{\partial x}{\partial \alpha}\right)^2 \text{var } \alpha + \left(\frac{\partial x}{\partial \beta}\right)^2 \text{var } \beta + \left(\frac{\partial x}{\partial \beta}\right)^2 \text{var } \gamma + 2\left(\frac{\partial x}{\partial \alpha}\right)\left(\frac{\partial x}{\partial \beta}\right) \text{cov } (\alpha,\beta)$$
$$+ 2\left(\frac{\partial x}{\partial \beta}\right)\left(\frac{\partial x}{\partial \gamma}\right) \text{cov } (\beta,\gamma) + 2\left(\frac{\partial x}{\partial \alpha}\right)\left(\frac{\partial x}{\partial \gamma}\right) \text{cov } (\alpha,\gamma) \qquad 3.287$$

The differential coefficients are obtained by computing x for 2 values of α, β, γ either side of the fitted values $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ and determining the slopes. The variance-covariance matric

$$\begin{bmatrix} \text{var } \alpha & \text{cov } (\alpha,\beta) & \text{cov } (\alpha,\gamma) \\ & \text{var } \beta & \text{cov } (\beta,\gamma) \\ & & \text{var } \gamma \end{bmatrix} \qquad 3.288$$

is obtained as the inverse of the symmetric matrix

$$\begin{bmatrix} -\frac{\partial^2 L}{\partial \alpha^2} & -\frac{\partial^2 L}{\partial \alpha \partial \beta} & -\frac{\partial^2 L}{\partial \alpha \partial \gamma} \\ & -\frac{\partial^2 L}{\partial \beta^2} & -\frac{\partial^2 L}{\partial \beta \partial \gamma} \\ & & -\frac{\partial^2 L}{\partial \gamma^2} \end{bmatrix} \qquad 3.289$$

where L is the likelihood function defined in Equation 3.244 so that

$$\frac{\partial^2 L}{\partial \alpha^2} = \frac{n\beta}{\alpha^2} - \frac{2}{\alpha^3} \sum_{i=1}^{n} (x_i - \gamma) \qquad 3.290$$

$$\frac{\partial^2 L}{\partial \alpha \partial \beta} = -n/\alpha \qquad 3.291$$

$$\frac{\partial^2 L}{\partial \alpha \partial \gamma} = -n/\alpha^2 \qquad 3.292$$

$$\frac{\partial^2 L}{\partial \beta^2} = -n\{\Gamma(\beta)\Gamma''(\beta) - \Gamma(\beta)^2\}\Gamma''(\beta)^{-2} \qquad 3.293$$

$$\frac{\partial^2 L}{\partial \beta \partial \gamma} = \sum_{i=1}^{n} (x_i - \gamma)^{-1} \qquad 3.294$$

$$\frac{\partial^2 L}{\partial \gamma^2} = (\beta-1) \sum_{i=1}^{n} (x_i-\gamma)^{-2} \qquad 3.295$$

Evaluation of Equations 3.290 to 3.295 substitution into matrix Equation 3.289 and inversion will give the required variances and covariances. Combining these with the differential coefficients in Equation 3.287 yields the standard error of estimate.

The return period, T, is incorporated directly in the expression for standard error since the differential coefficients are calculated as slopes at the event magnitude, x, corresponding to the return period, T.

Santos (36) has given useful tables and a numerical example of this method of estimating the standard error of estimate for a two-parameter Pearson Type III distribution.

3.3.6 Log-Pearson Type III

The U.S. Federal Water Resources Council has fairly recently recommended that the log-Pearson Type III distribution be adopted as the standard flood frequencv distribution by all U.S. government agencies. In a paper describing the investigations behind this recommendation Benson (3) explained that no rigorous statistical criteria exist on which a comparison of distributions can be based and therefore the choice of log-Pearson Type III is, to some extent, subjective.

The procedure used with the log-Pearson Type III distribution is identical to that described for the Pearson Type III except that the original variable is replaced with its logarithm. That is to say, first the logarithms of the sample data are taken and the mean, standard deviation and coefficient of skew of the logarithms are computed. Secondly the frequency factor corresponding to the computed coefficient of skew is found from Equation 3.279 or Table 3.21 and substituted, together with the mean and standard deviation of the logarithms, in the general frequency equation. Finally the antilog of the resulting figure is found. This is the event magnitude. The standard error, in log units is then computed from Equation 3.284 using σ_y, the standard deviation of logarithms instead of σ.

3.4 Comparison of Frequency Distributions

It has been mentioned previously in this chapter that no statistical test can ensure that any one distribution is the best one to use for a particular set of data. As an example of the variation possible between distributions, seven different distributions have been applied to one set of data.

The data chosen were annual maximum mean daily flows of the Saint John River at Fort Kent, New Brunswick, for the 37 years 1927 to 1963. These data have been described by Collier (7) who fitted several distributions in his study. For each distribution, event magnitudes at return periods of 2, 5, 10, 20, 50 and 100 years were computed using the standard frequency equation (e.g. Equation 3.10) and the tables of frequency factors provided in this chapter. Similarly values of the standard error of estimate were computed for each distribution for each return period using Equation 3.25 and the tables of δ provided in this chapter.

Tables 3.23 and 3.24 list the results and Figure 3.5 compares the fitted distributions with those recorded data points with assigned return periods greater than 2 years. It is apparent from the figure that the various distributions, while very close at low return periods, rapidly separate at higher return periods. The central group of distributions, 3-parameter lognormal, lognormal, log Pearson Type III and Pearson Type III appear to be among the best fitting. Because in this case, the coefficient of skew of the logarithms of the data is very close to zero (-0.118) the log Pearson Type III distribution is indistinguishable from the lognormal.

In Table 3.24 the standard errors for the truncated normal,

Table 3.23

Comparison of T-Year Event Magnitudes Using Various Frequency Distributions, Thousands of Cubic Feet per Second

Saint John River at Fort Kent, 1927-1963

Frequency		Cumulative Probability, P, %				
Distribution	50	80	90	95	98	99
		Corresponding Return Period, T, Years				
	2	5	10	20	50	100
Truncated Normal	81	100	110	118	128	134
Lognormal	78	98	111	123	137	148
3-Parameter Lognormal	77	99	112	124	140	151
Type I Extremal	77	100	115	129	148	162
Type II Extremal	75	99	120	144	182	216
Pearson Type III	80	100	111	123	134	140
Log Pearson Type III	78	99	112	123	137	148

Table 3.24

Comparison of Standard Errors of T-Year Events Using Various Frequency Distributions, Thousands of Cubic Feet per Second

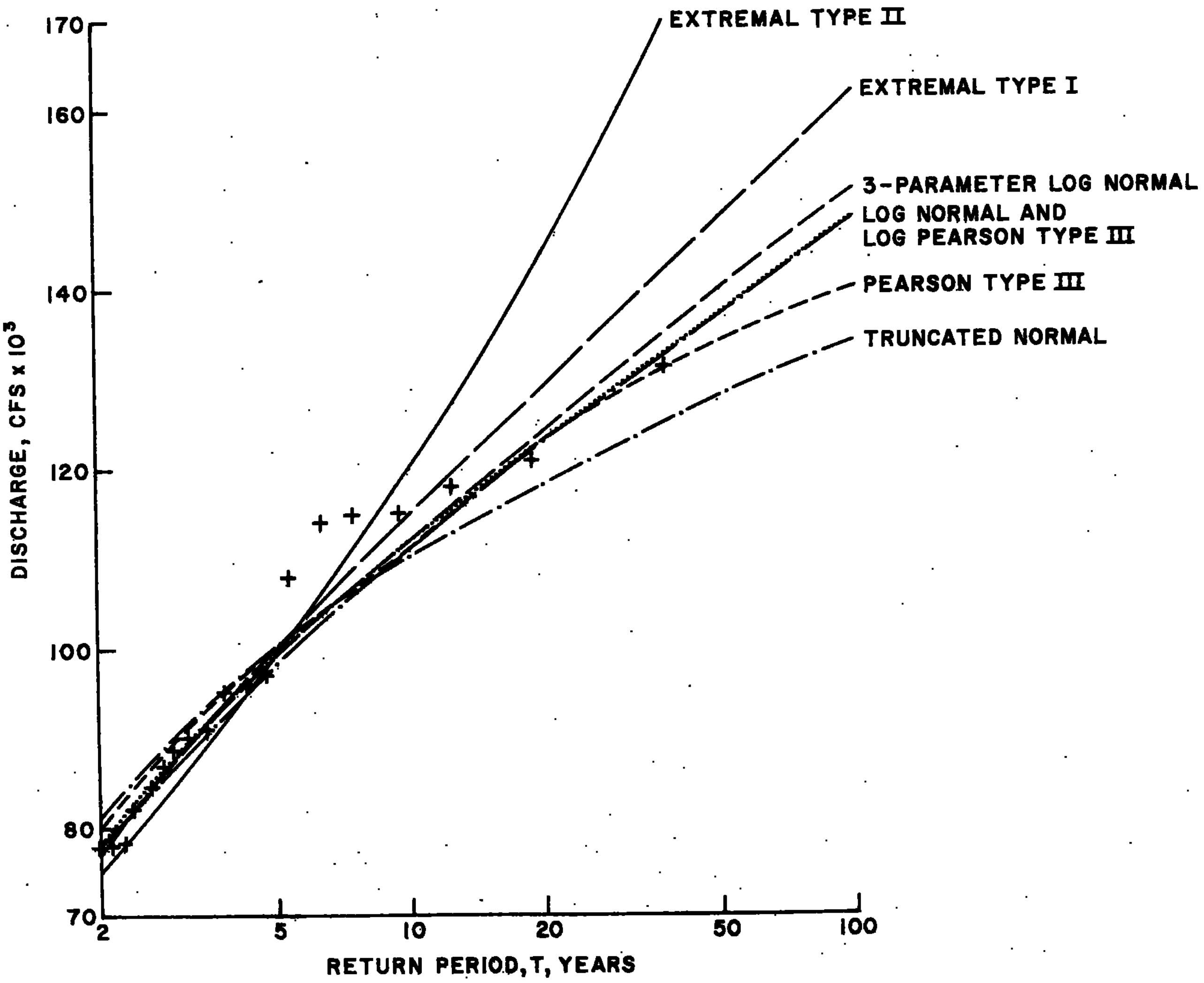
Saint John River at Fort Kent, 1927-1963

Frequency		Cumulative Probability, P, %				
Distribution	50	80	90	95	98	99
		Corresponding Return Period, T, Years				
	2	5	10	20	50	100
Truncated Normal	3.75	4.36	5.06	5.75	6.61	7.22
Lognormal	3.57	5.49	7.11	8.71	10.79	12.33
3-Parameter Lognormal	3.65	5.41	7.10	8.94	11.60	13.66
Type I Extremal	3.46	6.21	8.55	10.91	14.03	16.40
Type II Extremal	3.32	7.71	12.88	19.74	32.14	44.67
Pearson Type III	3.71	4.87	5.89	6.88	8.12	9.01
Log Pearson Type III	3.66	4.84	5.59	6.15	7.08	7.72

FIGURE 3.5

COMPARISON OF FREQUENCY CURVES FROM VARIOUS DISTRIBUTIONS

SAINT JOHN RIVER AT FORT KENT, 1927-1963



lognormal, Type I extremal and Pearson Type III were computed using Equation 3.23 and the relevant tables of δ contained in this chapter. For the 3 parameter lognormal, Type II extremal and log Pearson Type III the standard errors shown are averages of the positive and negative standard errors computed from Equations 3.100 and 3.101. For small coefficients of skew this approximation is accurate enough.

From an initial look at Figure 3.5 it was suspected that the extremal Type I and truncated normal distributions might be significantly different than the recorded distribution. To test this hypothesis a Chi-Square test was carried out. The statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(o_j - E_j)^2}{E_j} \qquad 3.296$$

is distributed assymptotically as Chi-Square with k-1 degrees of freedom where o_j is the observed number of events in the j_{th} class interval and E_j in the number of events that would be expected from the theoretical distribution. If the class intervals are defined such that each interval corresponds to an equal probability then E_j is n/k where n is the sample size and k is the number of class intervals and Equation 3.296 reduces to

$$\chi^2 = \frac{k}{n} \sum_{i=1}^{k} o_j^2 - n \qquad 3.297$$

The class intervals were computed for the various distributions as follows:

(a) Truncated Normal

$$CL = \bar{x} + tS \qquad 3.298$$

where $\bar{x}$ and S are the sample mean and standard deviation and t is the standard normal deviate corresponding to the probabilities of exceedence, P, listed in column 2 of Table 3.25.

(b) Lognormal

$$CL = \exp (\bar{x}_n + tS_n) \qquad 3.299$$

where $\bar{x}_n$ and S_n are the mean and standard deviation of the logarithms of the recorded events.

(c) 3-Parameter Lognormal

$$CL = a + \exp (\bar{x}_{na} + tS_{na}) \qquad 3.300$$

where a is the lower boundary of the distribution and $\bar{x}_{na}$ and S_{na} are the mean and standard deviation of the logarithms of the sample x-a.

(d) Type I Extremal

$$CL = \bar{x} + [\frac{y_m - \mu}{\sigma}] s \qquad 3.301$$

where y_m is -ln(-ln P) and μ and σ are the mean and standard deviation of the plotting positions.

(e) Type II Extremal

$$CL = \exp\ (\bar{x}_n + [\frac{y_m - \mu}{\sigma}]\ S_n) \qquad 3.302$$

(f) Pearson Type III

$$CL = \bar{x} + [\frac{\chi^2 \gamma_1}{4} - \frac{2}{\gamma_1}]\ S \qquad 3.303$$

where χ^2 is the value of Chi-Square at probability P and $8/\gamma_1^2$ degrees of freedom, γ_1 is the sample coefficient of skew.

(g) Log Pearson Type III

$$CL = \exp\ (\bar{x}_n + [\frac{\chi^2 \gamma_1}{4} - \frac{2}{\gamma_1}]\ S_n \qquad 3.304$$

The recorded events were then sorted in order of magnitude and the numbers of events within each class interval determined for each distribution.

Table 3.25 lists the computed class limits for each distribution together with the derived Chi-Square values. All the values of Chi-Square are significant at 95% for the appropriate degrees of freedom and so the original hypothesis cannot be proved.

The Kolmogorov-Smirnov test can give no further information in this case since it uses plotting positions as observed frequencies.

Table 3.25

Comparison of Class Limits and Chi-Square Values for Different Distributions

Class Interval	Probability	Truncated Normal	Lognormal	3-Parameter Lognormal	Type I Extremal	Type II Extremal	Pearson Type III	Log-Pearson Type III
0	0	0	0	0	0	0	0	0
1	0.14286	56.66	57.51	56.42	56.82	57.62	57.03	57.77
2	0.28571	68.10	66.32	65.24	65.67	64.34	67.14	65.53
3	0.42857	76.91	74.00	72.94	73.52	70.95	75.43	72.66
4	0.57143	85.11	81.97	80.94	81.85	78.71	83.56	80.40
5	0.71429	93.92	91.47	90.48	92.08	89.40	92.73	90.13
6	0.85714	105.36	105.48	104.58	107.76	108.68	105.37	105.49
7	1.0	∞	∞	∞	∞	∞	∞	∞
Chi-Square		3.68	4.43	3.68	4.43	2.54	2.54	3.68

3.5 Distribution of T-Year Events

As discussed earlier in this chapter, the empirical determination of confidence limits involves an assumption that the T-year event is distributed normally with mean x(K) and variance $s^2(K)$. In order to test the validity of this assumption, several experiments in data generation were conducted.

A set of 100 uniformly distributed random numbers were generated in the interval 0 - 1.0. Assuming that these numbers could represent probabilities, the corresponding event magnitudes were computed according to a particular distribution. These 100-year event magnitudes then represented a simulated 100-year record of annual maximum events from which the maximum or 100-year event was selected. By repeating this procedure n times a set of n 100-year events were generated and the distribution of these events was tested.

Tables 3.26, 3.27 and 3.28 show the results of these experiments using a lognormal distribution, a Type I extremal distribution and a Pearson Type III distribution. Data were generated using a range of values for standard deviation of the sample (and sample coefficient of skew in the case of the Pearson Type III distribution) and the resulting distributions of 100-year events were compared to normal and lognormal distributions using the Chi-Square and Kolmogorov-Smirnov tests. Comparison with tabulated Chi-Square and Kolomogorov-Smirnov statistics show that in nearly all cases the distributions of 100-year events are indistinguishable statistically from the normal or lognormal distributions. Out of 26 experiments only seven 100-year distributions are statistically different than normal and none are statistically different than a lognormal distribution. The assumption of normality of the T-year event distribution for sample size 100 would therefore seem reasonable.

Table 3.26

Results of Tests on the Distribution of 100-Year Events Generated from a Lognormal Distribution

Test Number	Sample Mean	Sample Standard Deviation	Mean of Generated 100-Year Events	Standard Deviation of Generated 100-Year Events	Chi-Square for Normal	Kolmogorov-Smirnov for Normal	Chi-Square for Lognormal	Kolmogorov-Smirnov for Lognormal
1	500	100	818	71	12.2	0.08	16.7	0.07
2		200	1236	237	27.5	0.12	12.8	0.07
3		250	1507	282	17.6	0.09	11.9	0.05
4		300	1819	477	29.9	0.11	18.5	0.06
5		400	2211	632	24.8	0.09	6.8	0.04
6		500	3341	1519	48.5	0.15	13.1	0.07

Notes:

(a) Tabulated value of Chi-Square at 95% significance for the number of degrees of freedom used is 26.3.

(b) Tabulated value of the Kolmogorov-Smirnov statistic at 95% significance for the sample size used is 0.14.

Table 3.27

Results of Tests on the Distribution of 100-Year Events Generated from a Type I Extremal Distribution

Test Number	Sample Mean	Sample Standard Deviation	Mean of Generated 100-Year Events	Standard Deviation of Generated 100-Year Events	Chi-Square for Normal	Kolmogorov-Smirnov for Normal	Chi-Square for Lognormal	Kolmorogov-Smirnov for Lognormal
1	500	100	862	89	23.6	0.11	15.5	0.08
2		200	1249	193	13.7	0.07	7.1	0.04
3		250	1451	264	14.6	0.08	9.8	0.07
4		300	1669	281	22.4	0.08	17.3	0.05
5		400	2035	391	28.4	0.07	11.0	0.05
6		500	2464	455	9.2	0.05	4.1	0.03

Notes:

(a) Tabulated value of Chi-Square at 95% significance and the degrees of freedom used is 26.3.

(b) Tabulated value of Kolmogorov-Smirnov statistic at 95% significance for the sample size used is 0.14.

Table 3.28

Results of Tests on the Distribution of 100-Year Events Generated from a Pearson Type III Distribution

Test Number	Sample Mean	Sample Standard Deviation	Sample Coefficient of Skew	Mean of Generated 100-Year Events	Standard Deviation of Generated 100-Year Events	Chi-Square for Normal	Kolmogorov-Smirnov for Normal	Chi-Square for Lognormal	Kolmogorov-Smirnov for Lognormal
1	500	100	1.0	823	80	26.6	0.11	16.1	0.07
2		200		1159	148	22.7	0.06	19.4	0.05
3		250		1327	219	27.2	0.08	15.8	0.06
4		300		1484	195	13.4	0.06	12.8	0.05
5		400		1790	272	12.2	0.07	18.8	0.05
6		500		2092	360	25.4	0.10	16.4	0.08
7	2600	800	1.0	5146	577	25.4	0.10	20.6	0.09

Notes:

(a) Tabulated value of Chi-Square at 95% significance for the number of degrees of freedom used is 26.3.

(b) Tabulated value of the Kolmogorov-Smirnov statistics at 95% significance for the sample size used is 0.14.

(c) Test number 7 is provided for comparison with the results of Matalas and Wallis (30).

Having shown that for the 100-year events an assumption of normality is not unreasonable for varying means and standard deviations but constant sample size (100 events) the next step was to investigate whether this assumption could be maintained for sample sizes smaller than 100. This was necessary since, in hydrology, a sample of 100 events would be the exception rather than the rule.

From a particular distribution (again, lognormal, Type I extremal and Pearson Type III distributions were used) with a population mean of 500 and population standard deviation 250, a sample of 100 was generated. From the computed mean and standard deviation of this sample the magnitude of the 100-year event was calculated. This procedure was repeated 100 times so that 100 100-year events were available and the distribution of these events was checked using, as before, Chi-Square and Kolmogorov-Smirnov tests. The sample size was then changed from 100 to 90 and the entire procedure repeated. Similarly sample sizes of 80, 70, 60, 50, 40, 30, 20 and 10 were used.

Figure 3.6 shows the results for the lognormal distribution. Examining this figure it is evident that as sample size decreases there is a very clear increase in the standard deviation of the distribution of 100-year events. This is as expected, as the available information is decreased this is reflected in an increased variance in the result. Similarly there is a small decrease in the magnitude of the mean of the 100-year events as sample size decreases. The significant result is that as sample size decreases the value of Chi-Square comparing the empirical distribution of the 100-year events to the normal distribution increases fairly rapidly. Thus for smaller samples the assumption of normality is not as valid as for larger samples. By taking the tabulated critical value of Chi-Square at 95% significance for the 100 events and plotting this value on Figure 3.6 a minimum acceptable sample size of 10 items results.

Thus at 95% significance the assumption of normality of the distribution of 100-year events derived from a lognormal distribution is valid for sample sizes greater than 10. Table 3.29 lists the detailed results for the lognormal distribution.

Similarly Tables 3.30 and 3.31 and Figures 3.7 and 3.8 show the results for the Type I Extremal and Pearson Type III distributions. In each case the validity of the normality assumption decreases with decreasing sample size.

Because of the nature of the "pseudo-random" numbers used in digital computer generating subroutines, the data generation for each distribution was repeated three times using different "seeds" for the random number generator.

It should be noted that the lines fitted to the plotted values of the mean and standard deviation of the extreme events in Figures 3.6, 3.7 and 3.8 represent only sample variations. In the cases of the lognormal and Pearson Type III distributions the theoretical value of the mean of extremes is constant and the sample variation indicated on the figures is not representative.

Further work is being carried out on this subject and results will be reported.

Figure 3.6

VARIATION OF MEAN, STANDARD DEVIATION AND CHI-SQUARE OF 100-YEAR EVENTS OF LOGNORMAL DISTRIBUTION WITH DIFFERENT SAMPLE SIZE.

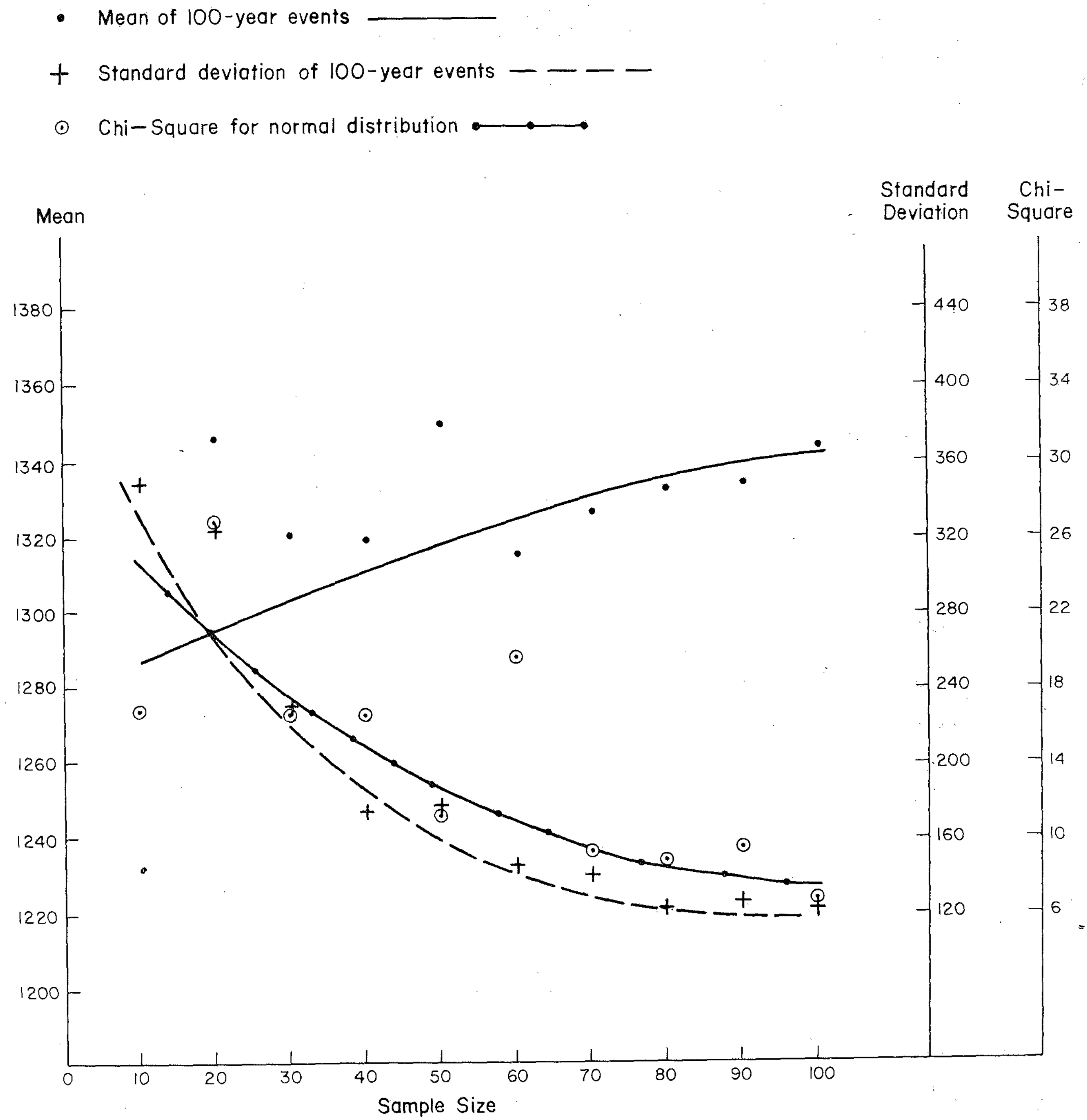


Figure 3.7

VARIATION OF MEAN, STANDARD DEVIATION AND CHI-SQUARE OF 100-YEAR EVENTS OF A TYPE I EXTREMAL DISTRIBUTION WITH DIFFERENT SAMPLE SIZES.

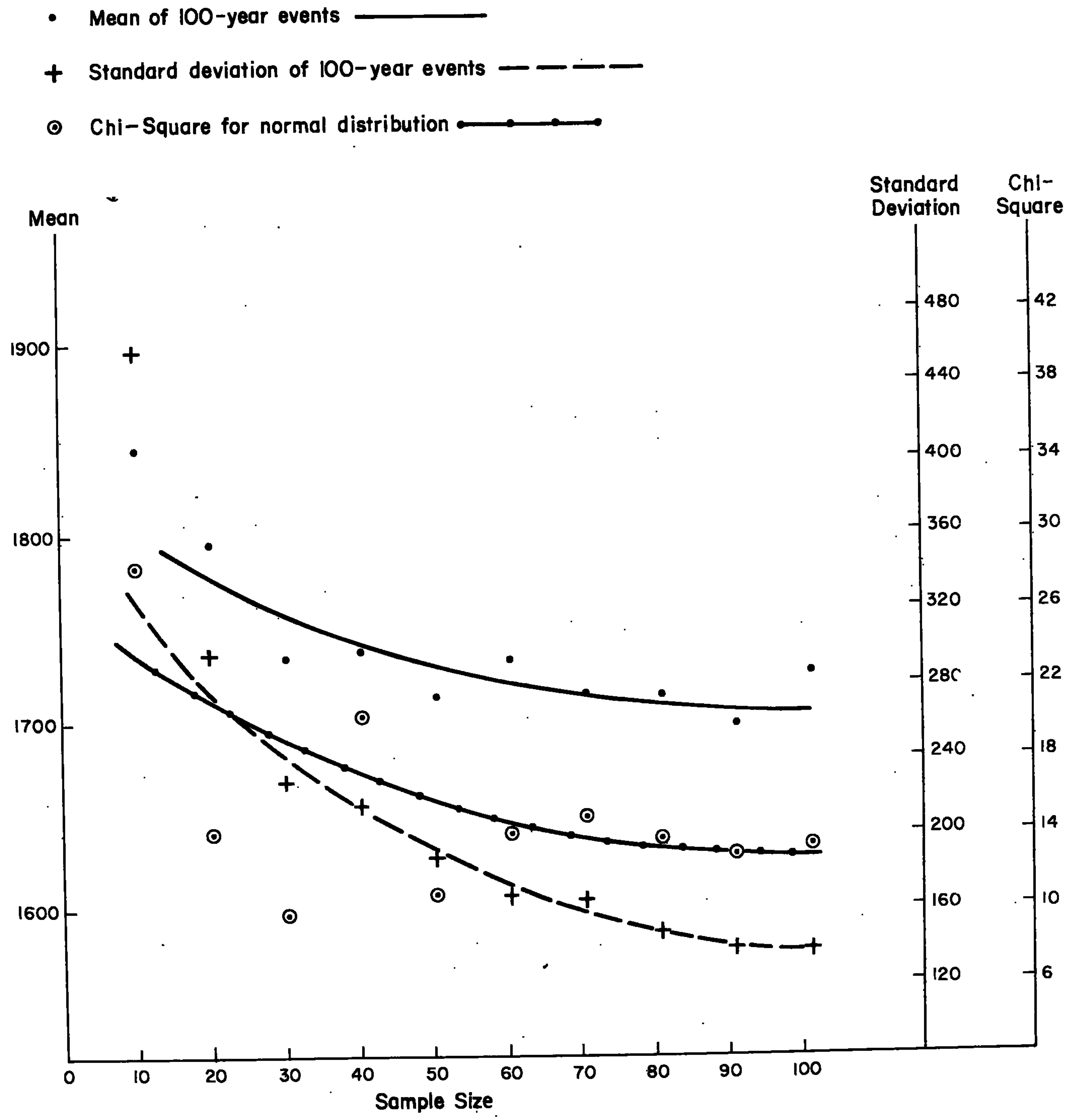


Figure 3.8

VARIATION OF MEAN, STANDARD DEVIATION AND CHI-SQUARE OF 100-YEAR EVENTS OF A PEARSON TYPE III DISTRIBUTION WITH DIFFERENT SAMPLE SIZES.

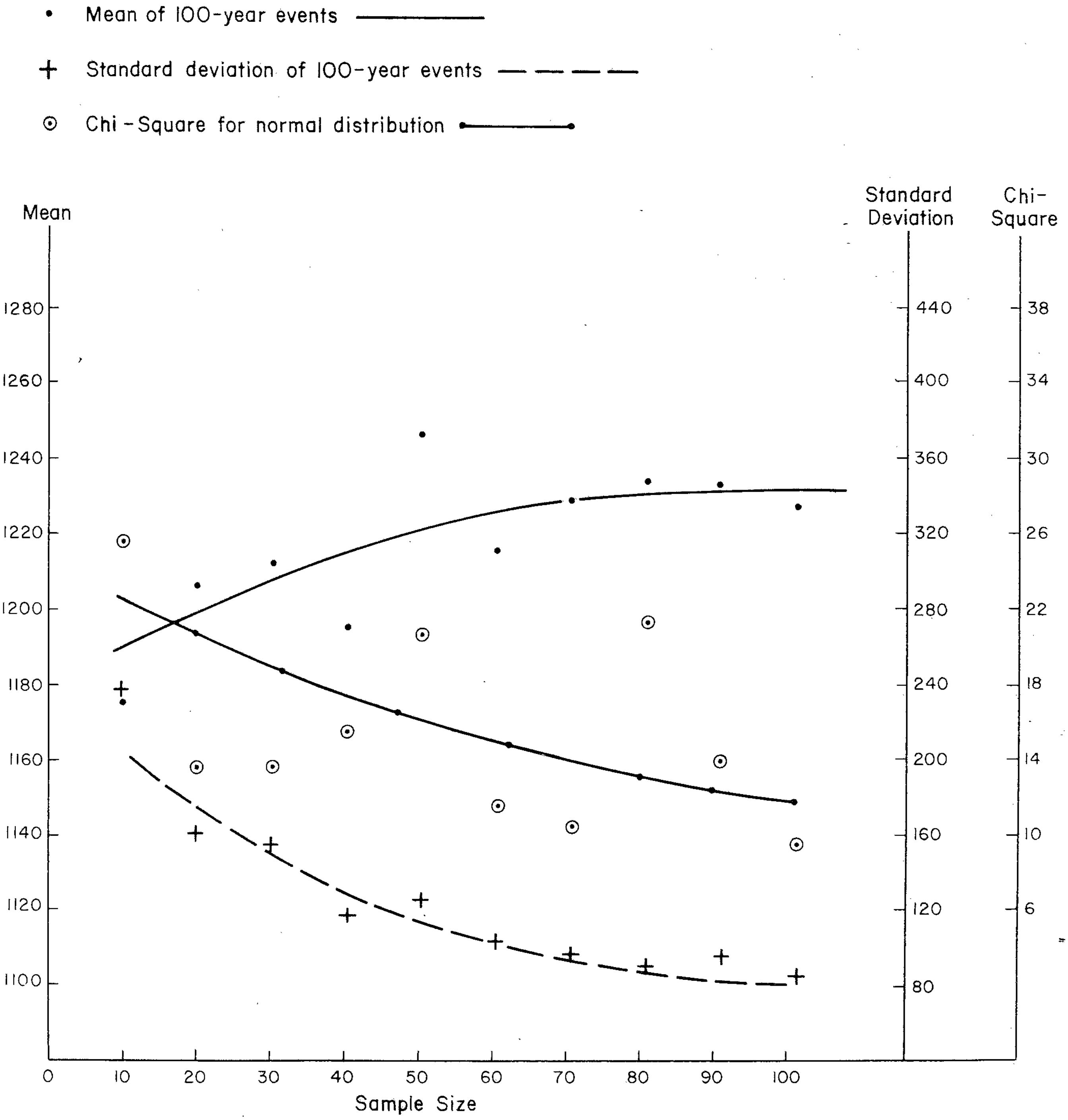


Table 3.29

Results of Tests on the Distribution of 100-Year Events Generated from Different Sized Samples of a Lognormal Distribution

Sample Size	Means of Generated 100-Year Events	Standard Deviation of Generated 100-Year Events	Chi-Square for Normal Distribution	Kolmogorov-Smirnov for Normal Distribution	Chi-Square for Lognormal Distribution	Kolmogorov-Smirnov for Lognormal Distribution
100	1343	124	6.80	0.023	6.80	0.040
90	1334	127	9.80	0.043	16.10	0.040
80	1331	124	8.90	0.070	5.00	0.043
70	1326	140	9.20	0.047	10.70	0.040
60	1314	145	19.70	0.047	16.40	0.037
50	1349	177	11.30	0.050	9.20	0.040
40	1298	174	16.10	0.083	15.20	0.073
30	1320	229	16.10	0.047	10.10	0.047
20	1346	322	26.30	0.120	16.40	0.073
10	1231	347	16.40	0.063	4.40	0.033

Notes:

(a) Tabulated value of Chi-Square at 95% significance for the number of degrees of freedom used is 26.3.

(b) Tabulated value of the Kolmogarov-Smirnov statistic at 95% significance for the sample size used is 0.14.

Table 3.30

Results of Tests on the Distribution of 100-Year Events Generated from Different Sized Samples of a Type I Extremal Distribution

Sample Size	Mean of Generated 100-Year Events	Standard Deviation of Generated 100-Year Events	Chi-Square for Normal Distribution	Kolmogorov-Smirnov for Normal Distribution	Chi-Square for Lognormal Distribution	Kolmogorov-Smirnov for Lognormal Distribution
100	1725	139	13.40	0.077	17.00	0.077
90	1698	139	12.80	0.070	7.40	0.053
80	1715	146	13.70	0.063	11.60	0.043
70	1716	162	14.90	0.047	11.00	0.033
60	1733	165	14.00	0.050	9.50	0.043
50	1715	185	10.70	0.047	17.30	0.057
40	1739	215	20.90	0.067	7.70	0.047
30	1734	228	9.80	0.057	5.60	0.030
20	1796	298	14.00	0.047	5.90	0.030
10	1845	445	28.10	0.103	16.40	0.063

Notes:

(a) Tabulated value of Chi-Square at 95% significance for the number of degrees of freedom used is 26.3.

(b) Tabulated value of the Kolmogarov-Smirnov statistic at 95% significance for the sample size used is 0.14.

Table 3.31

Results of Tests on the Distribution of 100-Year Events Generated from Different Sized Samples of a Pearson Type III Distribution

Sample Size	Mean of Generated 100-Year Events	Standard Deviation of Generated 100-Year Events	Chi-Square for Normal Distribution	Kolomogorov-Smirnov for Normal Distribution	Chi-Square for Lognormal Distribution	Kolmogorov-Smirnov for Lognormal Distribution
100	1227	83	9.50	0.063	12.50	0.063
90	1232	96	14.00	0.047	11.30	0.047
80	1233	88	21.50	0.053	22.10	0.077
70	1229	97	10.10	0.043	7.40	0.023
60	1216	103	11.60	0.047	12.50	0.060
50	1246	123	20.90	0.070	27.80	0.067
40	1195	119	15.50	0.043	7.40	0.037
30	1212	154	13.70	0.037	12.50	0.047
20	1207	160	13.70	0.050	10.40	0.050
10	1176	234	25.70	0.093	14.00	0.063

Notes:

(a) Tabulated value of Chi-Square at 95% significance for the number of degrees of freedom used is 26.3.

(b) Tabulated value of the Kolmogarov-Smirnov statistic at 95% significance for the sample size used is 0.14.

References for Chapter 3

1. Abramowitz, M., and I.A. Stegun, 1965, Handbook of Mathematical Functions, Dover Publications, New York.

2. Benson, M.A. 1962, Evolution of Methods for Evaluating the Occurrence of Floods, USGS Water Supply Paper No. 1580-A.

3. Benson, M.A., 1968, Uniform Flood-Frequency Estimating Methods for Federal Agencies, Water Resources Research, Vol. 4, No. 5, pp. 891-908.

4. Chin, W.O., 1967, Formulae and Tables for Computing and Plotting Drought Frequency Curves, Technical Bulletin No. 8, Inland Waters Branch, Ottawa.

5. Chow, V.T., 1954, The Log-Probability Law and its Engineering Applications, Proc. ASCE, Vol. 80, pp. 1-25.

6 Chow, V.T., 1964, Editor-in-Chief, Handbook of Applied Hydrology, McGraw-Hill.

7. Collier, E.P., 1965, Flood Frequency Curves - Single Station Analysis, unpublished paper, Water Resources Branch, Ottawa.

8. Condie, R., 1973, Unpublished notes, Environment Canada, Ottawa.

9. Coulson, A., 1966, Tables for Computing and Plotting Flood Frequency Curves, Technical Bulletin No. 3, Inland Waters Branch, Ottawa.

10. Cramer, H., 1946, Mathematical Methods of Statistics, Princeton University Press.

11. Cuthbert, D.R., and G.L. Latham, 1972, Program FLDFRQ, Flood Frequencies by the Method of Maximum Likelihood, unpublished report, Water Management Service, Ottawa.

12. Dalrymple, T., 1960, Flood Frequency Analysis, USGS Water Supply Paper 1543-A.

13. Deininger, R.A., and J.D. Westfield, 1969, Estimation of the Parameters of Gumbel's Third Asymptotic Distribution by Different Methods, Water Resources Research, Vol. 5, No. 6, pp. 1238 - 1243.

14. Frost, J. and R.T. Clarke, 1972, Estimating the T-year Flood by the Extension of Records of Partial Duration Series, Bull. IAHS, Vol. XVIII, No. 1, pp. 209 - 217.

15. Greenwood, J.A., and D. Durand, 1960, Aids for Fitting the Gamma Distribution, Technometrics, Vol. 2, No. 1, pp. 55-66.

16. Gumbel, E.J., 1958, Statistics of Extremes, Columbia University Press.

17. Gumbel, E.J., 1966, Extreme Value Analysis of Hydrologic Data, Proc. Hydrology Symp. No. 5, NRC, Ottawa, pp. 149 - 169.

18. Hall, W.A., and D.T. Howell, 1963, Estimating Flood Probabilities within Specific Time Intervals, J. Hydrology, Vol. 1, No. 1, pp. 265 - 271.

19. Hardison, C.H., 1969, Accuracy of Streamflow Characteristics, USGS Professional Paper No. 650-D, pp. D210 - D214.

20. Joseph, E.S., 1970, Frequency of Design Drought, Water Resources Research, Vol. 6, No. 4, pp. 1199 - 1201.

21. Kaczmarek, Z., 1958, Efficiency of the Estimation of Floods with a Given Return Period, Proc. AISH General Assembly of Toronto, 1957, Vol. III, 144 - 159.

22. Kalinin, G.P., 1960, Calculation and Forecasts of Streamflow from Scanty Hydrometric Readings, Trans. Interregional Seminar on Hydrologic Networks and Methods, Bangkok, 1959, WMO Flood Control Series No. 15, pp. 42 - 52.

23. Kendall, G.R. 1959, Statistical Analysis of Extreme Values, Proc. Hydrology Symp. No. 1, Spillway Design Floods, NRC, Ottawa, pp. 54 - 78.

24. Kendall, M.G., and A. Stuart, 1963, The Advanced Theory of Statistics, Vol. I, Griffin, London.

25. Kilmartin, R.F., and J.R. Peterson, 1972, Rainfall - Runoff Regression with Logarithmic Transforms and Zeros in the Data, Water Resources Research, Vol. 8, No. 4, pp. 1096 - 1099.

26. Kite, G.W., and R.L. Pentland, Data Generating Methods in Hydrology, Technical Bulletin No. 36, Inland Waters Branch, Ottawa.

27. Leese, M.N., 1973, The Use of Censored Data in Estimating T-Year Floods, Proceedings of the UNESCO/WMO/IAHS Symposium on the Design of Water Resources Projects with Inadequate Data, Madrid, Vol. 1, pp. 235 - 247.

28. Lowery, M.D., and J.E. Nash, 1970, A Comparison of Methods of Fitting the Double Exponential Distribution, Journal of Hydrology Vol. 10, No. 3, pp. 259 - 275.

29. Matalas, N.C., 1963, Probability Distribution of Low Flows, USGS Professional Paper No. 434-A.

30. Matalas, N.C., and J.R. Wallis, 1973, Eureka! It Fits a Pearson Type 3 Distribution, Water Resources Research, Vol. 9, No. 2, pp. 281 - 289.

31. Moran, P.A.P., 1957, The Statistical Treatment of Flood Flows, Trans. American Geophysical Union, Vol. 38, No. 4, pp. 519 - 523.

32. Nash, J.E., and J. Amorocho, 1966, The Accuracy of the Prediction of Floods of High Return Period, Water Resources Research, Vol. 2, No. 2, pp. 191 - 198.

33. Panchang, G.M., 1967, Improved Precision of Future High Floods, Proc. Symp. on Floods and their Computations, Leningrad, pp. 51 - 59.

34. Samuelsson, B., 1972, Statistical Interpretation of Hydrometeorological Extreme Values, Nordic Hydrology, Vol. 3, No. 4, pp. 199 - 214.

35. Sangal, B.P., and A.K. Biswas, 1970, The 3-Parameter Lognormal Distribution and its Applications in Hydrology, Water Resources Research, Vol. 6, No. 2, pp. 505 - 515.

36. Santos, A., 1970, The Statistical Treatment of Flood Flows, Water Power, Vol. 22, No. 2, pp. 63 - 67.

37. Shane, R.M., 1966, A statistical Analysis of Base-Flow Flood Discharge Data, Cornell University, Ph.D. Thesis.

38. Singh, K.P., and R.A. Sinclair, 1972, Two-Distribution Method for Flood-Frequency Analysis, Proc. ASCE, Vol. 98, No. HY1, pp. 29 - 45.

39. Soil Conservation Service, U.S. Dept. of Agriculture, 1968, New Tables of Percentage Points of the Pearson Type III Distribution, Technical Release No. 38, Central Technical Unit.

40. Spence, E.S., 1973, Theoretical Frequency Distributions for the Analysis of Plains Streamflow, Canadian Journal of Earth Sciences, Vol. 10, pp. 130 - 139.

41. Todorovic, P., and J. Rousselle, 1971, Some Problems of Flood Analysis, Water Resources, Vol. 7, No. 5. pp.

42. Todorovic, P., and D.A. Woolhiser, 1972, On the Time When the Extreme Flood Occurs, Water Resources Research, Vol. 8, No. 6, pp. 1433 - 1438.

43. Verma, R.D., and R.M. Advani, 1973 Flood Control Planning with Inadequate Hydrologic Data, Proceedings of Second International Symposium in Hydrology, Water Resources Publications, Fort Collins, Colorado 80521, pp. 259 - 268.

44. Weatherburn, C.E., 1962, A First Course in Mathematical Statistics, Cambridge University Press.

45. Weiss, L.L., 1955, A Nomogram Based on the Theory of Extreme Values for Determining Values for Various Return Periods, Monthly Weather Review, Vol. 83, No. 3, pp. 69 - 71.

46. WMO, 1969, Estimation of Maximum Floods, Technical Note No. 98, WMO No. 233.TP.126, Geneva.

47. Yevjevich, V., 1972, Probability and Statistics in Hydrology, Water Resources Publications, Fort Collins, Colorado.

48. Zelenhasic, E., 1970, Theoretical Probability Distributions for Flood Peaks, Hydrology Paper No. 42, Colorado State University, Fort Collins, Colorado.

CHAPTER 4

Regional Analysis

4.1 Introduction

Chapter 3 has described some of the probability distributions that can be used to carry out a frequency analysis on a set of observed or computed data. Using any one of those techniques the event magnitude corresponding to a given probability of occurrence can be determined. This event magnitude will, however, apply only to the exact location at which the original observations were made. Frequently in hydrology it is necessary to estimate event magnitudes at sites where no observations have been taken. As an example, the design of a highway culvert may require the estimation of a design flood for a small ungauged catchment area. Regional analysis is the term given to techniques which make this estimation possible.

In addition, as noted in Chapter 1, the use of more than one set of data tends to reduce the sampling error and, even for a gauged site, will produce more reliable event estimates than a single station frequency analysis.

The earliest approach to the regionalisation problem was to use empirical equations relating floodflow, Q, to drainage area, A, within a particular region (3) such as:

$$Q = cA^n \qquad 4.1$$

where c and n are constants. Other types of empirical equation (some still in use, such as the Rational Formula) related floodflow to rainfall intensity and area, as:

$$Q = ciA \qquad 4.2$$

where c is a runoff coefficient, i is the rainfall intensity and A is the area. The objective of all these equations was to extrapolate from gauged basins to ungauged basins by means of parameters which could be estimated (rainfall intensity) or measured from maps (area).

Other methods in use include that designed by Coulson (10) for Southern Ontario. From the records of 59 gauging stations in the area, together with an isohyetal map of mean annual precipitation, a map of lines of equal mean annual runoff was drawn. Re-writing the standard frequency equation (see Chapter 3) as

$$Q_T = \bar{Q}\ (K.z + 1) \qquad 4.3$$

where Q_T is the event magnitude at the required return period, T, $\bar{Q}$ is the mean annual runoff, K is a frequency factor depending on the probability distribution used and z is the coefficient of variation. For an ungauged drainage basin for which an estimate of Q_T is required, Coulson (10) obtained $\bar{Q}$ by planimetering the mean annual runoff isoline map for the particular basin. Using a Pearson Type III distribution the value of K, the frequency factor, depends upon the return period, T, and the coefficient of skew of the distribution, γ_1. It was found in Southern Ontario that the coefficient of skew could best be derived from the coefficient of variation as:

$$\gamma_1 = 2z \qquad 4.4$$

while the coefficient of variation could be obtained from the drainage area, A, as (14):

$$z = 0.35 - 0.03 \log (A + 1) \qquad 4.5$$

It should be noted that this method was used only for annual flows and not for instantaneous maxima or minima.

The USGS currently uses a method described as an "index-flood" technique (12). There are two major parts to this method. Firstly, basic dimensionless frequency curves are drawn representing the ratios of the floods at various frequencies to the mean annual flood for each gauged basin. Secondly, relationships are developed between the characteristics of drainage areas and the mean annual flood. Combining the mean annual flood with a regional frequency curve enables flood magnitudes to be estimated at any location within the region.

Many modifications of the original index-flood method have been made (3) (11) mainly to try and increase the number of independent variables used to transfer the hydrologic information. In general there are two types of variables used: (a) physiographic characteristics such as drainage area, elevation, slope, percent of basin covered by lakes, swamps, etc. and (b) hydrometeorologic variables such as mean annual precipitation and mean annual temperature.

One of the major developments since the index-flood technique is the "square-grid" method (26). Originally programmed for mean annual runoff this method has been adapted for frequency analyses, simulation and modelling (16), (25), (18).

Alternate methods include the use of standard single-station frequency distributions modified for use as regional distributions and the regional record maxima technique.

Many of the methods used in regional analysis depend upon inter-station correlation of streamflows in order to produce time series with a uniform period of record. The final section of this chapter discusses information transfer together with the concept of effective number of years of record and number of stations.

4.2 Index-Flood Method

The basic idea behind the index-flood method (12) is to increase the reliability of the frequency characteristics within a region. If, within a hydrologically homogeneous area, a number of hydrometric stations have been operating and recording the effects of the same meteorologic factors then a combination of these records will provide, not a longer record, but a more reliable record. The following brief description of the index-flood method does not include all the computation details of the procedure. These can be obtained, if required, from Dalrymple's (12) report.

Firstly, the data sets available within a region are listed, unsuitable stations eliminated, and a common period of record selected. Generally stations having less than 5 years of record of gauging regulated or controlled streams are excluded. Since streamflows are not pure random but contain trends and periodicities the period over which measurements are made becomes important when records are combined. A bar graph showing the period of record of each gauge is useful in determining which base period to use. The base period should be planned so as to include the maximum information content i.e. maximum number of station-years. Missing data points may be filled in by inter-station correlations (see section 4.6). Data points filled in in this way are not used directly but only as aids in assigning representative return periods to the recorded events.

The index-flood method next computes return periods, T, for each recorded event for each station in the region using the equation:

$$T = \frac{n + 1}{m} \qquad 4.6$$

where n is the sample size and m is the order number of an event; m = 1 for the maximum event and m = n for the minimum event. For each station a graph of T versus event magnitude is plotted and a smooth curve drawn through the points. No attempt is made to force a straight line fit or to fit any mathematical distribution. The mean annual event for the station is then picked off the smooth curve at the point T = 2.33. This is a theoretical result taken from the Type I extremal distribution (see Chapter 3). Benson (2) has confirmed experimentally that the mean annual event (i.e. the mean of all the observed annual maxima) does occur with a return period of 2.33 years. It is preferred in the index-flood method to derive the mean annual event graphically rather than arithmetically.

Dalrymple (12) has described a test which should be used at this stage of the index-flood procedure to check for regional hydrologic homogeneity. If the standard error of estimate of the reduced variable, y, in a Type I extremal distribution is given by:

$$\sigma_y = \frac{e^y}{\sqrt{n}}\sqrt{\frac{1}{T-1}} \qquad 4.7$$

then, assuming a normal distribution of the estimates, 95% of the estimates will lie within ±$2\sigma_y$ of the most probable value. If T, the return period of the estimate, is taken as 10 years, then

$$2\sigma_y = \frac{0.666e^y}{\sqrt{n}} \qquad 4.8$$

Since for T = 10 the reduced variable in a Type I extremal distribution is 2.25 (see Chapter 3) then the confidence limits are given by

$$2.25 \pm 6.33/\sqrt{n} \qquad 4.9$$

Table 4.1, after Dalrymple (12), gives the upper and lower confidence limits with the corresponding return periods for various values of n.

Table 4.1

Confidence limits for Index-Flood Homogeneity Test[1]

Sample size	lower limit		upper limit	
n	$y-2\sigma_y$	T_L	$y+2\sigma_y$	T_U
5	-0.59	1.2	5.09	160
10	0.25	1.8	4.25	70
20	0.83	2.8	3.67	40
50	1.35	4.4	3.15	24
100	1.62	5.6	2.88	18
200	1.80	6.5	2.70	15
500	1.97	7.7	2.53	13
1000	2.05	8.3	2.45	12

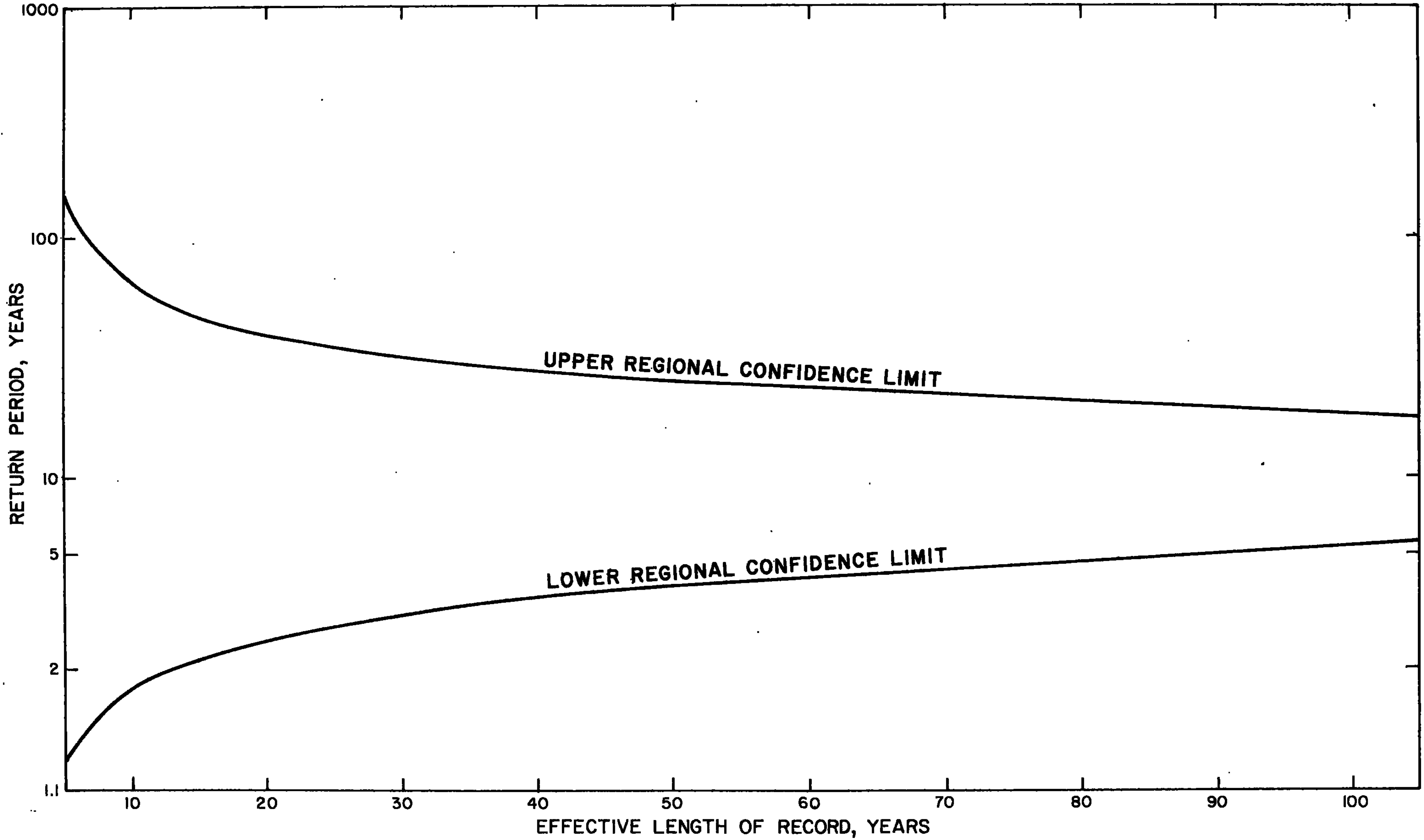
[1] From Dalrymple (12)

The procedure used for the test is to first of all plot T_L and T_U from Table 4.1 versus n on probability scale graph paper. Then, for each station in the region to be tested, the ratio of the 10-year event to the mean annual event is computed and an average ratio for the region calculated . Then the average ratio for the region is multiplied by the mean annual event for each station to give a modified 10-year event magnitude for each station. The return periods corresponding to these modified 10-year events are then found for each station from the individual station frequency curves, say T_E. The effective period of record of each gauging station is determined as the number of recorded annual events plus one half the number of events computed for that station by inter-station correlation , say N_E. Next, the coordinate pairs (T_E, N_E) for each station are plotted on the test graph showing curves of T_L and T_U. Any station for which the plotted point is outside the confidence limit curves is then excluded from the homogeneous region. Figure 4.1 is a base graph which could be photo-copied for use in this test.

For each station which remains in the hydrologically homogeneous region ratios of events of different return periods to the mean annual event are computed for T values of say 1.1, 1.5, 5, 10, 20, 50 and median values of these ratios are determined for the region. A plot of these median ratios versus return period is then the regional frequency curve and represents the most likely relationship for all parts of the region.

The next major step in the index-flood analysis is to plot drainage area versus mean annual event for those stations within

FIGURE 4.1
PLOT TO BE USED IN REGIONAL HOMOGENEITY TEST



the homogeneous region and graphically fit a smooth curve through the points. An alternative available at this stage is to develop a multiple regression equation between mean annual event and basin characteristics.

To define a frequency curve at any location within the homogeneous region the mean annual event is determined from the curve relating this event to drainage area. The mean annual event is then multiplied by the median ratios for the various return periods required, as determined from the regional frequency curve.

Regional index-flood studies have been carried out for most of the states in the U.S. and for the South Saskatchewan River Basin (6), New Brunswick - Gaspé (7) and Nova Scotia (9) areas in Canada.

Benson (3) has noted three deficiencies found in the index-flood method:

(1) The index-flood (mean annual flood) for stations with short periods of record may not be typical. This means that the ratios of floods of different return periods to the index-flood may vary widely between stations.

(2) The homogeneity test is used to determine whether the differences in slopes of frequency curves are greater than may be attributed to chance alone. This test uses the ratio of the 10-year flood to the mean annual flood as the slope. The test cannot practicably be applied at a level much higher than that of the 10-year flood because many individual records are too short to adequately

define the frequency curve at higher levels. It has been found in some studies that although homogeneity is apparently established at the 10-year level, the individual curves show wide and sometimes systematic differences at higher levels.

(3) In the use of the index-flood method, it has been accepted that within a flood-frequency region, frequency curves may be combined for all sizes of drainage areas, excluding only the largest. Although the variation in the slope of the frequency curve with drainage area had been investigated at the time of each study, it was studied at the 10-year point where the effect is small. The error of neglecting this drainage-area effect has been reduced by giving separate and special treatment to large streams. Recent studies by the U.S.G.S. for which ratios of less frequent floods were used have shown in all regions where such data are available that the ratios of any specified flood to the mean annual flood will vary inversely with the drainage area. In general, the larger the drainage area, the flatter the frequency curve. The effect of drainage area is relatively greater for floods of higher recurrence intervals.

In applying the index-flood technique in Canada, Collier (6) has provided comments on some of the problems he encountered. It has been found in the foothills area of Alberta that normally the annual flood is due to snowmelt in the high regions usually combined with rainfall from Pacific air masses moving over the mountains from the west. Occasionally, however, moist tropical air is sucked up from the southeast and heavy precipitation occurs on the foothills. Simple flood frequency analysis does not distinguish between these two types of events and, as a result, a plot of the floods on arithmetic-normal probability paper shows a distinct S-shape.

Other comments made by Collier (6) included a discussion of the steps to take when a group of stations fails the regional homogeneity test as well as a discussion of the validity of the index-flood approach.

4.3 Multiple Regression Techniques

Rather than plotting drainage area versus mean annual flood, as in the index-flood method, many investigators (see Benson (3) for a partial list) have studied the relationships between discharges at specified return periods and basin characteristics. In general, the relationships take the form

$$Q_T = f(A^a, B^b, C^c \ldots Z^z) \qquad 4.10$$

where A, B, C...Z are independent variables and a, b...z are constants derived by multiple regression analysis. Since a similar relationship, but with different exponents, will be found for the mean annual flood, the ratio of Q_T to the mean annual flood will not be a dimensionless constant as assumed in the index-flood method, but will be proportionate to the basin characteristics. A further advantage of this multiple regression modification to the index-flood method is that it obviates the necessity of assuming any underlying distribution for the flood peaks.

Many different procedures are available for determining the relevent parameters in Equation 4.10 including the simple linear regression, multiple linear regression, forward, backward and stepwise procedures. Packages of computer programs for these procedures are commonly available as library functions at computer centres. Wampler (29) has tested and compared more than 20 linear least squares computer programs.

De Coursey (13) used a canonical correlation procedure to select watershed characteristics and then derived a multiple regression matrix relationship

$$Q = aA + b \qquad 4.11$$

where Q is a column vector of peak flows at various return periods, A is a column vector of watershed characteristics and a and b are respectively a matrix and a vector of regression coefficients. This approach preserves the intercorrelation between the dependent variables in vector Q.

An alternate approach is to regress the basin characteristics not with flood magnitudes at given return periods, but with the parameters of a chosen probability distribution. As an example, if the lognormal probability distribution is used (16) then the mean and standard deviation of the annual events should be used in the correlation.

4.4 Square Grid Method

The definition of areal runoff in geographical areas exhibiting basically similar hydrological characteristics can be facilitated by employing the square grid technique as proposed by Solomon et al (26). As implied in the term square grid this technique entails dividing the study area into a uniform grid, the squares of which are identified in cartesian co-ordinates. Each square grid has associated with it a set of parameters such as the elevations of each corner and centre point, the percentage areas of the square covered by lakes, swamps, forests, urban areas etc., soil types, indicator of bedrock, etc., derived from topographic maps and other information sources such as soil surveys. From this basic data other physiographic characteristics such as mean slope, azimuth of slope, barrier height in different directions, distance to the sea in different directions can be easily derived for each square. Those square grids containing meteorologic or hydrometric stations will also have associated with them data on mean annual precipitation and temperature and streamflow respectively. Each grid square also carries identification of up to four streamflow directions into and out of adjacent squares so that drainage basin data can be accumulated and averaged automatically.

The grid interval determines to a large extent the accuracy of the results since the finer the grid the more basic data is available. Nevertheless, for a given set of conditions, the gain in accuracy obtained by further decreasing the grid interval diminishes greatly beyond a certain value of the interval, and a further increase of the number of squares is not warranted. In general, the optimum grid interval is determined by the size of the area, the size of the individual drainage basins considered, the details of the available data, the computer characteristics, the purpose and budget of the study, etc. For usual problems, grid intervals between about 1 and 10 km. can be considered. In Canada grid sizes of 10 km, 5 km and 4 km have been used in studies and for the mountainous areas of British Columbia, physiographic data have been abstracted using a 2 km grid interval.

The steps involved in estimating the distribution of mean annual runoff using the square grid method are as follows:

(1) Using data at selected meteorological stations a regression equation is established between mean annual temperature at the stations and corresponding square grid physiographic characteristics. This equation is used with the data file of square grid physiographic characteristics to estimate the mean annual temperature in the remaining squares. A similar analysis is used to make a preliminary estimate of mean annual precipitation in each square.

(2) Preliminary evaporation is estimated for each square using temperature, preliminary precipitation and Turc's equation (28)

$$E = P/(0.9 + P^2/L^2)^{\frac{1}{2}} \qquad 4.12$$

where E is evaporation in mm., P is precipitation in mm., T is temperature in degrees Centigrade, L is defined as

$$L = 300 + 25T + 0.5T^2 \qquad 4.13$$

Using this equation, E = P for P^2/L^2 less than 0.1 mm^2.

(3) The preliminary estimates of precipitation and evaporation as described above are used to calculate a preliminary value of mean annual runoff for each square of the watersheds having flow data as:

$$\text{Runoff} = \text{Precipitation} - \text{Evaporation} \qquad 4.14$$

(4) The square grid runoff as established under step (3) is used to compute preliminary average runoff for each basin having flow records, and, coefficients (K) representing the ratio between the recorded and computed average flow are established.

(5) A new precipitation value is computed for each square of the basins having flow data, using K as a correction factor:

$$\text{Precipitation (corrected)} = K \times \text{Runoff} + \text{Evaporation} \qquad 4.15$$

The entire error is attributed to precipitation.

(6) Using the corrected values of the precipitation in each square and the precipitation data at rain gauging stations, a new regression equation is established between precipitation and physiographic factors. Data at rain gauging stations is weighted 10 times larger than the precipitation estimates in each square.

(7) The procedure is then repeated as often as is required to obtain K values as close to 1 as is considered reasonable.

(8) Once square grid runoff, precipitation and evaporation values are finalized in gauged basin areas, the runoff and precipitation distributions in ungauged areas are estimated using regression equations between the final precipitation and runoff in gauged areas and physiographic data.

The end result of the square grid technique is a data file of mean annual temperature, precipitation and runoff for each grid square. Thus estimates of these parameters have been transferred from gauged to ungauged basins using physiographic characteristics as the transfer media. Solomon et al (26) originally applied the square grid method to Newfoundland but it has since been extended to cover all of Canada except for northern Ontario and the Arctic Archipelago (27).

Pentland and Cuthbert (25) have described a method by which the square grid technique was extended for the generation of synthetic streamflow traces. The operational hydrology model proposed by Young and Pisano (32) was used because of the small number of statistical parameters required. This model uses only estimates of a single variance-covariance matrix and a single lag 1 covariance matrix whereas most other models require monthly matrices.

Young and Pisano's model (32) is set up as follows:

(1) The available streamflow data are logarithmically

transformed.

(2) The transformed data are standardized on a monthly basis to eliminate the annual periodicity.

(3) The generating equation

$$X_{i+1} = AX_i + Be \qquad 4.16$$

is then used.

For m stations, X_{i+1} and X_i are standardized flows in successive time periods (m x 1 matrices). A and B are m x m matrices to be defined and e is an m x 1 matrix of random components.

(4) The matrix M_o is the variance-covariance matrix, and M_1 is the covariance matrix with a lag of 1 time period.

(5) The matrix A can be defined from the equation

$$M_1 = AM_o \qquad 4.17$$

(6) The matrix B can be calculated by solving the equation:

$$B\ B^T = M_o - M_1\ M_o^{-1}\ M_1^T \qquad 4.18$$

(7) After having generated standardized variables, the data are destandardized, and the inverse logarithmic transform applied.

If one or more of the stations has no recorded data, monthly means and standard deviations can be estimated by regression analysis with basin physiographic characteristics derived from the square grid data.

$$\mu = K_1 A^{a_1} B^{b_1} C^{c_1} \ldots Z^{z_1} \qquad 4.19$$

$$\sigma = K_2 A^{a_2} B^{b_2} C^{c_2} \ldots Z^{z_2} \qquad 4.20$$

where μ and σ are the monthly mean mean streamflow and standard deviation, A, B....Z are physiographic characteristics and K_1, K_2, a_1,b_1....z_1, a_2, b_2....z_2 are regression constants.

The other parameters required by the generating model are variances, covariances and lagged covariances. All recorded data is subjected to a logarithmic transform, and standardized on a monthly basis. In order to estimate covariances for the streams with no recorded data, a multiple regression equation is then established between covariances (representing cross correlations between stations) and the differences in physiographic characteristics for all pairs of gauged streams in the region.

In the covariance matrix with a lag of one time period, elements of the diagonal (representing serial correlation for each stream) for gauged stations can be calculated directly, and can be estimated for ungauged stations by regression analysis. The remainder of this matrix can be estimated as the product of its diagonal and the variance covariance matrix calculated earlier.

Having estimated the missing parameters with the square grid approach, and further regression analysis, the Young and Pisano (32) model was applied directly by imposing the estimated parameters at the appropriate time.

Pentland and Cuthbert (25) tested the generation procedure on five streams in the northeast of the Province of New Brunswick. Comparisons of simulated and recorded monthly means, standard deviations, serial and cross correlations and firm flows showed good agreements.

A procedure for using the square grid technique to estimate events at required return periods on ungauged streams has been described by Kite (16) for the Mackenzie River area, Northwest Territories. Basically, equations similar to 4.19 and 4.20 were developed relating the mean and standard deviation of the annual maximum instantaneous flows to the square grid physiographic data for those streams which are gauged. The relationships developed were then extended to ungauged streams and, assuming a lognormal probability distribution of annual extremes, event magnitudes at any required return period were calculated. Any other probability distribution thought suitable could have been used in place of the lognormal.

Kouwen (18) has described an advanced model, based upon square grid techniques, used for the simulation of complete watersheds. The model allows forecasts of hydrographs to be incorporated based on weather forecasts especially with regard to the prediction of flood peaks.

The basic input to the model consists of topographic data such as streambed elevations and landslope, drainage channel directions, watershed boundary coordinates, precipitation records, streamflow records and a soil permeability index.

The program is also set up in such a way that precipitation data from radar, and snow pack and soil moisture measurements from satellite can be included. The principle characteristic of the simulation is that runoff passes through successive 1 km x 1 km square elements from higher to lower elevations. For each element there exist relationships between channel capacity and drainage area, surface storage and channel inflow, surface storage and infiltration, subsurface storage and channel inflow, and channel storage and channel discharge.

4.5 Use of Standard Frequency Distributions

In order to remove the necessity for personal judgement in drawing the preliminary frequency curves and to provide a means of computing confidence limits on the regional frequency curve in the index-flood method, Collier (6) produced an alternative regional analysis procedure. This procedure is recommended for regions where the Gumbel (Type I extremal) distribution produces reasonably reliable individual flood frequency curves. The procedure is described briefly below:

1. All stations in the region with 10 or more years of record (either natural flow or with minor regulation only) are selected. Stations with less than 10 years of record would usually be discarded, and most of the selected stations should have at least 15 years of record.

2. A frequency curve covering the range up to the 100-year flood is constructed by the Gumbel (Type I extremal) method for each of the individual stations. For the purpose of this discussion these will be called the preliminary curves. Confidence limits are constructed on each of the preliminary curves using the degree of confidence required in the regional curve.

3. A homogeneity test is carried out exactly as in the index-flood method. For the purpose of this discussion it will be considered that each station passes the test and the region has therefore been

demonstrated to be homogeneous.

4. The preliminary curves are considered to be a sample composed of a number of different estimates of the same rɘgional curve. The estimates are averaged by the method described in the next paragraph to obtain the required estimate of the regional curve.

5. The averaging procedure can be carried out only if the preliminary curves are reduced to dimensionless terms (to remove the effect of the different sizes of the drainage basins concerned). This is accomplished by computing a set of flood ratios (ratio of flood to mean annual flood) for each of the stations over a range of arbitrarily selected recurrence intervals. The data for computing the ratios are read from the preliminary curves.

6. For each of the selected recurrence intervals, the mean of the ratios from all the stations is computed. The resulting means are the flood ratios for the regional curve. These are plotted on Gumbel paper (with arithmetic ordinate scale) and the best-fit straight line drawn through them. The resulting line is taken as the required regional frequency curve.

7. To compute confidence limits for the regional curve, a recurrence interval (say 50 years) is selected arbitrarily and the width of the confidence band at this interval is read off each of the preliminary curves. The width is taken as the vertical distance between the preliminary curve and the upper (or lower) band and it is expresses in cfs. The resulting figures are divided by the appropriate mean annual floods to produce a set of ratios, which

are defined as the "errors" in the individual curves. The errors are combined by computing the square root of the sum of their squares and dividing this square root by the number of stations. The resulting ratio is taken as the "error" in the regional curve or, in other words it is the width of the confidence bands for the regional curve at the selected recurrence interval (50 years in this case). The procedure is repeated at another recurrence interval (say 5 years). The "errors" from the two sets of computations are plotted on the regional curve by laying them off at the appropriate recurrence intervals in a vertical direction either side of the main curve. The resulting points are joined by straight lines to produce the required confidence bands for the regional curve. Note that these bands represent the same degree of confidence as was used in computing the confidence limits for the preliminary curves.

8. Having obtained the dimensionless regional frequency curve, complete with confidence limits, it is necessary to introduce a relationship between mean annual flood and basin characteristics so that estimates may be made for ungauged drainage basins. In a study of the Province of Nova Scotia using Collier's procedure, Coulson (9) ran stepwise linear regressions of mean annual flood versus drainage area, size and position of lakes and swamps, main channel slope, average basin elevation, mean barrier elevation and mean annual precipitation. He ended up with an equation of the form

$$\bar{Q} = f(A_u + \lambda^k A_c) \qquad 4.21$$

where A_u and A_c are the drainage areas uncontrolled and controlled by lakes and swamps respectively,

$$\lambda = \frac{A_c - A_L}{A_c} \qquad 4.22$$

where A_L is the total surface area of major lakes and swamps, and k is a constant optimised by minimising the standard error of $\bar{Q}$, the mean annual flood.

Collier and Nix (7) used a similar approach in a flood frequency study of the New Brunswick-Gaspé region.

The principal advantage of Collier's alternative procedure for the regional frequency curve is that since no personal judgement is involved, the entire procedure can be programmed for computer. Although Collier (6) described the alternative procedure utilising a Type I extremal distribution there is no reason why any other type of distribution thought suitable could not be used.

Cruff and Rantz (11) have described the adaptations made by U.S. agencies to use the lognormal, extremal Type I (Gumbel) and Pearson Type III distributions in regional anlaysis. Basically the procedures used consist of the following steps:

(1) The mean and standard deviation of the peak discharge data at each gauging station are computed for the available periods of record. In the case of the lognormal distribution the means and standard deviations of the logarithms of the peak discharge data are computed in the procedure described by Cruff and Rantz (11) but this is not strictly necessary (see chapter 3).

(2) The computed statistical parameters are then adjusted to a standard base period by computing linear correlations between concurrent peak discharges for a long-term station and the short-term stations. Then

$$\sigma_{1b} = \sigma_{1a} + (\sigma_{2b} - \sigma_{2a}) \cdot R^2 . \sigma_{1a}/\sigma_{2a} \qquad 4.23$$

and

$$\mu_{1b} = \mu_{1a} + (\mu_{2b} - \mu_{2a}) \cdot R^2 . \sigma_{1b}/\sigma_{2b} \qquad 4.24$$

where μ and σ are the means and standard deviations respectively; subscript 1 refers to the short-term station, 2 to the long-term station, subscript a refers to the short-term period and b refers to the base period; and R is the coefficient of correlation between 1 and 2.

If these equations are derived by standardising the means and standard deviations of the two periods of record at each station then the R^2 is not statistically correct and should be omitted.

In the case of the Pearson Type III distribution, for which skews are needed, Equations 4.23 and 4.24 are used to generate events at the short-term stations to complete the record for all years of the base period by using the relationship:

$$x_1 = \mu_{1b} + R\ (x_2 - \mu_{2b}) . \sigma_{2b}/\sigma_{1b} \qquad 4.25$$

where x_1 is the peak discharge to be estimated at a short-term station, x_2 is the peak discharge measured at the long-term station and the other parameters are as previously defined. When the full number of annual events is available for each of the short-term stations the coefficients of skew are computed for each station.

(3) The parameters of the distribution (mean, standard deviation and for Pearson Type III, skewness) are then related to the basin and climatologic characteristics by multiple linear regression equations as explained in section 4.3 of this chapter.

(4) For any site the mean, standard deviation (and, if necessary, coefficient of skew) can then be determined from the derived regression equations, and the event magnitude at return period T can be obtained from

$$x_T = \mu + K.\sigma \qquad 4.26$$

where K is the frequency factor. As explained in Chapter 3 the frequency factor can be developed in terms of T for each distribution and tables are commonly available.

4.6 Regional Record Maxima

Suppose that there exists a set of n independent identically distributed concurrent series each containing k extreme events x_{ij} i = 1,n; j=1,k . If the maximum event of each of the n series is abstracted and ordered from highest to lowest in a new series y_i, i = 1,n then the probability, $P(y>y_i)$, that another event y exceeds the ith event in the series of maxima y_i is given by Conover and Benson (8) as:

$$P(y>y_i) = \sum_{m=o}^{i-1} n!\ /[(n-m)!\ k \prod_{j=o}^{m} (n+1/k-j)] \qquad 4.27$$

As an example, Carrigan (5) has shown that for the three series of four events, x_{ij},

3	69	3
38	24	48
17	61	60
32	30	83

the series of maximum events, y_i, is (83, 69, 38) and the probability, P(y>69), that another event y exceeds the second largest event in the series of maxima is

$$P(y>69) = \sum_{m=o}^{1} 3!\ /[(3-m)!\ 4 \prod_{j=o}^{m} (3+1/4-j)] \qquad 4.28$$

$$P(y>69) = 0.180 \qquad 4.29$$

The analogy to be made here is with annual maximum streamflows recorded at a set of gauging stations within a hydrologically homogeneous region. Within this homogeneous region it is a reasonable assumption that the same probability distribution is applicable to the records of maximum events on each stream. Just as in the index-flood method the different records could be reduced to an identical distribution by normalising with the computed mean annual floods. The procedure outlined above then takes the n independent samples of k events and forms a sample size nk; thus probabilities can be computed associated with return periods of nk years instead of only n years. The catch is that streamflow records are not independent but are quite strongly cross-correlated. This reduces the maximum return period available from nk to f(R)nk where f(R) is some function of the correlation coefficient, R, between streamflow records. The expression f(R) varies between 1 when R = 0 (independent records) to 1/n when R = 1 (identical records).

The probability of another random event, y, exceeding one of the ordered record maxima, y_i, cannot be determined anlaytically when the records are not independent. By assuming that the exceedence probability is independent of the identical distribution of the records, Carrigan (5) has derived the probability by data generation. Using the normal distribution for simplicity the generation model used by Carrigan (5) is

$$x = B\varepsilon \qquad 4.30$$

where X is an $n \times k$ matrix of generated events, B is an $n \times n$ principal component matrix and ε is an $n \times k$ matrix of independent normally distributed random numbers with zero mean and unit variance. The principal component matrix B is derived from the correlation matrix R of the n records of k hydrologic events as follows:

$$B = E\lambda \qquad 4.31$$

where E is an $n \times n$ matrix of eigenvectors obtained from R and λ is an $n \times n$ matrix for which the diagonal elements are the square roots of the eigenvalues of R and the off-diagonal elements are zero.

After generation of X, the n maxima are selected and put in order of magnitude and the exceedence probabilities computed using a digital approximation to the normal distribution.

In summary, the method extracts from the records of a series of gauging stations a matrix of the inter-station correlation. This correlation is then incorporated into a large number of generated events from which extreme probabilities can be measured. In effect the method converts the spatially-distributed information into time-distributed information on extreme events.

4.7 Single Station and Regional Information Content

General

A time series may or may not consist of observed outcomes which are independent of one another. Streamflow is a hydrologic variable whose observations, equally spaced in time, are not necessarily randomly distributed. Because of natural storage such as groundwater, lakes, swamps and annual persistance as well as manmade factors such as reservoirs, the stochastic precipitation variable becomes modulated and serially correlated. This means that each unit of streamflow data does not contain totally new and independent information. A data tends to repeat some of the previously obtained information. A data set of N units may therefore only contain a lesser number, N_e, of effective data units.

The early stage of any regional analysis procedure calls for the examination of available data. It is usual that some gauging station records will be longer than others within the required region and often, after selection of a base period of record, it will be necessary to fill in gaps in some records and extend other records to the full base period. This provision of missing data can have two general purposes:- (a) To provide estimates of event magnitudes in order to better obtain plotting positions of recorded events. This procedure is used in the index-flood regional analysis technique where the estimated event magnitudes are not themselves used at all, they are there merely to improve estimates for the recorded events. (b) To improve estimates of the parameters of a theoretical distribution of recorded events such as the mean and standard deviation.

Data are not only serially correlated, but because streamflows in rivers within a region are affected by the same conditions of precipitation and radiation, simultaneous observations of streamflow in different rivers will not be independent observations but will contain an overlap of information. Thus a region of n gauging stations may reduce to a much smaller number, n_e, of effective stations or equivalent independent stations.

Single Station Information Content

The purpose of maintaining records of precipitation, stage, streamflow, etc., is to extract from the recorded observations information on the parameters of the underlying distribution. Matalas and Langbein (21) defined the amount of information given by a statistical estimate, I, as the reciprocal of the variance of the estimate. Considering the mean, μ, of a random series of N events, x_i, i = 1...N,

$$\mu = \sum_{i=1}^{N} x_i/N \qquad 4.32$$

an estimate of the variance of μ is given by

$$\text{var } \mu = \sigma^2/N \qquad 4.33$$

where σ^2 is an estimate of the population variance of the random time series. Defining the random series as the standard, the relative information content about the mean of any other time series with variance of the estimated mean, σ_μ^2, is

$$I_\mu = (\sigma^2/N)/\sigma_\mu^2 \qquad 4.34$$

referred to the random series. Since variances are always positive I_u can vary from zero to plus infinity. If I_u is less than unity the time series being tested conveys less information about the mean than a random series of the same length. If I_u is greater than unity then the time series contains more information about the mean than an equal length random series.

Many time series exhibiting persistance, such as streamflow, can be described by a simple first order linear Markov model (31) such as:

$$x_{i+1} = R_1 x_i + \varepsilon_{i+1} \qquad 4.35$$

where x_i and x_{i+1} are the variable values at time i and i+1 respectively, ε_{i+1} is a random component independent of x and R_1 is the first order serial or autocorrelation coefficient where, in general, the k-th order autocorrelation coefficient is defined (30) as:

$$R_k = \frac{\sum_{i=1}^{n-k} (x_i - \mu)(x_{i+k} - \mu)}{(N-K)\sigma^2} \qquad 4.36$$

where N is the number of observations and μ and σ are the sample estimates of the mean and standard deviation of the time series.

For a first order linear Markov model the variance of the mean, σ_μ^2, is given (21) by

$$\sigma_\mu^2 = \frac{\sigma^2}{N}\left[\frac{1+R_1}{1-R_1} - \frac{2}{N}\ \frac{R_1(1-R_1^N)}{(1-R_1)^2}\right] \qquad 4.37$$

From Equation 4.34 the relative information content on the mean is

$$I_\mu = \left[\frac{1+R_1}{1-R_1} - \frac{2}{N} \frac{R_1 \, (1-R_1^N)}{(1-R_1)^2} \right]^{-1} \qquad 4.38$$

which is less than unity for $R_1 > 0$.

If a number of independent observations of a random time series, N_e, contain the same information content about the mean as the number N observations of the Markov model then

$$I_\mu = Ne/N \qquad 4.39$$

and

$$N_e = N \left[\frac{1+R_1}{1-R_1} - \frac{2}{N} \frac{R_1 \, (1-R_1^N)}{(1-R_1^2)} \right]^{-1} \qquad 4.40$$

Two-Station Transfer of Information

Interstation transfer of information is commonly used in regional analysis to fill in missing data or to extend short time series to a longer common base period. The method used in the index-flood method of regional analysis (12) is as follows:

A graph is drawn of the flow at one station versus the flow at the other station for each year of the common period of record. A straight line is fitted by eye through the coordinate points and this line is then used to extend the shorter period of record. This process is a simplification of the least squares fitting of a linear regression equation of the type

$$y = mx + c \tag{4.41}$$

where x and y are the annual maximum flows at the two stations, m is the slope and c is the intercept of the straight line. The missing event magnitudes are then estimated from this regression line and the total events, recorded and estimated are placed in order of decreasing magnitude. Plotting positions (see Chapter 2) are assigned to the recorded events on the basis of the total number of events and the estimated events are then discarded and used no further.

Many investigators (19) (24) have found that the logarithms of hydrologic events are better correlated than the recorded events and have used equations such as

$$\ln y = m_1 \ln x + c_1 \tag{4.42}$$

where x and y may be the recorded events or the deviations of the recorded events from some mean value.

The question arises as to whether the estimated events actually increase the information content of the shorter time-series i.e. does the extended data provide better estimates of the population distribution parameters than the originally recorded series? Langbein (19) has shown that to improve the significance of the mean of a time series the effective period of record, N_e, of a combined recorded and estimated record must be greater than N_1, the number of years of recorded data, where

$$N_e = \frac{N_1 + N_2}{1 + \frac{N_2}{N_1 - 2}(1-R^2)} \qquad 4.43$$

N_2 is the number of years of estimated data and R^2 is the coefficient of determination of the simple linear regression used to provide the estimated data.

If two random normally distributed time series x, of length $N_1 + N_2$, and y, of length N_1, are linearly related with a simple linear correlation coefficient R, and the time series x is used to extend time series y by N_2 data points, then the variance, σ_μ^2, of the weighted mean of series y, μ_y, where

$$\mu_y = \frac{N_1\mu_1 + N_2\mu_2}{N_1+N_2} \qquad 4.44$$

is given (21) by

$$\sigma_\mu^2 = \frac{\sigma_y^2}{N_1}\left[1 - \frac{N_2}{N_1+N_2}\left\{R^2 - \frac{(1-R^2)}{(N_1-3)}\right\}\right] \qquad 4.45$$

where μ_1 and μ_2 are the sample means of time series y based on N_1 observations and N_2 regression estimates respectively and

σ_y^2 is the variance of time series y based on the N_1 observations.

Referring back to Equation 4.34 the information content on the mean of the extended time series y is seen to be:

$$I_\mu = \left[1 - \frac{N_2}{N_1+N_2} \left\{ R^2 - \frac{(1-R^2)}{(N_1-3)} \right\} \right]^{-1} \qquad 4.46$$

and the effective number of observations, N_e, is given by:

$$N_e = (N_1+N_2) \left[1 - \frac{N_2}{N_1+N_2} \left\{ R^2 - \frac{(1-R^2)}{(N_1-3)} \right\} \right]^{-1} \qquad 4.47$$

For the cross-correlation to provide additional information on the mean, $I_u > 1$ and from Equation 4.46

$$\frac{-N_2R^2}{N_1+N_2} + \frac{N_2}{(N_1+N_2)(N_1-3)} - \frac{N_2R^2}{(N_1-3)} < 0 \qquad 4.48$$

from which

$$R^2 > \frac{1}{N_1-2} \qquad 4.49$$

Similarly, Fiering (14) concluded that correlation should not be used to augment time series for estimation of the variance unless the computed information content on the variance,

$I_{\sigma 2}$, is greater than unity, where

$$I_{\sigma^2} = \left\{ 1 + \frac{N_2}{2(N_1+N_2-1)^2} \left[2A(N_1-1) + (N_2+2)(N_1-1)B + (N_1+N_2 - 1)(N_1-1)C + (N_1+1)(2N_1+N_2-2) \right] \right\}^{-1} \qquad 4.50$$

in which

$$A = (N_1-1)R^4 + (N_1+4)R^2(1-R^2) + \frac{N_1+1}{N_1-3} (1-R^2)^2 \qquad 4.51$$

$$B = R^4 + \frac{6R^2(1-R^2)}{N_1-3} + \frac{3(1-R^2)}{(N_1-3)(N_1-5)} \qquad 4.52$$

and

$$C = \frac{2(N_1-4)(1-R^2)}{N_1-3} \qquad 4.53$$

Fiering (14) concluded that, in general, the estimate of the population variance will be improved if $R > 0.85$.

If the two time series x and y are not random but are serially correlated then for an equal number of observations, N, the effective number of data points, N_e, has been given by

Yevjevich (31) as:

$$N_e = N/(1+2R_1R_1' + 2R_2R_2' + \dots 2R_{n-1} R'_{n-1}) \qquad 4.54$$

where R_k and R_k' are the kth order autocorrelation coefficients of the x and y time series respectively. This equation is only useful if all periodicities have been removed from both time series. If the equal-length time series x and y can be described by first order linear Markov models then the effective number of observations is given by (31):

$$N_e = N \left[\frac{1 - R_1R_1'}{1 + R_1R_1'} \right] \qquad 4.55$$

If x and y are two first order linear Markov models of length $N_1 + N_2$ and N_1 respectively and x is correlated with y to provide N_2 regression estimates for y then the relative information content for the mean of the augmented time series varies with N_1, N_2, R_1, R_1' and R in a complex fashion. Assuming that $R_1 = R_1'$, Matalas and Langbein (21) have tabulated values of I_u for different values of these variables.

Equations such as 4.41 and 4.42 yield event magnitudes on the regression line and, although this does not affect the estimate of the mean, it does induce a bias in the estimate of the variance. To overcome this bias it is necessary to introduce into the generating equation a random variable with mean zero and variance $(1-R^2)\sigma^2$ where

σ^2 is the varriance of the recorded series. The true regression equation thus becomes:

$$y = \mu_y + m(x-\mu_x) + (1 - R^2)^{\frac{1}{2}} \sigma_y \varepsilon \qquad 4.56$$

where ε is a normally distributed (0,1) random number. The term $(1-R^2)^{\frac{1}{2}}\sigma_y\varepsilon$ is the random component of the generated time series and in communications theory is referred to as noise.

Under the assumptions that (a) events are independently distributed in time, (b) the concurrent events for the two sequences have a joint normal distribution, (c) the relation between the concurrent events is defined by a linear regression, and (d) no changes occur in the hydrologic regimes with which the sequences are associated, Matalas and Jacobs (22) have evaluated the reliability of estimates of population parameters under conditions of noise and no-noise.

Matalas and Jacobs (22) recommended the use of the following equations to compute the mean, μ_y, and variance, σ_y^2, of an extended time series:

$$\mu_y = \bar{y}_1 + \frac{N_2}{N_1+N_2} \cdot m(\bar{x}_2 - \bar{x}_1) \qquad 4.57$$

and

$$\sigma_y^2 = \frac{1}{N_1+N_2+1} \left\{ (N_1 - 1)\, S_{y_1}^2 + (N_2-1) m^2 S_{x_2}^2 \right. \qquad 4.58$$

$$\left. + (N_2-1)\, \alpha^2\, (1-R^2) S_{y_1}^2 + \frac{N_1N_2}{(N_1+N_2)}\, m^2 (\bar{x}_2-\bar{x}_1)^2 \right\}$$

where $\bar{y}_1$ and $S_{y_1}^2$ are the mean and variance of the recorded events in the augmented series, $\bar{x}_1$ is the means of the concurrent augmenting series, $\bar{x}_2$ and $S_{x_2}^2$ are the mean and variance of the total number of events in the augmenting series, and

$$\alpha^2 = \frac{N_2(N_1-4)(N_1-1)}{(N_2-1)(N_1-3)(N_1-2)} \qquad 4.59$$

Equations 4.57 and 4.58 should only be used if the interstation correlation coefficient, R, is greater than the critical values given in Tables 4.2, 4.3 and 4.4

Table 4.2

Critical Minimum Values of R for Estimation of the Mean[1]

N	10	15	20	25	30
R	0.35	0.28	0.24	0.21	0.19

[1] From Matalas and Jacobs (22)

Table 4.3

Critical Minimum Values of R for Estimation of the Variance, Including Noise Component[1]

			N_1		
N_2	10	15	20	25	30
10	0.65	0.54	0.52	0.42	0.38
15	.65	.54	.51	.42	.39
20	.65	.54	.51	.42	.39
25	.65	.54	.50	.42	.39
30	.65	.54	.50	.42	.39

Table 4.4

Critical Minimum Values of R for Estimation of the Variance, Excluding Noise Component[1]

			N_1		
N_2	10	15	20	25	30
10	0.73	0.63	0.70	0.76	0.76
15	.75	.77	.79	.80	.80
20	.76	.79	.81	.81	.82
25	.78	.80	.84	.83	.81
30	.77	.80	.82	.83	.84

Regional Transfer of Information

As well as considering the transfer of information from a long-term station to an adjacent short-term station, hydrologists

[1] From Matalas and Jacobs (22)

are often interested in studying how a hydrologic variable, such as river discharge, varies with the physical parameters describing the drainage area. The variation may be studied by assembling the data for many gauging stations and using regression analysis to define a relationship betwween the hydrologic variable and the physiographic characteristics.

In a given region, however, rivers may rise in response to a rain storm that affects all the rivers in the region and, at another time, may be low due to a common lack of rainfall. Thus the flows of different streams are affected by common causes and are therefore not independent but cross-correlated.

If a number n of hydrometric gauging station records within a hydrologically homogeneous region are intercorrelated, then the effective number of stations or equivalent number of independent gauging stations, n_e can be derived as follows (31): If the mean and variance of the observations at the jth gauging station are given by:

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij} \qquad 4.60$$

$$\sigma_j^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_{ij} - \mu_j)^2 \qquad 4.61$$

where N is the number of observations at each station, then the estimates of the regional mean, μ, and the variance of the regional mean, σ_μ^2, are (28)

$$\mu = \frac{1}{N} \sum_{j=1}^{n} \mu_j \qquad 4.62$$

$$\sigma_\mu^2 = \frac{1}{n} \sum_{j=1}^{n} \sigma_j^2 + \frac{2}{n^2} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} R_{ij}\sigma_i\sigma_j \qquad 4.63$$

where R_{ij} is the cross-correlation coefficient between stations i and j.

Equation 4.63 can be simplified by defining the regional mean cross-correlation coefficient, $\bar{R}$, as:

$$\bar{R} = \frac{2 \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} R_{ij}}{n(n-1)} \qquad 4.64$$

If the time-series are standardised to a common mean of zero and variance σ^2 then Equation 4.63, incorporating Equation 4.64, becomes:

$$\sigma_m^2 = \frac{\sigma^2}{n} \quad [1+\bar{R}(n-1)] \qquad 4.65$$

The relative information content on the regional mean is therefore given by:

$$I_{\mu} = [1+\bar{R}(n-1)]^{-1} \qquad 4.66$$

and the effective number of stations, n_e, or equivalent number of independent stations is:

$$n_e = n/[1+\bar{R}(n-1)] \qquad 4.67$$

In a study of regional flood frequency relations for 164 basins in New England, Benson(4) found that using Equation 4.67 the effective number of gauging stations or equivalent number of independent stations was 3.8. In a later study, Matalas and Benson (20) point out that, assuming the same value of $\bar{R}$, if n were only 20 stations, n_e would be 3.4 and if n were 500, n_e would be 3.8. This illustrates the rapid arrival at the limiting number of independent records. No appreciable increase in information is attained by using 500 stations instead of 20, if they are all within the same region. As a further example, if there are an infinite number of stations and $\bar{R}$ = 0.1, the effective number of stations is only 10.

The theory of regression analysis is based on the assumption amongst others, that the values of the dependent variable are mutually independent. It is apparent, then, that in equations such as:

$$x = a_o + a_1A + a_2B + a_3C + \ldots\ldots \qquad 4.68$$

where x is some streamflow characteristic and A,B,C... are basin physiographic characteristics, the regression theory is impaired since the x values are not independent. Matalas and Benson (20) have investigated this problem very thoroughly and have concluded that the estimation of the regression constants, a_o, a_1.... and the subsequent estimated value $\hat{x}$ are not affected by interstation correlation. If interstation correlation is present, however, the variance of a_o will be larger than if there were no correlation, the variances of a_1, $a_2 \ldots a_n$ will be smaller and the variance of $\hat{x}$ may be larger or smaller.

Considering a set of n gauging stations each of N observations but which are both serially and cross-correlated then Equations 4.54 and 4.67 can be combined to give a total effective number of station-events defined as:

$$N_en_e = Nn/\left\{\left[1+\bar{R}(n-1)\right](1+2\bar{R}_1 + 2\bar{R}_2 + \ldots 2\bar{R}_n)\right\} \qquad 4.69$$

where $\bar{R}_1$, $\bar{R}_2$ are the average serial correlation coefficients of the n time series.

Developing this analysis, the procedure can be used as a means of defining homogeneous hydrologic regions (30), (17).

References for Chapter 4

1. Alexander, G.N., 1954, Some Aspects of Time Series in Hydrology, J. Inst. Engineers (Australia) p. 196.

2. Benson, M.A., 1960, Characteristics of Frequency Curves Based on a Theoretical 1,000-Year Record, USGS Water Supply Paper 1543-A, pp. 51-73.

3. Benson, M.A., 1962, Evolution of Methods for Evaluating the Occurrence of Floods, USGS Water Supply Paper 1580-A.

4. Benson, M.A., 1962, Factors Influencing the Occurrence of Floods in a Humid Region of Diverse Terrain, USGS Water Supply Paper 1580-B.

5. Carrigan, P.H., Jr., 1971, A Flood-Frequency Relation Based on Regional Record Maxima, USGS Professional Paper No. 434-F.

6. Collier, E.P., 1963, Regional Flood Frequency Analysis, unpublished paper, Water Resources Branch, Ottawa.

7. Collier, E.P., and G.A. Nix, 1967, Flood Frequency Analysis for the New Brunswick-Gaspé Region, Technical Bulletin No. 9, Inland Waters Branch, Ottawa.

8. Conover, W.J., and M.A. Benson, 1963, Long-term Flood Frequencies Based on Extremes of Short-term Records, USGS Professional Paper No. 450-E, pp. E159-E160.

9. Coulson, A., 1967, Flood Frequencies of Nova Scotia Streams, Technical Bulletin No. 4, Water Resources Branch, Ottawa.

10. Coulson, A., 1967, Estimating Runoff in Southern Ontario, Technical Bulletin No. 7, Inland Waters Branch, Ottawa.

11. Cruff, R.W., and S.E. Rantz, 1965, A Comparison of Methods Used in Flood Frequency Studies for Coastal Basins in California, USGS Water Supply Paper 1580-E.

12. Dalrymple, T., 1960, Flood Frequency Analyses, USGS Water Supply Paper 1543-A.

13. De Coursey, D.G., 1973, Objective Regionalisation of Peak Flow Rates, Proceedings of Second International Symposium in Hydrology, pp. 395-405, Water Resources Publications, Fort Collins, Colorado 80521.

14. Fiering, M.B., 1963, Use of Correlation to Improve Estimates of the Mean and Variance, USGS Professional Paper No. 434-C.

15. Kalinin, G.P., 1960, Calculation and Forecasts of Streamflow from Scanty Hydrometric Readings, Trans. Interregional Seminar on Hydrologic Networks and Methods, Bangkok, 1959, WMO Flood Control Series No. 15, pp. 42-52.

16. Kite, G.W., 1973, Flood Frequency for Mackenzie Highway Culverts, unpublished paper, Water Resources Branch, Ottawa.

17. Kite, G.W., 1973, Serial Correlation as a Measure of Regional Uniformity, unpublished notes, Water Resources Branch, Ottawa.

18. Kouwen, N., 1973, Watershed Modelling Using a Square Grid Technique, Proceedings of the 9th Canadian Hydrology Symposium, Edmonton, Alberta.

19. Langbein, W.B., 1960, Hydrologic Data Networks and Methods of Extrapolating or Extending Available Hydrologic Data, Trans.

Interregional Seminar on Hydrologic Networks and Methods, Bangkok, 1959, WMO Flood Control Series No. 15, pp. 13-41.

20. Matalas, N.C., and M.A. Benson, 1961, Effect of Interstation Correlation on Regression Analysis, Journal of Geophysical Research, Vol. 66, No. 10, pp. 3285-3293.

21. Matalas, N.C., and W.B. Langbein, 1962, Information Content of the Mean, Journal of Geophysical Research, Vol. 67, No. 9, pp. 344-348.

22. Matalas, N.C., and B. Jacobs, 1964, A Correlation Procedure for Augmenting Hydrologic Data, USGS Professional Paper No. 434-E.

23. Matalas, N.C., 1967, Time Series Analysis, Water Resources Research, Vol. 3, No. 3, pp. 817-830.

24. Pentland, R.L., 1967, Extending Streamflow Records, Program No. 20, unpublished paper, Water Resources Branch, Ottawa.

25. Pentland, R.L., and D.R. Cuthbert, 1971, Operational Hydrology for Ungauged Streams by the Grid Square Technique, Water Resources Research, Vol. 7, No. 2, pp. 283-291.

26. Solomon, S.I., T.P. Denouvilliez, C. Cadou and E.J. Chart, 1968, The Use of a Square-Grid System for Computer Estimation of Precipitation, Temperature and Runoff in a Sparsely Gauged Area, Water Resources Research, Vol. 4, No. 5, pp. 919-930.

27. Solomon, S.I., and A.S. Qureshi, 1972, Hydrologic Data Banks - Present Status and Potential, Engineering Journal, Vol. 55, No. 1/2, pp. 9-14.

28. Turc, L., 1954, Le Bilan D'eau des Sols: Relations entre les Precipitations, L'evaporation et L'écoulement, Annales Agronomiques, Vol. 4, pp. 491-595.

29. Wampler, R.H., 1969, An Evaluation of Linear least Squares Computer Programs, Journal of Research of the National Bureau of Standards - B. Mathematical Sciences, Vol. 73B, No. 2, pp. 59-90.

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

31. Yevjevich, V.M., 1972, Probability and Statistics in Hydrology Water Resources Publications, Fort Collins, Colorado.

32. Young, G.K., and Pisano, W.C., 1968, Operational Hydrology Using Residuals, Proc. ASCE, Vol. 94, No. HY4, pp. 909-923.

CHAPTER 5

Risk

5.1 The Need for Risk Analysis

The most important question facing the designer of any hydrologic structure is: what is the risk of failure? The price of failure of a major dam is high and the risk of this occurence must be minimised. A study of over 1600 dams (in(8)) has shown the following causes of failure:

Foundation problems	40%
Inadequate spillway	23%
Poor construction	12%
Uneven settlement	10%
High pore pressure	5%
Acts of war	3%
Embankment slips	2%
Defective materials	2%
Incorrect operation	2%
Earthquakes	1%
	100%

In a more recent study of over 300 dam disasters (8) it was found that roughly 35% of the failures were due to inadequate spillway design. Also of importance here is the study of dam failures noted by the AWWA (3). Inadequate spillway design is usually caused by inadequate design flood analysis and this is the direct concern of the hydrologist. Design floods are estimated either from frequency techniques or as the Probable Maximum Flood.

As noted in Chapter 1, the technique of Probable Maximum Flood, despite its name, is a totally deterministic concept and as such has no risk associated with it. Because there is no proof of the existence of extreme boundaries in the meterorological factors which cause floods (19) the concepts of maximum probable precipitation, maximum probable flood and other similarly named imaginary events may be considered as arbitrary. They are concepts of expediency.

Frequency analysis, on the other hand, accepts events of any magnitude as being possible although as the magnitude increases so the probability of occurrence decreases.

The simplest procedure in the frequency analysis estimation of spillway design floods is to select a return period and use either graphical techniques or a mathematical distribution to derive the corresponding event magnitude. Some of the return periods commonly used for different types of structure are (in(7)):

Major dams with probable loss of life

Earth dam	1000 years
Masonry or concrete dam	500 years

Costly dams with no likelihood of loss of life	500 years
Moderately costly dams	100 years
Minor dams	20 years

In addition, McCaig and Erickson (12) note that in the past it has been common practice to design major dams for floods having theoretical return periods of up to 10,000 years. The ASCE Hydraulics Division Committee on Hydrometeorology (2) has suggested that the Probable Maximum Flood is perhaps equivalent to a design return period of 10,000 years.

This elementary procedure takes no account of the increase of risk with increasing project life or of the economically optimum design.

5.2 Economic Design

A better procedure sometimes used in the design of hydraulic structures establishes the design spillway capacity not only on the magnitude and frequency of possible floods but also on the monetary value of the dam, the unit cost of the spillway and the value placed upon the lives and property of the people downstream of the dam. McCaig and Erickson (12) have provided a very clear description of this method of design using in their example lognormal distributions of fall and spring floods.

If the average annual losses for a particular structure can be expressed as:

$$C_1 = \Sigma \Delta L . P \qquad 5.1$$

where ΔL is the incremental average loss for a particular design flood, x, in dollars and P is the exceedence probability of that design flood; and if the average annual cost of the spillway is given by:

$$C_2 = \Delta x . Q \qquad 5.2$$

where Δx is the incremental cost, in dollars per cfs, of providing spillway capacity for flow Q cfs; then the optimum structure design will occur when

$$C = C_1 + C_2$$

is at a minimum. That is to say, for a particular structure and a set of flood flows there will result a particular value of C. By repeating the same set of flood flows with different structure capacities a graph of C versus capacity or design flood can be obtained.

McCaig and Ericson (12) assumed a lognormal probability distribution for flood events so that:

$$P = \frac{1}{\sqrt{2\pi}\,\sigma_y} \int_y^{\infty} e^{\frac{-(y-\mu_y)^2}{2\sigma_y}} \, dy \qquad 5.4$$

where y is the logarithm of the flood event, x, and μ_y and σ_y are respectively the population mean and standard deviation estimated from the logarithms of the recorded flood events.

Substituting Equations 5.4, 5.1 and 5.2 into Equation 5.3, differentiating and equating to zero, the optimum design capacity, Q_d, can be obtained.

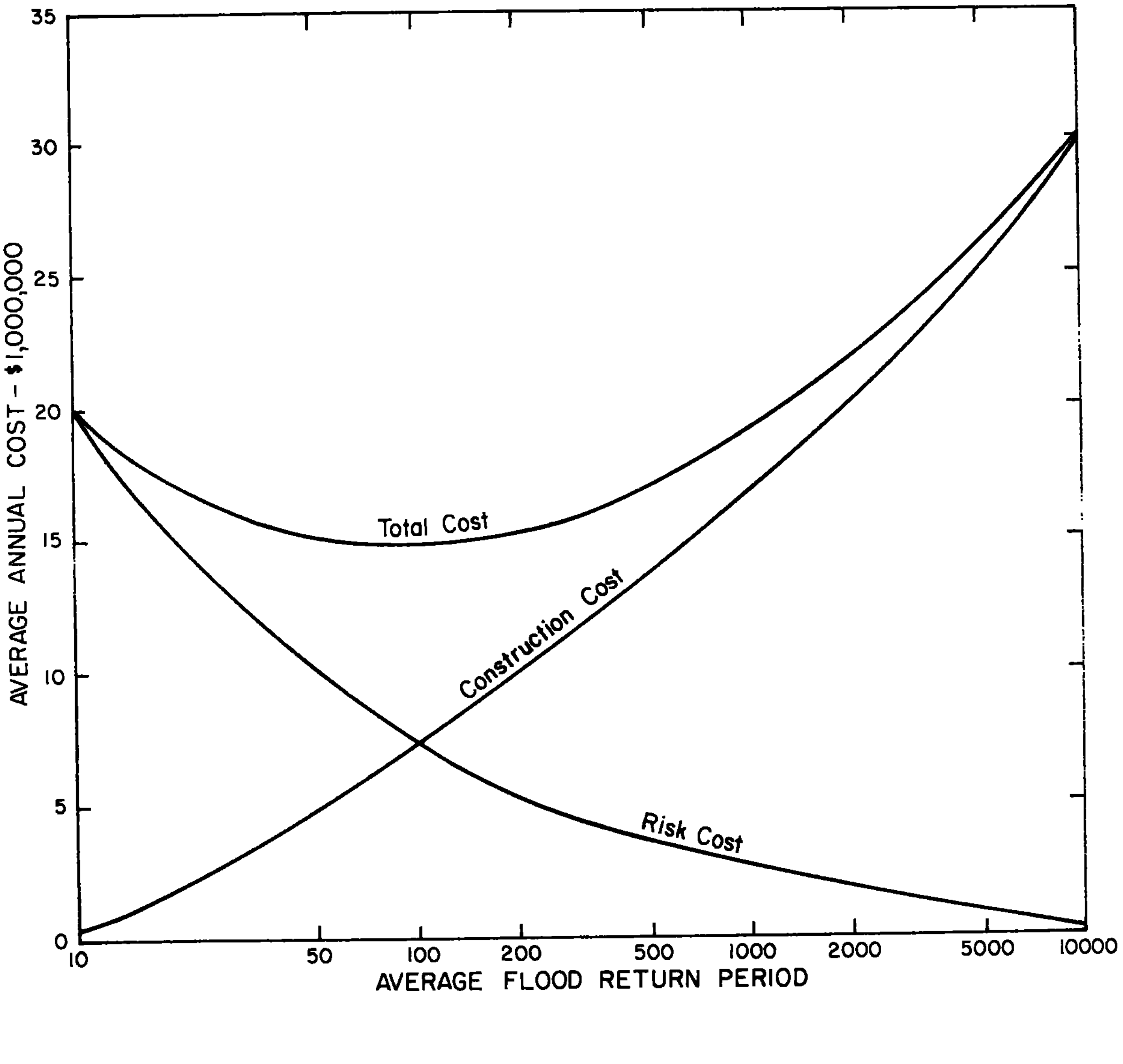
The ASCE (2) has recently described a similar procedure to McCaig and Ericson but designed for the re-evaluation of the spillway capacity of existing dams. A series of alternate project designs are identified by their spillway design floods e.g. the 500 year design project, the 1000 year design project, etc. This series would include the existing project. For each of the possible projects the costs associated with an array of floods with return periods varying from very low to very high are determined.

Damages caused by the various floods to each of the alternate project designs should include upstream damages (in the event of overtopping and subsequent failure of the dam) to recreation, piers, boats, buildings, loss of power, loss of water supply; to the structure itself including dam fill eroded, repair time, powerhouse losses, switchyard losses, etc., and damage downstream of the dam including deaths, injuries, property damage, compensation for loss of water supply, power supply, telephone, road access and lost employment. It is instructive to note that in the ASCE example (2) death was valued at \$150,000, permanent disabling injury at \$200,000 and a non-disabling injury at \$10,000. The property damages should be determined by carrying out a stage-damage analysis using measured flood profiles.

For each project the average annual risk can be calculated by arithmetic strip integration of the area beneath the return period-damage curve. The cost of each alternate project design is known and can be converted to an average annual cost. This cost, sometimes known as the "operating rate" (12), may include items for interest, taxes, depreciation, etc., and normally ranges between 8 and 10 per cent of the total capital cost. Curves of the type shown in Figure 5.1 can then be drawn and the optimum project design determined.

Note that the series of alternate projects might consist of one dam design with floods of successively longer return period being accomodated by a longer spillway, by downstream flood

FIGURE 5.1
AVERAGE ANNUAL COSTS FOR DIFFERENT DESIGNS
(EXAMPLE ONLY)



protection work, by paving the dam top and downstream dam surface to reduce erosion from overtopping, by construction of an upstream reservoir to reduce inflows or by other similar means.

5.3 Risk Design

Neither of the two techniques described so far include the concept of total risk. For any hydraulic structure there is a total risk of failure which can be broken down into the risk of failure of each project component i.e. hydrologic, hydraulic and structural. The risk within any component can then by broken down into true risk and uncertainty. Yen and Ang (18) have used the terms objective risk and subjective risk.

For the hydrologic component, risk is the calculable probability of failure e.g. occurrence of a certain flood, occurrence of a drought, etc. The calculation of risk is based on the assumption that the underlying event distribution is known. As an example, if it is known that flood magnitudes in a particular river valley location follow the lognormal distribution and that the time-distribution of the floods follow a Poisson distribution then the risk that the flood of a certain magnitude will occur in the next five years can be computed exactly.

Uncertainty occurs because the basic data available contains random measurement and computation errors, systematic errors, non-homogeneity in time, loss of information in changing from a continuous record to a discrete data set and so on. These imperfect data are then used to estimate the parameters of the assumed population distribution. Uncertainty generally increases as the variance of the sample data increases and decreases as the sample length increases.

Thomas (16) has evaluated the errors in streamflow estimates made from a continuous stage record while Moss (13) has related the standard error of discharge estimates to the number of streamflow measurments made per year and the associated costs of maintaining the station.

The effect of uncertainty on the parameters of the population distribution can be included in an analysis by computing the standard error of estimate of the particular distribution at the required probability level. Confidence limits around the expected event magnitude can then be calculated.

To summarize this concept, hydrologic risk is made up of basic risk and uncertainty both of which can be evaluated. What cannot be evaluated is the error caused by selecting the wrong distribution to fit the sample data. It is true that the goodness of fit of a distribution, once chosen, can be measured using the Chi-Square or Kolmogorov-Smirnov or similar tests and thus the best-fitting distribution can be selected. Generally, however, the sample data will occupy the central portion of a frequency distribution while the event magnitudes which it is required to compute will be in the extremes so that the best-fitting distribution may not necessarily be the best to use.

The computation of standard errors of estimates for various common distributions has been described in Chapter 3. The remainder of this chapter will cover the calculation of basic risk, the assumption being made that the underlying distribution is known.

Suppose that for a time invariant hydrologic system the probability of occurrence of an event, x, greater than the design event, x_o, during a period of n years is P. Then the probability of non-occurrence, Q, is 1-P.

If this design event has a return period of T years and a corresponding annual probability of exceedence of p then:

$$p = \frac{1}{T}, \qquad 5.5$$

the probability of non-occurrence in any one year is:

$$q = 1 - \frac{1}{T}, \qquad 5.6$$

the probability of non-occurrence in n years is:

$$Q = (1 - \frac{1}{T})^n \qquad 5.7$$

So that, finally, the probability that x will occur at least once in the n years is:

$$P = 1 - (1 - \frac{1}{T})^n \qquad 5.8$$

This is the risk of failure and is based on the assumption of independence of annual events. Yen (17) has tabulated values of T, the required design return period, for various expected project lives, n, and permissible risks of failure, P. Table 5.1 is adapted from Yen. Figure 5.2 is based on the solution of Equation 5.8.

If the series of recorded or measured events are not an annual series but a partial duration series with an average of K observations per year then the probability that the T year event will be equalled or exceeded in n consecutive years is:

$$P = 1 - [1 - 1/T.K]^{nK} \qquad 5.9$$

Figures similar to Figure 5.1 (for which K=1) can be drawn for all values of K, (1).

If failure is associated not with exceedence of the design event but with failure to reach the design event, e.g. a drought design, then the return period must be redefined.

Table 5.1

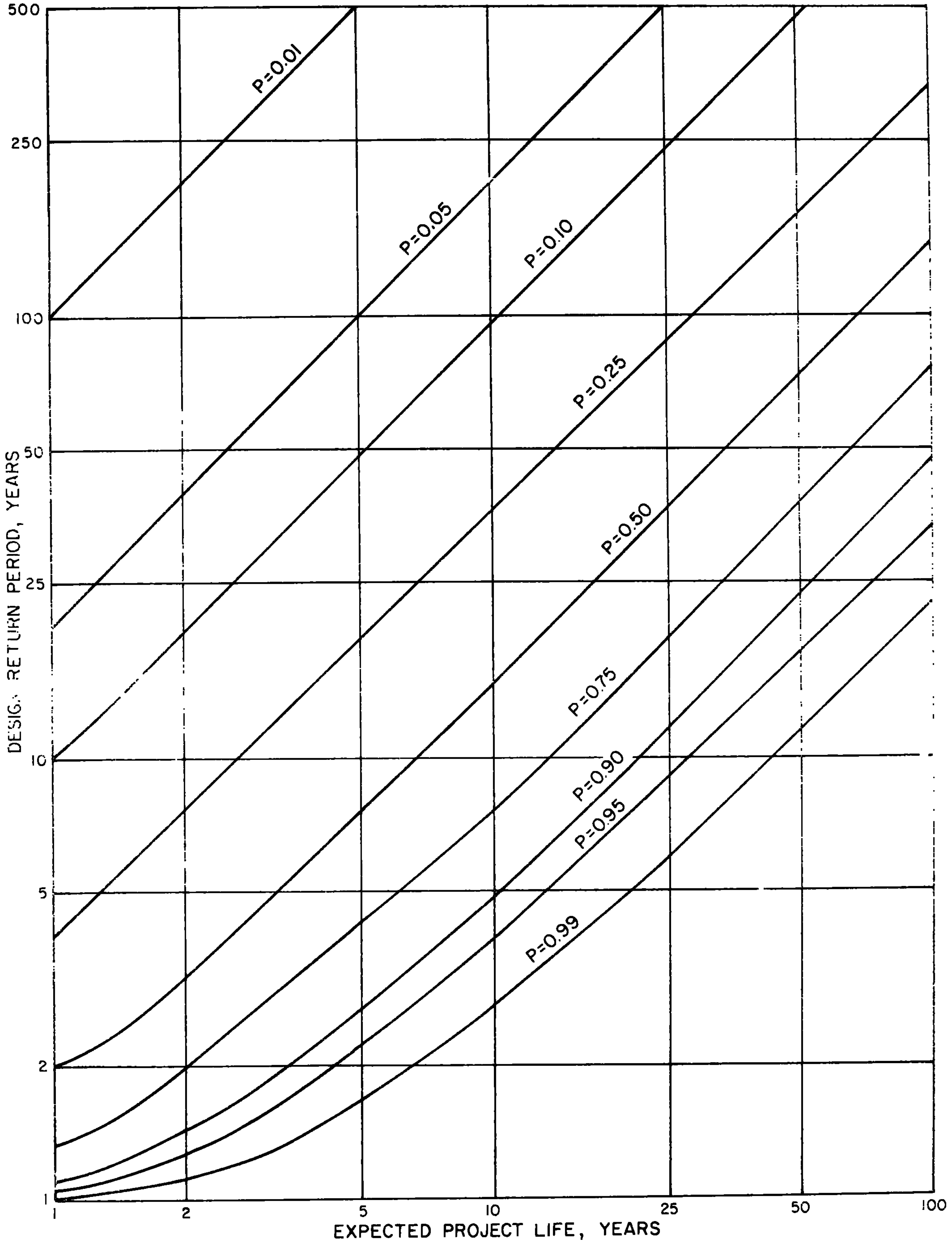
Design Return Period for Various Project Lives and Risks of Failure[1]

Permissible risk of failure	Expected Project Life, n, in years							
	1	2	5	10	20	25	50	100
0.99	1.01	1.11	1.66	2.71	4.86	5.95	11.4	22.2
0.95	1.05	1.29	2.22	3.86	7.16	8.85	17.2	33.9
0.90	1.11	1.46	2.71	4.86	9.19	11.4	22.2	43.9
0.75	1.33	2.00	4.13	7.73	14.9	18.6	36.6	72.6
0.50	2.00	3.41	7.73	14.9	29.4	36.6	72.6	145.
0.33	3.00	5.45	12.9	25.2	49.9	62.1	124.	247.
0.25	4.00	7.46	17.9	35.3	70.0	87.3	174.	348.
0.20	5.00	9.47	22.9	45.3	90.1	113	225.	449.
0.10	10.0	19.5	48.0	95.4	190.	238.	475.	950.
0.05	20.0	39.5	98.0	195.	390.	488.	975.	1,950.
0.02	50.0	99.0	248.	495.	990.	1,238.	2,476.	4,951.
0.01	100.	199.5	498.	995.	1,990.	2,488.	4,977.	9,953.

[1] From Yen (17)

FIGURE 5.2

THEORETICAL PROBABILITY OF FAILURE FOR GIVEN PROJECT LIFE AND DESIGN RETURN PERIOD



In the event that the design return period is made equal to the expected project life there is a 63.4% chance of failure of the project. This can be shown in Equation 5.8 by putting T = n:

$$P = 1 - (1 - \frac{1}{n})^n \qquad 5.10$$

In the limit as $n \rightarrow \infty$

$$(1 - \frac{1}{n})^n \rightarrow \frac{1}{e} = 0.368 \qquad 5.11$$

and so, for large n, P tends to 63%.

Similarly, supposing that a project has been designed against a hydrologic event of return period T years then the risk of failure after completion of n' years of the expected project life of n years can be calculated (14).

Writing Equation 5.7 as

$$Q = [(1 - \frac{1}{T})^T]^{T/n} \qquad 5.12$$

and using the same asymptotic approximation as in Equation 5.11, Gill (10) has shown that for a given value of P or Q there is a linear relationship between T and n, as:

$$Q = (\frac{1}{e})^{T/n} \qquad 5.13$$

$$n = T.\ln\ (1/Q) \qquad 5.14$$

A frequently used approximation resulting from Equation 5.8 is:

$$T \sim n/P \qquad 5.15$$

Gumbel (11) termed this the "design quotient".

The probabilities referred to above are all probabilities of occurrence of an event of a certain magnitude. Also of interest is the average probability of occurrence of all events above that certain magnitude. For example, in a series of n annual events the number, m, of events which equal or exceed the T-year event in (n+1)/T. The annual probability of occurrence of the maximum event is 1/(n+1), of the second largest event is 2/(n+1), of the third largest event is 3/(n+1), etc. so that, the average probability $\bar{p}$ of the n' events which exceed the T-year event is given by:

$$\bar{p} = (\frac{1}{n+1} + \frac{2}{n+1} + \frac{3}{n+1} + \ldots\ldots \frac{n'}{n+1})/n' \qquad 5.16$$

Benson (6) has shown that this expression reduces to:

$$\bar{p} = (n+T)/2.T.(n+1) \qquad 5.17$$

which, as n approaches infinity, becomes

$$\bar{p} \sim 1/2T \qquad 5.18$$

Thus, in general, the average probability of occurrence of all events above the T-year event is approximated by the probability of the 2T-year event. For example, the average probability of occurrence of all events greated than the 100-year event is approximately 0.005, which corresponds to the 200-year event.

The expressions developed so far in this chapter have all been distribution-free, that is, no assumptions have been made regarding the underlying event distribution. If it is required to estimate the event magnitude corresponding to the design return period computed from, for example, Equation 5.8 then a probability distribution must be assumed.

To show the wide variation possible in the results of this assumption of a distribution, Figure 5.3 is reproduced from Gumbel (11). This figure shows the relationship between design quotient and the reduced variable, z, where

$$Z = (x - \mu)/\sigma \qquad 5.19$$

for several commonly used distributions.

Those distributions shown on Figure 5.3 which have fixed skewness are the normal with $\gamma_1 = 0$ and the double exponential (or Type I extremal) with $\gamma_1 = 1.3$. The skewnesses have been arbitrarily chosen for the other distributions shown on the figure.

Alternatively, if the assumptions are made that events are independent and the mean number of events in unit time is constant then the Binomial and Poisson distributions can be used to evaluate risk, as described in Chapter 3. For a Poisson distribution of event occurrences and an extremal Type I distribution of event magnitudes, Shane (15) has defined the design event, x, as:

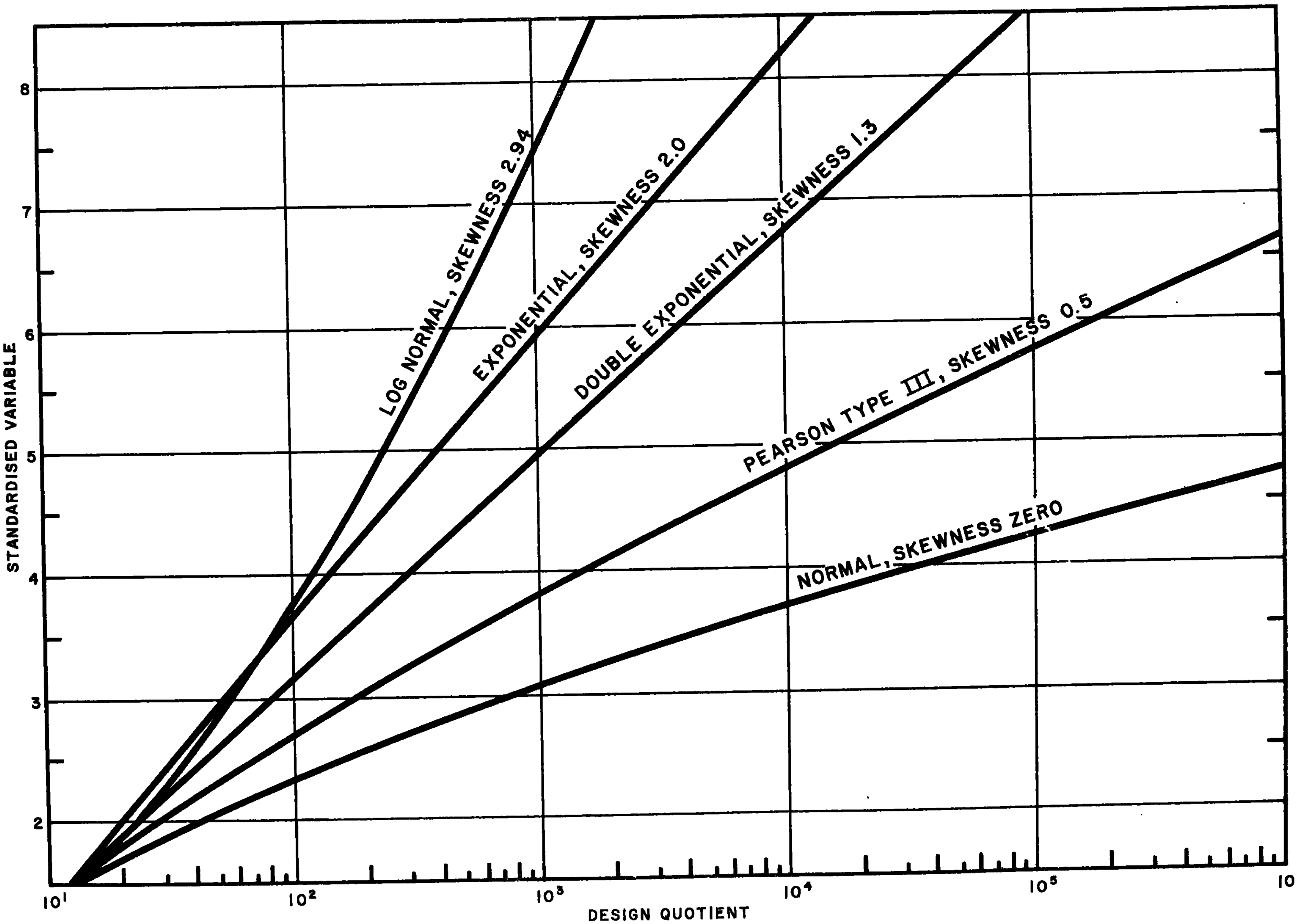


FIGURE 5.3 - STANDARDISED VARIABLE VS DESIGN QUOTIENT

$$x = v + \gamma \ln (\lambda.F) \qquad 5.20$$

where v is a base flow, γ is a parameter of the extremal distribution, λ is the expected rate of occurrence of events, λ = np, in the Poisson process and F is the risk factor. The maximum likelihood estimates of γ and λ are given (15) as:

$$\hat{\gamma} = \bar{x} - v \qquad 5.21$$

and

$$\hat{\lambda} = n_e/n \qquad 5.22$$

where n_e is the number of events recorded and n is the period of record.

Benson (5) investigated the variation which occurs when small samples are used to estimate a frequency distribution for which the parameters are known exactly. Starting with a known frequency curve Benson (5) constructed short random data sets, drew best-fitting curves and estimated events at different return periods from those curves. From a basic set of 1000 points, 100 records of ten points, 40 records of twenty-five points, 20 records of fifty points and 10 records of one hundred points each were drawn. It was found that records of up to twenty-five points cound not define satisfactorily even short-term events. Long-term records (forty to fifty points) were found to define events magnitudes up to the length of record with reasonable accuracy.

Yen and Ang (18) have described a procedure for designing hydraulic structures on the basis of a risk analysis. Using as an example the design of an urban sewer system, an overall project risk was chosen on the basis of possible property damage. The hydraulic and hydrologic risks are combined as α_h and are related to the structural risk, α_s, and overall risk, α, by

$$(1 - \alpha) = (1 - \alpha_s)(1 - \alpha_h) \qquad 5.23$$

Yen and Ang then defined the combined hydraulic and hydrologic risks as:

$$\alpha_h = P(x > Q_c).P(N > \nu) \qquad 5.24$$

where $P(x>Q_c)$ is the probability of an event X exceeding a design event, Q_c, (the hydrologic risk) and $P(N>\nu)$ is the probability that N, a random variable, will exceed ν, a safety factor, (hydraulic risk) where

$$\nu = Q_b/Q_c \qquad 5.25$$

and Q_b is the discharge actually used in design. N was assumed to be distributed lognormally with unit mean and a variance, σ_N^2, equal to the total of the variances of the uncertainties such as inaccuracy of measurement, systematic errors in computation, etc.

$$\sigma_N^2 = \sigma_1^2 + \sigma_2^2 + \ldots . \qquad 5.26$$

as discussed earlier in the chapter.

If α and α_s are known, then α_h can be determined from Equation 5.23 and for various values of α, the safety factor, corresponding values of $P(X>Q_c)$, the hydrologic risk, can be found. The equivalent design return period can be found from Equation 5.8 and, assuming a probability distribution to fit the observed data, the corresponding event magnitude, Q_c, is found. Yen and Ang (18) used a Type I extremal distribution although any other suitable distribution could equally well have been used. By plotting values of Q_c versus α (or $Q_b = \alpha.\ Q_c$ versus α) the optimum discharge can be found.

Thus by defining rigorously the hydrologic risk the common hydraulic practice of using a safety factor to include the effects of hydraulic risk is provided with a scientific basis.

In the event that no streamflow records are available at the design site, Davis et al (9) have described a method of evaluating uncertainty by considering the distribution of rainfall events. If the number of rainfall events per season, N, is Poisson distributed with mean λ, i.e.

$$f_N(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad 5.27$$

and if the amount of rainfall, R, per event is exponentially distributed

$$f_R\ (k|u)\ -\ ue^{-uk} \qquad 5.28$$

where 1/u is the mean rainfall per event, then the return period of k units of rain in a season, T, is

$$T_R\ (k|\lambda,u)\ =\ [1\ -\ \exp(-\lambda e^{-uk})] \qquad 5.29$$

By using a linear rainfall-runoff relationship

$$Q = C\ (R-A) \qquad 5.30$$

where C is a coefficient depending upon the rainfall characteristics of a given watershed and A is a measure of initial abstractions, also depending on the watershed, then an expression for the probability density distribution of the flood return period, T_Q, can be given as

$$T_e\ (y|\lambda,u)\ =\ [1\ -\ \exp\left\{ -\lambda+\lambda F_Q(y|u) \right\}]^{-1} \qquad 5.31$$

where $F_Q(y|u)$ is the distribution function of runoff per event.

Uncertainty is included in the analysis (9) by considering the parameters λ, u and c as variables. It was assumed that λ and u could be described by a two parameter gamma distribution while a beta distribution was used for c.

The results of this approach provide design flows relying only on rainfall data for watersheds with ungauged streams by taking into account the uncertainty of the site parameters. It was found (9) that a closed form solution was not possible and so data generation was used to derive the distribution of the flood return period.

To conclude this chapter on hydrologic risk, a final note on the accumulation of risk. On an individual basis a design return period of 1000 years is often considered safe. When it is considered, however, that there are now approximately 10,000 large dams in the world (7), 1000 of which can be thought of as in independent basins, then the probability is 0.001 x 1000 = 1.0 that the 1000-year event will be equalled or exceeded each year at at least one of the dam sites. Similarly, Alexander (1) has shown that in Japan, where there are about 1700 dams, the average design return period for spillway design floods is of the order of 200 years. It would be expected, therefore, that on the average 8 or 9 dams will incur design floods annually. Quite a risk!

References for Chapter 5

1. Alexander, G.N., 1969, Application of Probability to Spillway Design Flood Estimation, Proc. IASH Symposium on Floods and their Computations, Leningrad, 1967, IASH-UNESCO-WMO Studies and Reports in Hydrology No. 3, pp. 536-543.

2. ASCE, 1973, Re-evaluating Spillway Adequacy of Existing Dams, Report of the Task Committee on the Re-evaluation of the Adequacy of Spillways of Existing Dams of the Committee on Hydrometeorology of the Hydraulics Division, Proc. ASCE, Vol. 99, No. HY2, pp. 337-372.

3. American Water Works Association, 1966, Spillway Design Practice, AWWA Manual No. M13, New York.

4. Banerji, S., and D.K. Gupta, 1969, On a General Theory of Duration Curve and its Application to Evaluate the Plotting Position of Maximum Probable Precipitation or Discharge, Proc. IASH Symposium on Floods and their Computations, Leningrad, 1967, IASH-UNESCO-WMO Studies and Reports on Hydrology No. 3, pp. 183-193.

5. Benson, M.A., 1960, Characteristics of Frequency Curves Based on a Theoretical 1,000-Year Record, USGS Water Supply Paper No. 1543-A, pp. 51-73.

6. Benson, M.A., 1967, Average Probability of Extreme Events, Water Resources Research, Vol. 3, No. 1, p. 225.

7. Biswas, A.K., 1971, Some Thoughts on Estimating Spillway Design Flood, Bull. IASH, Vol. XVI, No. 4, pp. 63-72.

8. Biswas, A.K., and S. Chatterjee, 1971, Dam Disasters: An Assessment, J. Engineering Institute of Canada, Vol. 54, No. 3, pp. 3-8.

9. Davis, D.R., L. Duckstein, C.C. Kisiel and M.M. Fogel, 1973, A Decision - Theoretic Approach to Uncertainty in the Return Period of Maximum Flow Volumes Using Rainfall Data, Proceding of UNESCO-WMO-IASH Symposium on the Design of Water Resources Projects with Inadequate Data, Madrid, Vol. 1, pp. 63-74.

10. Gill, M.A., 1972, Analysis of Probability and Risk Equations, Proc. ASCE, Vol. 98, No. HY5, pp. 969-971.

11. Gumbel, E.J., 1955, The Calculated Risk in Flood Control, Appl. Sci. Res., Section A, Vol. 5, pp. 273-280.

12. McCaig, I.W., and O.M. Erickson, 1959, Spillway Capacity and Flood Flows, Proc. Symposium No. 1, Spillway Design Floods, NRC, Ottawa, pp. 262-287.

13. Moss, M.E., 1969, Maximisation of Net Benefit from a Streamgauge, Proceedings of Fiftieth Annual Meeting of the American Geophysical Union, Washington, D.C.

14. Prasad, T., 1971, Discussion of "Risks in Hydrologic Design of Engineering Projects", Proc. ASCE, Vol. 97, No. HY1, pp. 201-202.

15. Shane, R., 1966, A Statistical Analysis of Base-Flow Flood Discharge, Ph.D. Thesis, Cornell University.

16. Thomas, R.B., 1971, Errors in Streamflow Estimates from Continuous Stage Records, Proceedings of Symposium on Statistical Hydrology, Tucson, Arizona.

17. Yen, B.C., 1971, Risks in Hydrologic Design of Engineering Projects, Proc. ASCE, Vol. 96, No. HY4, pp. 959-966.

18. Yen, B.C., and A.H.S. Ang, 1971, Risk Analysis in Design of Hydraulic Projects, Proc. Symp. on Stochastic Hydraulics, Univ. Pittsburgh, pp. 694-709.

19. Yevjevich, V., 1968, Misconceptions in Hydrology and their Consequences, Water Resources Research, Vol. 4, No. 2, pp. 225-232.

CHAPTER 6

Conclusions

6.1 General

The magnitude and frequency of occurrence of extreme hydrologic events is of every day importance in most parts of the world. Since man has, for reasons of communication, water supply, agriculture, etc., built most of his communities on the flood plains of large rivers his life-style is extremely susceptible to flood damage. Today's pressure of population increases the density of development along the river banks. The flood of 1948 on the Fraser River in British Columbia caused $20 million of damange. It has been estimated that if the same magnitude of flood occurred today the damages would be over $200 million.

At the opposite end of the water spectrum, the production of sufficient food to feed the world's rapidly increasing population necessitates the increasing use of irrigation. Mankind thus becomes ever more susceptible to disaster through drought.

Proper use of existing hydrologic techniques could, through flood plain zoning and efficient design techniques, eliminate much of the present loss of life and damage caused by floods and droughts. The available hydrologic methods can be divided into the deterministic techniques of empirical equations, unit hydrograph, storm transposition etc., and the stochastic technique of flood frequency analysis.

In the 1920's and 1930's the introduction of simple statistical analysis gave an impetus to the science of hydrology.

Very soon, however, a general distrust of probability methods began to grow because too many users knew too little about statistical analysis and they apparently expected the methods to overcome the lack of data (10).

A growing use of deterministic methods, replacing the fall in popularity of probability techniques, led to the development in the late 1930's of the unit hydrograph principle. Advances in meteorology enabled the conditions producing storm rainfall to be analysed, with the result that maximum rain producing storms could be synthesised. This technique produced a large number of new, rather vague, technical terms such as Probable Maximum Precipitation, Maximum Possible Precipitation, Standard Project Storm, Maximum Probable Flood, etc., based on the premise that some definite limit existed for all the variables responsible for flood events and that, subsequently, some limit must apply to the flood runoff itself. The drawback to this method is that no probability level can be assigned to the "probable" events, because of their deterministic origins. Similarly no confidence limits can be applied to these events and the non-specialist is left with the impression that these estimates are 100% accurate with no risk involved.

The philosophical error in the Maximum Probable argument has been described by Yevjevich (28) and others. It is not reasonable to say that a precipitation of 30 inches in one hour can occur but a precipitation of 30.1 inches cannot. The probability approach states that any variable has a finite probability of reaching any value between zero and infinity. As the variable increases in magnitude so the associated probability of occurence decreases and for very large events approaches zero. However, given enough time, even the improbable becomes certain. The essential stochasticity of precipitation has been recognised by Yevjevich (29) as being mainly due to the random nature of atmospheric variables such as opacity and transmitted radiation. The stochastic precipitation events are then somewhat attenuated by the water and energy storages of the oceans and continents to produce runoff events of mixed deterministic/stochastic nature.

Today, the necessity of producing economically designed projects has produced a need for hydrologic risk analysis and a corresponding upsurge in the use of probabilistic methods in hydrology. Compared to the 1920's and 1930's, however, more data is available today and the theory of sampling errors and risk analysis is better understood. Problems still exist in the statistical techniques, however, and a certain amount of subjectivity is still involved, particularly in the choice of a frequency distribution to fit to the observed data.

Users of hydrologic data can be placed in several categories. There are users in water planning who assess the potential for development in a basin or region. There are users in water project design interested in data related to a specific project or water operation system; there are users in construction and in administration.

There are also users in the operations and management area working on the planning of project operations such as ship traffic control, hydroelectric systems operation, hydrometric network planning and flow forecasting operations. These users all have differing data requirements but one factor that all have in common is a frequent use of hydrologic probability analyses. Because of the large number of different government and commercial data users and the large number of available methods of frequency analysis, if each user obtains the same basic data from the collecting agency and carries out their own data processing then widely differing results may be expected.

In addition to the scientific inaccuracy, this procedure involves a wasteful duplication of time and manpower. These inefficiencies could be eliminated if the data collection agency processed the data and issued, on request, frequency analyses in a form suitable for direct application by the user. The data collection agency usually has the data on magnetic tape or other computer-compatible medium and can carry out the processing of the data very much more quickly and efficiently than many of the data users. In addition, the data collection agency is better aware of the accuracy limitations of the raw data than are many of the data users and so can tailor the anlaysis procedure to the degree of sophistication of the data. After all, applying a complex and time-consuming analysis technique to data containing large inaccuracies is, at best, wasteful and, at worst, misleading. There is a need, therefore, for a method which makes full use of the information content of the original data without introducing a false degree of security resulting from mathematical sophistication.

6.2 Data Abstraction, Graphs, and Plotting Positions

Data for frequency analyses may be abstracted from the recorded data using either annual series or partial duration series. Annual series consist of one event per year; partial duration series consist of all events above a base magnitude, regardless of time of occurrence. The partial-duration series method would initially seem advantageous in that more data, and hence information, is incorporated. As shown in Table 1.1, however, this additional data increases the definition of event magnitudes only in the lower part of the frequency curve which is the area of least interest. Use of the partial-duration series always involves the arbitrary establishment of a base flow and sometimes requires subjective decisions regarding the independence of adjacent events.

For these reasons and because the annual series is simpler to abstract and analyse it is to be preferred in frequency studies.

Riggs (20) and Benson (3) have detailed many other reasons, but to summarise, it is usually stated that mathematical fitting of a standard probability distribution is preferable to plotting a graph and fitting a curve by eye, because:-

(a) mathematical fitting is theoretically better,

and (b) mathematical fitting eliminates the subjectivety of individual judgement.

In fact, curve fitting by eye and by probability function are equally empirical since, as discussed later in this chapter, the true distribution of the recorded events is not known. In addition the very lack of subjectivity of the mathematical procedure is sometimes a disadvantage; the inclusion or exclusion of one or two events may result in large changes in the resulting frequency relationship. The mathematical procedure incorporates all data whereas the individual drawing in a curve by eye may elect to ignore some events in order to get a better-fitting curve. A method is not better simply because it leads to uniform answers, if those answers are uniformly unsound (2).

An intermediary or semi-graphical method exists in which graphs are used to fit curves of standard probability distributions to the data. The procedure recommended is to use a mathematical method of fitting a standard probability distribution (which one will be discussed later) but to arrange for the output of the method always to include a plot of the data points and the fitted curve. In this way the procedure can be standardised and automated for machine computation while retaining the option of looking at a graph and reviewing the fitted line on the basis of engineering experience.

Any type of plot of extreme events requires the consideration of plotting positions. When all analysis computations were done by hand, graphical techniques of analysis were extremely attractive because of their brevity and simplicity and the choice of plotting positions was therefore of great importance. Today, when all calculations are performed by computer, graphs are used only as a pictorial form of output presentation and the problem of plotting position has faded somewhat in importance.

With this in mind it would be difficult to justify use of any plotting position other than the mean frequency

$$p = m/(n+1) \tag{6.1}$$

where m is the order of the event in the sorted series of n observed events. For the largest event in the series m = 1 and for the smallest event in the series, m = n. This method gives conservative results in that the return period conforms closely to the period of record. Benson (4) has demonstrated that this is the best plotting position to use for economic studies of hydrologic design.

In certain cases it is found that one or two events in a series will plot well above or well below the other events. These are known as "outliers" and pose a very sticky statistical problem. Assuming that no physical reason can be found (any obvious error in measurement or computation, for example) then the decision to accept or reject the outliers must be made on the basis of statistical tests. Anscombe (1) has discussed the history and background of rejection rules and has explained the theory behind two commonly used rules. These rules have been formulated on the assumption that the population variance is known. Since this is not the case in hydrology, the rules should be used with care.

6.3 Frequency Distributions

6.3.1 Selection of a Distribution

The primary objectives of frequency analysis are to determine the return periods of recorded events of known magnitude and then to estimate the magnitude of events for design return periods beyond the recorded range. The intermediary between these two objectives is the theoretical probability distribution. The sample data is used as an estimate of an unknown population to calculate the parameters of the selected probability distribution. The fitted distribution is then used to estimate event magnitudes corresponding to return periods greater than or less than those of the recorded events.

There is no general agreement among hydrologists as to which of the various theoretical distributions available should be used. The present state of the art is also such (5) that no general agreement has been reached as to preferable techniques, and no standards have been established for design purposes. As examples of this divergence of choice, Spence (26) compared the fit of the normal, lognormal, Type I extremal and log-Type I extremal distributions to annual maximum flows on the Canadian Prairies and found that the lognormal was the best fitting; Cruff and Rantz (11) compared six probability distributions in California and found that the Pearson Type III was the most desirable. In other studies, Santos (22) has found the lognormal distribution better than the Pearson Type III, Gumbel (12) has explained that "It seems that the rivers know the [extreme value] theory. It only remains to convince the engineers...of the validity of this analysis", and Benson (3) has found in a study of 100 long term flood records that no one type of frequency distribution gives consistently better results. In the U.S., Reich (19) conducted a survey of engineers and hydrologists and found that of the Extremal Type I, log Extremal Type I and log Pearson Type III, the log Pearson Type III was preferred. In Italy, Cicioni et al (9) tested the lognormal, 3-parameter lognormal, 2-parameter gamma, Pearson Type III and Extremal Type I distributions on 108 data sets and found the lognormal to be the most suitable.

In short, no one distribution is acceptable to all hydrologists.

A problem with some distribution comparison studies is that improper statistical techniques are used to judge the performance of the different probability distributions. As an example, a recent study fitted various distributions to sample data by a simple linear least squares regression on suitable transformed data. The distribution which gave the highest simple correlation coefficien was concluded to be the correct distribution for flood flows. However, the goodness of fit of a particular distribution to a sample set of data in no way guarantees that that distribution is correct for the population of events. This is particularly true when the purpose of the distribution fitting is to extrapolate from the measured data.

Acceptance of a certain model for analysis of flood peaks must be based on the goals and conditions that are to be fulfilled and satisfied by the model (30). Goodness of fit is a necessary but not a sufficient condition for acceptance. Goodness of fit tests often used include Chi-Square and Kolmogorov-Smirnov (27) as well as Cramer-Von Mises and Anderson-Darling statistics (9). If goodness of fit were the only criterion, then high order polynomials would often provide a much better fit than any of the standard distributions, and yet this method is not used because there is no hydrologic justification. The most important criteria in the selection of a model are that there be a sound theory describing the phenomenon and that the model should abstract the maximum information from the data using proper estimation techniques.

This was realised by the recent U.S. Water Resources Council Work Group (5) who wrote that "no single method of testing the computed results against the original data was acceptable to all those on the Work Group, and the statistical consultants could not offer a mathematically rigorous method." The Work Group concluded that a frequency distribution could not be chosen solely on statistical grounds but recommended that the log-Pearson Type III distribution be used because as a 3 parameter distribution it offers considerable flexibility, for a zero skewness it reduces to the lognormal distribution, and finally because it is in common use by U.S. government agencies.

The problem of choosing a distribution is not restricted to hydrology by any means. In the field of biosciences,Katti and Sly (13) summarised their findings as:

(a) No single theoretical distribution has been found to describe any large scale data.

(b) For a number of data there could be two or more theoretical distributions that fit equally well and there is no way to choose between them based on fit alone.

(c) Two or more physical models could lead to the same final statistical distribution and hence the estimation of the parameters of the distribution may not have unique meaning.

(d) ...different methods of estimation lead to widely differing estimates when the methods are consistent...there are a number of empirical frequencies to which the same theoretical frequency function has been fitted by different consistent methods..."

Although statistical methods cannot by themselves determine the correct frequency distribution, they can, in some cases, provide reasons why distributions may not be suitable. As an example some distributions such as the Type III Extremal, Pearson Type III and log Pearson Type III require the estimation of the coefficient of skewness from the sample data. It is well known that the variability of sample estimates of the coefficient of skew is large (11), (21) and this may be sufficient reason to prefer some other distribution.

As a second example of this process of elimination the following objections have been raised (3), to the use of the Type I extremal distribution (Gumbel) for flood flows:

(a) It is assumed that the treatment derived for daily discharges can also be applied to instantaneous flows.

(b) The daily discharges are not independent events.

(c) The 365 daily discharges in a year do not constitute a large number as required by the theory of extreme values.

(d) An assumption underlying the extreme value theory is that all the events are part of the same statistical population. Yet, in many cases, the annual maximum event may be due to a variety of causes such as normal rainfall, snowmelt or hurricane. There are different physical factors controlling each of these types of events. The assumption of one population therefore may not be valid.

Whether or not these objections to the use of the Type I extremal distribution are sufficiently serious to deter use, is a matter of opinion.

A third example of the use of statistics to eliminate distributions might also be mentioned; the normal distribution can be eliminated because it ranges from $-\propto$ to $+\infty$ and can thus give real probabilities to negative flows. In some cases a truncated normal distribution has been used in which the probability of events being less than or equal to zero is replaced by a probability mass at $x = 0$.

Of lower importance than the theoretical background is the ease of computation. Matalas and Wallis (15) found several computational difficulties in using the Pearson Type III and log-Pearson Type III distributions and recommended that the use of other distributions warrants consideration. Similarly Pentland and Cuthbert (18) found that use of the log-Pearson Type III distribution for the Fraser River Basin, in British Columbia, led to large discontinuties and unnatural flood frequencies. They substituted the lognormal distribution in place of the log Pearson Type III.

On the basis of this study of available literature it is recommended that, for data sets of annual maxima containing less than 100 items, the lognormal distribution be used. The lognormal has as much theoretical justification as any other distribution (8) and, at the same time, it is computationally easier than many distributions. In a case like this where many methods compete it is always better to use the easiest and simplest method until another is proved to be superior.

The preferred procedure is to compute the T-year event using the mean and standard deviation of the recorded events rather than the mean and standard deviation of the event logarithms.

For those cases in which a single lognormal probability distribution does not provide a suitable fit then the procedure of using two lognormal distributions as described by Singh and Sinclair (23) may be useful.

For data sets containing more than 100 events (this figure is only an estimate (21) and should not imply any certain cutoff) then the coefficient of skew may be computed with sufficient accuracy to justify use of the three parameter log Pearson Type III distribution. If the skewness of the logarithms of the recorded events is zero then the log Pearson Type III is identical to the lognormal distribution.

For drought flow analysis the situation is less critical since a known lower limit to event magnitude exists. The two distributions commonly used are Type III extremal and Pearson Type III and there is little basis for any choice between these.

As noted earlier it is considered necessary in all cases to produce a graph of the observed events and the fitted frequency curve. There will always be times when the fitted curve will be poor and, by checking a graph plot, indications may be available as to a more suitable distribution. A good procedure is to incorporate a plotting routine into any computer program producing frequency curves.

6.3.2 Estimation of Parameters

There are four main methods of estimating distribution parameters: moments, maximum likelihood, least squares and graphical. It is now generally accepted (15) that the method of maximum likelihood is the most efficient method and should be used wherever possible. By computing relative efficiencies, Matalas (14) has shown that for low flow analysis the method of moments used only one-half of the sample information extracted by maximum likelihood.

For the recommended lognormal distribution the maximum likelihood estimation method is very simple and the results are identical to those of the method of moments. The maximum likelihood method is more involved and time-consuming for the Pearson Type III and log Pearson Type III distributions but solutions are available and have been described.

6.3.3 Frequency Factors

For any distribution the T-year event magnitude can be computed from a general equation of the form:

$$x(K) = \mu + K\sigma \qquad 6.2$$

where μ and σ are sample estimates of the population mean and standard deviation and K is a frequency factor specific to the chosen distribution. For each probability distribution the frequency factor K, can be derived from the sample size, sample parameters, etc.

6.3.4 Confidence Limits

Once a suitable probability distribution has been chosen and the magnitude of the event, x(K), at the required return period, T, has been computed from the general frequency equation then upper and lower confidence limits should be established for this event magnitude.

From the frequency curve, two methods are available to compute confidence limits. The analytical method uses numerical techniques to integrate the probability density function of the sample quantile. For practical distributions, however, this is very difficult and an empirical method is often used instead. The empirical method uses moments to compute the standard error of x(K), s(K), from the variances of μ and σ then assumes that the T-year event is normally distributed with mean x(K) and standard deviation s(K) so that the confidence interval is

$$x(K) \pm t.s(K) \qquad 6.3$$

where t is the standard normal deviate at the required confidence level.

The validity of the empirical method rests on the assumption of normality of the distribution of T-year events. This has been tested by data generation and the results have been described. On the basis of these tests the assumption of normality is certainly reasonable.

Nash and Amorocho (16) have shown that, for normal and double exponential distribution, the standard error, s(K), tends to become a fixed proportion of the T-year event magnitude as T tends to infinity. Thus estimating the 10,000-year event is possible with no greater relative error than occurs in an estimate of the 100-year event, provided that one is certain that the assumed form of the probability distribution is correct.

6.4 Regional Analysis

Regional analysis techniques provide a means of combining records from many gauges. This provides the two advantages of reducing standard errors of estimates at gauge sites and enabling estimates to be prepared for ungauged sites.

The description of regional analysis techniques given in Chapter 4 divided the techniques into five main methods: index-flood, multiple regression, square grid, modified single station probability distributions, and regional record maxima. Each of the first four methods contains several variations, however, and these variations tend to overlap between methods so that the division becomes somewhat artificial. The index-flood method, as originally proposed and put into practice suffers from many defects. The most serious defect is that the method is totally empirical, the distribution of peak events is not known and all relationships are derived by graphical curve-fitting.

The square grid method has a logical ordering of basic data which, when combined with the convenience of automatic data processing, provides a data file of regularly spaced physiographic, meteorologic and hydrologic data easily and rapidly accessible for many types of study. The data file can be regularly updated and may be enlarged by the addition of more parameters as they become available. Using this data file as a base, an extremely versatile series of analyses becomes possible. Originally the square grid method was used to define the areal distribution of mean annual runoff for preliminary province-wide water resources studies (24), but additional steps now available include the generation of synthetic monthly flows at ungauged sites (17), and, with the addition of meteorologic forecasts, the use of a parametric model to provide monthly forecasts of streamflow (25), as well as the provision of flood frequency analyses at ungauged sites.

Using the square grid method for frequency analyses combines the use of a standard frequency distribution with a more efficient data base. In recommending this square grid approach for regional frequency analysis the probability distribution to be used should be selected on the basis of the comments and recommendations previously made.

Hydrologic series frequently consist of observations that are dependent upon one another, they are serially correlated. Thus each observation contains some information which has already been contained in previous observations. A serially or autocorrelated time series therefore contains less information than would an equal number of pure random observations. If a time series of a given length is correlated with a shorter time series, then the two-station correlation can be used to increase the information content of the shorter series. If either or both of the time series are autocorrelated then the increase in information will be less than if the time series were pure random. Similarly if a number of gauging stations are used to estimate regional parameters then the information content of those parameters will be a maximum when all the station records are independent and will decrease as the dependence between records increases.

6.5 Risk

Most hydraulic structures are designed on the basis of deterministic Probable Maximum Floods for which the risk is totally unknown. For those structures designed on the basis of frequency analysis, the hydrologic risk is generally accounted for by a simple choice of return period for the design flood. For any given return period however, the risk of project failure is proportional to the expected project life.

A better procedure is to initially assign an acceptable overall risk to the project. This overall risk should be computed on the basis of the consequences and costs of failure. This total risk can then be subdivided into structural, hydraulic and hydrologic risk and the allowable hydrologic risk can be determined. On the basis of this allowable hydrologic risk and the expected project life the required return period can be found and by assuming a probability distribution the corresponding T-Year event magnitude can be calculated. In this manner, Probable Maximum Floods or other deterministic estimates can often be assigned a risk and used in further analysis.

To account for the uncertainties in estimation of distribution parameters it is usual to compute the upper 95% confidence limit for the T-Year event and design for that discharge.

It is well to keep in mind, however, that no matter how sophisticated a risk analysis is undertaken, the unexpected tends to occur with remarkable regularity. As an example (6), the

Rincon de Bonete hydroelectric project on the Rio Negro in Uraguay was desinged for the 1000-year flood of 325,000 cfs on the basis of 27 years of streamflow records during which the peak flow was 135,000 cfs. Fourteen years after construction (1959) as a result of prolonged heavy rainfall a flood peak 605,000 cfs was measured. Using the original frequency analysis the 1959 flood has a theoretical return period of 500,000 years. Fortunately, the dam held and about one-half of the flood peak was absorbed by the reservoir which rose 15 feet above its designed maximum level.

References for Chapter 6

1. Anscombe, F.J., 1960, Rejection of Outliers, Technometrics, Vol. 2, No. 2, pp. 123-147.
2. American Water Works Association, 1966, Spillway Design Practice, AWWA Manual No. M13, New York.
3. Benson, M.A., 1962, Evolution of Methods for Evaluating the Occurrence of Floods, USGS Water Supply Paper 1580-A.
4. Benson, M.A., 1962, Plotting Positions and the Economics of Engineering Planning, Proc. ASCE, Vol. 88, No. HY6, pp. 57-72.
5. Benson, M.A., 1968, Uniform Flood Frequency Estimating Methods for Federal Agencies, Water Resources Research, Vol. 4, No. 5, pp. 891-908.
6. Biswas, A.K., 1971, Some Thoughts on Estimating Spillway Design Floods, Bull. IAHS, Vol. XVI, No. 4, pp. 63-72.
7. Blench, T., 1959, Empirical Methods, Proc. Symposium No. 1, Spillway Design Floods, NRC, Ottawa, pp. 36-48.
8. Chow, V.T., 1954, The Log-Probability Law and Its Engineering Applications, Proc. ASCE, Vol. 80, pp. 1-25.
9. Cicioni, G., G. Giuliano and F.M. Spaziani, 1973, Best Fitting of Probability Functions to a Set of Data for Flood Studies, Proc. Second International Symposium in Hydrology, Floods and Droughts, Water Resources Publications, Ft. Collins, Colorado, pp. 304-314.

10. Clark, R.H., 1959, Opening Address, Symposium No. 1, Spillway Design Floods, NRC, Ottawa, pp. 1-3.

11. Cruff, R.W., and S.E. Rantz, 1965, A Comparison of Methods Used in Flood Frequency Studies for Coastal Basins in California, USGS Water Supply Paper 1580-E.

12. Gumbel, E.J., 1966, Extreme Value Analysis of Hydrologic Data, Proc. Hydrology Symposium No. 5, NRC, Ottawa, pp. 147-169.

13. Katti, S.K., and L.E. Sly, 1965, Analysis of Contagious Data Through Behaviouristic Models, in: Classical and Contagious Discrete Distributions, ed. G.P. Patil, Statistical Publishing Society, Calcutta, pp. 303-319.

14. Matalas, N.C., Probability Distribution of Low Flows, USGS Professional Paper 434-A.

15. Matalas, N.C. and J.R. Wallis, 1973, Eureka! It fits a Pearson Type III Distribution, Water Resources Research, Vol. 9, No. 2, pp. 281-289.

16. Nash, J.E., and J. Amorocho, 1966, The Accuracy of the Prediction of Floods of High Return Period, Water Resources Research, Vol. 2, No. 2, pp. 191-198.

17. Pentland, R.L., and D.R. Cuthbert, 1971, Operational Hydrology for Ungauged Streams by the Grid Square Technique, Water Resources Research Vol. 7, No. 2, pp. 283-291.

18. Pentland, R.L. and D.R. Cuthbert, 1973, Multisite Daily Flow Generator, Water Resources Research, Vol. 9, No. 2, pp. 470-473.

19. Reich, B.M., 1973, Log Pearson Type III and Gumbel Analysis of Floods, Proc. Second International Symposium in Hydrology, Floods and Droughts, Water Resources Publications, Ft. Collins, Colorado, pp. 291-303.

20. Riggs, H.C., 1968, Frequency Curves, Chapter A2, Book 4, in Techniques of Water Resources Investigations of the USGS.

21. Sangal, B.P., and A.K. Biswas, 1970, The 3-Parameter Lognormal Distribution and Its Applications in Hydrology, Water Resources Research, Vol. 6, No. 2, pp. 505-515.

22. Santos, A., 1970, The Statistical Treatment of Flood Flows, Water Power, Vol. 22, No. 2, pp. 63-67.

23. Singh, K.P., and R.A. Sinclair, 1972, Two-Distribution Method for Flood Frequency Analysis, Proc. ASCE, Vol. 98, No. HY1, pp. 29-45.

24. Solomon, S.I., T.P. Denouvilliez, C. Cadou and E.J. Chart, 1968, The Use of a Square Grid System for Computer Estimation of Precipitation, Temperature and Runoff in a Sparesly Gauged Area, Water Resources Research, Vol. 4, No. 5, pp. 919-930.

25. Solomon, S.I. and A.S. Quershi, 1972, Application of a Parametric Model for Estimating Snow Accumulation and Flow Forecasting, Proc. IHD/UNESCO/WMO Symposia on the Role of Snow and Ice in Hydrology, Banff.

26. Spence, E.S., 1973, Theoretical Frequency Distributions for the Analysis of Plains Streamflow, Canadian Journal of Earth Sciences, Vol. 10, pp. 130-139.

27. Yevjevich, V., 1964, Fluctuations of Wet and Dry Years, Part II, Analysis by Serial Correlation, Hydrology Paper No. 4, Colorado State University, Ft. Collins, Colorado.

28. Yevjevich, V., 1968, Misconceptions in Hydrology and Their Consequences, Water Resources Research, Vol. 4, No. 2, pp. 225-232.

29. Yevjevich, V., 1971, Stochasticity in Geophysical and Hydrological Time Series, Nordic Hydrology, Vol. II, pp. 217-242.

30. Zelenhasic, E., 1970, Theoretical Probability Distributions For Flood Flows, Hydrology Paper No. 42, Colorado State University, Fort Collins, Colorado.