Section 8-I

STATISTICAL AND PROBABILITY ANALYSIS OF HYDROLOGIC DATA

PART I. FREQUENCY ANALYSIS

VEN TE CHOW, *Professor of Hydraulic Engineering, University of Illinois.*

I. INTRODUCTION

A. Importance of Statistical and Probability Analysis

Quantitative scientific data may be classified into two kinds: experimental data and historical data. The *experimental data* are measured through experiments and usually can be obtained repeatedly by experiments. The *historical data*, on the other hand, are collected from natural phenomena that can be observed only once and then will not occur again. Most hydrologic data are historical data which were observed from natural hydrologic phenomena.

Since hydrologic data are the only source of information upon which quantitative hydrologic investigations are generally based, their measurements have been continuously expanding and resulting in ever-increasingly large amounts of sampled data. *Statistics* deals with the computation of sampled data, and *probability* deals with the measure of chance or likelihood based on the sampled data. The mounting quantities of hydrologic data can suitably be expressed in statistical terms and be treated with probability theories. Furthermore, natural hydrologic phenomena are highly erratic and commonly stochastic in nature, and therefore are amenable to statistical interpretation and probability analysis. Section 8 covers some fundamental principles and methods of statistics and probability that are useful in the solution of hydrologic problems.

One of the important problems in hydrology deals with interpreting a past record of hydrologic events in terms of future probabilities of occurrence. This problem arises

in the estimates of frequencies of floods, droughts, storages, rainfalls, water qualities, waves, etc; the procedure involved is known as *frequency analysis*. The methods of frequency analysis and some fundamentals in statistics and probability are discussed in this Part I of Section 8.

For general discussions on statistical and probability analysis of hydrologic data, reference may be made to Refs. 1-4. For frequency analysis of hydrologic data in particular, Refs. 5-8 may be found to be useful.

B. Hydrologic Frequency Studies

1. Flood and Streamflow Studies. The frequency analysis of streamflow data is believed to have been first applied to flood studies by Herschel and Freeman (see [9]) in 1880 to 1890 by means of a graphical procedure of using flow-duration curves (Subsec. 14-V-A). According to Fuller [10], the use of probability methods in runoff studies had been suggested to him in 1896 by George W. Rafter. Owing to the dearth of long-period records on American rivers at that time, the use of probability methods for flood frequency analysis was apparently hindered until later years.

The Gaussian law of probability, or the normal law of errors, is the basic and simplest tool for frequency analysis. It was therefore used for flood studies in the very early days. For such studies, Horton [11] discussed briefly in 1913 the earlier applications of the Gaussian law, and in 1914 Fuller [10] gave a full account of the first really comprehensive study of statistical methods applied to floods in the United States.

However, Hazen [12] soon discovered that if the logarithms representing the annual floods are used instead of the numbers themselves, the agreement with the normal law of errors is closer. This is true because the frequency distributions of annual floods are usually skewed or asymmetrical and the distribution can be suitably represented by such frequency distribution laws as the Galton, or lognormal-probability, law. He proposed the use of lognormal-probability paper [13] and developed a procedure of analysis [14]. Hazen's method requires a table of factors for computing theoretical frequency curves by means of the coefficients of variation and skewness. The table [13, p. 219; 14, pp. 49, 188] was originally obtained by empirical methods and hence has been found to be inaccurate. A corresponding table of exact factors based on a mathematical procedure was later prepared by Chow [15] (Table 8-I-1). For the study of streamflow variability, Lane and Lei [16] made use of the lognormal-probability plotting of flood flows to determine the variability index (Subsec. 14-V-A).

Other laws of frequency distribution and methods of frequency analysis of floods were also proposed by many hydrologists. Type 1 and Type 3 of Karl Pearson's curves of frequency distribution were put in a form convenient for use in flood studies by Foster [17]. A table of frequency factors similar to Hazen's table was given by Foster and extended by Switzer and Miller [18]. Hall [19] proposed a special "hydraulic probability paper" in which the probability scale was obtained empirically from flow-duration curves of 35 California streams. Goodrich [20] proposed a special skew-frequency paper which was later tested and refined by Harris [21]. Up to 1934, Slade [22] derived various skew probability functions to which was introduced an ultimate limiting magnitude of flood flow or the limiting flood potentialities of the drainage basin.

In 1941, Gumbel [23] published the first of a great number of papers (e.g., Refs. 24–29) on the application of the Fisher-Tippett theory of extreme values to flood frequency analysis. The use of extreme-value theory has been further extended by other hydrologists. Powell [30] derived an extremal probability paper for graphical application of the method. Cross [31] soon applied it to the study of flood frequencies in Ohio. As the extremal distribution assumes a constant skewness, the variate of a given recurrence interval should theoretically depend on the coefficient of variation and the mean. Potter [32] applied this assumption to 370 extremal probability curves and derived practical graphic relationships between variate, mean, and coefficient of variation. Benson [33] developed a synthetic "1,000-year record" of peak floods based on a straight-line plotting on the extremal probability paper.

Both the lognormal-probability law and the extreme-value law have been used

extensively in recent years. From a theoretical point of view, Chow [15] has shown that the extreme-value law is practically a special case of the lognormal-probability law, or it is practically identical with the latter for a skewness coefficient of 1.139 and a coefficient of variation equal to 0.364. He also proposed a flexible straight-line fitting of flood data based on the merits of both methods. See Sections 14 and 25-I and Refs. 4–8 and 34–49 for other discussions of flood frequencies.

2. Rainfall Studies. Many frequency studies on rainfalls and other meteorological events have been made. An extensive rainfall frequency study in the United States was first made by Yarnell [50] in 1935 (Subsec. 9-VI). This study produced the well-known *Yarnell rainfall frequency data*, which are a set of 56 isohyetal maps of the continental United States, covering the range of durations of 5, 10, 30, 60, and 120 min for 2-year frequency and of the same durations plus 4, 8, 16, and 24 hr for 5, 10, 25, 50, and 100-year frequencies. For longer durations of 1, 2, 3, 4, 5, and 6 days, the Miami Conservancy District [51] published the data for 15, 25, 50, and 100-year frequencies in 1936, covering that part of the continental United States east of the 103d meridian.

Since Gumbel proposed the Type I extremal distribution for flood frequency analysis, Chow [52] applied it to the study of the rainfall-intensity frequency in Chicago, Illinois in 1953, and also published a design chart [53] for approximate determination of rainfall frequencies in the continental United States.

As more rainfall data were collected, extended and detailed rainfall frequency analyses were made by the U.S. Weather Bureau. In 1961, a rainfall frequency atlas of the United States was published by the Weather Bureau, which completely revises and supersedes the Yarnell data ([54]; Subsec. 9-VI).

Comparisons of several methods of rainfall frequency analysis have been made by Huff and Neill [55] and by Hershfield [56].

3. Drought and Low Streamflow Studies. Type III extremal distribution was first proposed by Gumbel [57] for drought frequency analysis in 1954. The method was later applied to actual problems [29, 58], including graphical applications to Michigan streams [59] and to streams in eastern United States [60].

Other frequency studies of droughts and low streamflows are discussed in Section 18 and Refs. 61–63.

4. Water Quality Studies. Frequency analysis has been applied to virus, bacteria, alkalinity, salinity, chlorides, sulfates, and other dissolved and undissolved materials in water. Some recent studies are discussed in Ref. 64.

5. Water Wave Studies. Frequency analysis of water waves constitutes its own unique field, as it has its special purposes in oceanography. Important developments in this field may be found in Refs. 65–69. The Type I extremal distribution also has been applied to wave frequency analysis first by Gumbel [70] and Jasper [71], and later by Bennet [72] and others.

II. FUNDAMENTALS

A. Statistical Variables

Hydrologic data can be treated as statistical variables. In statistics, the whole collection of objects under consideration is called a *population*, or *universe*. A segment of a population may have one or more characteristics associated with them. Their characteristics are called *variables*, usually designated as X. An individual observation or the value x of any variable X is known as a *variate*. In hydrologic phenomena, for example, the variable X may be the depth of rainfall, and it may have a value, say, $x = 1.45$ in.

Variables may be obtained by an experiment consisting of random operations known as *trials*. The result of an unspecified trial is called a *random variable*. The collection of all possible values for the random variables associated with an experiment is called a *sample space*. In hydrologic phenomena, the observations for a certain period may be considered as a trial. By this trial, the rainfall depth, for example, is obtained as a random variable. Since the value of the rainfall depth can have all possible nonnegative values, the sample space is infinite.

Random variables are of two kinds: *discrete* and *continuous*. The discrete random variable has finite sample space, whereas the sample space of a continuous variable has an interval of real numbers or a union of such intervals. For example, the number of rainy days is a discrete random variable, while the depth of rainfall is a continuous random variable. For practical purposes, however, it is sometimes necessary to treat arbitrarily the discrete variables as continuous variables by fitting a continuous function to the variates, or vice versa by breaking down the continuous variable into intervals and then grouping them as discrete numbers.

B. Frequency, Probability, and Statistical Distributions

1. For Discrete Random Variables. For discrete random variables, the number of occurrences of a variate is generally called *frequency*. When the number of occurrences, or the frequency, is plotted against the variate as the abscissa, a pattern of distribution is obtained. This pattern is called *frequency distribution* (Fig. 8-I-1*a*).

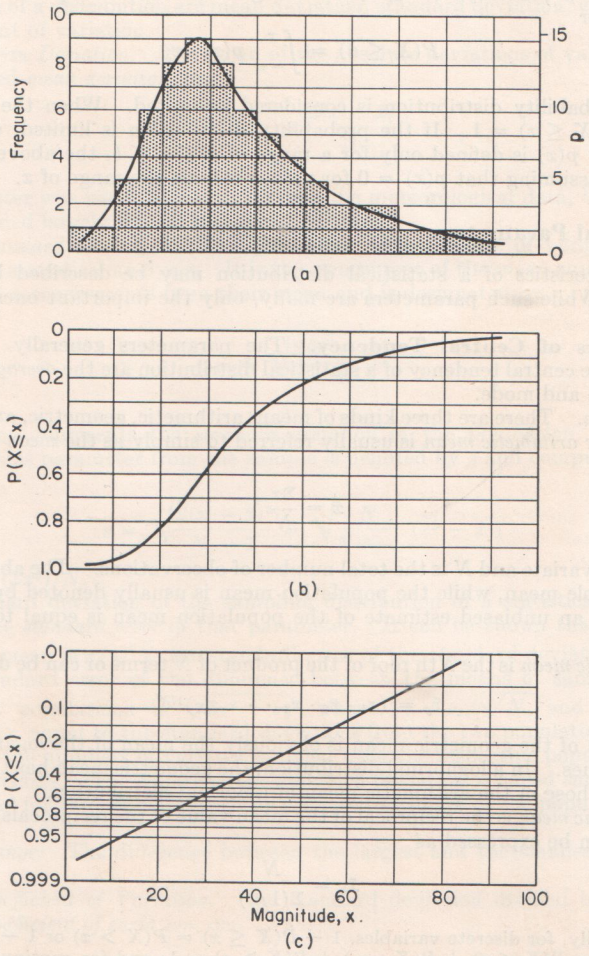


FIG. 8-I-1. Linearization of a statistical distribution. (*a*) Frequency or probability distribution curve; (*b*) cumulative probability curve plotted on rectangular coordinates; (*c*) cumulative probability curve linearized on probability paper.

When the number of occurrences of a discrete variate is divided by the total number of occurrences, the result is a probability p of the variate. The total probability for all variates should be equal to unity, or $\Sigma p = 1$. Distribution of the probabilities of all variates, instead of their frequencies, is called *probability distribution* (Fig. 8-I-1*a*).

The ordinates of the frequency distribution and its corresponding probability distribution are obviously proportional to each other. Both distributions may be called *statistical distributions*.

The *cumulative probability* of a variate (Fig. 8-I-1*b*) is the probability that the random variable has a value equal to or less than certain assigned value, say x. This cumulative probability may be designated as $P(X \leq x)$. Thus, the probability of being equal to or greater than x is equal to $1\text{-}P(X \leq x)$ or designated by $P(X \geq x)$.*

2. For Continuous Random Variables. For continuous random variables, the probability of a variate can be considered as the probability $p(x)$ of a discrete value grouped in the range from x to $x + \Delta x$. As x is a continuous value or Δx becomes dx, the probability $p(x)$ becomes a continuous function called *probability density*. The cumulative probability $P(X \leq x)$ is an integral function of the probability density (Fig. 8-I-1*b*), or

$$P(X \leq x) = \int_{-\infty}^{x} p(x)\, dx \tag{8-I-1}$$

where the probability distribution is considered unlimited. When the upper limit of $x = \infty$, $P(X \leq x) = 1$. If the probability distribution is limited, or the probability density $p(x)$ is defined only for a range of $a \leq x \leq b$, the above equation is also valid by assuming that $p(x) = 0$ for values outside the range of x.

C. Statistical Parameters

The characteristics of a statistical distribution may be described by *statistical parameters*. While such parameters are many, only the important ones are defined below:

1. Measures of Central Tendency. The parameters generally representing measures of the central tendency of a statistical distribution are the *averages*, including mean, median, and mode.

a. The Mean. There are three kinds of mean: arithmetic, geometric, and harmonic.

The familiar *arithmetic mean* is usually referred to simply as the *mean* and is designated by

$$\bar{x} = \frac{\Sigma x}{N} \tag{8-I-2}$$

where x is the variate and N is the total number of observations. The above equation gives the sample mean, while the population mean is usually denoted by μ. It may be noted that an unbiased estimate of the population mean is equal to the sample mean.

The *geometric mean* is the Nth root of the product of N terms or can be designated by

$$\bar{x}_g = (x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_N)^{1/N} \tag{8-I-3}$$

The logarithm of the geometric mean is obviously the mean of the logarithms of the individual values. In a lognormal distribution, the geometric mean has the properties analogous to those of the arithmetic mean of a normal distribution.

The *harmonic mean* is the reciprocal of the mean value of the reciprocals of individual values. It can be expressed as

$$\bar{x}_h = \frac{N}{\Sigma(1/x)} \tag{8-I-4}$$

* Theoretically, for discrete variables, $1 - P(X \leq x) = P(X > x)$ or $1 - P(X < x) = P(X \geq x)$, since $P(X < x) + P(X = x) + P(X > x) = 1$; and for continuous variables, $1 - P(X < x) = P(X > x)$, since $P(X = x)$ is infinitesimal. For practical purposes, $P(X \leq x) + P(X \geq x) = 1$ is acceptable for both discrete and continuous variables.

b. The Median. The *median* is the middle value or the variate which divides the frequencies in a distribution into two equal portions.

The arithmetic mean is more commonly used than other measures of central tendency as a "normal" or "standard" on account of its computational simplicity and, in general, its greater sampling stability. For example, the U.S. Weather Bureau has a long-established practice of using the mean as the precipitation normal. However, in extremely skew distributions the mean may be misleading. In such cases, the median will provide a better indication, particularly for a continuous variable because all variates greater or less than the median always occur half the time. Also, the use of median makes it easy to locate an internal estimate by adding and subtracting a specified amount from this central value so that the portions of the distribution outside the interval have the same probability.

c. The Mode. In a distribution of discrete variables, the *mode* is the variate which occurs most frequently. In a distribution of continuous variables, this is the variate which has a maximum probability density, i.e., $dp/dx = 0$ and $d^2p/dx^2 < 0$.

2. Measures of Variability. The important parameters representing variability or dispersion of a distribution are mean deviation, standard deviation, variance, range, and coefficient of variation.

a. The Mean Deviation. The mean of the absolute deviations of values from their mean is called *mean deviation*, or

$$\text{M.D.} = \frac{\Sigma|x - \bar{x}|}{N} \tag{8-I-5}$$

This parameter was used frequently to describe meteorological data, but it has been now superseded largely by the standard deviation.

b. The Standard Deviation. This parameter as a measure of variability is most adaptable to statistical analysis. It is the square root of the mean-squared deviation of individual measurements from their mean and is designated by

$$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}} \tag{8-I-6}$$

This equation represents the *standard deviation* of the population. An unbiased estimate of this parameter from the sample is denoted by s and computed by

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N - 1}} = \sqrt{\frac{N}{N - 1}\,(\overline{x^2} - \bar{x}^2)} \tag{8-I-7}$$

where $\overline{x^2} = (\Sigma x^2)/N$.

The standard deviation of the sampling distribution of a statistical parameter is known as the *standard error* of that parameter. It can be shown that the standard error of the mean is $\sigma/\sqrt{N}$, the standard error of the standard deviation is $\sigma/\sqrt{2N}$, and the standard error of the difference between the means of samples from two independent populations is $\sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2}$ where $\sigma_{\bar{x}} = \sigma_x/\sqrt{N_1}$ and $\sigma_{\bar{y}} = \sigma_y/\sqrt{N_2}$ with σ_x and σ_y equal to the standard deviations from the two populations and N_1 and N_2 equal to the numbers of variates sampled from the respective populations.

c. The Variance. The square of the standard deviation is called *variance*, which is denoted by σ^2 for the population. The unbiased estimate of the population variance is s^2.

d. The Range. The difference between the largest and the smallest values is the *range*.

e. The Coefficient of Variation. The standard deviation divided by the mean is called the *coefficient of variation*, or

$$C_v = \frac{\sigma}{\mu} \approx \frac{s}{\bar{x}} \tag{8-I-8}$$

3. Measures of Skewness. The lack of symmetry of a distribution is called *skewness* or *asymmetry*. The statistical parameter to measure this property is the *skewness* defined as

$$\alpha = \frac{1}{N}\sum (x - \mu)^3 \tag{8-I-9}$$

This equation represents the skewness for the population. An unbiased estimate of this parameter from the sample is

$$\begin{aligned} a &= \frac{N}{(N-1)(N-2)}\sum (x - \bar{x})^3 \\ &= \frac{N^2}{(N-1)(N-2)}\,(\overline{x^3} - 3\overline{x^2}\bar{x} + 2\bar{x}^3) \end{aligned} \tag{8-I-10}$$

where $\overline{x^3} = (\Sigma x^3)/N$ and other notations have been defined previously.

One commonly used measure of skewness is the *coefficient of skewness* represented by

$$C_s = \frac{\alpha}{\sigma^3} \approx \frac{a}{s^3} \tag{8-I-11}$$

For a symmetrical distribution, $C_s = 0$. A distribution with $C_s > 0$ is said to be skewed to the right (with a long tail on the right side), while a distribution with $C_s < 0$ is said to be skewed to the left.

Another measure of skewness often used in practice is *Pearson's skewness*, or

$$S_k = \frac{\mu - \text{mode}}{\sigma} \approx \frac{\bar{x} - \text{mode}}{s} \tag{8-I-12}$$

D. Statistical Moments

In a statistical distribution, the rth moment about the origin $x = 0$ of the variates $x_1, x_2, \ldots, x_k$, having a weighted mean $\bar{x}$, is

$$\nu_r = \frac{1}{N}\sum_{i=1}^{k} p_i x_i^r \tag{8-I-13}$$

where p_i is the frequency or probability of x_i and $N = \Sigma p_i$ with $i = 1, \ldots, k$.

The rth (central) moment about the weighted mean $\bar{x}$ of the variates $x_1, x_2, \ldots, x_k$, is

$$\mu_r = \frac{1}{N}\sum_{i=1}^{k} p_i (x_i - \mu)^r \tag{8-I-14}$$

For the first three moments, with $r = 1$, 2, and 3, it can be shown that

$$\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= \nu_2 - \nu_1^2 = \sigma^2 \\ \mu_3 &= \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3 = \sigma^3\alpha \end{aligned} \tag{8-I-15}$$

and

$$\begin{aligned} \nu_1 &= \bar{x} \\ \nu_2 &= \mu_2 + \nu_1^2 \\ \nu_3 &= \mu_3 + 3\mu_2\nu_1 + \nu_1^3 \end{aligned} \tag{8-I-16}$$

The above equations show that the mean is equal to the first moment about the origin, the standard deviation is the square root of the second moment about the mean, and the skewness is the third moment about the mean divided by the cube of the standard deviation. For a detailed discussion of the statistical moments, see Refs. 8, 73, and 74.

Moments of order higher than three are not commonly used in the statistical analy-

sis of hydrologic data because most hydrologic data do not have sufficiently long length of record and thus cannot warrant reliable estimates of the moments of higher order.

E. Hydrologic Models, Processes, and Systems

Hydrologic models considered here are mathematical formulations to simulate natural hydrologic phenomena which are considered as processes or as systems.

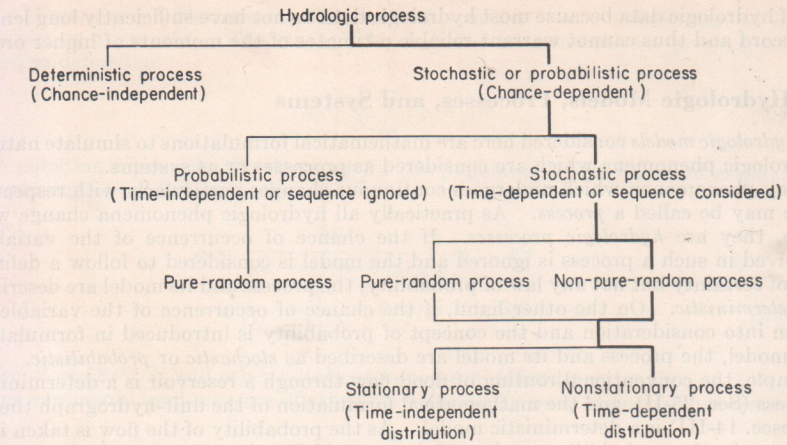
Any phenomenon which undergoes continuous changes particularly with respect to time may be called a *process*. As practically all hydrologic phenomena change with time, they are *hydrologic processes*. If the chance of occurrence of the variables involved in such a process is ignored and the model is considered to follow a definite law of certainty but not any law of probability, the process and its model are described as *deterministic*. On the other hand, if the chance of occurrence of the variables is taken into consideration and the concept of probability is introduced in formulating the model, the process and its model are described as *stochastic* or *probabilistic*. For example, the conventional routing of flood flow through a reservoir is a deterministic process (Sec. 25-II), and the mathematical formulation of the unit-hydrograph theory (Subsec. 14-III) is a deterministic model. As the probability of the flow is taken into account in the probability routing (Subsec. 14-V-C), the process and the queuing model employed to simulate the process are considered as stochastic or probabilistic.

Strictly speaking, a stochastic process is different from a probabilistic process, as the former is generally considered as time-dependent and the latter as time-independent. For the time-independent *probabilistic process*, the sequence of occurrence of the variates involved in the process is ignored and the chance of their occurrence is assumed to follow a definite probability distribution in which variables are considered pure-random. For the time-dependent *stochastic process*, the sequence of occurrence of the variates is observed and the variables may be either pure-random or non-pure-random, but the probability distribution of the variables may or may not vary with time. If *pure-random*, the members of the time series are independent among themselves and thus constitute a random sequence. If *non-pure-random*, the members of the time series are dependent among themselves, are composed of a deterministic component and a pure-random component, and thus constitute a nonrandom sequence. For example, the flow-duration-curve procedure (Subsec. 14-V-A) is probabilistic, whereas the probability routing mentioned above is stochastic.

In reality, all hydrologic processes are more or less stochastic. They have been assumed deterministic or probabilistic only to simplify their analysis. Mathematically speaking, a *stochastic process* is a family of random variables $X(t)$ which is a function of time (or other parameters) and whose variate x_t is running along in time t within a range T. Quantitatively, the stochastic process, which may be discrete or continuous, can be sampled continuously or at discrete or uniform intervals of $t = 1, 2, \ldots$, and the values of the sample form a sequence of $x_1, x_2, \ldots$, starting from a certain time and extending for a period of T. This sequence of sampled values is known as a *time series*, which may be discrete or continuous. For example, a hydrograph is a continuous time series. Daily, monthly, and annual discharges represent a discrete time series.

The random variable $X(t)$ has a certain probability distribution. If this distribution remains constant throughout the process, the process and the time series are said to be *stationary*. Otherwise, they are *nonstationary*. For example, a virgin flow (Subsec. 14-I) with no significant change in river-basin characteristics or climatic conditions for the period of record is considered as a stationary time series. If it is affected by man's activities in the river basin or nature's large accidental or slow modifications of the rainfall and runoff conditions, the recorded or historical flow is a nonstationary time series. Since a nonstationary process is very complicated mathematically, hydrologic processes are generally treated as stationary.

For clarity, the classification of hydrologic processes may be shown below. It may be noted that actual hydrologic processes are processes following the path of the heavy line while the processes following the thin lines are only approximations which may be assumed in order to simplify the analysis.



In this section, the probabilistic models and the frequency analysis of the probabilistic process are mainly discussed, since the stochastic process and various stochastic models are covered in several other sections, including Section 8-III (moving averages, sum of harmonics, autoregression, correlograms), Section 8-IV (Markov process), Section 18 (queuing theory), Section 14 (ranges, queuing theory, theory of storage), and Section 26-II (stochastic programming models).

A *system* is an aggregation or assemblage of objects united by some form of regular interaction or independence. The system is said to be *dynamic* if there is a process taking place in it. If the process is considered probabilistic or stochastic, the system is said to be *stochastic*. Otherwise, it is a *deterministic system*. Furthermore, the system is called *sequential* if it consists of *input*, *output*, and some working fluid (matter, energy, or information) known as *throughput* passing through the system. The hydrologic cycle or a drainage basin is a sequential, dynamic system in which water is a major throughput. Since a stochastic system is very complicated analytically, the hydrologic system has been generally treated as deterministic and its formulation by deterministic models, such as instantaneous unit hydrographs, has been proposed (Subsec. 14-III).

For detailed mathematical discussions on stochastic processes and on systems, see Refs. 75–82.

F. Statistical Homogeneity

The nature of homogeneity in hydrologic processes can be examined statistically with respect to time and space.

1. Time Homogeneity—Trend, Periodicity, and Persistence. A process or time series may be considered *time-homogeneous* if the identical events under consideration in the series are equally likely to occur at all times. Thus, purely random and stationary processes or time series are time-homogeneous. In hydrology, strictly time-homogeneous data are practically nonexistent because various kinds of variations of natural or artificial causes exist in most hydrologic phenomena. However, such variations, if appreciable, may be analyzed by various techniques.

Types of departure from true time homogeneity in hydrologic data may be roughly classified as trend, periodicity, and persistence.

a. Trend. This is a unidirectional diminishing or increasing change in the average value of a hydrologic variable, such as the trend of annual precipitation that is often plainly visible on a plotted graph. A number of statistical techniques may be used to determine the trend. A commonly used method is to analyze the trend by the method of moving averages (Subsec. 8-III-II-B).

b. Periodicity. This represents a regular or oscillatory form of variations, such as diurnal, seasonal, and secular changes that exist frequently in hydrologic phenomena. Such variations are of nearly constant length and they may be assumed sinusoidal and determined by harmonic analysis.

In the *harmonic analysis* a Fourier series is used to represent the time series x_1, x_2, . . . , x_N of a total period of length T:

$$x_t = \tfrac{1}{2}A_0 + \sum_{j=1}^{T/2}\left(A_j \cos\frac{360jt}{T} + B_j \sin\frac{360jt}{T}\right) \tag{8-I-17}$$

where A_0 is a constant, t is the time, and the coefficients A_j and B_j are *amplitudes* being expressed by

$$A_j = \frac{2}{N}\sum_{t=1}^{N} y_t \cos\frac{360jt}{T} \tag{8-I-17a}$$

$$B_j = \frac{2}{N}\sum_{t=1}^{N} y_t \sin\frac{360jt}{T} \tag{8-I-17b}$$

where y_t is the deviation of x_t from the arithmetic straight-line trend for the period selected, with $j = 1, 2, \ldots$, and N being the number of years of record used in the analysis. The sum of the squared amplitudes is

$$R_j^2 = A_j^2 + B_j^2 \tag{8-I-17c}$$

If no periodic fluctuations are present in the series; that is, if the series is a pure-random (nonautocorrelated) series of N terms having a normal distribution, the mean-squared amplitude of the series is

$$R_m^2 = \frac{4\sigma^2}{N} \tag{8-I-17d}$$

where σ^2 is the variance of the series y_t.

If the series has periodic fluctuations, three tests for periodicity are available [83–85]:

(1) Schuster Test. According to Schuster [86], the probability P_s in per cent that the squared amplitude R_j^2 is k times the mean-squared amplitude R_m^2 is

$$P_s = e^{-k} \tag{8-I-18}$$

where

$$k = \frac{R_j^2}{R_m^2} = -\ln P_s \tag{8-I-18a}$$

The value of R_j^2 for a given series can be tested to see if it differs from R_m^2 derived from a pure-random series. It is apparent that the higher the probability P_s the more likely the series is pure-random since the hypothesis being tested is that the series is pure-random. Generally $P_s = 10$ per cent may be taken as the level of significance. The corresponding value of $k = 2.303$. Thus, $R_j^2 = 2.303R_m^2 = 9.212\sigma^2/N$. Computing this value and substituting it in Eqs. (8-I-17*a* to *c*), the value of j can be computed. The possible hidden periodicity is equal to T/j.

(2) Walker Test. According to Walker [87], the probability that *at least* one squared amplitude R_j^2 will be k times R_m^2 is

$$P_w = 1 - (1 - e^{-k})^{N/2} \tag{8-I-19}$$

which may be used for a periodicity test as in the Schuster test.

(3) Fisher Test. Let R_J^2 be the largest of the squared amplitudes R_j^2. According to Fisher [88], the probability P_f that $R_J^2/2s^2$, where s^2 is the unbiased estimate of

σ^2, is greater than a given value g is

$$P_f = \sum_{i=0}^{m} (-1)^i \binom{j}{i} (1 - ig)^{j-1} \qquad (8\text{-I-}20)$$

where m is the greatest integer less than $1/g$, and $j = 1, 2, \ldots$ is the number of periods. This probability may be used for a periodicity test as in the Schuster test.

It must be noted that the above tests are based on normal distribution of the deviations and they apply only to strict periodicity and to nonautocorrelated data. Therefore, before using these tests, the effect of persistence or autocorrelation should be eliminated and the deviations be known as reasonably normally distributed.

Periodicity of secular nature is a matter of controversy [89]. Although the 11-year sunspot cycle is generally believed to have effect of various degrees upon hydrologic phenomena through the corresponding variations in solar radiant energy, statistical tests of possible astronomical effects on hydrologic phenomena have failed to show any statistical significance. Huntington [90] believed that very long secular variations are really not truly cyclic and therefore described them as *pulsations*.

Because of the seasonal effect on most hydrologic phenomena, *water year* instead of calendar year is usually adopted for the analysis of annual data. The water year will vary somewhat materially with the climatic conditions in various parts of the world. Water year usually starts when the ground and surface storage are both reduced to a minimum. The U.S. Geological Survey arranges the runoff data for a water year from October 1 to September 30. In England, the water year from September 1 to August 31 is sometimes used. Brakensiek [91] suggested that an optimum water year for tabulating water yield data can be determined by correlation analysis.

c. Persistence. This means that the successive members of a time series are linked among themselves in some persistent manner, resulting in non-pure-randomness. Due to meteorological and climatic causes, it has been found that wet years tend to occur in groups and dry years to occur together likewise. This tendency in grouping having the carryover effect of the immediate antecedent hydrologic conditions is the indication of the presence of persistence in hydrologic phenomena.

Since the carryover effect plays a significant part in hydrologic phenomena, it and hence the persistence are inversely related to the time interval between observations of such effects. When the time interval is shorter, the carryover effect or the persistence becomes more pronounced. As the effect of persistence exists, the degree of pure-randomness of the hydrologic data reduces. The magnitude of persistence may be determined by serial correlation analysis and correlograms (Subsec. 8-IV-II). It has been found that the magnitude depends on the type of hydrologic data; for example, it is higher in streamflows than in rainfalls.

Leopold [92] has described the nature of persistence with reference to probability analysis applied to a water-supply problem. He pointed out that Hurst [93] analyzed the longest record of river stage in the world (1,050 years of recorded stage of the Nile at the Roda gage) and obtained the evidence that the tendency for wet years to occur together and dry years together increased variability of means of various periods. In other words, the variability of groups of streamflows in their natural order of occurrence is actually larger than if the same flows occurred in random sequence. To illustrate this point, Leopold prepared Fig. 8-I-2 to show the variability of mean values of streamflow for records of various lengths. The dashed curve was plotted with grouped data taken from some longest streamflow records in the United States and Europe. If the annual streamflows were to occur in random sequence, the variability of means of groups would decrease inversely as the square root of the number of years comprising the group. Thus, the means of 100-year groups would be $1/\sqrt{100}$ or 1/10 as variable as 1-year values. The solid curve represents this random-sequence data. The difference between the dashed and solid curves represents the effect of persistence.

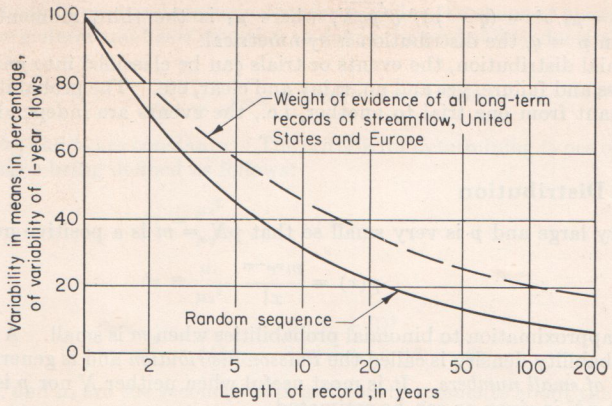


FIG. 8-I-2. Variability of mean values of streamflow for various lengths. (*Leopold* [92].)

2. Space Homogeneity. *Statistical meteorological homogeneity,* or *statistical hydrologic homogeneity, in space* implies that the occurrences of a particular meteorological, or hydrologic, event at all places within a so-called *statistically homogeneous area* are equally likely within a tolerable statistical difference. Because of the changes in geographical environment, statistically homogeneous areas are limited and can be delineated by statistical *regional analysis* (Subsec. IV-J).

III. PROBABILITY DISTRIBUTIONS

There are many probability distributions that have been found to be useful for hydrologic frequency analysis. Theoretical derivations and detailed discussions of such distributions can be found in many standard textbooks on statistics [73–75, 94–95].

A. Rectangular Distribution

The rectangular distribution is a uniform distribution of a continuous variable X between two constants a and b. The probability density of this distribution is

$$\begin{aligned} p(x) &= 0 && \text{for } x < a \\ p(x) &= \frac{1}{b-a} && \text{for } a \leq x \leq b \\ p(x) &= 0 && \text{for } b < x \end{aligned} \qquad \text{(8-I-21)}$$

and

The statistical parameters are: Mean $= (b + a)/2$; and variance $= (b - a)^2/12$.

B. Binomial Distribution

This is one of the most commonly used discrete distributions. It represents the distribution of probabilities in Bernoulli trials, say tossing a coin. The probability density is

$$p(x) = C_x{}^N p^x q^{N-x} \qquad \text{(8-I-22)}$$

where p is the probability of occurrence of an event, for example, a success in tossing a coin; $C_x{}^N$ is the number of combinations of N things taken x at a time; q is the probability of failure or $1 - p$; N is the total number of trials; and x is the variate or the number of successful trials.

The statistical parameters are: Mean $= pN$; standard deviation, $\sigma = \sqrt{pqN}$; and

skewness, $\alpha = \mu_3/\sigma^3 = (q - p)/\sqrt{pqN}$, where μ_3 is the third moment about the mean. When $p = q$, the distribution is symmetrical.

In a binomial distribution, the events or trials can be classified into only two catagories: success and failure, yes and no, rainy and clear, etc. The probabilities p and q remain constant from one trial to another, i.e., the events are independent to each other.

C. Poisson Distribution

If N is very large and p is very small so that $pN = m$ is a positive number, then

$$p(x) = \frac{m^x e^{-m}}{x!} \tag{8-I-23}$$

gives a close approximation to binomial probabilities when m is small. A distribution with this probability density is called the *Poisson distribution* and is generally referred to as the *law of small numbers*. It is most useful when neither N nor p is known but their product pN is given or can be estimated.

The statistical parameters are: Mean $= m$; standard deviation $= m$; and skewness $= 1/\sqrt{m}$.

D. Normal Distribution

This is a symmetrical, bell-shaped, continuous distribution, theoretically representing the distribution of accidental errors about their mean, or the so-called *Gaussian law of errors*. The probability density is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{8-I-24}$$

where x is the variate, μ is the mean value of the variate, and σ is the standard deviation. In this distribution, the mean, mode, and median are the same. The total area under the distribution is equal to 1.0. The cumulative probability of a value being equal to or less than x is

$$P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\mu)^2/2\sigma^2}\, dx \tag{8-I-25}$$

This represents the area under the curve between the variates of $-\infty$ and x. Areas for various values of x have been calculated by statisticians, and tables for such areas are available in many textbooks and handbooks on statistics.

E. Gamma Distribution

The probability density of this distribution is

$$p(x) = \frac{x^a e^{-x/b}}{b^{a+1}\,\Gamma(a+1)} \tag{8-I-26}$$

with $b > 0,\ a > -1$ for $x = 0$
and $p(x) = 0$ for $x \leq 0$

where a and b are constant and $\Gamma(a + 1) = a!$ is a *gamma function*. The cumulative probability being equal to or less than x $(< \infty)$ is known as the *incomplete gamma function*.

The statistical parameters are: Mean $= b(a + 1)$; and variance $= b^2(a + 1)$.

F. Pearson Distributions

Karl Pearson [96] has derived a series of probability functions to fit virtually any distribution. Although these functions have only slight theoretical basis, they have

been used widely in practical statistical works to define the shape of many distribution curves. The general and basic equation to define the probability density of a Pearson distribution is

$$p(x) = e^{\int_{-\infty}^{x} (a+x)/(b_0+b_1x+b_2x^2)dx} \tag{8-I-27}$$

where a, b_0, b_1, and b_2 are constants. The criteria for determining types of distribution are β_1, β_2, and κ being defined as follows:

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} \tag{8-I-28}$$

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2} \tag{8-I-29}$$

and

$$\kappa = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)} \tag{8-I-30}$$

where μ_2, μ_3, and μ_4 are the second, third, and fourth moments about the mean (Sheppard's corrections may be made if necessary).

With $\beta_1 = 0$, $\beta_2 = 3$, and $\kappa = 0$, the resulting Pearson distribution is identical with the normal distribution. Types I and III distributions are often used in the hydrologic frequency analysis.

1. Type I Distribution. For Type I, $\kappa < 0$. This is a skew distribution with limited range in both directions, usually bell-shaped but may be J-shaped or V-shaped. Its probability density is

$$p(x) = p_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2} \tag{8-I-31}$$

with $m_1/a_1 = m_2/a_2$ and the origin at the mode. The values of m_1 and m_2 are given by

$$m_1 \text{ or } m_2 = \frac{1}{2}\left[r-2 \pm r(r+2)\frac{\sqrt{\mu_2\beta_1}}{2(a_1+a_2)}\right] \tag{8-I-31a}$$

When μ_3 is positive, m_2 is the positive root and m_1 is the negative root; and vice versa in signs. The other values are

$$r = \frac{6(\beta_2-\beta_1-1)}{6+3\beta_1-2\beta_2} \tag{8-I-31b}$$

$$a_1+a_2 = \tfrac{1}{2}\sqrt{\mu_2[\beta_1(r+2)^2+16(r+1)]} \tag{8-I-31c}$$

and

$$p_0 = \frac{N}{a_1+a_2}\frac{{m_1}^{m_1}{m_2}^{m_2}}{(m_1+m_2)^{m_1+m_2}}\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)} \tag{8-I-31d}$$

where N is the total frequency.

The statistical parameters are: Mean = mode $-$ $(\mu_3/2\mu_2)[(r+2)/(r-2)]$; standard deviation = $\sqrt{\mu_2}$; and Pearson's skewness = $(\sqrt{\beta_1}/2)[(r+2)/(r-2)]$.

2. Type III Distribution. For Type III, $\kappa = \infty$ or $2\beta_2 = 3\beta_1 + 6$. This is a skew distribution with limited range in the left direction, usually bell-shaped but may be J-shaped. Its probability density with the origin at the mode is

$$p(x) = p_0\left(1+\frac{x}{a}\right)^{c} e^{-cx/a} \tag{8-I-32}$$

where

$$c = \frac{4}{\beta_1} - 1 \tag{8-I-32a}$$

$$a = \frac{c}{2}\frac{\mu_3}{\mu_2} \tag{8-I-32b}$$

$$p_0 = \frac{N}{a}\frac{c^{c+1}}{e^c\Gamma(c+1)} \tag{8-I-32c}$$

The statistical parameters are: Mean = mode $- \mu_3/2\mu_2$; standard deviation = $\sqrt{\mu_2}$; and Pearson's skewness = $\sqrt{\beta_1}/2$.

G. Extremal Distributions

Fréchet (on Type II) in 1927 [97] and Fisher and Tippett (on Types I and III) in 1928 [98] independently studied the distribution of extreme values and found that the distribution of the N largest (or the N smallest) values, each of which values is selected from one of m values contained in each of N samples, approaching a limiting (asymptotic) form as m is increased indefinitely. The type of the limiting form depends on the type of the initial distribution of the Nm values. For three different types of initial distribution, three asymptotic *extremal distributions* can be derived. A systematic study of the three asymptotes to the corresponding types of initial distributions was made by von Mises [99]. For detailed discussions on these distributions, see Refs. 100–102.

1. Type I Distribution. This distribution results from any initial distribution of *exponential type* which converges to an exponential function as x increases. Examples of such initial distributions are the normal, the chi-square, and the lognormal distributions. The probability density of Type I distribution is

$$p(x) = \frac{1}{c} e^{-(a+x)/c - e^{-(a+x)/c}} \tag{8-I-33}$$

with $-\infty < x < \infty$, where x is the variate, and a and c are parameters. The cumulative probability is

$$P(X \leq x) = e^{-e^{-(a+x)/c}} \tag{8-I-34}$$

By the method of moments, the parameters have been evaluated as

$$a = \gamma c - \mu \tag{8-I-34a}$$

and

$$c = \frac{\sqrt{6}}{\pi} \sigma \tag{8-I-34b}$$

where $\gamma = 0.57721 \ldots$ a Euler's constant, μ is the mean, and σ is the standard deviation. The distribution has a constant coefficient of skewness equal to $C_s = 1.139$.

2. Type II Distribution. This distribution results from an initial distribution of *Cauchy type* which has no moments from a certain order and higher. The cumulative probability is

$$P(X \leq x) = e^{-(\theta/x)^k} \tag{8-I-35}$$

with $0 \leq x < \infty$, where the parameter θ is the expected largest value defined for a sample of size n and increases with n, and k is an order of moments and independent of n.

3. Type III Distribution. This distribution results from a type of initial distribution in which x is limited by $x \leq \epsilon$. The cumulative probability is

$$P(X \leq x) = e^{-[(x-\epsilon)/(\theta-\epsilon)]^k} \tag{8-I-36}$$

with $-\infty < x \leq \epsilon$. The parameter k is the order of the lowest derivative of the probability function that does not vanish at $x = \epsilon$, and θ is the expected largest value.

In application, Type I distribution is sometimes known as *Gumbel distribution* since Gumbel [23] first applied it to flood frequency analysis. Type III is known as *Weibull distribution* since Weibull [103–104] first applied it to the description of the strength of brittle materials although Gumbel [57] also applied it later to drought frequency analysis.

H. Logarithmically Transformed Distributions

Many probability distributions can be transformed by replacing the variate with its logarithmic value. Three transformed distributions commonly used in hydrologic studies are as follows:

1. Lognormal Distribution. This is a transformed normal distribution in which the variate is replaced by its logarithmic value. This distribution represents the so-called *law of Galton* because it was first studied by Galton [105] as early as 1875. Its probability density is

$$p(x) = \frac{1}{\sigma_y e^y \sqrt{2\pi}} e^{-(y-\mu_y)^2/2\sigma_y^2} \tag{8-I-37}$$

where $y = \ln x$, x is the variate, μ_y is the mean of y, and σ_y is the standard deviation of y. This is a skew distribution of unlimited range in both directions.

Chow [15] has derived the statistical parameters for x as

$$\mu = e^{\mu_y + \sigma_y^2/2} \tag{8-I-37a}$$

$$\sigma = \mu(e^{\sigma_y^2} - 1)^{1/2} \tag{8-I-37b}$$

$$\alpha = (e^{3\sigma_y^2} - 3e^{\sigma_y^2} + 2)C_v^3 \tag{8-I-37c}$$

$$M = e^{\mu_y} \tag{8-I-37d}$$

$$\frac{\mu}{M} = e^{\sigma_y^2/2} \tag{8-I-37e}$$

$$C_v = (e^{\sigma_y^2} - 1)^{1/2} \tag{8-I-37f}$$

$$C_s = 3C_v + C_v^3 \tag{8-I-37g}$$

where μ is the mean, σ is the standard deviation, C_s is the coefficient of skewness, M is the median, and C_v is the coefficient of variation. Chow [15] has also shown that the Type I extremal distribution is essentially a special case of the lognormal distribution when $C_v = 0.364$ and $C_s = 1.139$. For other discussions, see Refs. 106–109.

2. Logextremal Distributions. Let x be replaced by y in Eq. (8-I-34) and then equate Eq. (8-I-34) to Eq. (8-I-35) and to Eq. (8-I-36). It can be found that for Type II extremal distribution, y is a linear function of $\ln x$, and for Type III extremal distribution, y is a linear function of $\ln (x - \epsilon)$. In other words, if the variate x in Type I distribution is replaced by a linear function of the logarithm of x and $x - \epsilon$, the resulting logarithmically transformed distributions become Type II and Type III distributions respectively.

3. Truncated Lognormal Distributions. Slade [22] introduced two truncated and shifted logarithmically transformed normal distributions for hydrologic frequency analysis. One is called the *partly bounded distribution* which has an unlimited range only in the positive direction of the variate. Its probability density is

$$p(x) = ae^{-c^2[\ln d(x+b)]^2} \tag{8-I-38}$$

with $-b \leq x < \infty$, where a, b, c, and d are parameters which can be derived from the first three statistical moments.

The other is called the *totally bounded distribution* which has the maximum and minimum limits of fluctuations from the mean. Its probability density is

$$p(x) = ae^{-p^2c^2\{\ln d[(x+b)/(g-x)]\}^2} \tag{8-I-39}$$

with $-b < x < g$, where the parameters a, p, c, and d are determined empirically.

IV. PROCEDURE OF ANALYSIS

Frequency analysis of hydrologic data starts with the treatment of raw hydrologic data and finally determines the frequency or probability of a hydrologic design value. Since the time sequence of hydrologic phenomena is not considered primarily in this

section, probabilistic frequency analysis is mainly discussed here. In such analyses, a probability distribution is assumed as a mathematical model to which the hydrologic frequency data are to be fitted without considering the sequence of occurrence of the data. See Subsec. II-E.

A. Treatment of Raw Data

1. Data Sampling. Large masses of hydrologic data are unwieldy and uneconomical to analyze and their population is infinite or nearly infinite in size. For the use in analysis, the data must be sampled. For probabilistic frequency analysis, it is required that samples be pure-random. In other words, they should be unbiased, independent, and homogeneous.

A sample for which the sampling procedure is entirely by chance is called an *unbiased*, or a *random*, sample. In order to prevent the sample's being biased, the sample must be as representative as possible of the total population. In the collection of rainfall data in a drainage basin, for example, the stations should be so located that a large part of the basin would be covered and that various types of basin conditions would be represented. Such a representative sample is called a *stratified* sample, which is the opposite of a *spot* sample that is taken only from one small area or class of population.

Dependence of data may be referred to on either time or space. *Time dependence* is the major cause for non-pure-randomness of the data. For example, two successive floods occurring very closely may result in a high degree of dependence as the storm producing the first flood may effectively affect the meteorological condition that produces the second flood. *Space dependence* may be a major reason to produce unstratified data. For example, two rainfall stations placed closely together will produce practically identical data and should be considered only as one station in computing mean rainfall.

Lack of *homogeneity* means that the samples are taken from two different populations. For example, temperatures taken under the sun should not be averaged with those taken in the shade if the two conditions are considered as constituting different populations in the analysis.

2. Observation Errors. Nowadays vast amounts of hydrologic data are being collected. The basic form of such data is generally a continuous record in time, which is too bulky for publication. Usually, only selected or processed data are published.

Measurement and publication often involve instrumental and human errors. Such errors may be considered of two kinds, namely accidental and systematic errors, although it is sometimes difficult to distinguish between them and many errors are a combination of the two kinds. *Accidental errors* are usually due to the observer and sometimes due to the uncertain nature of the measuring instrument. Such errors may be considered random errors; they are disordered in their incidence and variable in magnitude, positive and negative values occurring in repeated measurements in no ascertainable sequence. On the other hand, *systematic errors* may arise from the observer or the instrument. Such errors are not random; they may be constant and create a trend, or vary in some regular way and produce periodicity.

3. Inherent Defectiveness. Major defectiveness of hydrologic data, such as non-pure-randomness, nonstationarity, missing data, etc., should be investigated. If they affect measurably the basic assumptions required in probabilistic frequency analysis, the raw data should be adjusted accordingly by various methods, such as serial correlation analysis for persistence, moving averages for trend, and Fourier or harmonic analysis for periodicity (Subsec. II-F-1). Missing data may sometimes be estimated by regional analysis by correlation with other hydrologic data in the neighborhood (Subsec. 9-V).

Statistical properties of hydrologic phenomena may also depend on inferences derived from long-term nonhydrologic natural data. Examples of such data which may possibly be used for this purpose include widths of tree rings, pattern of fossil pollen, distribution of clay varves, fluctuations in levels of closed lakes, glacial movement, and very long-range historical records of extraterrestrial phenomena. These

nonhydrologic phenomena contain the intrinsic record of the nature of non-pure-randomness, nonstationarity, and other characteristics of time-series events. They may be used to improve the quality of hydrologic data through statistical inferences. However, they are available only in limited number. Furthermore, their use in statistical inferences requires the understanding of the processes involved in these natural phenomena and the correct interpretation of the results obtained from the inferences.

It may be noted that hydrologic phenomena seldom completely satisfy the requirements of the statistical theory. Before the raw data are used for frequency analysis, they should be examined for possible observation errors and inherent defectiveness. If such errors and defectiveness are appreciable, they should be analyzed and corrected before the frequency analysis can be suitably applied.

B. Selection of Data Series

The available hydrologic data are generally presented in chronological order. Figure 8-I-3 exhibits a hypothetical set of such data for a certain period of observation, say 20 years as shown in the figure. The magnitude of data is expressed in an arbitrary unit. Since all available data are shown, they constitute a *complete-duration series.*

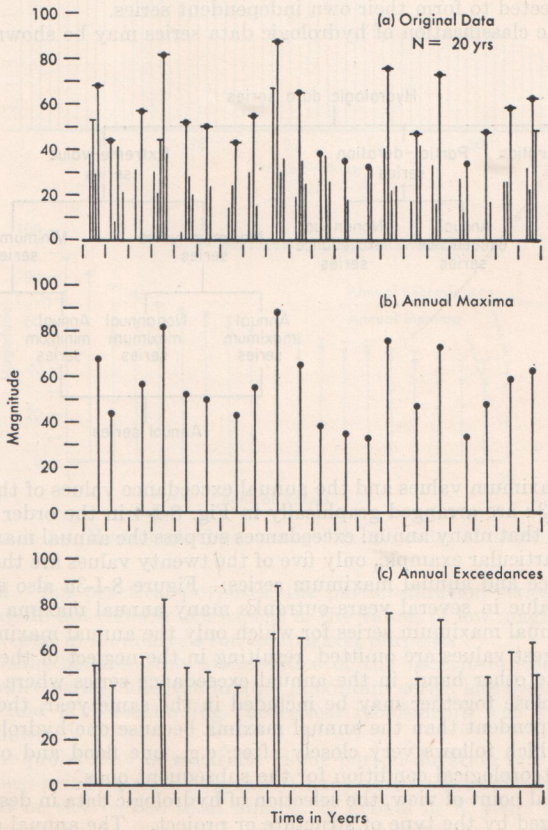


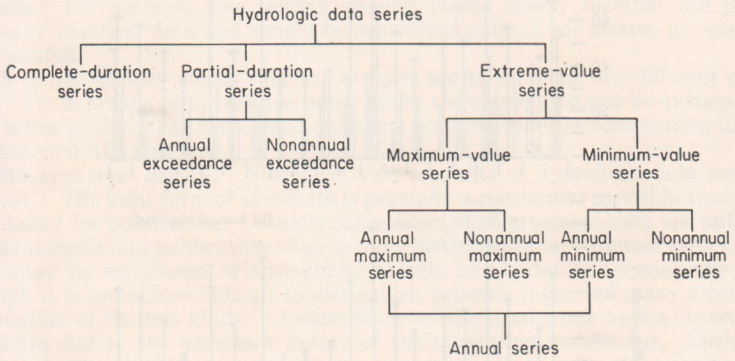
FIG. 8-I-3. Hydrologic data arranged in the order of occurrence.

Experience has shown that many of the original data have practically no significance in the analysis because the hydrologic design of a project is usually governed by a few critical conditions only. Thus, except sometimes in a few cases such as in the analysis by duration curves and mass curves, the complete-duration series is not always used. In order to save labor and time in the publication and analysis of the data, the data of insignificant magnitude should be excluded. For this purpose, two types of data are generally selected from the complete-duration series: the partial-duration series and the extreme-value series.

The *partial-duration series*, or *partial series*, is a series of data which are so selected that their magnitude is greater than a certain *base value*. If the base value is selected so that the number of values in the series is equal to the number of the record, the series is called *annual exceedance series* as shown in Fig. 8-I-3*c*.

The *extreme-value series* includes the largest or smallest values with each value selected from an equal time interval in the record. The time interval is usually taken as one water year and the series so selected is the *annual series*. For largest annual values, it is an *annual maximum series* as shown in Fig. 8-I-3*b*. For smallest annual values, it is an *annual minimum series*. When the time interval decreases, the dependence between observations and the number of selected values increase. If the time interval is less than one year, the seasonal variation will further introduce nonhomogeneity to the data. However, homogeneity of the data may be maintained at least for practical purposes if the data are selected only from a particular season, month, or other definite duration within a year [48]. For example, summer storms and spring floods can be selected to form their own independent series.

For clarity, the classification of hydrologic data series may be shown as follows:



The annual maximum values and the annual exceedance values of the hypothetical data in Fig. 8-I-3*a* are arranged graphically in Fig. 8-I-4 in the order of magnitude. The figure shows that many annual exceedances surpass the annual maxima in magnitude. In this particular example, only five of the twenty values are the same in both annual exceedance and annual maximum series. Figure 8-I-3*a* also shows that the second largest value in several years outranks many annual maxima in magnitude. Thus, in the annual maximum series for which only the annual maxima are selected these second largest values are omitted, resulting in the neglect of their effect in the analysis. On the other hand, in the annual exceedance series where several values which occurred close together may be included in the same year, the selected data may be less independent than the annual maxima because one hydrologic event can affect another which follows very closely after, e.g., one flood and one storm may influence the meteorological condition for the subsequent ones.

From the logical point of view, the selection of hydrologic data in designing a structure may be judged by the type of structure or project. The annual exceedances or the partial-duration series should be used if the second largest values in the year would

affect the design. For instance, the damage to bridge foundations caused by flooding sometimes results from the repetition of flood occurrence rather than from a single peak flow. A culvert subject to flood damage or destruction may be rapidly and economically repaired or restored and then soon again exposed to future damage. Similarly, in highway drainage, the loss due to traffic interruption as a result of flooding will be weighed by the number of flood peaks and the extent of flooding which are largely caused by associated peak flows. In other cases where the design is governed by the most critical condition, such as the design of a spillway, the annual maxima

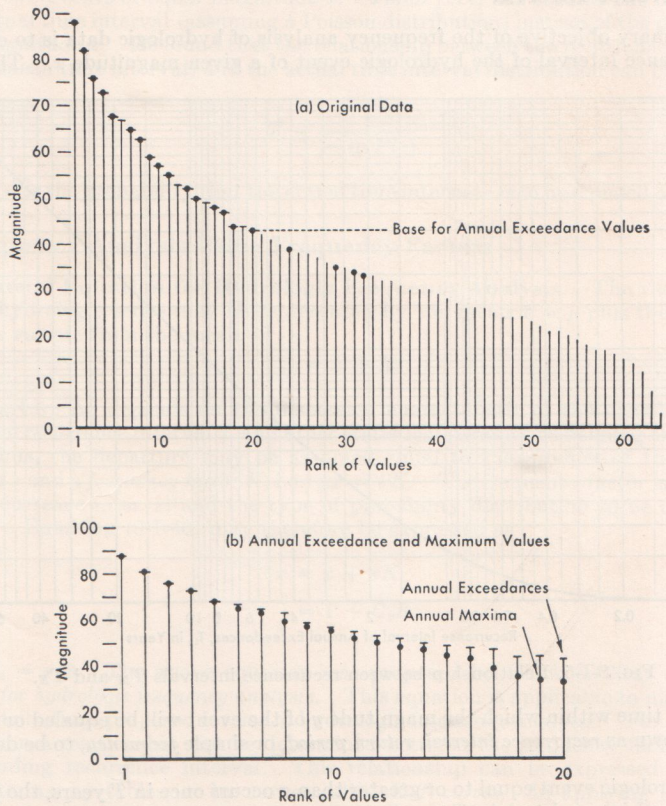


FIG. 8-I-4. Hydrologic data arranged in the order of magnitude.

should be used. For practical purposes, the partial-duration series and the annual maximum series do not differ much except in the values of low magnitude. Usually, both series are used in an analysis for comparison purposes.

The relationship between the probabilities of the partial-duration series (or the annual exceedance series) and the annual maximum series has been investigated by Langbein [37] and a corresponding theoretical relationship was derived by Chow [52, 110] as follows:

Let P_E be the probability of a variate in the partial-duration series (or the annual exceedance series) being equal to or greater than x, and let m be the average number of events per year or mN be the total number of events in N years of record. Then P_E/m is the probability of an event being equal to x or greater, and $1 - P_E/m$ is the probability of an event being less than x. Thus the probability of an event of magni-

tude x becoming a maximum of the m events in a year is $(1 - P_E/m)^m$. This probability approaches e^{-P_E} when P_E is small compared with m, which is true for most cases. Therefore, the probability P_M of an annual maximum of magnitude x being equaled or exceeded is equal to

$$P_M = 1 - e^{-P_E} \tag{8-I-40}$$

or

$$P_E = -\ln (1 - P_M) \tag{8-I-41}$$

It can be shown that $P_M \approx P_E$ as both P_M and P_E become large.

C. Recurrence Interval

The primary objective of the frequency analysis of hydrologic data is to determine the recurrence interval of the hydrologic event of a given magnitude x. The *average*

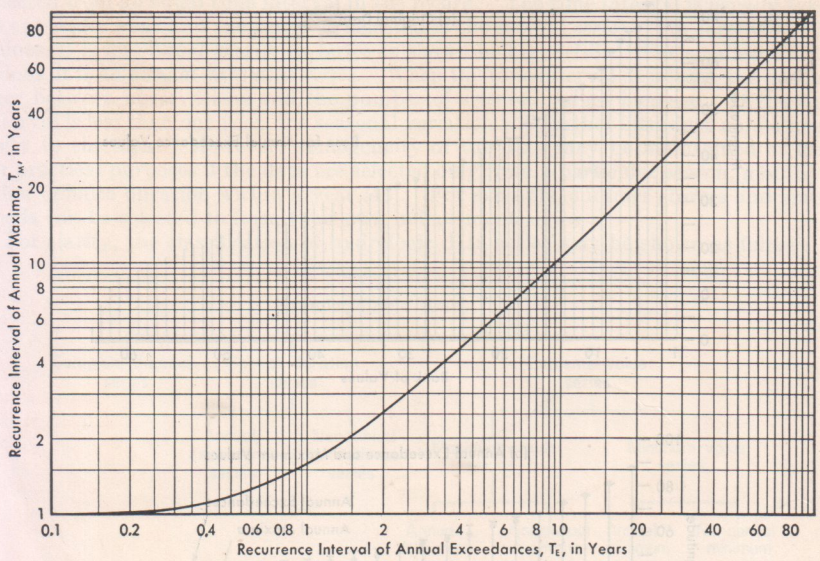


FIG. 8-I-5. Relationship between recurrence intervals T_M and T_E.

interval of time within which the magnitude y of the event will be equaled or exceeded once is known as *recurrence interval, return period,* or simple *frequency,* to be designated by T.

If a hydrologic event equal to or greater than x occurs once in T years, the probability $P(X \geq x)$ is equal to 1 in T cases, or

$$P(X \geq x) = \frac{1}{T} \tag{8-I-42}$$

Hence,

$$T = \frac{1}{P(X \geq x)} = \frac{1}{1 - P(X \leq x)} \tag{8-I-43}$$

If T_M and T_E are the recurrence intervals of the annual maximum series and the partial-duration series (or the annual exceedance series) respectively, and P_M and P_E are their corresponding probabilities of being equal to and greater than the magnitude x, Eq. (8-I-42) gives $P_M = 1/T_M$ and $P_E = 1/T_E$. Substituting these expressions of P_M and P_E in Eq. (8-I-40) and simplifying, the relationship between the two recurrence intervals is

$$T_E = \frac{1}{\ln T_M - \ln (T_M - 1)} \tag{8-I-44}$$

This relationship is plotted as shown in Fig. 8-I-5. Langbein [37] has plotted a number of actual cases which were found to be very close to this theoretical curve. It can be shown from this curve that a difference between P_M and P_E is equal to about 10 per cent when P_E is 5 years, and the difference becomes about 5 per cent when P_E is 10 years. In ordinary hydrologic analysis, a 5 per cent difference may be considered tolerable. Therefore, it may be concluded that the two recurrence intervals are practically identical for recurrence intervals greater than 10 years.

It should be noted that the recurrence interval as usually defined is a mean time interval based on the distribution of the variate x; it is not the actual time interval between two events of equal magnitude x. Thom [111] has studied the distribution of the actual time interval (assuming a Poisson distribution) instead of the distribution of the magnitude x. He found that the relationship between the recurrence interval T (i.e., the mean time interval) and the actual time interval distribution can be expressed as

$$T = \frac{\tau}{-\ln P(\tau)} \tag{8-I-45}$$

where $P(\tau)$ is the probability that the actual time interval τ is to be equaled or exceeded.

D. Frequency Analysis Using Frequency Factors

1. General Equation for Hydrologic Frequency Analysis. The variate x of a random hydrologic series may be represented by the mean $\bar{x} \approx \mu$ plus the departure Δx of the variate from the mean, or

$$x = \bar{x} + \Delta x \tag{8-I-46}$$

The departure Δx depends on the dispersion characteristic of the distribution of x and on the recurrence interval T and other statistical parameters defining the distribution. Thus, the departure may be assumed equal to the product of the standard deviation σ and a *frequency factor* K, i.e., $\Delta x = \sigma K$. The frequency factor is a function of the recurrence interval and the type of probability distribution to be used in the analysis. Equation (8-I-46) may therefore be expressed as

$$x = \bar{x} + \sigma K \tag{8-I-47}$$

or

$$\frac{x}{\bar{x}} = 1 + C_v K \tag{8-I-48}$$

where $C_v = \sigma/\bar{x}$. The above equation was proposed by Chow [112] as the *general equation for hydrologic frequency analysis*. This equation is applicable to many probability distributions proposed for use in hydrologic frequency analysis. For a proposed distribution a relationship can be determined between the frequency factor and the corresponding recurrence interval. This relationship can be expressed in mathematical terms, by tables, or by curves called *K-T curves*. In applying the general equation, the statistical parameters required in the proposed distribution are first computed from the random hydrologic data series. For a given recurrence interval, the frequency factor can be determined from the *K-T* relationship for the proposed distribution and the magnitude x for the recurrence interval can be computed by Eq. (8-I-47) or (8-I-48), using the corresponding frequency factor and the computed statistical parameters.

2. The *K-T* Relationship. Based on the observations of many streams, Fuller [10] derived the earliest empirical formula for the frequency analysis of annual maximum daily flow as

$$x = \bar{x}(1 + 0.8 \log T) \tag{8-I-49}$$

Comparing this formula with Eq. (8-I-48), the frequency factor can be easily found as

$$K = \frac{0.8}{C_v} \log T \tag{8-I-50}$$

which is a function of C_v and T. The value of C_v varies from 0.1 to 2.0, having an average of 0.50. The Fuller formula is actually based on an empirical statistical distribution and thus the K-T relationship so derived is also empirical. Since Fuller's

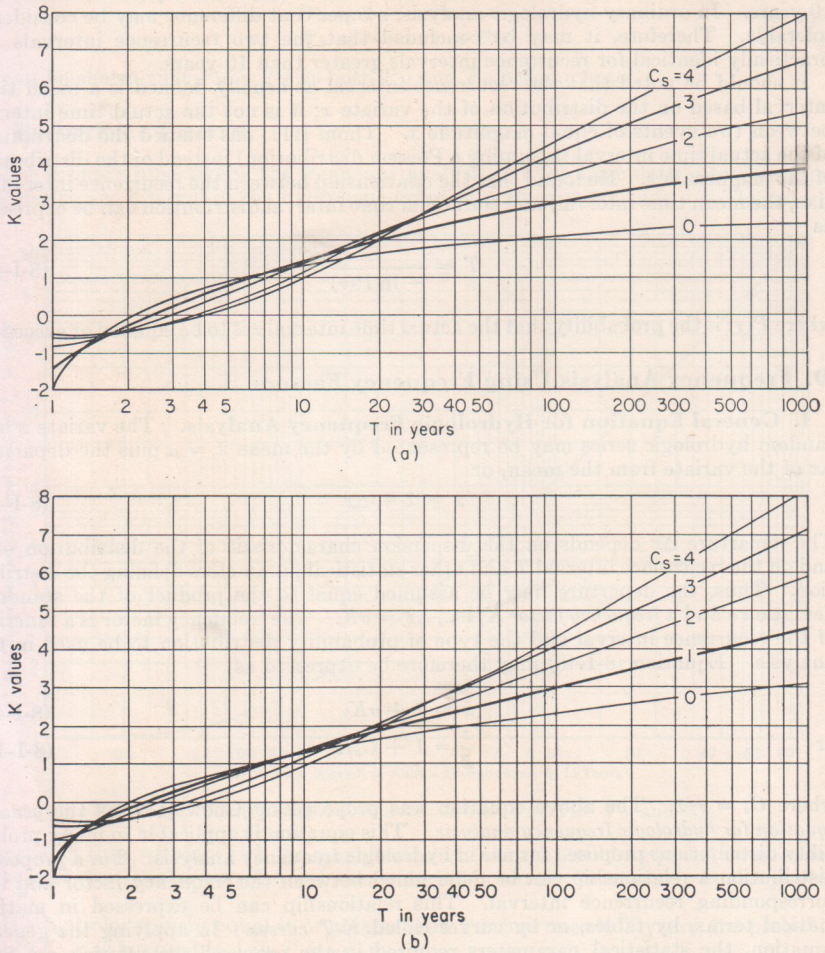


FIG. 8-I-6. K-T curves for Pearson distributions. (*a*) Pearson Type I distribution. (*b*) Pearson Type III distribution.

time, many methods using theoretical distributions have been proposed. The K-T relationships for some important theoretical distributions are discussed below:

a. Normal Distribution. Taking $\bar{x} = \mu$, the frequency factor can be expressed from Eq. (8-I-47) as

$$K = \frac{x - \mu}{\sigma} \tag{8-I-51}$$

Substituting this expression in Eq. (8-I-25),

$$P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{K} e^{-K^2/2}\, dK \tag{8-I-52}$$

Values of $P(X \leq x)$ for various values of K can be found from normal probability tables in many textbooks and handbooks on statistics. The corresponding values of T can be obtained from Eq. (8-I-43).

b. Pearson Distributions. Foster [17] proposed a method in which the Pearson Type I and Type III distributions are used. From Foster's derivation, the frequency factor of these distributions can be shown by K-T curves in Fig. 8-I-6. Foster suggested that the coefficient of skewness computed by Eq. (8-I-11) should be multiplied

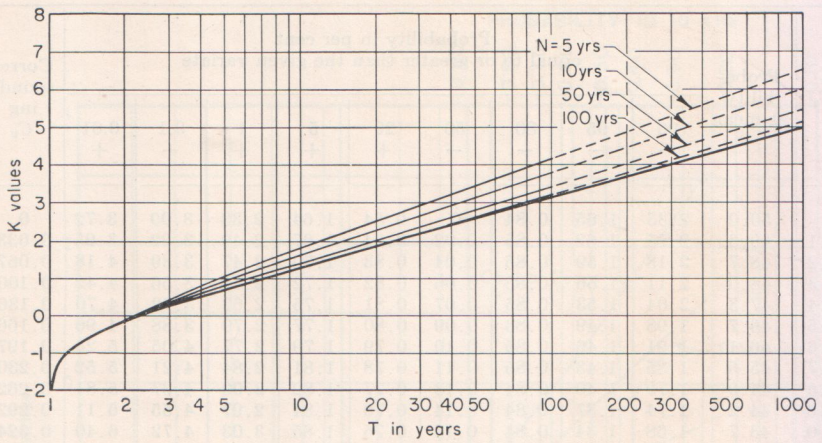


FIG. 8-I-7. K-T curves for Type I extremal distributions.

by $1 + 8.5/N$ for Type I distribution and by $1 + 6/N$ for Type III distribution in order to adjust the influence due to the length of hydrologic records.

c. Extremal Distribution. The frequency factor for the Type I extremal distribution can be derived from Eq. (8-I-33) in a manner similar to that described above for the normal distribution. It has been given by Chow [52, 112] as

$$K = -\frac{\sqrt{6}}{\pi}\left[\gamma + \ln \ln \left(\frac{T}{T-1}\right)\right] \tag{8-I-53}$$

which is plotted as the heavy line in Fig. 8-I-7. Potter [32] studied 370 Type I extremal distributions for maximum rainfall intensities of various durations, monthly and annual rainfall amounts, and peak rates of surface runoff. The result of his study can be plotted in K-T curves as shown in Fig. 8-I-7. It can be seen that the K-T relationship so obtained depends on the number of years of record, N. These curves are shown by thin lines with the dashed portions extrapolated. As N increases, the K-T relationship approaches the theoretical relationship which is derived for the population. The curve for $N = 100$ years is practically identical with the theoretical curve.

When $x = \bar{x}$, Eq. (8-I-48) gives $K = 0$ and thus Eq. (8-I-53) results in $T = 2.33$ years. This is the recurrence interval of the mean of the Type I extremal distribution. It is taken by U.S. Geological Survey as the *recurrence interval of a mean annual flood.*

d. Lognormal Distribution. Substituting $x = e^y$, $\bar{x} \approx \mu$ by Eq. (8-I-37*a*), and σ by Eq. (8-I-37*b*) in Eq. (8-I-47), Chow [15, 106] has derived the frequency factor for the lognormal distribution as

$$K = \frac{e^{\sigma_y K_y - \sigma_y^2/2} - 1}{(e^{\sigma_y^2} - 1)^{1/2}} \tag{8-I-54}$$

where $K_y = (y - \bar{y})/\sigma_y$ and can be expressed in a form similar to Eq. (8-I-47) as

$$y = \bar{y} + \sigma_y K_y \tag{8-I-55}$$

Equation (8-I-37) shows that y is normally distributed while its antilogarithm x is lognormally distributed. For a given recurrence interval T, or probability $P(X \geq x)$ or $P(X \leq x)$, the value of K_y can be computed in a manner similar to that described above for the normal distribution. When K_y is known, the value of K can be computed by Eq. (8-I-54) for any given value of σ_y. The value of σ_y and the corresponding

Table 8-I-1. Frequency Factors for Lognormal Distribution

C_s	Probability at mean	Probability in per cent equal to or greater than the given variate									Corresponding C_v
		99 −	95 −	80 −	50 −	20 +	5 +	1 +	0.1 +	0.01 +	
0	50.0	2.33	1.65	0.84	0	0.84	1.64	2.33	3.09	3.72	0
0.1	49.3	2.25	1.62	0.85	0.02	0.84	1.67	2.40	3.22	3.95	0.033
0.2	48.7	2.18	1.59	0.85	0.04	0.83	1.70	2.47	3.39	4.18	0.067
0.3	48.0	2.11	1.56	0.85	0.06	0.82	1.72	2.55	3.56	4.42	0.100
0.4	47.3	2.04	1.53	0.85	0.07	0.81	1.75	2.62	3.72	4.70	0.136
0.5	46.7	1.98	1.49	0.86	0.09	0.80	1.77	2.70	3.88	4.96	0.166
0.6	46.1	1.91	1.46	0.85	0.10	0.79	1.79	2.77	4.05	5.24	0.197
0.7	45.5	1.85	1.43	0.85	0.11	0.78	1.81	2.84	4.21	5.52	0.230
0.8	44.9	1.79	1.40	0.84	0.13	0.77	1.82	2.90	4.37	5.81	0.262
0.9	44.2	1.74	1.37	0.84	0.14	0.76	1.84	2.97	4.55	6.11	0.292
1.0	43.7	1.68	1.34	0.84	0.15	0.75	1.85	3.03	4.72	6.40	0.324
1.1	43.2	1.63	1.31	0.83	0.16	0.73	1.86	3.09	4.87	6.71	0.351
1.2	42.7	1.58	1.29	0.82	0.17	0.72	1.87	3.15	5.04	7.02	0.381
1.3	42.2	1.54	1.26	0.82	0.18	0.71	1.88	3.21	5.19	7.31	0.409
1.4	41.7	1.49	1.23	0.81	0.19	0.69	1.88	3.26	5.35	7.62	0.436
1.5	41.3	1.45	1.21	0.81	0.20	0.68	1.89	3.31	5.51	7.92	0.462
1.6	40.8	1.41	1.18	0.80	0.21	0.67	1.89	3.36	5.66	8.26	0.490
1.7	40.4	1.38	1.16	0.79	0.22	0.65	1.89	3.40	5.80	8.58	0.517
1.8	40.0	1.34	1.14	0.78	0.22	0.64	1.89	3.44	5.96	8.88	0.544
1.9	39.6	1.31	1.12	0.78	0.23	0.63	1.89	3.48	6.10	9.20	0.570
2.0	39.2	1.28	1.10	0.77	0.24	0.61	1.89	3.52	6.25	9.51	0.596
2.1	38.8	1.25	1.08	0.76	0.24	0.60	1.89	3.55	6.39	9.79	0.620
2.2	38.4	1.22	1.06	0.76	0.25	0.59	1.89	3.59	6.51	10.12	0.643
2.3	38.1	1.20	1.04	0.75	0.25	0.58	1.88	3.62	6.65	10.43	0.667
2.4	37.7	1.17	1.02	0.74	0.26	0.57	1.88	3.65	6.77	10.72	0.691
2.5	37.4	1.15	1.00	0.74	0.26	0.56	1.88	3.67	6.90	10.95	0.713
2.6	37.1	1.12	0.99	0.73	0.26	0.55	1.87	3.70	7.02	11.25	0.734
2.7	36.8	1.10	0.97	0.72	0.27	0.54	1.87	3.72	7.13	11.55	0.755
2.8	36.6	1.08	0.96	0.72	0.27	0.53	1.86	3.74	7.25	11.80	0.776
2.9	36.3	1.06	0.95	0.71	0.27	0.52	1.86	3.76	7.36	12.10	0.796
3.0	36.0	1.04	0.93	0.71	0.28	0.51	1.85	3.78	7.47	12.36	0.818
3.2	35.5	1.01	0.90	0.69	0.28	0.49	1.84	3.81	7.65	12.85	0.857
3.4	35.1	0.98	0.88	0.68	0.29	0.47	1.83	3.84	7.84	13.36	0.895
3.6	34.7	0.95	0.86	0.67	0.29	0.46	1.81	3.87	8.00	13.83	0.930
3.8	34.2	0.92	0.84	0.66	0.29	0.44	1.80	3.89	8.16	14.23	0.966
4.0	33.9	0.90	0.82	0.65	0.29	0.42	1.78	3.91	8.30	14.70	1.000
4.5	33.0	0.84	0.78	0.63	0.30	0.39	1.75	3.93	8.60	15.62	1.081
5.0	32.3	0.80	0.74	0.62	0.30	0.37	1.71	3.95	8.86	16.45	1.155

value of C_v for an assigned value of C_s can be computed by means of Eqs. (8-I-37b) and (8-I-37g). Table 8-I-1 is a list of frequency factors so computed for assigned values of C_s and of various probabilities $P(X \geq x)$. The table also lists the probabilities at mean, which occur when $K = 0$ or $K_y = \sigma_y/2$. The recurrence interval T is related to $P(X \geq x)$ by Eq. (8-I-43). Thus, this table can be used to plot K-T curves using C_s as the parameter.

E. Probability Paper

The cumulative probability of a distribution may be represented graphically on a *probability paper* which is designed for the distribution. On such paper the ordinate usually represents the value of x in certain scale and the abscissa represents the probability $P(X \geq x)$ or $P(X \leq x)$, or the recurrence interval T. The ordinate and abscissa scales are so designed that the distribution plots as a straight line and the

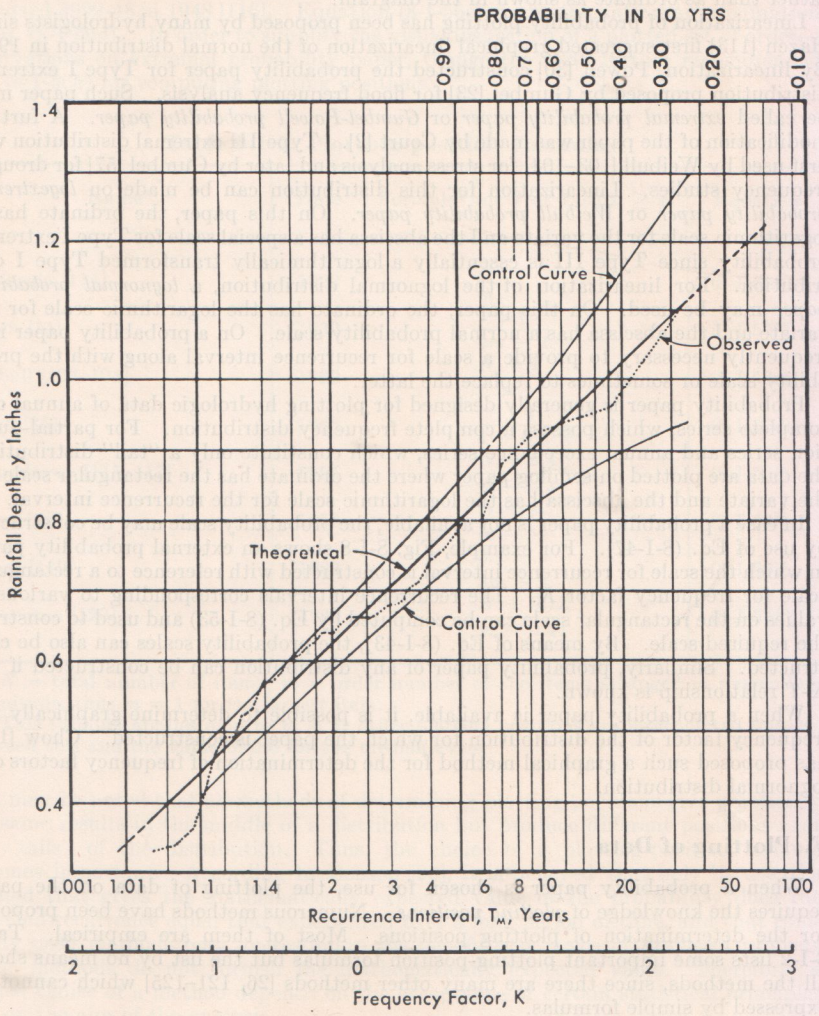


FIG. 8-I-8. Probability plotting of annual maxima of 10-min rainfall depth.

data to be fitted appear close to the straight line. The objective of using the probability paper is to linearize the distribution so that the plotted data can be easily analyzed for extrapolation or comparison purposes. In case of extrapolation, however, the effect of sampling errors is often magnified. Therefore, hydrologists should be warned against such practice if no consideration is paid to this effect in interpreting the extrapolated information.

Figure 8-I-1 shows the linearization of a distribution and the formation of a probability paper. The top diagram represents a frequency distribution and its probability distribution of the data is shown in Fig. 8-I-3*a*. The center diagram represents the cumulative probability curve plotted on a paper with rectangular scales. The lower diagram shows the straight-line plotting of the cumulative probability curve on a probability paper having a special scale for the probability designed for the given distribution. It should be noted that in practice the probability is plotted as abscissa rather than as ordinate as shown in the diagram.

Linearization of probability plotting has been proposed by many hydrologists since Hazen [113] first suggested graphical linearization of the normal distribution in 1914. By linearization, Powell [30] constructed the probability paper for Type I extremal distribution proposed by Gumbel [23] for flood frequency analysis. Such paper may be called *extremal probability paper* or *Gumbel-Powell probability paper*. A further modification of the paper was made by Court [2]. Type III extremal distribution was first used by Weibull [103–104] for stress analysis and later by Gumbel [57] for drought frequency studies. Linearization for this distribution can be made on *logextremal probability paper* or *Weibull probability paper*. On this paper, the ordinate has a logarithmic scale for the variate and the abscissa has a special scale for Type I extremal probability since Type III is essentially a logarithmically transformed Type I distribution. For linearization of the lognormal distribution, a *lognormal probability paper* may be used. On this paper, the ordinate has the logarithmic scale for the variate and the abscissa has a normal probability scale. On a probability paper it is frequently necessary to provide a scale for recurrence interval along with the probability scale or sometimes to replace the latter.

Probability paper is generally designed for plotting hydrologic data of annual and complete series, which possess a complete frequency distribution. For partial-duration series and annual exceedance series, which constitute only a "tail" distribution, the data are plotted on semilog paper where the ordinate has the rectangular scale for the variate and the abscissa has the logarithmic scale for the recurrence interval.

In case a probability paper is not available, the probability scale may be constructed by use of Eq. (8-I-47). For example, Fig. 8-I-8 shows an external probability paper in which the scale for recurrence interval is constructed with reference to a rectangular scale for frequency factor K. The recurrence intervals corresponding to various K values on the rectangular scale can be computed by Eq. (8-I-53) and used to construct the required scale. By means of Eq. (8-I-43), the probability scales can also be constructed. Similarly, probability paper of any distribution can be constructed if the K-T relationship is known.

When a probability paper is available, it is possible to determine graphically the frequency factor of the distribution for which the paper is constructed. Chow [107] has proposed such a graphical method for the determination of frequency factors of a lognormal distribution.

F. Plotting of Data

When a probability paper is chosen for use, the plotting of data on the paper requires the knowledge of *plotting positions*. Numerous methods have been proposed for the determination of plotting positions. Most of them are empirical. Table 8-I-2 lists some important plotting-position formulas but the list by no means shows all the methods, since there are many other methods [26, 121–125] which cannot be expressed by simple formulas.

Equation (8-I-56*a*) is believed to be the earliest formula for computing plotting positions. Use of this formula is known as the *California method*, since it was first employed to plot flow data of the California streams. Chow [52] has demonstrated theoretically that this method is suitable for plotting annual exceedance series or partial-duration series. However, this simple formula plots data at the edges of group intervals and produces a probability of 100 per cent which cannot be plotted on a probability paper. Thus, it was gradually replaced by the *Hazen formula*, Eq. (8-I-56*b*), which plots data at the centers of group intervals. As the extremal

distribution was later introduced to frequency analysis, the *Weibull formula*, Eq. (8-I-56*c*), was soon found to be very satisfactory. Chow [52] has shown that this formula is theoretically suitable for plotting the annual maximum series. A comparative study of the Beard, Hazen, and Weibull methods by Benson [126] has also revealed that, on the basis of theoretical sampling from extreme values and normal distributions, the Weibull formula provides the estimates that are consistent with experience. The *Chegodayev formula*, Eq. (8-I-56*e*), is an empirical formula commonly used in the U.S.S.R., but Eq. (8-I-56*c*) has been recommended as the All-Union Standard 3999-48 in 1948 [116]. Equation (8-I-56*e*) is a mathematical approximation of Eq. (8-I-56*d*). In order to simplify the visual inspection of a plotted set of ordered observations on extremal probability paper, Gringorten [120] further recommended Eq. (8-I-56*h*) for computing plotting positions.

Table 8-I-2. Plotting-position Formulas

Name	Date	Formula* for T or $1/P(X \geq x)$	Equation
California [114]	1923	$\frac{N}{m}$	(8-I-56*a*)
Hazen [14]	1930	$\frac{2N}{2m - 1}$	(8-I-56*b*)
Weibull [103–104]	1939	$\frac{N + 1}{m}$	(8-I-56*c*)
Beard [35]	1943	$\frac{1}{1 - 0.5^{1/N}}$	(8-I-56*d*)†
Chegodayev [115–117]	1955	$\frac{N + 0.4}{m - 0.3}$	(8-I-56*e*)
Blom [118]	1958	$\frac{N + \frac{1}{4}}{m - \frac{3}{8}}$	(8-I-56*f*)
Tukey [119]	1962	$\frac{3N + 1}{3m - 1}$	(8-I-56*g*)
Gringorten [120]	1963	$\frac{N + 0.12}{m - 0.44}$	(8-I-56*h*)

* N = total number of items; m = order number of the items arranged in descending magnitude, thus $m = 1$ for the largest item.

† This formula applies only to $m = 1$; other plotting positions are interpolated linearly between this and the value of 0.5 for the median event.

It may be noted that all methods of determining plotting positions give practically the same results in the middle of a distribution but produce different positions near the "tails" of the distribution. Thus, the choice of a plotting-position formula becomes important. According to Benson [26], it is believed that only by use of a method that gives the mathematically expected value of the probability does the expected recurrence equal that experienced over a long period of time, and that commonly used methods may overestimate the benefit-cost ratios of proposed projects if the methods do not furnish the mathematically expected value. Therefore, a refined choice of a method depends on the acceptance of certain statistical principles and on the aim of the analysis.

G. Curve Fitting

After the hydrologic data are plotted on a probability paper, a curve may be fitted to the plotted points. The curve is a straight line if linearization of the distribution is attempted. The straight line can be essentially represented by Eq. (8-I-47). Curve fitting may be done either mathematically or graphically. In general, a

mathematical curve fitting can be achieved by three methods: the method of moments, the method of least squares, and the method of likelihood. Of course, the mathematical fitting does not necessarily require data plotting on a probability paper. By graphical fitting, a straight line is simply drawn to fit the plotted data by eye-fit, and this method is the simplest but involves human error.

Table 8-I-3. Frequency Analysis of Annual Maximum Values of 10-min Duration Rainfall Depth at Chicago, Illinois

m	x	x^2	T_M	$y = K$	y^2	xy	Δx
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	1.11	1.2321	36.000	2.332	5.4382	2.5885	0.177
2	0.96	0.9216	18.000	1.783	3.1791	1.7117	0.124
3	0.94	0.8836	12.000	1.455	2.1170	1.3677	0.091
4	0.92	0.8464	9.000	1.218	1.4835	1.1206	0.079
5	0.88	0.7744	7.200	1.033	1.0671	0.9090	0.070
6	0.80	0.6400	6.000	0.878	0.7709	0.7024	0.064
7	0.80	0.6400	5.143	0 745	0.5550	0.5960	0.059
8	0.76	0.5776	4.500	0.627	0.3931	0.4765	0.056
9	0.74	0.5476	4.000	0.522	0.2725	0.3863	0.052
10	0.71	0.5041	3.600	0.425	0.1806	0.3018	0.050
11	0.70	0.4900	3.272	0.337	0.1136	0.2359	0.047
12	0.68	0.4624	3.000	0.255	0.0650	0.1734	0.045
13	0.68	0.4624	2.769	0.177	0.0313	0.1204	0.044
14	0.66	0.4356	2.571	0.102	0.0104	0.0673	0.042
15	0.66	0.4356	2.400	0.032	0.0010	0.0211	0.041
16	0.66	0.4356	2.250	−0.035	0.0012	−0.0231	0.039
17	0.65	0.4225	2.118	−0.100	0.0100	−0.0650	0.038
18	0.64	0.4096	2.000	−0.164	0.0269	−0.1050	0.037
19	0.64	0.4096	1.895	−0.225	0.0506	−0.1440	0.037
20	0.63	0.3969	1.800	−0.286	0.0818	−0.1802	0.036
21	0.62	0.3844	1.715	−0.346	0.1197	−0.2145	0.035
22	0.61	0.3721	1.636	−0.405	0.1640	−0.2471	0.035
23	0.60	0.3600	1.565	−0.464	0.2153	−0.2784	0.034
24	0.58	0.3364	1.500	−0.523	0.2735	−0.3033	0.034
25	0.57	0.3249	1.440	−0.582	0.3387	−0.3317	0.033
26	0.57	0.3249	1.385	−0.643	0.4134	−0.3665	0.033
27	0.53	0.2809	1.333	−0.704	0.4956	−0.3731	0.033
28	0.52	0.2704	1.285	−0.768	0.5898	−0.3994	0.033
29	0.49	0.1401	1.242	−0.834	0.6956	−0.4087	0.033
30	0.49	0.2401	1.200	−0.904	0.8172	−0.4430	0.033
31	0.47	0.2209	1.162	−0.980	0.9604	−0.4606	0.033
32	0.41	0.1681	1.125	−1.064	1.1321	−0.4362	0.034
33	0.36	0.1296	1.092	−1.159	1.3433	−0.4172	0.035
34	0.34	0.1156	1.058	−1.277	1.6307	−0.4342	0.037
35	0.33	0.1089	1.029	−1.445	2.0880	−0.4769	0.042
$\Sigma = 22.71$		15.8049		−0.987	27.1261	4.6705	

$\bar{x} = 0.6489$ $\bar{x}^2 = 0.4516$ $\bar{y} = -0.0282$ $\bar{y}^2 = 0.7750$ $\overline{xy} = 0.1334$
$B = 0.1960$ $A = 0.6544$ $\sigma = 0.1775$
Line of best-fit: $x = 0.1960K + 0.6544$

1. Method of Moments. By this method, the statistical parameters or moments are computed from the data and then substituted in the probability function of the given distribution. This method gives a theoretically exact fitting but the accuracy can be substantially affected by any errors involved in the data at the tails of the distribution where the moment arms are long and the errors are thus magnified. The method originally proposed by Gumbel [23] to fit Type I extremal distribution is a method of moments. Lieblein [121] modified this method by order statistics and

developed a procedure which maintains the original time order of the extreme-value series, divides the values into subgroups, and then weighs each observation according to its ordered rank in the subgroup which in turn is a function of the sample size. Hershfield [56] made a comparison of the two procedures and concluded that the Gumbel procedure gives a better estimate beyond the range of data for the areally independent data tests, but overestimates the longer recurrence-intervals in the dependent data tests.

2. Method of Least Squares. By this method, a regression line is computed to fit the plotted data (Sec. 8-II). The curve so obtained may not represent the exact theoretical distribution but it gives a better overall fit than the method of moments. For extremal distributions, Gumbel [101] introduced a modified least-squares method by minimizing both vertical and horizontal deviations and taking the geometric mean of the parameters obtained from the two minimizations. Based on the general equation for hydrologic frequency analysis, Eq. (8-I-47), proposed by Chow [112], a least-squares procedure for fitting a normal, lognormal, or extremal distribution was developed by Brakensiek [127].

Table 8-I-3 shows the computation for fitting annual maximum values plotted on an extremal probability paper in Fig. 8-I-8. In this table, m is the rank number, x is the variate or the rainfall depth, and y is the frequency factor K. The recurrence interval T is computed by Eq. (8-I-56c), and the frequency factor by Eq. (8-I-53). The coefficients A and B of the least-squares equation are computed by Eqs. (8-II-8) and (8-II-9).

3. Method of Maximum Likelihood. By this method, the value of a parameter is determined to make the probability of obtaining the observed outcome as high as possible. Mathematically, $\partial \log p(x)/\partial u = 0$, where $p(x)$ is probability density and u is a statistical parameter. This method provides the best estimate of the parameters but it is usually very complicated for practical application. Kimball [128–129] has suggested this method for fitting extremal distributions, and a practical procedure was later developed by Panchang and Aggarwal [130].

H. Reliability of Analysis

1. Sampling Reliability. The fact that observed data may exhibit a straight-line trend on a suitable probability paper but do not exactly follow the theoretical curve to be fitted leads to the belief that singular sampled events cannot be represented with perfect confidence by the theory of probability. It is therefore important to know the reliability of results obtained by the frequency analysis; i.e., to know how well the individual event agrees with the theoretical prediction derived from the sampled data.

The curve or distribution function fitted to the hydrologic data can be considered only to represent either the mean or sometimes the mode of the data at a given cumulative probability or recurrence interval. The distribution of the data for the given cumulative probability or recurrence interval can be described by the so-called *confidence limits* established on both sides of the fitted curve. Such confidence limits define the probability density areas on both sides of the mean, or of the mode of an assumed distribution of the data for the given cumulative probability or recurrence interval. *Control curves*, which join the equal confidence limits, can then be drawn to show the *confidence bands*. The reliability of any plotted point lying within the confidence band is thus indicated by the probability on which the confidence limits are based.

Gumbel [28] has proposed a method for establishing the confidence limits in the plotting of annual maximum values. This method is based on the principle that the theoretical value of rank m situated on the straight line and corresponding to a given recurrence interval is the approximation to the most probable mth value. Considering a probability equal to 68.269 per cent (i.e., the probability for a deviation of $\pm\sigma$ from the predicted value in a normal distribution), the control curves can be constructed, respectively, above and below the fitted theoretical straight line on the extremal probability paper with a vertical distance of Δx from the line. In other words, the mth

observation is contained in the confidence band defined by $x - \Delta x < \hat{x} < x + \Delta x$. It is expected that 68.269 per cent of all possible values would fall within the band Gringorten [131] has developed a set of graphs for use in contructing such confidence limits. For practical purposes, the method can be simplified by the approximate

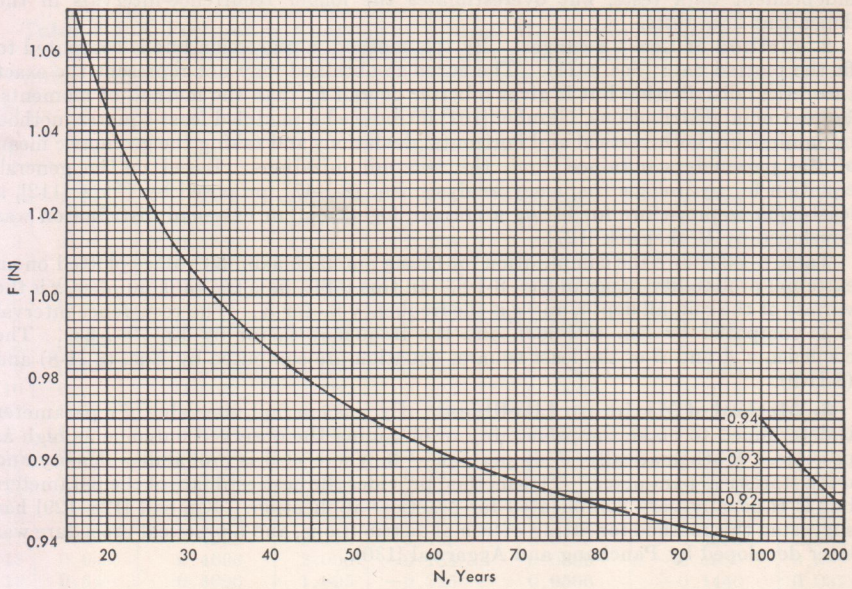


FIG. 8-I-9. Relation between N and $F(N)$.

procedure given below [52]. In Fig. 8-I-8, the control curves are constructed by this method and the computation is given in Table 8-I-3.

a. For the largest value with $m = 1$, the half vertical width of the confidence band is

$$\Delta x_1 = sF(N) \tag{8-I-57}$$

where s is the standard deviation of the observed data and $F(N)$ is a function of the N years of record as expressed graphically in Fig. 8-I-9.

b. For the second largest value with $m = 2$,

$$\Delta x_2 = \frac{0.661(N + 1)}{N - 1} \Delta x_1 \tag{8-I-58}$$

c. For intermediate values of rank m,

$$\Delta x_m = \frac{0.877}{\sqrt{N}} \Delta x_1 F(T_M) \tag{8-I-59}$$

where $F(T_M)$ is a function of the recurrence interval T_M as expressed graphically in Fig. 8-I-10. When T_M is greater than 10 years, $F(T_M) = \sqrt{T_M}$.

d. For very small values, control curves are generally not necessary. For extrapolation beyond the largest value, Gumbel suggested that the control curves be drawn as two lines parallel to the extrapolated straight line. This suggestion, however, will result in sudden breaks on the control curves and in narrowing down the growing width of the confidence band. Therefore, it is generally not followed. Kimball [12]

has suggested a procedure for constructing the control curves by a method of order statistics, which avoids such difficulties in extrapolation.

In the above discussion, the confidence limits, which represent the probable errors of estimate, are computed on the basis of an assumed distribution of errors, since the actual distribution is unknown. In the case of regional analysis (Subsec. IV-J), an approximate actual distribution of the sampling errors may be established from the data obtained from a group of stations in the region of statistical homogeneity. The

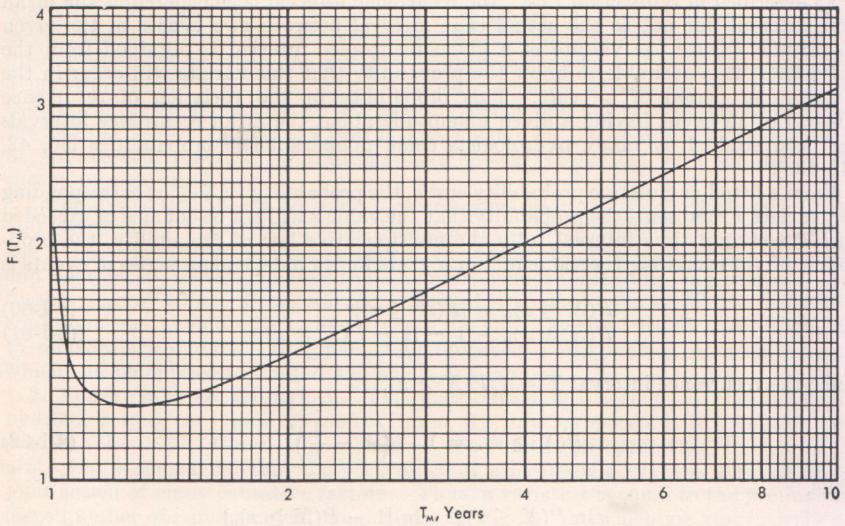


FIG. 8-I-10. Relation between T_M and $F(T_M)$.

sampling errors from the station mean for a given recurrence interval can be computed and analyzed to develop a distribution function. From this function, the per cent of errors to be equaled or exceeded among a given per cent, say 68.269 per cent, of the sampled stations can be determined. This per cent of errors multiplied by the variate x of the given recurrence interval will give the range Δx which can be used to construct the confidence limits. For other recurrence intervals, the confidence limits can be computed similarly.

Occasionally, the plotted point or points near the two ends of the distribution depart markedly from the general path indicated by the curve of best-fit. The departure may be so great that the plotted points fall far outside the confidence band, becoming the so-called *off-control data*. It has been explained that such off-control data follow some other type of distribution which applies to events that may occur at very long recurrence intervals. In other words, such events are statistically incompatible with the events with which they are associated in a given sample. Therefore, it is logical to recognize that these data may have an actual recurrence interval many times greater than the length of record. Whether or not these data follow the usual pattern or some other pattern of distribution than that represented by the rest of the data, it would seem obvious that they cannot be assumed to have a recurrence interval equal to the value computed by the plotting-position formula for which the available length of record is used. Consequently, inclusion of the off-control data in the computation of a theoretical probability curve would produce a different result from the one the sample should indicate. Since the off-control data are, in a sense, nonhomogeneous with the rest of the sample, they should be excluded from the fitting of the data. If the theoretical curve first computed using all data produces off-control data, the curve should be recomputed by excluding the latter unless the off-control data are *assumed* to be homogeneous with the total sample.

2. Prediction Reliability. In the above discussion on sampling reliability, the probable error in the estimate of the variate by frequency analysis for a given recurrence interval is considered. Such errors which would affect the reliability of the result are largely due to sampling defects. There is another problem which relates to the probability of an event of a given average recurrence interval occurring during a given period of time. This probability would affect the reliability of prediction on the basis of the recurrence interval obtained from data fitting.

As described in Subsection IV-C, the recurrence interval is considered as the mean time interval but not as the actual time interval between two events of the given magnitude. Once the variate of a given recurrence interval is obtained from the fitted data, it is desirable to know the probability that this variate will occur in the forthcoming period of n years. This probability or the variation of recurrence interval of an event having a given magnitude about the mean recurrence intervals has been studied by many hydrologists using nonparametric probabilities [36, 42, 132–133].

From a fitted cumulative probability curve, the probability $P(X \leq x)$ corresponding to a variate x represents the probability that the value x will not be equaled or exceeded during a certain time interval. By the multiplicative law of probability, the probability of not exceeding the value of x in n years for an independent series of events is

$$P(X \leq x)_n = [P(X \leq x)]^n \tag{8-I-60}$$

or

$$P(X \geq x)_n = 1 - [1 - P(X \geq x)]^n \tag{8-I-61}$$

Since the recurrence interval $T = 1/P(X \geq x)$,

$$P(X \geq x)_n = 1 - \left(1 - \frac{1}{T}\right)^n \tag{8-I-62}$$

Thus,

$$n = \frac{\ln P(X \leq x)_n}{\ln P(X \leq x)} = \frac{\ln [1 - P(X \geq x)_n]}{\ln [1 - P(X \geq x)]} = \frac{\ln [1 - P(X \geq x)_n]}{\ln [(T - 1)/T]} \tag{8-I-63}$$

The value of $P(X \leq x)$, $P(X \geq x)$, or T of a given variate x can be obtained from the fitted data. The probability that this variate will occur in a period of n years can be computed by Eq. (8-I-61) or (8-I-62). If this probability $P(X \geq x)_n$ is given according to a design policy, the value of n, known as a *design period*, can be computed by Eq. (8-I-63). As an illustration, an additional scale is shown on top of the diagram in Fig. 8-I-8 for $P(X \geq x)_n$ in $n = 10$ years. The scale is computed by means of Eq. (8-I-63). Thus, the probability that a 10-min rainfall depth of 0.91 in. with a recurrence interval of 10 years will have a chance of 65 per cent to occur in the next 10 years. If a chance of 50 per cent occurrence in the next 10 years is considered in the design, the design rainfall should be 0.97 in. and have a recurrence interval of 15 years. Similarly, another scale for 20 years can be drawn and it can be shown that this rainfall of 0.91 in. will have a chance of 88 per cent to occur in the next 20 years.

I. Theoretical Justifications

From a practical point of view, the frequency analysis is only a procedure to fit the hydrologic data to a mathematical model of distribution. It is only experience and verification of data that decide the use of a certain distribution. However, there are several theoretical interpretations or reasonings for the preference of one distribution to another. Such interpretations would describe a physical process of the hydrologic phenomena and thus help to understand the procedure of frequency analysis and the significance of the results, but they are usually based on a number of assumptions which may not be readily satisfied in the real world. Most theoretical distributions recommended for hydrologic frequency analysis are asymptotic. The asymp-

totic condition is valid only when the number of variates becomes indefinitely large. Most of these distributions also assume independent variables. In actual hydrologic phenomena, however, the number of variates is always limited and mostly of small size, and the variables are likely to be interdependent to a certain extent.

1. Type I Extremal Distribution. This distribution was first proposed by Gumbel [23] for the analysis of flood frequencies. Gumbel considered the daily flow as a statistical variable unlimited to the positive end of the distribution, and defined a flood as being the largest value of the 365 daily flows. The flood flows are therefore the largest values of flows. According to the theory of extreme values, the annual largest values of a number of years of record will approach a definite pattern of frequency distribution when the number of observations in each year becomes large. Thus, the annual maximum floods constitute a series which can be fitted in the theoretical extremal distribution of Type I. Although it has been questioned whether the number of observations in a year is large enough for the asymptotic distribution to be approached, practical applications have shown satisfaction with the use of this theory to many problems.

In applying this theory to some meteorological data, Barricelli [134] and Brooks and Carruthers [1] have found some defects and therefore modified the theory in their use. They found that for temperatures, the Gumbel approximation overestimates, and for rainfall, underestimates, the maximum values reached in long periods. As a further improvement, Jenkinson [135] derived a general solution of the functional equation which should satisfy the extreme values of all types of distributions applicable to meteorological data. Borgman [136] proposed a distribution of near extremes which can be applied to limited and small-size samples.

2. Lognormal Distribution. This distribution has been used empirically for hydrologic frequency analysis since Hazen [12] first proposed it in 1914 for flood studies. In 1955, Chow [15] offered a theoretical interpretation to justify its use. Chow considered that the occurrence of a hydrologic event is a result of the joint action of many causative factors. Thus, a variate x is equal to the product of a large number of r independent magnitudes $x_1, x_2, \ldots x_r$ which are respectively due to the r causative factors. The logarithm of x is therefore equal to the sum of logarithms of a very large number of independent variates. By the central limit theorem, it can be shown that the logarithm of x is normally distributed when r becomes infinitely large.

The lognormal distribution contains three interdependent parameters. When hydrologic data are plotted on a lognormal probability paper, a straight-line trend is possible only for one value of C_s when C_v is given. Figure 8-I-11 shows that the plot is theoretically a straight line only for $C_s = 1.139$ when $C_v = 0.364$. For other values of C_s, the plots are curved for $C_v = 0.364$. Since the value of C_s computed from ordinary hydrologic data is not so reliable, it has been suggested that if the plot shows curvature, the value of C_s should be modified so that a straight line is obtained. Otherwise, a special probability paper may be constructed if a straight-line plot for the original value of C_s is desired [15].

3. Exponential Distribution. The exponential distribution has been applied empirically to partial-duration series. However, Chow [52] reasoned that the probability $p(x)$ of occurrence of a variate is the product of the probabilities of r number of causative factors. Thus, $p(x) = p^r$ where p is the geometric mean probability of all causative factors. When r is infinitely large and x is of high magnitude, it can be shown mathematically that the distribution of x is exponential.

4. Logextremal Distribution. The Type III extremal distribution was first proposed by Gumbel [57] for drought frequency analysis. Gumbel defined the drought as the smallest annual values of the mean daily discharges of a river. Since there is always a limit to the drought with a minimum of zero, Type III extremal distribution is assumed to be suitable. In this distribution, Eq. (8-I-36), three parameters are involved. The parameter ϵ is the lower limit called the *minimum drought*. This is the drought for which the probability of a value equal to or greater than it is unity and the recurrence interval is infinite. The parameter θ is called the *drought characteristic*, which has a recurrence interval of 1.58198 years. The parameter k has no particular

name with reference to the drought, but its reciprocal is a scale parameter which defines skewness.

In applying Gumbel's method to drought solution, the droughts can be plotted on logextremal probability paper. If $\epsilon = 0$, the plot should have a straight-line trend. If $\epsilon > 0$, the plot will appear curved at the side of the long recurrence interval.

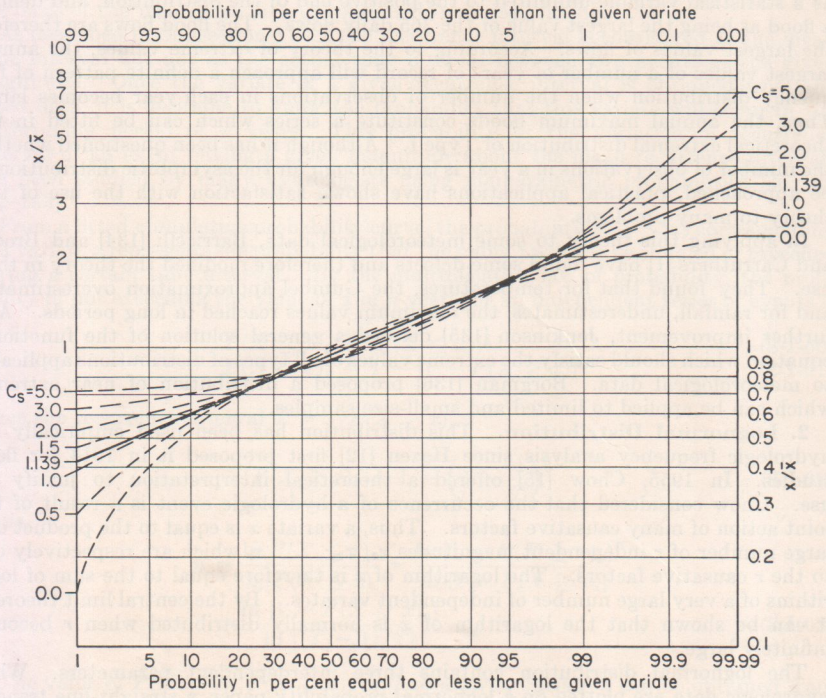


FIG. 8-I-11. Lognormal probability plotting.

The parameters of the distribution may be computed by statistical fitting of the distribution function. For extrapolation, the computed distribution function can be used since graphical extension of the curve may not be easy.

J. Regional Analysis

The observations at a geographical point, such as at the site of a rain gage or a stream gaging station, are *point data*. Extension of the results of the frequency analysis of the point data to an area requires *regional analysis*. Various methods of regional analysis have been developed, including the station-year method for rainfall analysis (Sec. 9) and the regional methods for flood analysis [38] (Subsec. 25-I-III-B-I). For all these methods, a statistically homogeneous region is defined. Within such regions, the results of point-data analysis can be averaged to best represent the frequency characteristics of the whole region. Usually, an average probability curve so obtained is applicable throughout the region.

In order to define a homogeneous region, a test has been developed by Langbein [7, 137] for regional flood-frequency analysis practiced by the U.S. Geological Survey (Subsec. 25-I-III-B-1). This *homogeneity test* requires a study of the 10-year flood as estimated from the probability curve at each station within a region. These 10-year floods expressed as ratios to mean annual floods (which have a recurrence

interval of 2.33 years according to the extremal distribution) are averaged to obtain the mean 10-year ratio for the area. The recurrence interval corresponding to the mean annual flood times the averaged 10-year ratio is determined from the probability curve of each station and plotted against the number of years of record on a test graph (Fig. 8-I-12). If the points for all of the stations lie between the two control curves (indicating 95 per cent reliability) on the graph, they are considered homogeneous.

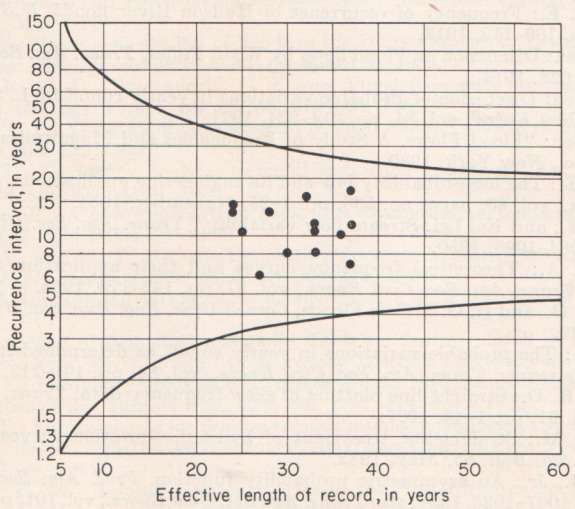


FIG. 8-I-12. Homogeneity test graph. (*U.S. Geological Survey.*)

Points outside the curves indicate that the region should be subdivided for homogeneity. Repeat the procedure until all the subdivided areas pass the homogeneity test.

The principle of the above test is to determine in a statistical sense whether the records in a group differ from one another by amounts that cannot reasonably be expected by chance. If the differences are found to be no more than those due to the operations of chance, they can be considered to represent merely different aspects of the same event and thus can be grouped. The test graph is constructed on the basis of the extremal distribution. The control curves represent a range of variation equal to two standard deviations of the reduced variate [i.e., $(a + x)/c$ in Eq. (8-I-33)] on the 10-year flood. This means that 95 per cent of the estimate will lie within 2σ of the most probable value of a recurrence interval of 10 years. The 10-year flood is used in this test because this is the longest recurrence interval for which most flood records will give dependable estimates.

V. REFERENCES

1. Brooks, C. E. P., and N. Carruthers: "Handbook of Statistical Methods in Meteorology," Great Britain Meteorological Office, H.M.S.O., London, 1953.
2. Court, Arnold: Some new statistical techniques in geophysics, in "Advances in Geophysics," ed. by H. E. Landsberg, vol. 1, Academic Press Inc., New York, 1952, pp. 45–85.
3. Johnstone, Don, and W. P. Cross: "Elements of Applied Hydrology," The Ronald Press Company, New York, 1949, pp. 236-264.
4. Foster, E. E.: "Rainfall and Runoff," The Macmillan Company, New York, 1948.
5. Beard, L. R.: Statistical methods in hydrology, U.S. Army Engineer District, Corps of Engineers, Sacramento, Calif., January, 1962.
6. Benson, M. A.: Evolution of methods for evaluating the occurrence of floods, *U.S. Geol. Surv. Water-Supply Paper* 1580-A, 1962.

7. Dalrymple, Tate (ed.): Flood-frequency analysis, Manual of Hydrology, pt. 3, Flood-flow techniques, *U.S. Geol. Surv. Water-Supply Paper* 1543-A, 1960.
8. Foster, H. A.: "Methods of Analyzing Hydrological Records," privately published, 1955.
9. Foster, H. A.: Duration curves, *Trans. Am. Soc. Civil Engrs.*, vol. 99, pp. 1213–1235, 1934.
10. Fuller, W. E.: Flood flows, *Trans. Am. Soc. Civil Engrs.*, vol. 77, pp. 564–617, 1914.
11. Horton, R. E.: Frequency of recurrence of Hudson River floods, *U.S. Weather Bur. Bull. Z*, pp. 109–112, 1913.
12. Hazen, Allen: Discussion on Flood flows by W. E. Fuller, *Trans. Am. Soc. Civil Engrs.*, vol. 77, p. 628, 1914.
13. Hazen, Allen: Discussion of Probable variations in yearly runoff by L. S. Hall, *Trans. Am. Soc. Civil Engrs.*, vol. 84, pp. 214–224, 1921.
14. Hazen, Allen: "Flood Flows, A Study of Frequencies and Magnitudes," John Wiley & Sons, Inc., New York, 1930.
15. Chow, V. T.: The log-probability law and its engineering applications, *Proc. Am. Soc. Civil Engrs.*, vol. 80, paper no. 536, pp. 1–25, November, 1954.
16. Lane, E. W., and Kai Lei: Stream flow variability, *Trans. Am. Soc. Civil Engrs.*, vol. 115, pp. 1084–1098, 1950.
17. Foster, H. A.: Theoretical frequency curves and their application to engineering problems, *Trans. Am. Soc. Civil Engrs.*, vol. 87, pp. 142–173, 1924.
18. Switzer, F. G., and H. G. Miller: Floods, *Cornell Univ. Eng. Exp. Sta. Bull.* 13, December 15, 1929.
19. Hall, L. S.: The probable variations in yearly runoff as determined from a study of California streams, *Trans. Am. Soc. Civil Engrs.*, vol. 84, pp. 191–213, 1921.
20. Goodrich, R. D.: Straight-line plotting of skew-frequency data, *Trans. Am. Soc. Civil Engrs.*, vol. 91, pp. 1–43, 1927.
21. Harris, R. M.: Straight-line treatment of hydraulic duration curves, *Univ. Wash. Eng. Exp. Sta. Bull.* 65, May, 1932.
22. Slade, J. J., Jr.: An asymmetric probability function, *Proc. Am. Soc. Civil Engrs.*, vol. 60, pp. 1007–1023, 1934; also *Trans. Am. Soc. Civil Engrs.*, vol. 101, pp. 35–61, 1936.
23. Gumbel, E. J.: The return period of flood flows, *Ann. Math. Statist.*, vol. XII, no. 2, pp. 163–190, June, 1941.
24. Gumbel, E. J.: Probability interpretation of the observed return periods of floods, *Trans. Am. Geophys. Union*, vol. 21, pp. 836–850, 1941.
25. Gumbel, E. J.: Statistical control-curves for flood-discharges, *Trans. Am. Geophys. Union*, vol. 23, pp. 489–500, 1942.
26. Gumbel, E. J.: On the plotting of flood discharges, *Trans. Am. Geophys. Union*, vol. 24, pp. 699–719, 1943.
27. Gumbel, E. J.: Floods estimated by probability methods, *Eng. News-Rec.*, vol. 134, no. 24, pp. 97–101, June 14, 1945.
28. Gumbel, E. J.: The statistical forecast of floods, *Bull.* 15, The Ohio Water Resources Board, Columbus, Ohio, 1949.
29. Gumbel, E. J.: Statistical theory of floods and droughts, *J. Inst. Water Engrs.*, vol. 12, no. 3, pp. 157–184, May, 1958.
30. Powell, R. W.: A simple method of estimating flood frequencies, *Civil Eng.*, vol. 13, pp. 105–106, February, 1943.
31. Cross, W. P.: Floods in Ohio, magnitude and frequency, *Bull.* 7, The Ohio Water Resources Board, Columbus, Ohio, 1946.
32. Potter, W. D.: Simplifications of the Gumbel method for computing probability curves, *U.S. Dept. Agr. Soil Conserv. Serv.* SCS-TP-78, May, 1949.
33. Benson, M. A.: Characteristics of frequency curves based on a theoretical 1,000-year record, in Ref. 7, pp. 57–74; also *U.S. Geol. Surv. Open-file Report*, 1952.
34. Jarvis, C. S., and others: Floods in the United States, *U.S. Geol. Surv. Water-Supply Paper* 771, 1936.
35. Beard, L. R.: Statistical analysis in hydrology, *Trans. Am. Soc. Civil Engrs.*, vol. 108, pp. 1110–1160, 1943.
36. Thomas, H. A., Jr.: Frequency of minor floods, *J. Boston Soc. Civil Engrs.*, vol. 34, pp. 425–442, October, 1948.
37. Langbein, W. B.: Annual floods and the partial-duration flood series, *Trans. Am. Geophys. Union*, vol. 30, pp. 879–881, 1949.
38. Dalrymple, Tate: Regional flood frequency, in "Surface Drainage," *Highway Res. Board Res. Rept.* 11-B, pp. 4–20, December, 1950.
39. Bodhaine, G. S., and W. H. Robinson: Floods in Western Washington—frequency and magnitude in relation to drainage basin characteristics, *U.S. Geol. Surv. Cir.* 191, 1952.

40. Review of flood frequency methods, Final Report of the Subcommittee of the Joint Division Committee on Floods, *Trans. Am. Soc. Civil Engrs.*, vol. 118, pp. 1220–1230, 1953.
41. Mitchell, W. D.: Floods in Illinois—magnitude and frequency, Illinois Division of Waterways in cooperation with U.S. Geological Survey, 1954.
42. Gumbel, E. J.: The calculated risk in flood control, *Appl. Sci. Res.* (The Hague), sec. A, vol. 5, pp. 273–280, 1955.
43. "Studies of Floods and Flood Damages 1952–1955," American Insurance Association, New York, May, 1956.
44. Chow, V. T.: Hydrologic studies of floods in the United States, in "Symposia Darcy," *Intern. Assoc. Sci. Hydrology Pub.* 42, pp. 134–170, 1956.
45. Rowe, R. R., G. L. Long, and T. C. Royce: Flood frequency by regional analysis, *Trans. Am. Geophys. Union*, vol. 38, pp. 879–884, 1957.
46. Moran, P. A. P.: The statistical treatment of flood flows, *Trans. Am. Geophys. Union*, vol. 38, pp. 519–523, 1957.
47. Chow, V. T.: Frequency analysis in small watershed hydrology, *Agr. Eng.*, vol. 39, pp. 222–225, and 231, April, 1958.
48. Rangarajan, R.: A new approach to peak flow estimation, *J. Geophys. Res.*, vol. 65, no. 2, pp. 643–650, February, 1960.
49. Hall, W. A., and D. T. Howell: Estimating flood probabilities within specific time intervals, *J. Hydrology*, vol. 1, no. 3, pp. 265–271, 1963.
50. Yarnell, D. L.: Rainfall intensity-frequency data, *U.S. Dept. Agr. Misc. Publ.* 204, 1936.
51. Engineering Staff of Miami Conservancy District: Storm rainfall of eastern United States, *Tech. Rept.*, pt. V, Miami Conservancy District, Dayton, Ohio, 1917.
52. Chow, V. T.: Frequency analysis of hydrologic data with special application to rainfall intensities, *Univ. Illinois Eng. Exp. Sta. Bull.* 414, July, 1953.
53. Chow, V. T.: Design charts for finding rainfall intensity frequency, *Water and Sewage Works*, vol. 99, no. 2, pp. 86–88, February, 1952; also *Concrete Pipe News*, vol. 4, no. 6, pp. 8–10, June, 1952.
54. Hershfield, D. M.: Rainfall frequency atlas of the United States for durations from 30 minutes to 24 hours and return periods from 1 to 100 years, *U.S. Weather Bur. Tech. Rept.* 40, May, 1961.
55. Huff, F. A., and J. C. Neill: Comparison of several methods for rainfall frequency analysis, *J. Geophys. Res.*, vol. 64, no. 5, pp. 541–547, May, 1959.
56. Hershfield, D. H.: An empirical comparison of the predictive value of three extreme-value procedures, *J. Geophys. Res.*, vol. 67, no. 5, pp. 1535–1542, April, 1962.
57. Gumbel, E. J.: Statistical theory of droughts, *Proc. Am. Soc. Civil Engrs.*, vol. 80, sep. no. 439, pp. 1–19, May, 1954.
58. Gumbel, E. J.: Statistical forecast of droughts, *Bull. Intern. Assoc. Sci. Hydrology*, 8th year, no. 1, pp. 5–23, April, 1963.
59. Velz, C. J., and J. J. Gannon: Drought flow characteristics of Michigan streams, Department of Environmental Health, School of Public Health, University of Michigan, in cooperation with Michigan Water Resources Commission, 1960.
60. Hardison, C. H., and R. O. R. Martin: Low-flow frequency curves for selected long-term stream-gaging stations in eastern United States, *U.S. Geol. Surv. Water-Supply Paper* 1669-G, 1963.
61. Hudson, H. E., Jr., and W. J. Roberts: 1952–55 Illinois drought with special reference to impounding reservoir designs, *Illinois State Water Surv. Bull.* 43, 1955.
62. Stall, J. B., and J. C. Neill: A partial duration series for low-flow analysis, *J. Geophys. Res.*, vol. 66, no. 12, pp. 4219–4225, 1961.
63. Matalas, N. C.: Probability distribution of low flows, *U.S. Geol. Surv. Profess. Paper* 434-A, 1963.
64. Ledbetter, J. O., and E. F. Gloyna: Predictive techniques for water quality inorganics, *Proc. Am. Soc. Civil Engrs., J. Sanitary Eng. Div.*, vol. 90, no. SA1, pp. 127–151, February, 1964.
65. Seaway, chap. 1 in B. V. Korvin-Kroukovsky (ed.): "Theory of Seakeeping," The Society of Naval Architects and Marine Engineers, 1961.
66. "Ocean Wave Spectra," Proceedings of a Conference arranged by the National Academy of Sciences, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.
67. Longuet-Higgins, M. S.: On the statistical distribution of sea waves, *Sears Foundation: J. Marine Res.*, vol. XI, no. 3, pp. 245–266, December, 1952.
68. Putz, R. R.: Ocean wave record analysis, ordinate distribution and wave heights, *Univ. Calif. Inst. Eng. Res. Tech. Rept.* ser. 3, issue 351, 1953; also summarized as Measurement and analysis of ocean waves, in chap. 5 of J. W. Johnson (ed).: "Ships

and Waves," Council on Wave Research and Society of Naval Architects and Marine Engineers, 1955, pp. 63–72.
69. Cartwright, D. E., and M. S. Longuet-Higgins: The statistical distribution of the maxima of a random function, *Proc. Royal Soc.* ser. A, vol. 237, pp. 212–232, 1956.
70. Gumbel, E. J.: Statistical distribution patterns of ocean waves, *Trans. Soc. Naval Arch. Marine Engrs.*, vol. 8, p. 427, 1956.
71. Jasper, N. H.: Statistical distribution patterns of ocean waves and of wave induced ship stresses and motions, with engineering applications, a doctorate dissertation, The Catholic University of America Press, Washington, D.C., 1956; also summarized as Distribution patterns of wave heights, and ship motions and hull stress, chap. 34 in J. W. Johnson (ed.): "Ships and Waves," Council on Wave Research and Society of Naval Architects and Marine Engineers, 1955, pp. 489–503.
72. Bennet, Rutger: Stress and motion measurements on ships at sea, The Swedish Shipbuilding Research Foundation, *Rept.* 13, Goteburg, Sweden, 1963.
73. Elderton, W. P.: "Frequency Curves and Correlation," Cambridge University Press, London, 4th ed., 1953.
74. Burington, R. S., and D. C. May: "Handbook of Probability and Statistics with Tables," Handbook Publishers, Inc., Sandusky, Ohio, 1953; reprinted, McGraw-Hill Book, Company, Inc., New York, 1958.
75. Feller, William: "An Introduction to Probability Theory and Its Applications," vol. 1, John Wiley & Sons, Inc., 2d ed., 1957.
76. Doob, J. L.: "Stochastic Processes," John Wiley & Sons, Inc., New York, 1953.
77. Takács, Lajos: "Stochastic Processes," Methuen and Co., Ltd., London, 1960.
78. Goode, H. H., and R. E. Machol: "System Engineering," McGraw-Hill Book Company, Inc., New York, 1957.
79. Riordan, John: "Stochastic Service Systems," John Wiley & Sons, Inc., New York, 1962.
80. Brown, R. G.: "Smoothing, Forecasting and Prediction of Discrete Time Series," Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.
81. Rosenblatt, Murray (ed.): "Time Series Analysis," John Wiley & Sons, Inc., New York, 1963.
82. Hannan, E. J.: "Time Series Analysis," Methuen and Co., Ltd., London, 1960.
83. Davis, H. T.: "The Analysis of Economic Time Series," The Principia Press, Inc., Bloomington, Ind., 1941.
84. Tinter, Gerhard: "Econometrics," John Wiley & Sons, Inc., New York, 1952.
85. Kendall, M. G.: "The Advanced Theory of Statistics," vol. 2, Charles Griffen and Co., Ltd., London, 1948.
86. Schuster, A.: On the investigation of hidden periodicities with application to a supposed 26-day period meteorological phenomena, *Terrestrial Magnetism*, vol. 3, no. 1, pp. 13–41, March, 1898.
87. Walker, Sir Gilbert: On periodicity, *Roy. Meteorol. Soc. J.*, vol. 51, pp. 337–345; disc. 345–346, October, 1925.
88. Fisher, R. A.: Tests of significance in harmonic analysis, *Royal Soc. London Proc.*, ser. A, vol. 125, no. 796, pp. 54–59, 1929.
89. Chow, V. T.: Do climatic variations follow definite cycles? *Civil Eng.*, vol. 20, no. 7, p. 470, July, 1950.
90. Huntington, Ellsworth: "Civilization and Climate," Yale University Press, 3d ed., 1924.
91. Brakensiek, D. L.: Selecting the water year for small agricultural watersheds, *Trans. Am. Soc. Agr. Engrs.*, vol. 2, no. 1, pp. 5–8 and 10, 1959.
92. Leopold, L. A.: Probability analysis applied to a water supply problem, *U.S. Geol. Surv. Cir.* 410, 1959.
93. Hurst, H. E.: Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civil Engrs.*, vol. 116, pp. 770–799, 1951.
94. Dixon, W. J., and F. J. Massey, Jr.: "Introduction to Statistical Analysis," McGraw-Hill Book Company, Inc., New York, 2d ed., 1957.
95. Mood, A. M., and F. A. Graybill: "Introduction to the Theory of Statistics," McGraw-Hill Book Company, Inc., New York, 2d ed., 1963.
96. Pearson, Karl: "Tables for Statisticians and Biometricians," Part I, The Biometric Laboratory, University College, London; printed by Cambridge University Press, London, 3d ed., 1930.
97. Fréchet, Maurice: Sur la loi de probabilité de l'écart maximum (On the probability law of maximum error), *Ann. Soc. Polonaise Math.* (Cracow), vol. 6, pp. 93–116, 1927.
98. Fisher, R. A., and L. H. C. Tippett: Limiting forms of the frequency distribution of the smallest and largest member of a sample, *Proc. Cambridge Phil. Soc.*, vol. 24, pp. 180–190, 1928.

99. von Mises, R.: La distribution de la plus grande de n valeurs (The distribution of the largest of n values), *Revue Math. l'Union Interbalcanique* (Athens), vol. 1, pp. 1–20, 1936.
100. Gumbel, E. J.: "Statistics of Extreme Values," Columbia University Press, New York, 1958.
101. Statistical theory of extreme values and some practical applications, *U.S. Nat. Bur. Stds. Appl. Math. Ser.* 33, 1954.
102. Probability tables for analysis of extreme-value data, *U.S. Nat. Bur. Stds. Appl. Math. Ser.* 22, 1953.
103. Weibull, W.: A statistical theory of the strength of materials, *Ing. Vetenskaps Akad. Handl.* (Stockholm), vol. 151, p. 15, 1939.
104. Weibull, W.: The phenomenon or rupture in solids, *Ing. Vetenskaps Akad. Handl.* (Stockholm), vol. 153, p. 17, 1939.
105. Galton, Francis: Statistics by intercomparison with remarks on the law of frequency of error, *The London, Edinburgh, Dublin Phil. Mag. J. Sci.*, 4th ser., vol. XLIX, p. 33, January–June, 1875.
106. Chow, V. T.: On the determination of frequency factor in log-probability plotting, *Trans. Am. Geophys. Union*, vol. 36, pp. 481–486, 1955.
107. Chow, V. T.: Determination of hydrologic frequency factor, *Proc. Am. Soc. Civil Engrs., J. Hydraulics Div.*, vol. 85, no. HY7, pp. 93–98, July, 1959.
108. Aitchison, J., and J. A. C. Brown: "The Lognormal Distribution, with Special Reference to the Uses in Economics," Cambridge University Press, London, 1957.
109. Alekseyev, G. A.: Determination of standard parameters of a logarithmically normal distribution curve by three reference ordinates, *Soviet Hydrology: Selected Papers*, American Geophysical Union, no. 6, pp. 637–684, 1962.
110. Chow, V. T.: Discussion on Annual floods and the partial duration flood series, by W. B. Langbein, *Trans. Am. Geophys. Union*, vol. 31, pp. 939–941, 1950.
111. Thom, H. C. S.: Time interval distribution for excessive rainfalls, *Proc. Am. Soc. Civil Engrs., J. Hydraulics Div.*, vol. 85, no. HY7, pp. 83–91, July, 1959.
112. Chow, V. T.: A general formula for hydrologic frequency analysis, *Trans. Am. Geophys. Union*, vol. 32, pp. 231–237, 1951.
113. Hazen, Allen: Storage to be provided in impounding reservoirs for municipal water supply, *Trans. Am. Soc. Civil Engrs.*, vol. 77, pp. 1539–1659, 1914.
114. Flow in California streams, *Calif. State Dept. Pub. Works Bull.* 5, chap. 5, 1923.
115. Alekseyev, G. A.: O formule dlya vychisleniya obespechenosti gidrologicheskikh velichin (Formulas for the calculation of the confidence of hydrological quantities), *Metrologiia i Gidrologiia* (Leningrad), no. 6, pp. 40–43, November–December, 1955.
116. Leĭvikov, M. L.: "Meterologiia, gidrologiia i gidrometriia" (Meteorology, Hydrology and Hydrometry), Sel'khozgiz, Moscow, 2d ed., 1955.
117. Benard, A., and E. C. Bos-Levenbach: The plotting of observations on probability paper (Dutch), *Statistica* (Rijkswijk), vol. 7, pp. 163–173, 1953.
118. Blom, Gunnar: "Statistical Estimates and Transformed Beta-Variables," John Wiley & Sons, Inc., New York, 1958.
119. Tukey, J. W.: The future of data analysis, *Ann. Math. Statist.*, vol. 33, no. 1, pp. 1–67, 1962.
120. Gringorten, I. I.: A plotting rule for extreme probability paper, *J. Geophys. Res.*, vol. 68, no. 3, pp. 813–814, 1963.
121. Lieblein, Julius: A new method of analyzing extreme-value data, *Natl. Advs. Comm. Aero. Tech. Note* 3053, Washington, D.C., 1954.
122. Chernoff, Herman, and G. J. Lieberman: Use of normal probability paper, *J. Am. Statist. Assoc.*, vol. 49, pp. 778–785, 1954.
123. Kimball, B. F.: The bias in certain estimates of the parameters of the extreme-value distribution, *Ann. Math. Statist.*, vol. 27, pp. 758–767, 1956.
124. Kimball, B. F.: On the choice of plotting positions on probability paper, *J. Am. Statist. Assoc.*, vol. 55, no. 291, pp. 546–560, 1960.
125. Ferrell, E. B.: Plotting experimental data on normal or lognormal probability paper, *Indus. Qual. Control*, vol. 15, no. 1, pp. 12–15, July, 1958.
126. Benson, M. A.: Plotting positions and economics of engineering planning, *Proc. Am. Soc. Civil Engrs., J. Hydraulics Div.*, vol. 88, no. HY6, pt. 1, November, 1962.
127. Brakensiek, D. L.: Fitting generalized lognormal distribution to hydrologic data, *Trans. Am. Geophys. Union*, vol. 39, pp. 469–473, 1958.
128. Kimball, B. F.: Sufficient statistical estimation functions for the parameters of the distribution of maximum values, *Ann. Math. Statist.*, vol. 17, no. 3, pp. 299–309, September, 1946.
129. Kimball, B. F.: An approximation to the sampling variation of an estimated maximum

value of a given frequency based on fit of doubly exponential distribution of maximum values, *Ann. Math. Statist.*, vol. 20, no. 1, pp. 110–113, March, 1949.
130. Panchang, G. M., and V. P. Aggarwal: Peak flow estimation by method of maximum likelihood, *Tech. Memo.* HLO2, Government of India, Central Water and Power Research Station, Poona, India, March, 1962.
131. Gringorten, I. I.: Envelopes for ordered observations applied to meteorological extremes, *J. Geophys. Res.*, vol. 68, no. 3, pp. 815–826, 1963.
132. Kendell, G. R.: Statistical analysis of extreme values, First Canadian Hydrology Symposium, National Research Council of Canada, November 4 and 5, 1959.
133. Riggs, H. C.: Frequency of natural events, *Proc. Am. Soc. Civil Engrs., J. Hydraulics Div.*, vol. 87, no. HY1, pt. 1, pp. 15–26, January, 1961.
134. Barricelli, N. A.: Les plus grands et les plus petits maxima ou minima annuels d'une variable climatique (The largest and smallest of annual maxima or minima of a climatic variable), *Archiv for Mathematick og Naturvidenskab* (Oslo), vol. XLVI, no. 6, 1943.
135. Jenkinson, A. F.: The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quart. J. Roy. Meteorol. Soc.*, vol. 81, no. 348, pp. 158–171, April, 1955.
136. Borgman, L. E.: The frequency distribution of near extremes, *J. Geophys. Res.*, vol. 66, no. 10, pp. 3295–3307, 1961.
137. Langbein, W. B., and others: Topographic characteristics of drainage basins, *U.S. Geol. Surv. Profess. Paper* 968-C, 1947, pp. 125–157.