PROBABILITY FUNCTIONS OF BEST FIT TO DISTRIBUTIONS OF ANNUAL PRECIPITATION AND RUNOFF

By

Radmilo D. Markovic

August, 1965

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

8

PROBABILITY FUNCTIONS OF BEST FIT TO DISTRIBUTIONS OF ANNUAL PRECIPITATION AND RUNOFF

By

Radmilo D. Markovic

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO

August, 1965

No. 8

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Figures

Page

Tables

ABSTRACT

Distributions of annual precipitation and annual river flow are studied on 2506 selected precipitation and river gaging stations in the Western United States and Southwestern Canada.

Five probability functions - Normal, Log-normal with 2, Log-normal with 3, Gamma with 2 and Gamma with 3 parameters - are fitted to each individual observed distribution. The maximum likelihood method is used for estimating the functions' parameters from observed data. The probability of chi-square is used as a measure of goodness of fit of each function to every observed sample distribution. These five functions are then tested on all station samples grouped into four large ensembles: homogeneous precipitation, non-homogeneous precipitation, river flow, and river flow corrected for the change in carryover.

As results of this study, it has been found that all five probability functions investigated are applicable. No one function is more suitable than the other in fitting an observed individual station sample precipitation or river flow distribution. However, distributions of annual precipitation in homogeneous ensemble (1141 samples) and nonhomogeneous ensemble (473 samples) are best fitted by the Log-normal 2 parameter function. Distributions of annual runoff in river flow ensemble (446 samples) and river flow ensemble corrected for the change in carryover (446 samples) are best fitted by the Gamma 2 parameter function. The difference in goodness of fit in ensemble analysis between these two functions is negligible for all practical purposes, and both could be used interchangeably for all four ensembles.

PROBABILITY FUNCTIONS OF BEST FIT TO DISTRIBUTIONS OF ANNUAL PRECIPITATION AND RUNOFF

By: Radmilo D. Markovic

CHAPTER I

INTRODUCTION

1. General. The variability of precipitation and river flows has long been recognized as an important factor related to water resources use and development. In the past, this variableness has led to an extensive study of precipitation and river flows, especially with respect to their dependence on a large number of climatic and physiographic factors.

Precipitation and river flow are governed by chance phenomena, that is, there are so many causes at work that the influence of each cannot be readily identified. Therefore, statistical and probability methods must be applied to adequately describe these hydrologic phenomena.

2. Subject. The purpose of this study is to find theoretical probability functions of best fit to distributions of annual precipitation and annual river flow as exemplified by fitting theoretical curves to observed data.

In addition to the main purpose of this study, answers have been sought to the following questions:

(1) Are there significant regional characteristics of annual precipitation and annual runoff which would indicate a better fit to observed data using a particular theoretical function?

(2) Is a particular theoretical distribution function as compared to another distribution function specifically advantageous in fitting the observed data?

(3) Are theoretical functions described by three parameters more suitable in fitting the observed data than those defined by two parameters?

(4) Does the nonhomogeneity of data significantly affect the fitting of annual precipitation values and if so, what is the resulting effect?

(5) Is there any significant difference in fitting distributions of annual flow in comparison with annual flow corrected for carryover?

3. Significant aspects of this study. To achieve the objective of this study and to answer the aforementioned questions, this study will investigate or make use of the following:

(1) Research data from a very large area, involving different climatic and physiographic conditions;

(2) A large number of precipitation and river gaging stations, 2506 station samples.

(3) A minimum of 30 years of observation of all hydrologic data; and,

(4) Three theoretical distribution functions are investigated simultaneously: normal, lognormal and gamma; as these latter two functions have each two cases, with two and with three parameters, practically five different functions are studied.

CHAPTER II

SELECTION OF RESEARCH DATA

1. Area under consideration. For the purpose of this study, the data from the western part of the United States and the southwestern part of Canada only is used. This large area was selected so that it would include many river basins of different sizes, a range of climatic areas from arid to humid regions of varied physiographic conditions ranging from plains to mountains. This large variety of natural conditions provides the basis for a generalization of the theoretical probability distributions for both annual precipitation and annual runoff. The selected area includes 21 states in the United States, as shown in fig. 1.

2. Basic research material. The basic material used in this investigation is constituted from two broad categories of data: annual precipitations from a large number of precipitation gaging stations; and annual river flows from numerous river gaging stations.

From these two categories of data, four large ensembles are formed with the following variables and notations:

(1) The homogeneous annual precipitation, P_1 - ensemble;

(2) The nonhomogeneous annual precipitation, P_2 - ensemble;

(3) The annual river flow, Q_1 - ensemble; and,

(4) The annual river flow corrected for carryover*, Q_2 - ensemble.

P_1 - ensemble consists of station samples of annual precipitation having homogeneous data. P_2 - ensemble includes station samples of annual precipitation having nonhomogeneous data, with the nonhomogeneity being proven or with the obvious evidence of nonhomogeneity. Q_1 - ensemble includes station samples of annual river flows. Q_2 - ensemble consists of the same station samples as Q_1 - ensemble, with the significant difference being that flows are corrected for the change in carryover.

Each station sample of annual observations for each of four ensembles has a size equal to the total length of observation, but not less than 30 years. The minimum 30-year period of observation is from 1931 to 1960, being chosen because it coincides with the standard climatological reference period. This period was adopted by the World Meteorological Organization as a standard reference period for all countries.

3. Criteria for selection of stations. Precipitation gaging stations were selected according to the following criteria:

(1) Minimum length of continuous period of observation of precipitation data is 30 years;

(2) Change of station location during the period of observation is less than one mile in horizontal direction and less than 100 feet in elevation for P_1 - ensemble; and the change is more than one mile in horizontal direction and more than 100 feet in elevation, and likewise not more than 5 miles and 500 feet, respectively, for P_2 - ensemble.

(3) No more than one year of missing data is estimated by regression analysis with neighboring stations and during the standard period of observations.

River gaging stations were selected according to the following criteria:

(1) Minimum length of continuous period of flow data observations is 30 years;

(2) No change in station location, or the change is negligible, or the flows are corrected for the change;

(3) No unaccounted transmountain diversions into the river basin or out of it, or diversions for irrigation do not exceed 2-3 percent of annual runoff; in case of large diversions, corrections are made in the river flows;

(4) No large storage reservoirs in river basin (in which their net capacity providing significant regulations);

(5) For large storage reservoirs the river flows are corrected for the differences in storage at the beginning and the end of water years;

(6) No more than one year of missing data during standard period of observation is estimated by regression analysis with neighboring stations; and

(7) Stations are independent among themselves; If more than one station is selected from the same river basin, the annual runoff at the downstream station(s) is (are) reduced for annual runoff at the upstream station(s).

4. Selected stations and their characteristics. On the basis of the aforementioned criteria, the following precipitation stations were selected:

* The synonymous expression used in Colorado State University Hydrology Papers Nos. 1, 4, and 7, is "Annual effective precipitation."

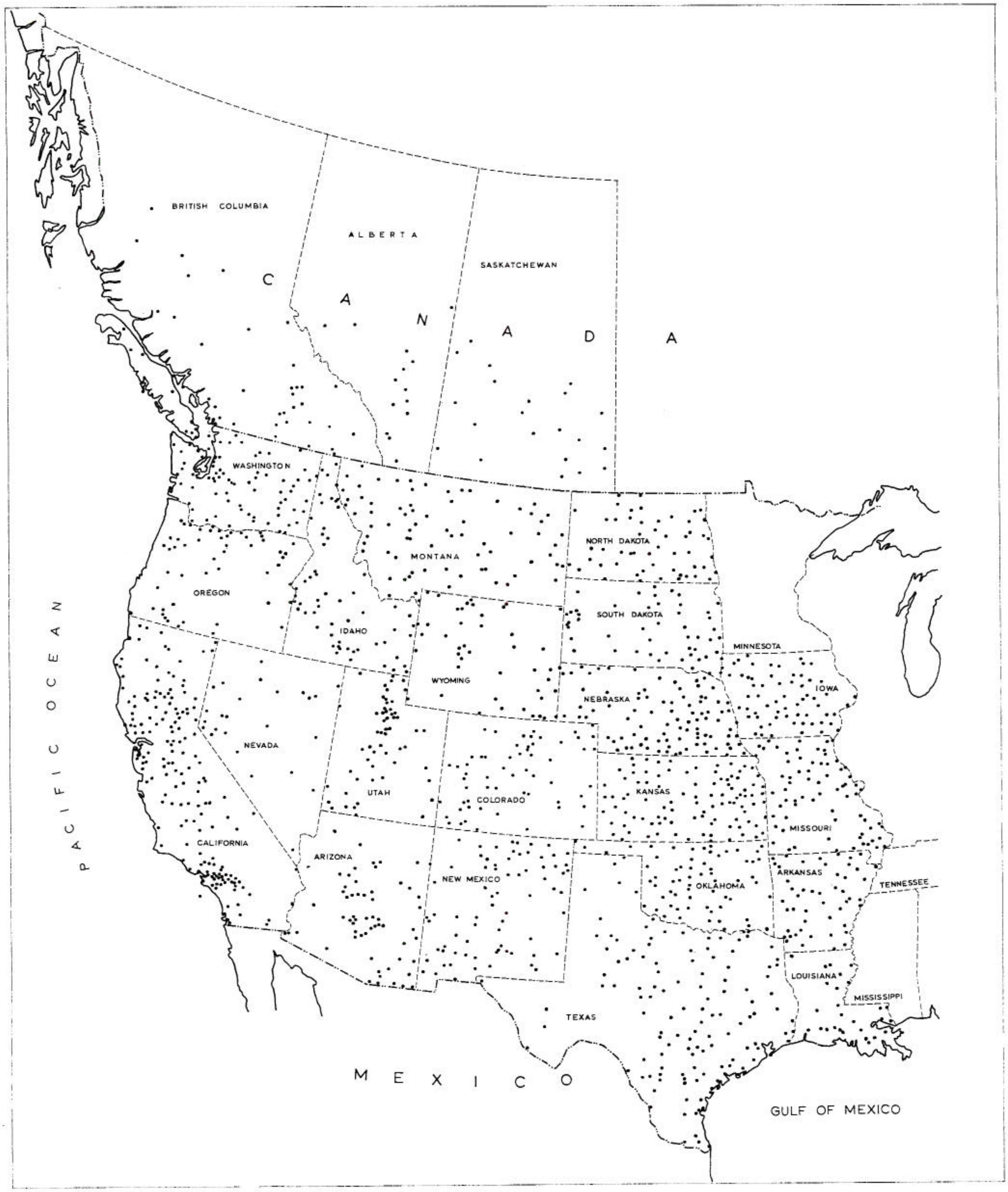

Fig. 1 General location of selected precipitation stations

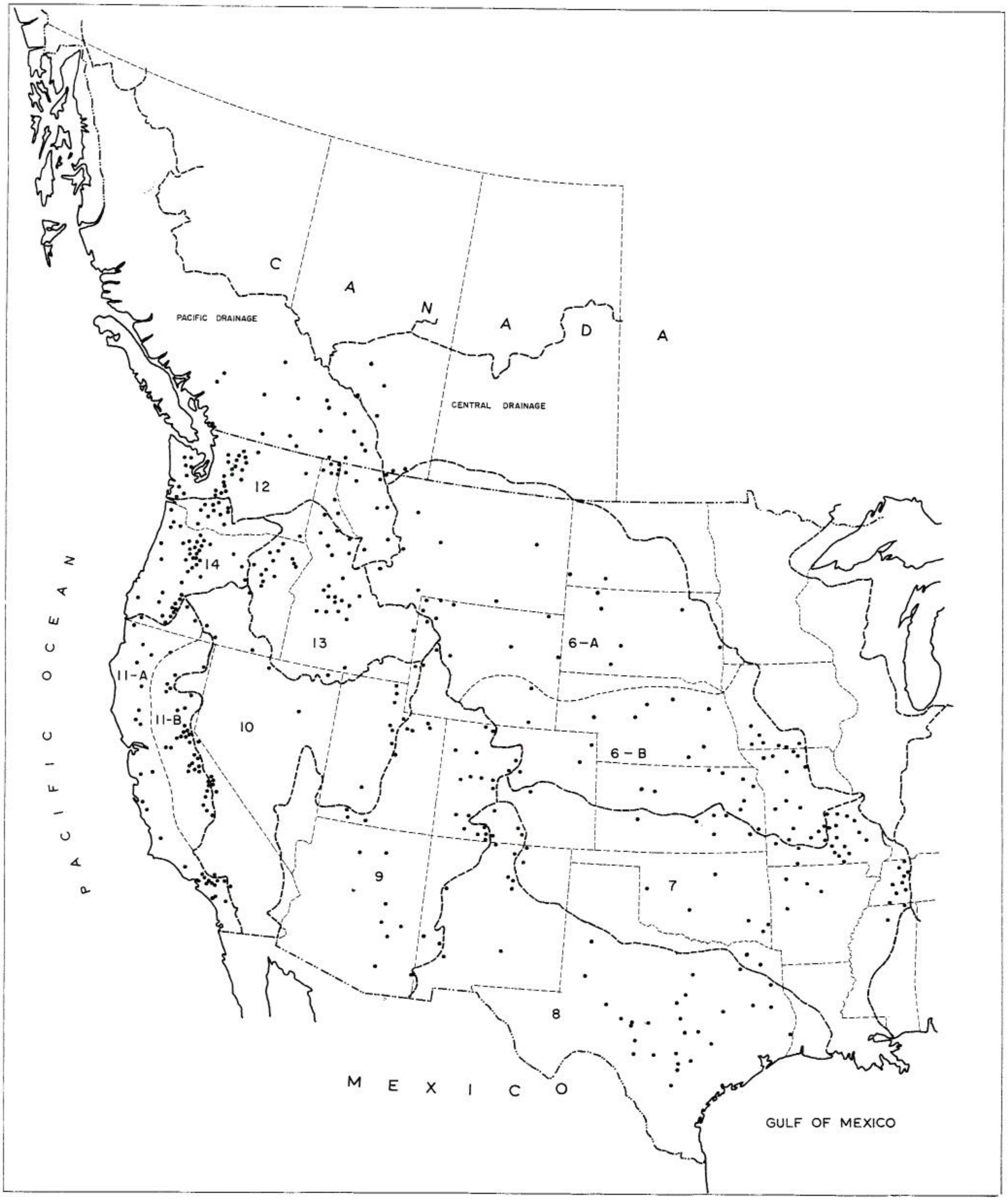

Fig. 2 General location of centroids of river basins controlled by selected river gaging stations

State or Province	Number of Stations P_1 ensemble	 P_2 ensemble
United States		
1) Washington	47	36
2) Oregon	39	14
3) California	153	28
4) Nevada	14	7
5) Idaho	40	20
6) Utah	43	15
7) Arizona	50	9
8) New Mexico	46	24
9) Colorado	41	23
10) Wyoming	26	21
11) Montana	60	17
12) North Dakota	48	14
13) South Dakota	44	17
14) Nebraska	83	24
15) Kansas	62	47
16) Oklahoma	42	26
17) Texas	85	41
18) Louisiana	22	7
19) Arkansas	34	17
20) Missouri	46	28
21) Iowa	34	38
Canada		
22) British Columbia	48	
23) Alberta	17	
24) Saskatchewan	17	
TOTAL	1141	473

The locations of selected precipitation stations are shown in fig. 1.

The lengths of the period of observation range from 30 years (majority of P_1- and P_2-ensemble stations) to 114 years (New Orleans WB City, Louisiana, USA) for P_1-ensemble, and 125 years (Leavenworth, Kansas, USA) for P_2-ensemble. The average station period of observation for all is 53.8 years for P_1-ensemble and 57.4 years for P_2-ensemble.

In accordance with the established criteria a total of 446 river gaging stations were selected from the considered area with the number of stations for both the Q_1- and Q_2-ensembles as follows:

Location	Number
(1) Part 14, Pacific slope basin in Oregon and the lower Columbia River basin	60
(2) Part 13, the Snake River basin	32
(3) Part 12, Pacific slope basins in Washington and the upper Columbia River basin	55
(4) Part 11, Pacific slope basins in California	
(a) Outside Central Valley	22
(b) Central Valley	40
(5) Part 10, The Great Basin	21
(6) Part 9, The Colorado River Basin	40
(7) Part 8, Western Gulf of Mexico basin	41
(8) Part 7, the Lower Mississippi River basin	49
(9) Part 6, The Missouri River basin	
(a) Above Sioux City, Iowa	26
(b) Below Sioux City, Iowa	45
(10) Pacific Drainage basin, Canada	9
(11) Central Drainage basin, Canada	6
TOTAL	446

The locations of these selected river gaging stations are shown in fig. 2.

Drainage areas controlled by selected stations range from 1.90 square miles (The Little Santa Anita Creek near Sierra Madre, California, USA) to 34,000 square miles (the Columbia River at Birchank, British Columbia, Canada). The lengths of the period of observation range from 30 years (majority of stations) to 72 years for the Verde River below Bartlett Dam, Arizona, USA. The Arizona station also represents the longest uninterrupted period of flow observation, from 1889 to 1960. The average station period of observation is 37 years.

5. Compilation of annual river flows corrected for carryover (Q_2-ensemble). While the three ensembles P_1, P_2 and Q_1 are observed data, Q_2-ensemble is derived. Basically, it is the Q_1-ensemble in which each annual river flow is corrected for the change in water carryover from year to year. The correction is done by applying the following equation:

$$Q_{2,i} = Q_{1,i} + \left(W_{1,i} - W_{1,i-1}\right)\frac{1}{T} \qquad (1)$$

in which,

$Q_{1,i}$ = Annual observed river flow;

$Q_{2,i}$ = Annual river flow corrected for the change in carryover in time T;

i, i-1 = Indices referring to the i-th and (i-1)th member of samples; and

$W_{1,i}$, $W_{1,i-1}$ = Total stored volume in the river basin at the end of the i-th and (i-1)-th water year, respectively.

Details of this correction are explained in reference [12]. With this correction the Q_2-ensemble is obtained from the Q_1-ensemble.

6. Properties of observed data. The most important properties of observed data in fitting the probability functions to observed distributions are the sample size, range, frequency property and comparability of data.

As it has been shown, the station sample sizes vary from 30 years of observation up to 114 for P_1, 125 for P_2 and 72 for Q_1 and Q_2 ensembles. Statistically speaking, all observed samples can be treated as samples of small sizes.

Considering the range of independent variables, it has been found that the average precipitation - the station sample mean - ranges from 1.66 inches per year (Greenland Ranch, California) to

173.21 inches per year (Ocean Falls, British Columbia, Canada). The annual precipitation as the sample members for the same stations vary from 0.01 to 4.62 and from 109.69 to 235.94 inches per year respectively. On the other hand, the average annual flow range between 0.723 cubic feet per second (Aliso Creek at El Toro, California) and 70,697 cubic feet per second (Columbia River at Birchank, British Columbia, Canada). The mean annual flows or sample members, range from 0.001 to 3,520 and from 52,300 to 88,700 cubic feet per second respectively. As it can be seen, the ranges of annual precipitation and annual river flow are very considerable, indicating the large variety of the climatic and physiographic conditions of the area under consideration. The independent variables range practically from very small values close to zero up to very high values which are not defined. They can go physically from zero as lower limit (dry) to very high values (flood), which can be theoretically considered as unlimited, i.e., infinity as upper limit. Thus, the theoretical range of the annual precipitation and annual river flow is from zero to plus infinity.

The frequency distributions that are generally found have the characteristic typical form. They usually start with the zero frequency, then rise to a maximum value, and again decrease finally to zero. They are generally tangent to the base at both lower and upper ends. The basic shape of frequency curves of observed data is thus the bell type, two-tailed curve. They are either slightly to very skewed or asymmetrical, having the following order of characteristics of central tendency: mean, median and mode. The brief inspection of raw data, however, indicates that this usual order of measures of central tendency changes in some extreme cases of natural conditions. Hence, the large variety of skewness of frequency curves of observed data could be expected.

From the comparability point of view, the annual data, as it is published, collected and classified, does not provide a comparison between stations because of high variability of means and standard deviations, In order to bridge this difficulty, actual observed values are transformed to dimensionless variables. At the same time, they should be simplified for ease in making comparisons. It has been shown that the most suitable form is to transform the annual values of P_1-, P_2-, Q_1- and Q_2-ensembles into modular coefficients (K_i), as

$$K_i = \frac{P_i}{\overline{P}} \quad \text{or} \quad K_i = \frac{Q_i}{\overline{Q}} \qquad (2)$$

in which $\overline{P}$ or $\overline{Q}$ denotes the sample mean of each selected station. The transformed annual values into dimensionless form are given in the example of river gaging station Weldon River at Mill Grove, Missouri, USA, in the Appendix.

CHAPTER III

SELECTION OF THEORETICAL DISTRIBUTION FUNCTIONS AND ESTIMATION OF PARAMETERS

1. Criteria for selection. According to properties of observed data, the theoretical distribution functions of best fit to observed distributions of annual precipitation and annual runoff should have the following characteristics: (1) the function is continuous and defined for all positive values of the observed variable K; (2) the lower tail is bounded by zero value or by a positive value, K_o; (3) the upper tail is unbounded; (4) the density curve is asymptotic to the axis for large values of K; (5) the basic shape is one peak bell-shaped two-tailed curve, with a large variety of skewness; and (6) the number of parameters which describe theoretical functions is limited to three.

2. Applicable functions. The general class of functions, originally studied by Karl Pearson, may be represented by the differential equation

$$\frac{df(K)}{dK} = \frac{f(K)(K + K_m)}{\phi(K)} \qquad (3)$$

with f (K) a density function, ϕ (K) a function of the independent variable K, and K_m the distance from the origin to the mode.

With ϕ (K) expanded in power series form, the general equation rewritten is

$$\frac{1}{f(K)} \frac{df(K)}{dK} = \frac{K + K_m}{C_o + C_1K + C_2K^2 + \dots} \qquad (4)$$

in which C_o, C_1, C_2, ... are constants. Their values determine the shape of the curve. Equation (4) is the differential equation of density functions for various values of K_m, C_o, C_1 and C_2. Thus, for the particular case with C_1, C_2, ... zeros, eq. (4) results in the normal (Gaussian) probability density function. The log-normal probability density function, as an example of transformation of the normal function, has been found to provide a significant goodness of fit to many observed distributions of hydrologic variables. Likewise, the Gamma function has very convenient properties for application to hydrologic data, and it can be defined with two or three parameters. The latter one is often known as Pearson Type III function. From several functions obtainable from eq. (4), by assigning various values to constants C_1, C_2, ... , and K_m, the following functions only have been selected for investigation.

3. Selected functions. Screening of the applicable functions with respect to the criteria required, their convenience for use in mass computation, and the experience already obtained in applying them in hydrology, lead to the following selection:

(1) Normal density function, or Normal;

(2) Log-normal density function with two parameters, or Log-normal 2;

(3) Log-normal density function with three parameters, or Log-normal 3;

(4) Gamma density function with two parameters, or Gamma 2; and

(5) Gamma density function with three parameters, or Gamma 3.

The expressions and parameters of these five functions are [8]:

(1) Normal with the classical form:

$$f(K) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(K-\mu)^2}{2\sigma^2}} \qquad -\infty \leq K \leq +\infty \qquad (5)$$

with K - the variable values; μ - the population mean; and σ - the population standard deviation.

(2) Log-normal 2 with the form:

$$f(K) = \frac{1}{K\sigma\sqrt{2\pi}} e^{-\frac{(\ln K - \ln\mu)^2}{2\sigma^2}} \qquad o \leq K \leq \infty \qquad (6)$$

with μ - the population geometric mean; and σ - the population standard deviation of the ln K values.

(3) Log-normal 3 with the form:

$$f(K-K_o) = \frac{1}{(K-K_o)\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\ln^2\left(\frac{K-K_o}{\mu}\right)} \qquad K_o \leq K \leq \infty \qquad (7)$$

with μ - the population geometric mean of $(K-K_o)$; K_o - the lower boundary of the distribution of the variable K; and σ - the standard deviation of the ln $(K-K_o)$ values.

(4) Gamma 2 with the form:

$$f(K) = \frac{1}{\beta^\alpha \Gamma(\alpha)} K^{\alpha-1} e^{-K/\beta} \qquad o \leq K \leq \infty \qquad (8)$$

with α - the shape parameter; β - the scale parameter; and $\Gamma(\alpha)$ the gamma function of α. It is skewed to the right for all values of parameters α and β.

(5) Gamma 3 with the form

$$f(K-K_o) = \frac{1}{\beta\,\Gamma(\alpha)} \left(\frac{K-K_o}{\beta}\right)^{\alpha-1} e^{-\frac{K-K_o}{\beta}}$$

$$K_o \leq K \leq \infty \qquad (9)$$

with K_o - the location parameter of the lower boundary, and α, β, and $\Gamma(\alpha)$, as previously defined.

4. Estimation of parameters of selected functions. In estimating the values of parameters of the parent population the following properties of estimators are desirable [2, 5, 6,]:

(a) Consistent estimators, meaning that the probability of the absolute value of the deviation of estimator $\hat{\theta}$ from the population parameter θ is less than any small quantity ϵ, tends to unity as sample size n tends to infinity, i. e. $P_r\ (|\hat{\theta} - \theta| < \epsilon) \longrightarrow 1$ as $n \longrightarrow \infty$;

(b) Unbiased estimators, or the expected value of estimator is equal to the population parameter, $E(\theta) = \theta$ with biasedness being defined as $E(\theta) - \theta$; and

(c) Efficient estimators, or among the class of consistent estimators, the minimum unbiased estimator $\hat{\theta}$ has the smallest variance.

According to desirable properties of estimators, the maximum likelihood method is chosen as the most suitable for the estimation of parameters in this investigation.

Maximum likelihood method, developed by R. A. Fisher, is based upon likelihood function L. This function is maximized by setting the first derivative of $\ln K$ with respect to θ equal to zero, and solving the resulting equation for θ:

$$\frac{\partial(\ln L)}{\partial\theta} = \frac{\partial\{\sum_{i=1}^{n} \ln[f(K_i;\theta)]\}}{\partial\theta} = 0 \qquad (10)$$

This yields a single equation for the solution of θ in terms of K's. For m parameters, m equations of eq. (10) give m estimators or unknown parameters. Maximum likelihood estimators are consistent, asymptotically normal and asymptotically efficient under general conditions. The method is completely numerical, applicable to all selected functions and convenient for mass computation. The maximum likelihood method gives the following equations for parameter estimators [8]:

Normal. Based on eq. (5) and the concept of eq. (10), the maximum likelihood function produces:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} K_i \qquad (11)$$

as estimator of the population mean, and

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(K_i - \hat{\mu})^2} \qquad (12)$$

as the estimator of population standard deviation.

Log-normal 2. According to eq. (6) and using the maximum likelihood equation, the maximum likelihood estimator of the population mean is:

$$\widehat{\ln\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln K_i \qquad (13)$$

and the estimator of the population standard deviation:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln K_i - \ln\hat{\mu})^2} \qquad (14)$$

Log-normal 3. Equations (7) and (10) yield the maximum likelihood equation with respect to parameters $\ln\mu$, σ and K_o. The maximum likelihood estimator of the population mean is:

$$\widehat{\ln\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln(K_i - \hat{K}_o) \qquad (15)$$

of the population standard deviation is:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\ln(K_i - \hat{K}_o) - \widehat{\ln\mu}]^2} \qquad (16)$$

and of the lower boundary K_o is

$$\sum_{i=1}^{n} \frac{1}{K_i - K_o}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln^2(K_i - \hat{K}_o) - [\frac{1}{n}\sum_{i=1}^{n}\ln(K_i - \hat{K}_o)]^2 - \frac{1}{n}\sum_{i=1}^{n}\ln(K_i - \hat{K}_o)\right\} + \sum_{i=1}^{n}\frac{\ln(K_i - \hat{K}_o)}{K_i - \hat{K}_o} = 0 \qquad (17)$$

in which $\hat{K}_o$ as the maximum likelihood estimator of the population lower boundary may be solved only by an iteration procedure.

Gamma 2. Applying the same technique to eq. (8), the maximum likelihood equation gives the two maximum likelihood partial differential equations for parameters α and β, and from them it follows

$$\ln\hat{\alpha} - \frac{\partial[\ln\Gamma(\hat{\alpha})]}{\partial\hat{\alpha}} = \ln\bar{K} - \frac{1}{n}\sum_{i=1}^{n}\ln K_i$$

with $\hat{\alpha}$ the estimator of α, and

$$\hat{\beta} = \frac{1}{\hat{\alpha}}\ \frac{1}{n}\sum_{i=1}^{n} K_i = \frac{1}{\hat{\alpha}}\ \bar{K} \qquad (18)$$

with $\hat{\beta}$ the estimator of β. The equation for α involves the digamma function

$$\frac{\partial[\ln\Gamma(\hat{\alpha})]}{\partial\hat{\alpha}} = \emptyset\ (\hat{\alpha})$$

and it is solved by a simplified technique. Nörlund [11] shows that

$$\phi(\hat{\alpha}) = \ln\hat{\alpha} - \frac{1}{2\hat{\alpha}} - \sum_{i=1}^{n} \frac{(-1)^{i-1} B_i}{2i\hat{\alpha}^{2i}} + R_n$$

is an asymptotic expansion in which B_i are the Bernouli numbers $B_1 = \frac{1}{6}$, $B_2 = \frac{1}{30}$, etc., and R_n is the remainder after n terms; for $n = 1$ it becomes

$$\phi(\hat{\alpha}) = \ln\hat{\alpha} - \frac{1}{2\hat{\alpha}} - \frac{1}{12\hat{\alpha}^2}$$

Substituting this in the above expression, a quadratic equation is obtained

$$12\left[\ln\left(\frac{1}{n}\sum_{i=1}^{n} K_i\right) - \frac{1}{n}\sum_{i=1}^{n} \ln K_i\right]\hat{\alpha}_1^{\,2} - 6\hat{\alpha}_1 - 1 = 0$$

whose only pertinent root is

$$\hat{\alpha} = \frac{1 + \sqrt{1 + \frac{4}{3}\left(\ln\bar{K} - \frac{1}{n}\sum_{i=1}^{n} \ln K_i\right)}}{4\left(\ln\bar{K} - \frac{1}{n}\sum_{i=1}^{n} \ln K_i\right)}$$

The error in $\hat{\alpha}$ resulting from using only one term in Nörlund's expansion is not readily expressed in mathematical form. Hence, $\hat{\alpha}$ should be corrected, that is, the correction factor $\Delta\hat{\alpha}$ which takes care of the error. is subtracted from the estimator, and giving

$$\hat{\alpha} = \frac{1 + \sqrt{1 + \frac{4}{3}\left(\ln\bar{K} - \frac{1}{n}\sum_{i=1}^{n} \ln K_i\right)}}{4\left(\ln\bar{K} - \frac{1}{n}\sum_{i=1}^{n} \ln K_i\right)} - \Delta\hat{\alpha} \qquad (19)$$

which is the final maximum likelihood estimator of α. The correction factor $\Delta\hat{\alpha}$ is given in Table 1. [11]. As $\bar{K} = 1$ for modular coefficients, then $\ln\bar{K} = 0$, so that in this case eqs. (18) and (19) can be simplified.

Gamma 3. In accordance with eq. (9), the maximum likelihood equation produces three partial differential equations which give the maximum likelihood estimator of the shape parameter α as [8]:

$$\hat{\alpha} = \frac{1 + \sqrt{1 + \frac{4}{3}\left[\ln(\bar{K}-\hat{K}_o) - \frac{1}{n}\sum_{i=1}^{n} \ln(K_i-\hat{K}_o)\right]}}{4\left[\ln(\bar{K}-\hat{K}_o) - \frac{1}{n}\sum_{i=1}^{n} \ln(K_i-\hat{K}_o)\right]} - \Delta\hat{\alpha} \qquad (20)$$

the maximum likelihood estimator of the scale parameter β

$$\hat{\beta} = \frac{1}{\hat{\alpha}}\,\frac{1}{n}\sum_{i=1}^{n}(K_i-\hat{K}_o) = \frac{1}{\hat{\alpha}}\,(\bar{K}-\hat{K}_o), \qquad (21)$$

and the maximum likelihood estimator of the lower boundary parameter, K_o, obtained in implicit form to be solved by an iteration procedure:

$$\frac{1+\sqrt{1+\frac{4}{3}\left[\ln(\bar{K}-\hat{K}_o) - \frac{1}{n}\sum_{i=1}^{n}\ln(K_i-\hat{K}_o)\right]}}{1+\sqrt{1+\frac{4}{3}\left[\ln(\bar{K}-\hat{K}_o) - \frac{1}{n}\sum_{i=1}^{n}\ln(K_i-\hat{K}_o)\right]} - 4\left[\ln(\bar{K}-\hat{K}_o) - \right.} \; *$$

$$\frac{*}{\left. - \frac{1}{n}\sum_{i=1}^{n}\ln(K_i-\hat{K}_o)\right]} - (\bar{K}-\hat{K}_o)\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{K_i-\hat{K}_o}\right) = 0 \qquad (22)$$

5. Computation of maximum likelihood estimates. By using the annual values expressed in dimensionless form K_i and stored on magnetic tape, the maximum likelihood estimates for each of the five selected functions and for each station sample separately are computed on a CDC 3600 electronic computer.

Parameters of the normal function are estimated by eqs. (11) and (12); those of Log-normal 2, eqs. (13) and (14); those of Log-normal 3, eqs. (15), (16) and (17). Parameters of Gamma 2 are estimated by eqs. (18) and (19), and those of Gamma 3 by eqs. (20), (21) and (22).

Having computed estimates of parameters, the five selected functions are then completely defined.

As an example, the computation of the maximum likelihood estimates for all five functions is shown for the case of the river gaging station of the Weldon River at Mill Grove, Missouri (Appendix).

TABLE 1

CORRECTION FACTOR $\Delta\hat{\alpha}$ FOR COMPUTATION OF MAXIMUM LIKELIHOOD ESTIMATES OF THE SHAPE PARAMETERS OF GAMMA FUNCTIONS WITH 2 AND 3 PARAMETERS

$\hat{\alpha}$	$\Delta\hat{\alpha}$	$\hat{\alpha}$	$\Delta\hat{\alpha}$
0.200	0.034	1.400	0.006
0.300	0.029	1.500	0.005
0.400	0.025	1.600	0.005
0.500	0.021	1.700	0.004
0.600	0.017	1.800	0.004
0.700	0.014	1.900	0.003
0.800	0.012	2.200	0.003
0.900	0.011	2.300	0.002
1.000	0.009	3.100	0.002
1.100	0.008	3.200	0.001
1.200	0.007	5.500	0.001
1.300	0.006	5.600	0.000

CHAPTER IV

TECHNIQUES FOR TEST OF GOODNESS OF FIT

To test the theoretical probability distribution functions for goodness of fit to observed data, as in any other frequency analysis, the distribution of a random variable should be classified into mutually exclusive and exhaustive categories or class intervals. It is not desirable to make a frequency distribution for fewer than about 20 to 25 observations, since a smaller number of observations may be studied in an array. In classifying the observed data, it is necessary to decide upon the number and the length of class intervals.

1. Number and length of class intervals. No satisfactory hard-and-fast rule has been established for the number of class intervals to be used. It is obvious, however, that if too many classes are used, some of them would have few or no frequencies and the resulting frequency distribution would be irregular. Likewise, if there are too few classes, the observed data would be very compressed, a large proportion of the frequencies would fall in one or two classes, and much information would be lost. In addition, different classifications for a given set of observations and for a continuous variable lead to different observed distributions and hence to different values of departures from postulated distribution.

Since there is no generally accepted method for determination of the number of class intervals, numerous rules have been suggested by many statisticians. According to these rules, the number of class intervals should not be smaller than about 10 and not larger than 20, but these practical rules have no theoretical basis. Nevertheless, a guide for the systematic choice of the number of class intervals has been developed. Based upon some theoretical considerations, several mathematical formulas have been suggested for the number of class intervals for different sample sizes and levels of significance. In such a situation, a practical rule commonly used by many statisticians will be applied in this analysis. The rule states that the number of class intervals should be chosen so that the average expected frequency of any class intervals is at least five. Since the observed sample sizes used in this study range from 30 to 72, with an average of about 37 for river flows, the total number of class intervals selected is seven.

The choice of the length of class intervals should be done in such a manner that the main characteristic features of the observed distribution are emphasized and chance variations are obscured [4]. Basically, there are two concepts for choice of the length of class intervals: (a) equal lengths, and (b) equal probabilities.

Equal lengths of class intervals are extensively used even though there is no theoretical foundation for it. However, it has some advantages in graphical representation of observed distributions, since the comparability is difficult to carry out by inspection when there exist inequalities in class intervals. Also, two arbitrary actions must be introduced: the choice of the size of equal interval, and the beginning of the first interval limit. Each of these actions directly affects the observed distribution, which is a disadvantage of this method.

Equal probabilities of class intervals, which can be considered as special case of unequal lengths, has some advantages over the previous method [7]. The arbitrary steps for equal lengths may be avoided by choosing intervals of equal probabilities instead of intervals of equal lengths. The required intervals are obtained from the probability integral transformation. The probabilities are uniformly distributed. Thus, the comparison of the observed distributions with any continuous theoretical distribution is reduced to the comparison of an observed with a theoretical uniform distribution. This method is more convenient and much simplier for numerical analysis than the previous one, and it is used in this report. According to this method, with the total number of class intervals already chosen in the above discussion, and with the fact that the total value of the probability integral is unity, the probability of each class interval is determined by

$$p_j = \frac{1}{k} \text{ with } j = 1, 2, \ldots k. \qquad (23)$$

For this value of probability, the required length of any class interval can be obtained from the probability integral transformation.

2. Test of fit. The well-known and frequently applied Chi-square test is used here as a measure of goodness of fit of the theoretical probability distributions to observed ones. Other similar tests to be noted, but were not used, include: the likelihood ratio (observed over expected maximum likelihood function), which is asymptotically equivalent to Chi-square test; Smirnov statistics (all of observations involved); and Kolmogorov statistics (only maximum departure involved) as function of cumulative distribution of the sample.

The problem of testing the goodness of fit of a hypothesized probability distribution to observed sample distribution was solved in the main by K. Pearson in 1900, who developed the Chi-square test. Later, R. A. Fisher contributed the significant idea of "degrees of freedom" by which proper account is taken of parameters estimated from the observed data [3].

The basic concept of the Chi-square test can be summarized as follows: The total range of sample observations is divided into k mutually exclusive and exhaustive class intervals, each having the observed class frequency O_j and corresponding expected class probability E_j $(j = 1, 2, \ldots, k)$. Using the expected value E_j as the norm of any class interval, it is reasonable to choose the quantity $(O_j - E_j)^2$ as a measure of departure from the norm. However, the magnitudes of the squared deviations $(O_j - E_j)$ would not be comparable from one class to another, since the scale of each is nearly proportional to the expected value. Therefore, a suitable measure is expressed by $(O_j - E_j)^2/E_j$ and the measure of total discrepancy between observations and expectations, X^2, becomes

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

This statistic is distributed asymptotically as Chi-square (χ^2) with k - 1 degrees of freedom, if the population parameters have not been estimated from the sample observations. Since in this study only the general form of the probability distribution is hypothesized, the parameters of the selected functions should be estimated from observed data. In such a case, the number of degrees of freedom is decreased for the number of parameters estimated from observations. For ν parameters, the total number of degrees of freedom is

$$f = k - 1 - \nu \tag{24}$$

The Chi-square statistic as previously given is a convenient form for representing the direct function of the differences between the observed frequencies and their hypothetical expectations, so that the comparability is possible by direct inspection. Furthermore, due to the large volume of computations involved in this investigation, this statistic is simplified for computational purposes. Thus, expanding the quadratic in the numerator

$$X^2 = \sum_{j=1}^{k} \frac{O_j^2}{E_j} - 2 \sum_{j=1}^{k} O_j + \sum_{j=1}^{k} E_j$$

and noting that $\Sigma O_j = \Sigma E_j = n$ (sample size), and $E_j = p_j n$ but $p_j = 1/k$, the following equivalent expression to be used is obtained

$$X^2 = \frac{k}{n} \sum_{j=1}^{k} O_j^2 - n \tag{25}$$

3. Expected and observed class frequencies. As the total number of class intervals is seven and the probability of each interval is the same, for given sample size, the expected class probability for any interval should be the same and independent of the type of probability function, i. e., it is dependent only on the sample size n, or

$$E_j = p_j n = \frac{n}{k} . \tag{26}$$

Therefore, the computation of expected class probabilities is simplified by choosing the constant number of class intervals of the same probability.

The observed class frequencies depend upon sample size; the class limits depend upon the type of probability function applied. Since the computational procedure is identical for all observed samples, as an example the five selected probability functions are applied to one station sample. (See Appendix).

First, the sample observations should be arranged in an array in increasing order. Then, to determine how many observations will fall in each of the seven chosen class intervals, six class interval limits must be computed for each of five selected functions separately.

Normal. Knowing the probability of any class interval, p_j (equal for all intervals), which represents the area under the probability curve, any class interval limit, K_j, can be evaluated from the corresponding cumulative distribution obtained by integrating eq. (5) in the limits which produce the same probabilities, provided that the lower limit of integral is previously known, as well as the parameters of the function. The solution of the integral of eq. (5) can be simplified by standardizing the variable, or

$$F(U) = jp_j = \int_{-\infty}^{U_j} \frac{1}{\sqrt{2\pi}} e^{-\frac{U^2}{2}} dU \tag{27}$$

with j = 1, 2, . . . , 7, and with the lower integral limit $-\infty$, the mean zero and the variance unity. This is a well known probability integral, the value of which is generally given in tabulated form. The class interval limits as expressed in terms of U_j are determined and given in Table 2. From the values of U_j, and the estimates of population mean and standard deviation, $\hat{\mu}$ and $\hat{\sigma}$, the particular class interval limits K_j of the variable K_i are

$$K_j = \hat{\mu} + U_j \hat{\sigma} \tag{28}$$

in which U_j are class interval limits of the variable U_i of eq. (27).

Log-normal 2. Similar to the previous case, the class interval limits of log-normal 2 are computed by using eq. (6), which is first transformed into a normal probability integral form. The class interval limits are then computed from the expression

$$K_j = \exp \ [\ln\hat{\mu} + U_j \hat{\sigma}] \tag{29}$$

in which K_j are class interval limits for the variable K_i, $\widehat{\ln\mu}$ is the mean of $\ln K_i$, and $\hat{\sigma}$ is the standard deviation of $\ln K_i$, while U_j are class interval limits of the variable U_i from eq. (27).

Log-normal 3. The class interval limits are determined by using eq. (7) and transforming the variable $(K_i - \hat{K}_o)$ first into normal probability integral form. Then, the class interval limits are obtained as

$$K_j = \hat{K}_o + \exp[\ln\hat{\mu} + U_j \hat{\sigma}] \tag{30}$$

where K_j are class interval limits for the variable K_i of eq. (7), $\widehat{\ln\mu}$ is the mean of $\ln(K_i - \hat{K}_o)$, and $\hat{\sigma}$ is the standard deviation of $\ln(K_i - \hat{K}_o)$, while U_j are class interval limits of the variable U_i from eq. (27). Since parameters of this function are determined earlier and the values of U_j are given in Table 2, eq. (30) gives class interval limits.

TABLE 2

NORMAL DENSITY FUNCTION
FOR COMPUTATION OF CLASS INTERVAL LIMIT VALUES

No. of class interval limits, j	1	2	3	4	5	6
Probability, F (U)	0.14286	0.28571	0.42857	0.57143	0.71429	0.85714
Abscissa, U_j	-1.068	-0.566	-0.180	+0.180	+0.566	+1.068

Gamma 2. The class interval limits are computed by using eq. (8) with the lower integral limit zero. In order to use the existing Tables of Incomplete Gamma Function, the integral of eq. (8) is first expressed in terms of the shape parameter only, by using the value of scale parameter of eq. (18) as follows:

$$F(K) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} \int_{o}^{K_j} K^{\alpha - 1} e^{-\alpha K} dK \qquad (31)$$

Then, introducing the replacements $\alpha - 1 = p$ and $\alpha K = x$ from which $\alpha = p + 1$ and $K = x/\alpha$, or $dK = \frac{dx}{\alpha} = \frac{1}{p+1} dx$, the above integral takes the form of:

$$I(x, p) = \frac{(p+1)^{p+1}}{\Gamma(p+1)} \int_{o}^{x_j} \frac{1}{(p+1)^{p}} x^{p} e^{-x} \frac{1}{p+1} dx$$

$$= \frac{1}{\Gamma(p+1)} \int_{o}^{x_j} x^{p} e^{-x} dx \qquad (32)$$

Because the argument x theoretically runs from 0 to $+\infty$, the more workable argument u is used in its place, therefore, the range is considerably decreased in the existing tables and determined by

$$u = \frac{x}{\sqrt{p+1}} \qquad (33)$$

The final form for which the values are tabulated, represents practically the familiar probability integral expressed as the ratio of incomplete to complete gamma function for arguments u and p:

$$I(u, p) = \frac{\Gamma_u(p+1)}{\Gamma_{\infty}(p+1)} = \frac{\int_o^u u^p e^{-u} du}{\int_o^{\infty} u^p e^{-u} du} \qquad (34)$$

Standard tables [9] give I (u, p) with the argument u proceeding by increments of 0.1 from zero up to that value of u which gives I (u, p) equal unity to the seventh decimal place. The argument p increases from -1.00 to 1.00 by increments of 0.05, from 1.0 to 5.0 by increments of 0.1 and from 5.0 to 50.0 by increments of 0.2.

For the purpose of this report the standard tables, previously mentioned, are recomputed and are presented in Table 3. For the negative values of argument p, $p \leq 0$, and $u \leq 0.800$ the graphical procedure was applied, since the linear interpolation would result in an error (fig. 3). For $0 < p \leq 50$ the existing tables were used, and for $p > 50$ the corresponding values were extrapolated by using the numerical procedure explained in reference [9].

The class interval limits are computed by using eq. (33) for x and replacement $K = x/\alpha$ in eq. (32), so that

$$K_j = \frac{u_j}{\sqrt{\hat{\alpha}}} \qquad (35)$$

with u_j selected for given value of $\hat{\alpha}$ from Table 3.

Gamma 3. The computation of the class interval limits is similar to that of the gamma function with two parameters. From eq. (8) the cumulative distribution is obtained and by means of transformation is reduced to that of eq. (35). Then, the class interval limit equation is [8]

$$K_j = \hat{K}_o + \frac{(1 - \hat{K}_o)\ u_j}{\sqrt{\hat{\alpha}}} \qquad (36)$$

By knowing the estimated parameters, the value of u_j can be selected for given $\hat{\alpha}$ from Table 3. Then, class limits are computed by eq. (36).

4. Computation of station sample chi-squares. The computational procedure is identical for all station samples. To each of them, five selected probability functions are fitted. Since seven class intervals are already chosen, six class interval limits for each function and every station sample are determined according to the following equations: for Normal function by eq. (28); Log-normal 2 by eq. (29); Log-normal 3 by eq. (30); Gamma 2 by eq. (35); and Gamma 3 by eq. (36).

Knowing the class interval limits, the corresponding observed class frequencies are determined, squared and summed and then station sample chi-squares computed by eq. (25). Since five

TABLE 3

INCOMPLETE GAMMA FUNCTION
FOR COMPUTATION OF CLASS INTERVAL LIMIT VALUES

Class interval, j		1	2	3	4	5	6
$I(u, p) = \frac{\Gamma_u(p+1)}{\Gamma_\infty(p+1)}$		0.14286	0.28571	0.42857	0.57143	0.71429	0.85714
$p = \hat{\alpha} - 1$	$\hat{\alpha}$	u_1	u_2	u_3	u_4	u_5	u_6
-0.8	0.2	0.007	0.015	0.036	0.092	0.303	0.932
-0.6	0.4	0.021	0.060	0.147	0.335	0.675	1.381
-0.4	0.6	0.048	0.140	0.299	0.540	0.919	1.630
-0.2	0.8	0.094	0.240	0.434	0.708	1.103	1.806
0.0	1.0	0.153	0.338	0.559	0.850	1.254	1.947
0.5	1.5	0.313	0.557	0.819	1.131	1.546	2.218
1.0	2.0	0.468	0.748	1.033	1.357	1.774	2.430
1.5	2.5	0.614	0.919	1.217	1.549	1.967	2.610
2	3	0.749	1.074	1.382	1.786	2.136	2.770
3	4	1.000	1.349	1.670	2.013	2.429	3.049
4	5	1.224	1.591	1.921	2.267	2.682	3.291
5	6	1.429	1.810	2.145	2.494	2.907	3.508
6	7	1.620	2.010	2.350	2.700	3.112	3.707
7	8	1.799	2.196	2.540	2.891	3.302	3.891
8	9	1.966	2.370	2.717	3.070	3.480	4.065
9	10	2.126	2.535	2.884	3.238	3.647	4.228
10	11	2.278	2.692	3.043	3.397	3.805	4.383
11	12	2.420	2.838	3.191	3.568	3.953	4.528
12	13	2.563	2.985	3.339	3.694	4.101	4.674
13	14	2.696	3.120	3.476	3.831	4.238	4.808
14	15	2.828	3.255	3.612	3.968	4.374	4.942
15	16	2.952	3.382	3.740	4.096	4.502	5.067
16	17	3.076	3.508	3.867	4.223	4.629	5.192
17	18	3.194	3.627	3.987	4.344	4.748	5.310
18	19	3.311	3.746	4.107	4.464	4.868	5.429
19	20	3.422	3.859	4.220	4.578	4.981	5.541
20	21	3.532	3.972	4.334	4.691	5.094	5.653
21	22	3.638	4.080	4.442	4.832	5.202	5.760
22	23	3.744	4.187	4.550	4.974	5.310	5.867
23	24	3.846	4.290	4.654	5.044	5.414	5.969
24	25	3.947	4.393	4.757	5.114	5.517	6.071
25	26	4.044	4.492	4.856	5.214	5.616	6.169
26	27	4.142	4.590	4.955	5.313	5.715	6.267
27	28	4.236	4.685	5.051	5.408	5.810	6.362
28	29	4.331	4.780	5.147	5.504	5.906	6.457
29	30	4.422	4.872	5.239	5.596	5.998	6.548
30	31	4.514	4.964	5.331	5.689	6.090	6.639
31	32	4.602	5.053	5.420	5.778	6.180	6.978
32	33	4.690	5.142	5.510	5.868	6.269	6.816
33	34	4.775	5.228	5.596	5.954	6.356	6.904
34	35	4.860	5.315	5.683	6.041	6.442	6.991
35	36	4.944	5.398	5.767	6.125	6.566	7.073
36	37	5.027	5.482	5.851	6.209	6.689	7.155
37	38	5.108	5.564	5.932	6.291	6.731	7.236
38	39	5.189	5.645	6.014	6.373	6.773	7.316
39	40	5.268	5.744	6.094	6.452	6.872	7.396
40	41	5.346	5.843	6.174	6.532	6.932	7.475
41	42	5.423	5.901	6.252	6.610	7.010	7.552
42	43	5.501	5.959	6.329	6.688	7.087	7.629
43	44	5.576	6.035	6.405	6.764	7.163	7.704
44	45	5.650	6.111	6.481	6.840	7.239	7.780
45	46	5.724	6.184	6.555	6.914	7.313	7.854
46	47	5.799	6.258	6.629	6.988	7.387	7.927
47	48	5.860	6.331	6.702	7.061	7.460	8.000
48	49	5.941	6.404	6.775	7.134	7.532	8.072
49	50	6.012	6.474	6.846	7.205	7.603	8.142
50	51	6.083	6.545	6.917	7.276	7.674	8.213
55	56	6.306	6.791	7.181	7.558	7.976	8.541
60	61	6.583	7.089	7.496	7.889	8.325	8.924
65	66	6.854	7.380	7.804	8.214	8.667	9.600
70	71	7.124	7.672	8.112	8.538	9.010	9.648

TABLE 3 - continued

$p = \hat{\alpha} - 1$	$\hat{\alpha}$	u_1	u_2	u_3	u_4	u_5	u_6
80	81	7.605	8.190	8.660	9.114	9.618	10.300
90	91	8.080	8.701	9.201	9.684	10.218	10.943
100	101	8.507	9.161	9.687	10.195	10.758	11.521
110	111	8.922	9.607	10.159	10.692	11.283	12.083
120	121	9.319	10.035	10.611	11.168	11.785	12.620

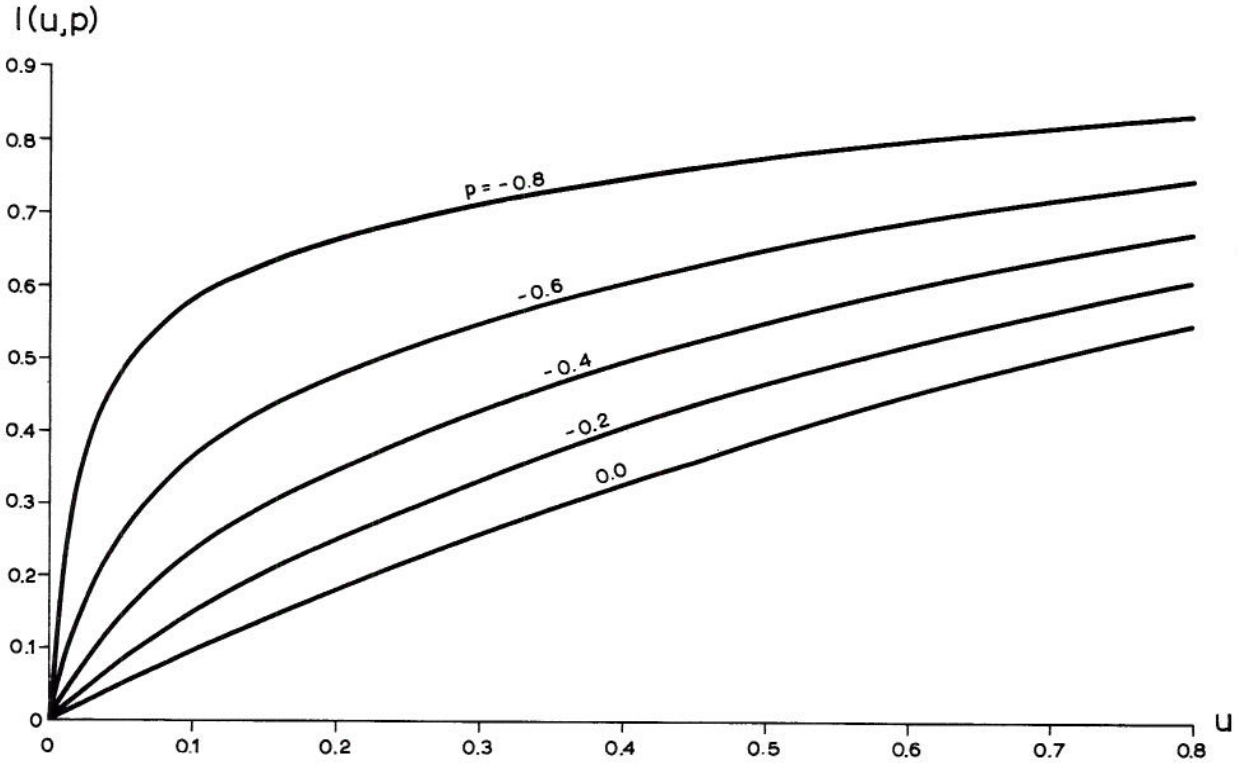

Fig. 3 Incomplete Gamma Function
Used only for $p \leq 0$ and $u \leq 0.800$

functions are fitted to annual observations of every station, each station sample is represented by five chi-square values. These five computed values for Normal, Log-normal 2 and Gamma 2 are distributed as Chi-square (χ^2) with four degrees of freedom ($f =$ 4 d. f.), while for Log-normal 3 and Gamma 3 distributed as Chi-square (χ^2) with three degrees of freedom ($f =$ 3 d. f.). These five chi-square values per station, one for each of five probability density functions, give automatically the measure of goodness of fit of a particular theoretical function to observed data for each individual station sample. However, this is not the only purpose of this investigation, but also includes the ensemble analysis involving all samples of the same population pooled together.

Class interval limits, observed class frequencies, and chi-squares for all five functions and all 2506 station samples are computed by the CDC 3600 electronic computer.

To check the program for computer and to show the computation, one example is presented. For that purpose Chi-squares with three and four degrees of freedom and different level of significance are given in Table 4, and their cumulative distributions plotted in fig. 4. The example used is the analysis of data for Weldon River at Mill Grove, Missouri, USA, and is given in the Appendix.

5. Transformation of station sample chi-square. As it is evident, the station sample chi-squares are not of the same degrees of freedom and thus, they are beyond comparison among themselves. In order to avoid this difficulty, to facilitate the further analysis and to insure their comparability, the computed station sample chi-squares are transformed into their corresponding probabilities. This transformation was performed on the CDC 3600 electronic computer by using the Chi-square cumulative distribution function

$$F(\chi^2) = \frac{1}{2^{\frac{1}{2}f}\,\Gamma(\frac{1}{2}f)} \int_0^{X^2} (\chi^2)^{\frac{1}{2}(f-2)}\, e^{-\frac{1}{2}\chi^2}\, d\chi^2 \qquad (37)$$

in which f stands for the number of degrees of freedom, and X^2, the upper integral limit, stands for the computed station sample chi-square. In this way, the probabilities of station sample chi-squares instead of chi-squares themselves, are used as a unique measure of goodness of fit of theoretical functions to observed distributions.

TABLE 4

CHI SQUARE DISTRIBUTION

	χ^2 for	
F (χ^2)	f = 3 d.f.	f = 4 d.f.
0.001	0.019	0.074
0.005	0.072	0.207
0.010	0.115	0.297
0.020	0.185	0.429
0.025	0.216	0.484
0.050	0.352	0.711
0.075	0.468	0.890
0.100	0.584	1.064
0.150	0.808	1.360
0.200	1.005	1.649
0.250	1.213	1.923
0.300	1.424	2.195
0.350	1.640	2.460
0.400	1.875	2.740
0.450	2.110	3.040
0.500	2.366	3.357
0.550	2.650	3.680
0.600	2.950	4.040
0.650	3.290	4.430
0.700	3.665	4.878
0.750	4.108	5.385
0.800	4.642	5.989
0.850	5.296	6.725
0.900	6.251	7.779
0.925	6.920	8.480
0.950	7.815	9.488
0.975	9.348	11.143
0.980	9.837	11.668
0.990	11.345	13.277
0.995	12.838	14.860
0.999	16.268	18.465

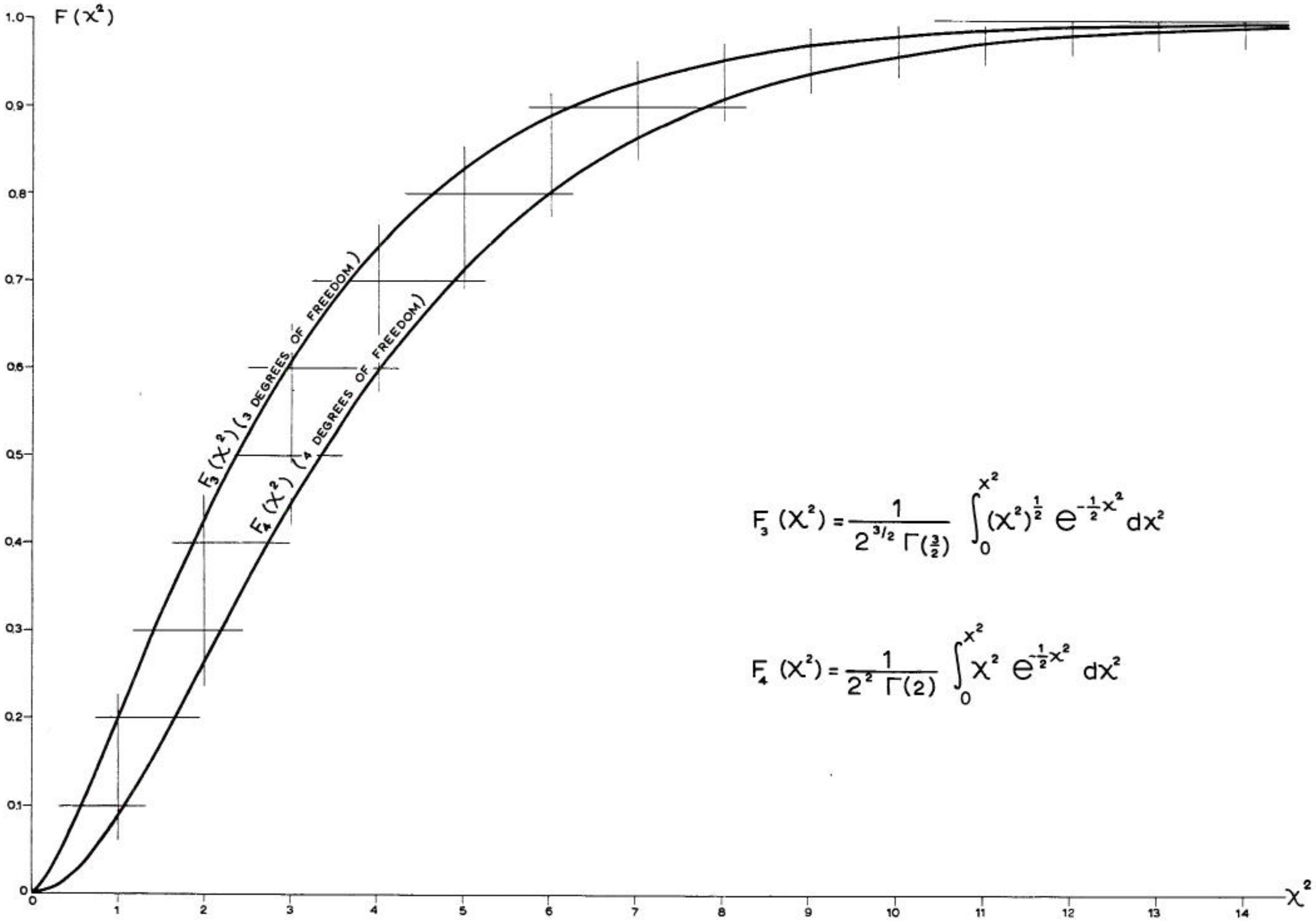

Fig. 4 Chi-square Cumulative Distribution

CHAPTER V

ANALYSIS OF RESULTS

1. Individual stations. As stated in the previous chapter, the probability of station sample chi-square is chosen as a measure of the goodness of fit of a theoretical function to an observed station sample distribution. If this probability of a hypothesized function is less than an assigned level of significance, this function would be acceptable as a good approximation to the distribution of a considered station sample. The departures between the theoretical and observed distribution are considered as not being significant. If the reverse is true, the departures would be significant and the postulated function would be rejected for a selected level of significance. For the purpose of this study, a significant level of 95 percent is used. For instance, if a normal function was fitted to the distribution of annual precipitation observed at the precipitation station at Anacortes, Washington, the departures between normal and observed distribution would give the probability of chi-square of 0.882. This value is less than 0.95. The departures are not significant, and the normal function is a good fit. This conclusion is supported by a relatively small skewness coefficient. This coefficient at this station is only 0.061 indicating that annual precipitation distribution is close to normal.

On the other hand, if the normal function is fitted to the distribution of annual precipitation observed at the precipitation station San Diego WB AP, California, the conclusion is somewhat different. The probability of sample chi-square is 0.984. It is greater than 0.95, indicating a high departure between the theoretical and the observed distribution. Hence, the normal function does not satisfactorily fit the observed data and it is rejected at 95 percent level of significance. This conclusion is supported by a relatively high skewness coefficient which is 1.304, indicating that the observed distribution is highly positively skewed and is far from normal.

These two precipitation stations offer two extreme examples. Though both stations are located at low elevations and on the coast of the Pacific Ocean close to the moisture source, their precipitation characteristics differ considerably. The average annual precipitation at Anacortes is 26.52 in./year, while at San Diego WB AP, California, it is 9.86 in./year. Their coefficients of variation are 0.186 and 0.408, and their skewness coefficients are 0.061 and 1.304 respectively. This data illustrates that it is not likely that either the altitude of station or its distance from the ocean could explain differences between distributions of annual precipitation of these two stations. The other factors, such as ocean currents, latitude, temperature, evaporation, prevailing wind direction of moist air masses, environmental orographic and other conditions are certainly governing factors in creating the difference in distribution.

The normal function was the only function considered in the above example. However, since five theoretical functions have been applied to the same observed distribution of any individual station, the five values of station sample chi-square probabilities were obtained. All these five values are of the same nature, namely, all of them are dimensionless and are comparable among themselves. The smaller the value of this probability, the smaller are departures between the theoretical and observed distributions and the better theoretical function fits an observed distribution. In the previous examples, the probabilities of chi-squares are:

	Anacortes, Washington	San Diego WB AP, California
Normal	0.882	0.984
Log-normal 2	0.302	0.063
Log-normal 3	0.302*	0.063*
Gamma 2	0.654	0.476
Gamma 3	0.894	0.442

It follows that distributions of annual precipitation at Anacortes and San Diego WB AP precipitation stations are best fitted by a Log-normal function with 2 parameters. Parameters which describe this function are different at each station.

The determination of the lower boundary was a problem in both the Log-normal 3 and Gamma 3 functions. Namely, the maximum likelihood eqs. (17) and (22) produce often negative values of the lower boundary parameter estimate. This was particularly true for distributions approaching a normal function with the following characteristics: (a) slightly positively or negatively skewed; (b) highly concentrated or with small range distributions; and (c) a relatively large natural logarithm of the geometric mean. Negative estimates for the lower boundary parameter in the cases of Log-normal 3 and Gamma 3 functions need clarifications. First, from the physical point of view, neither precipitation nor river flows can be negative. Second, from the mathematical point of view, both Log-normal 3 and Gamma 3 functions are defined only for the positive range of an independent variable. Therefore, the estimates of the lower boundary parameter have been constrained to a positive range, $K_o \geq 0$.

When a variable of the type of flow or precipitation can have zero values, then for $Q = 0$ or $P = 0$ there is a finite probability mass. In this case, the probability distribution is composed from a discrete part (probability mass at the value zero) and a continuous part for all values greater than zero. The negative value of the lower boundary and the negative values of the variable of probability density curve can be conceived. However, the area under the probability density curve between the lower boundary and the value zero should be approximately equal to the observed discrete probability of the value zero. However, there was no zero value of annual precipitation or annual runoff for stations considered by this study. This fact means that the above concept of negative (though immaginary) values of precipitation and runoff cannot be applied to cases when annual values constitute the samples. Therefore, it is necessary to replace negative lower

* Denotes that the lower boundary parameter of the function is zero, and, hence, Log-normal 3 reduces to Log-normal 2.

boundary estimates by zero values for Log-normal 3 and Gamma 3 functions when the maximum likelihood equations produce the above values. As a consequence, Log-normal 3 and Gamma 3 were automatically reduced to Log-normal 2 and Gamma 2, respectively. This was the reason why some station sample probabilities of chi-squares had the same value for two and three parameter functions of the same family.

As stated earlier, 2506 individual station samples have been used in this study. It is nearly impossible to analyze them individually. For the purpose of this report, only a few characteristic station samples are discussed and some conclusion advanced.

Two precipitation and two river-gaging station samples are selected for this discussion. The precipitation stations are: Ocean Falls, British Columbia, Canada, and Greenland Ranch, California. The first has the highest and the latter the lowest average of annual precipitation. The river-gaging stations are: Frenchman Creek near Hamlet, Nebraska, and Arroyo Trabuco near San Juan Capistrano, California. The first has the lowest and the latter the highest coefficient of variation of annual river flows. Basic data illustrating station characteristics are listed below. They include: altitude (H), latitude (Y), longitude (X), drainage area (A), the sample mean ($\overline{P}$, $\overline{Q}$), the average specific yield of a river basin ($\overline{q}$), standard deviation (s), coefficient of variation (C_v) and skewness coefficient (C_s).

		H	Y	X	A	$\bar{P}$ or $\bar{Q}$	$\overline{q}$	s	C_v	C_s
	STATION	ft.	digr.	digr.	sq. mi.	in. or (cfs)	cfs/sq. mi.	in. or (cfs)		
Precipitation	Ocean Falls	16	52.35	127.67		173.21		24.61	0.140	0.338
	Greenland Ranch	-168	36.45	116.87		1.66		1.17	0.700	0.768
Run-off	Frenchman Creek	2798	40.38	101.21	1480.0	98.83	0.067	9.29	0.092	-0.245
	Arroyo Trabuco	180	33.53	117.67	36.5	5.04	0.138	9.21	1.796	2.198

Time series of annual values, their duration, (cumulative frequency) and frequency curves are graphed for precipitation stations in fig. 5 and for river-gaging stations in fig. 6.

The precipitation station Ocean Falls is located at a low altitude and relatively high latitude. It is close to the Pacific Ocean and in a wet region. The precipitation station Greenland Ranch is located in a land depression - Death Valley - below sea level, at a lower latitude, several hundreds of miles inland, and in a very dry region. A big difference in annual precipitation and in its time distribution for two locations is mainly caused by the general air circulation patterns, and ocean currents. The station at Ocean Falls is under the influence of the warm North Pacific Current, which is closely associated with a high cyclonic activity and hence has frequent and high amounts of precipitation. The Southern California Coast is predominantly under the influence of the cold California Current. It is related to a high anticyclonic activity with infrequent and generally low precipitation. Besides, Greenland Ranch station is farther inland than Ocean Falls station, hence, it is affected by an additional decrease of precipitation which comes with an increase in the distance from the moisture source. Frequent rainfall causes the annual amounts of precipitation to be more uniformly distributed in time and more concentrated around the sample mean. The reverse is true for the infrequent rainfall. This fact is best illustrated by comparing the stations in fig. 5. Hence, the observed frequency distribution at Ocean Falls is best fitted by the Normal function, showing the lowest probability of station sample chi-square of 0.491. The Normal function seems to offer a satisfactory fitting within the observed range of annual precipitation values. The observed distribution at Greenland Ranch station is positively skewed and is better fitted by the Gamma 2 function, showing the lowest probability of station sample chi-square, 0.188, of all five functions investigated.

Distributions of annual river flows are affected by physiographic factors of a river basin apart from its precipitation. In the case of Frenchman Creek which has a relatively large drainage area, low average annual flow and remarkably low specific yield, the annual flow distribution is highly uniform. The frequency distribution of annual flows has a slight negative skewness though the annual precipitation over this basin is slightly positively skewed and follows the Log-normal 2 distribution. The explanation of these facts is closely related to drainage basin characteristics. The river basin in a relatively smooth topography and moderate relief is composed of Ogallala and Sandborn formations consisting mainly of gravel, sand, silt and clay. This hugh aquifier, averaging hundreds of feet in thickness, is underlain by an impermeable barrier of upper cretaceous shale and partially overlain by sand dunes. Such a very permeable surface structure provides for high infiltration resulting in water recharging the large underlaying groundwater reservoir. This large aquifier, connected with surrounding basins, is mainly responsible for extremely high participation of groundwater in the total runoff, on one side, and groundwater exchange between adjacent basins or watershed leakage on the other side. The topographic and phreatic divide of the watershed do not coincide because of the plain topography and geological structure. As a result, the distribution of annual river flows is highly uniform. This observed distribution is approximated by Normal function with $P(X^2) = 0.931$, the other four functions being positively skewed and hence, of worse fitting. The annual flow distribution at Arroyo Trabuco is highly nonuniform in time and of a very skewed distribution. Observed frequency distribution at this station is best fitted by Log-normal 2 function, having the probability of station sample chi-square of 0.133. Differences in annual flow distributions observed at these two river gaging stations are best illustrated by fig. 6.

These few examples of individual station sample analyses show a large variety of climatic and physiographic conditions which influence distributions of annual precipitation and runoff. As a consequence,

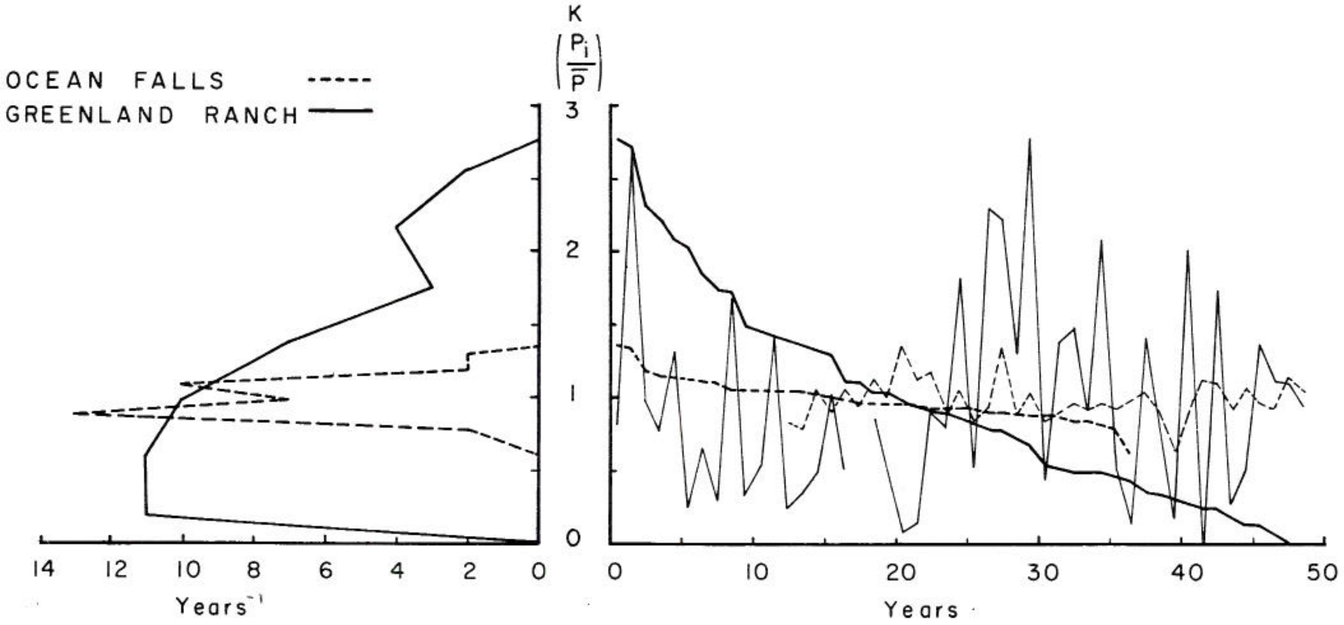

Fig. 5 Time series, cumulative distribution and frequency curves of annual precipitation at Ocean Falls, British Columbia, and at Greenland Ranch, California

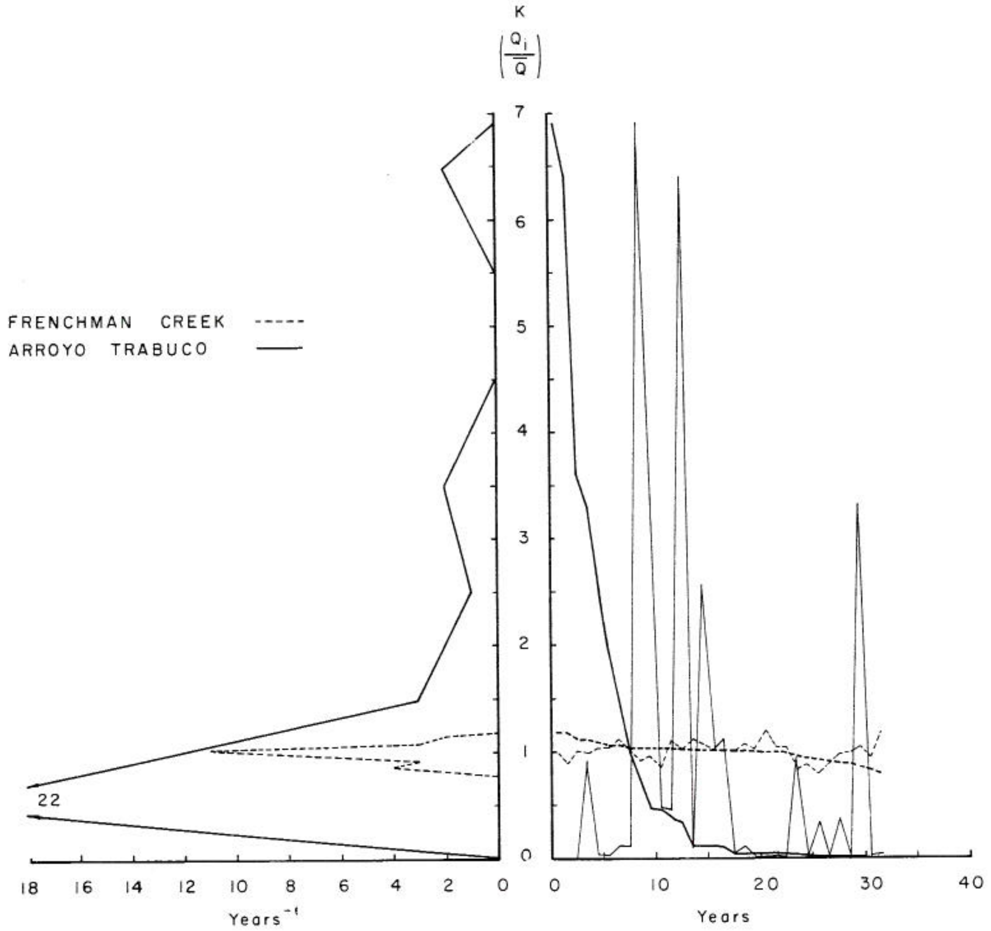

Fig. 6 Time series, cumulative distribution and frequency curves of annual river flows at Frenchman Creek near Hamlet, Nebraska, and at Arroyo Trabuco near San Juan Capistrano, California

a large variety of observed distributions is obtained. They have a very small and very high range, from extremely positive through symmetrical up to slightly negative skewness, and from a flat to a very high pickadness. These facts greatly complicate any generalization concerning the distributions at individual stations.

All five functions are applicable individually and can staisfactorily fit the observed frequency distributions of annual precipitation and annual runoff. In some particular cases this fit is exceptionally good:

Precipitation Stations	Best Fit By Function	$P(X^2)$
Scott, Saskatchewan, Canada	Normal	0.090
Galveston WB City, Texas	Log-normal 2	0.047
Parma, Missouri	Log-normal 3	0.093
Superior, Nebraska	Gamma 2	0.074
Hat Creek PH No. 1, California	Gamma 3	0.104
River Gaging Stations		
Middle Fork John Day River at Ritter, Oregon	Normal	0.067
Martin Creek near Paradise Valley, Nevada	Log-normal 2	0.094
Chevelon Fork near Winslow, Arizona	Log-normal 3	0.019
Blk River at Clark, Colorado	Gamma 2	0.045
Hatchie River at Bolivar, Tennessee	Gamma 3	0.057

It seems from these limited number of examples that no function studied has a particular advantage in fitting the observed distributions of individual station samples. Often all five functions fit the same observed distribution very well, however, some of them better than the others, as it is shown for the following four stations:

			$P(X^2)$		
Precipitation Stations	N	LN2	LN3	G2	G3
Rochelle 3E, Wyoming	0.342	0.052	0.235	0.187	0.133
Hudson, Kansas	0.098	0.159	0.299	0.098	0.529
River Gaging Stations					
North River near Raymond, Washington	0.487	0.187	0.335	0.114	0.335
Trapper Creek near Oakley, Idaho	0.623	0.280	0.167	0.582	0.505

On the other hand, there are cases where no function fits the observed distributions at the 95 percent level, and some of them do not fit at the 99 percent level of significance. Some examples are:

			$P(X^2)$		
Precipitation Stations	N	LN2	LN3	G2	G3
Fort Bidwell, California	0.999	0.999	0.999	0.999	0.999
Marlow 1WSW, Oklahoma	0.995	0.997	0.997	0.975	0.997
River Gaging Stations					
Chowchilla River at Buchanan Dam Site, California	0.999	0.999	0.999	0.999	0.999
Comal River at New Braunfels, Texas	0.998	0.999	0.999	0.999	0.999

These results indicate the need for an additional mathematical function or functions in order to cover the whole range of observed individual sample distributions. One of the properties of additional function(s) should be the ability to be negatively skewed, since all five functions used in this study are either symmetrical or positively skewed.

Considering the fitting of individual station sample distributions, in general, the two parameter functions are simplier and easier to work with than the three parameter functions. This is mainly due to the difficulty in estimating the lower boundary parameters. Besides, among two parameter functions, Normal and Log-normal 2 have some practical and computational advantages over Gamma 2: (1) They are familiar functions and have tables of normal integral; and (2) They can be easily transformed from one to another and have graphical scales to be plotted as straight lines.

The probability of station sample chi-square, as a measure of the goodness of fit of a theoretical function to an observed distribution, is more universal and more convenient for use in mass computation than the sample chi-square itself. The probability of chi-square is dimensionless and, hence, is very useful for comparison of distributions, regardless of the number of degrees of freedom and the physical units of basic data involved in the statistical analysis.

2. Ensembles of stations. The probabilities of station sample chi-squares, previously arranged by variables and grouped into P_1, P_2, Q_1 and Q_2 - ensembles, represent the basic material for ensemble analyses. Several ways can be used to test which one of the five selected functions is of the best fit to each of the four ensembles. One way is simply to count the number of station samples from the same

ensemble, which are either satisfactorily or unsatisfactorily fitted by theoretical function at an assigned level of significance. In other words, to count successes and failures in fitting tests. The greater the number of successes or the smaller the number of failures, the better is the fitting of observed distributions by a theoretical function. If the 95 and 99 percent levels of significance are applied, as commonly used, then the results are as follows:

ENSEMBLE	TOTAL NUMBER OF SAMPLES	FUNCTION	NUMBER OF STATION SAMPLES			
			95% Level of Significance		99% Level of Significance	
			Success	Failure	Success	Failure
P_1	1141	Normal	1053	88	1111	30
		Log-normal 2	1054	87	1126	15
		Log-normal 3	1044	97	1124	17
		Gamma 2	1036	105	1115	26
		Gamma 3	929	212	1085	56
P_2	473	Normal	429	44	463	10
		Log-normal 2	435	38	465	8
		Log-normal 3	431	42	464	9
		Gamma 2	434	39	461	12
		Gamma 3	396	77	455	18
Q_1	446	Normal	322	124	356	90
		Log-normal 2	398	48	430	16
		Log-normal 3	393	53	430	16
		Gamma 2	400	46	432	14
		Gamma 3	391	55	430	16
Q_2	446	Normal	321	125	360	86
		Log-normal 2	399	47	430	16
		Log-normal 3	397	49	429	17
		Gamma 2	398	48	428	18
		Gamma 3	401	45	428	18

According to this success-failure test results, P_1 and P_2 ensembles are best fitted by Log-normal 2 parameter function at both 95 and 99 percent levels of significance. Q_1-ensemble is best fitted by the Gamma 2 function at both 95 and 99 percent levels of significance, while Q_2-ensemble by Gamma 3 at 95 and Log-normal 2 at 99 percent levels of significance. It should be noted at this point that some differences between functions are negligible and that the above conclusions may be misleading, particularly in the case of the Q_2-ensemble. It should be noted that Log-normal 3 and Gamma 3 functions are not consistently three parameter functions in this ensemble analysis, but rather combinations of two and three parameter functions of the same family of functions. This is the consequence of the previous restriction upon the lower boundary parameter to be equal to or greater than zero. Whenever the lower boundary parameter is considered to be zero, the three parameter functions automatically have been reduced to two parameter functions. This happened a surprising number of times, so that the feasibility of the use of three parameter functions can be seriously questioned. The following table, which contains the number of station samples with a boundary zero for three parameter functions, illustrates this problem:

Ensemble	LOG-NORMAL 3		GAMMA 3	
	Number of Samples	Percentage	Number of Samples	Percentage
P_1	942	82.56	192	16.83
P_2	385	81.40	77	16.28
Q_1	375	84.08	283	63.45
Q_2	377	84.53	283	63.45

This success-failure test is relatively unrefined. It only takes care of the cumulative frequency of successes at a particular level of significance.

Another way of testing the fits of ensembles is by determining the maximum deviation of probabilities of station sample chi-squares for various functions and ensembles from a given standard distribution of probabilities of chi-squares. The idea of this test is a comparison of absolute maximum deviation, D, between the observed cumulative frequency, F_o, of probabilities of chi-squares, and a hypothesized uniform cumulative distribution, F_u, of these chi-square probabilities. The smaller the maximum deviation the better the fitting of a theoretical function to observed station sample distributions grouped into an ensemble. Applying this concept to probabilities of station sample chi-squares, $P(X^2)$, the above maximum deviation can be determined from the expression:

$$D = \max \left| F_o [P(X^2)] - F_u [P(X^2)] \right| \qquad (38)$$

For this purpose, the probabilities of station sample chi-squares are classified into 40 equal class intervals, the observed class frequencies determined, the relative and cumulative relative class frequencies computed. For the sake of brevity, computations are omitted here, but the result in the form of frequency and cumulative frequency distributions are graphed in figs. 7 through 10. From these figures, the maximum absolute deviations, D, between the observed and the hypothesized uniform cumulative distribution are obtained for each ensemble and for all five functions, as follows:

ENSEMBLE	FUNCTION	$F_o[P(X^2)]$ %	$F_u[P(X^2)]$ %	D %
P_1	Normal	36.19	52.50	16.31
	Log-normal 2	52.58	65.00	12.42
	Log-normal 3	27.86	42.50	14.64
	Gamma 2	41.91	55.00	13.09
	Gamma 3	32.61	62.50	29.89
P_2	Normal	45.66	65.00	19.34
	Log-normal 2	29.38	45.00	15.62
	Log-normal 3	26.00	45.00	19.00
	Gamma 2	25.16	42.50	17.34
	Gamma 3	37.00	70.00	33.00
Q_1	Normal	42.58	70.00	27.42
	Log-normal 2	36.56	52.50	15.94
	Log-normal 3	34.97	52.50	17.53
	Gamma 2	20.19	35.00	14.81
	Gamma 3	30.74	52.50	21.76
Q_2	Normal	52.47	80.00	27.53
	Log-normal 2	29.84	45.00	15.16
	Log-normal 3	46.62	62.50	15.88
	Gamma 2	38.78	52.50	13.72
	Gamma 3	30.95	52.50	21.55

P_1 - ensemble is best fitted by Log-normal 2, since for this function the deviation between the observed and the hypothetical distribution is the smallest. Close fitting to this ensemble could be obtained by the Gamma 2 function. The order of best goodness of fit then follows: Log-normal 3, Normal and Gamma 3 (fig. 7).

P_2 - ensemble is fitted exactly the same way as P_1. Probability functions follow the same order according to goodness of fit as in the P_1 - ensemble. The only difference is that the P_1 - ensemble is better fitted in general then the P_2 - ensemble. Though the random errors and the inconsistency in data of annual precipitation are partly involved in both P_1 and P_2 ensembles, most inconsistency appears in the latter. It seems that non-homogeneity is the prevailing factor for the above difference between these two ensembles. The effect of non-homogeneity in data is manifested in higher maximum deviations between observed and hypothesized distributions, and hence, in worse fitting (fig. 8).

Q_1 - ensemble is approximated better by Gamma 2, then by Log-normal 2 which shows almost the same result as Gamma 2. These two functions are then followed by Log-normal 3, Gamma 3 and Normal. The latter two functions show considerable deviations which represent bad fits (fig. 9).

Q_2 - ensemble, as fitting results show, does not differ from the Q_1 - ensemble. What was said for the Q_1 - ensemble is valid for the Q_2 - ensemble. It indicates that the correction of observed annual river flows for the change in water carryover from year to year does not significantly affect the distribution of annual river flows (fig. 10).

It is interesting to note in figs. 7 through 10, left side graphs, that the frequency curves of $P(X^2)$ are approximately linear. The zig-zag relative class frequencies clearly oscillate around straight lines. These relative class frequencies increase linearly with an increase of $P(X^2)$ from zero to unity. Conclusions are that ensembles of station samples of annual precipitation and annual river flow have probabilities of chi-squares which are more frequent for greater values of probabilities than for smaller values. Fits of straight lines to relative frequencies of $P(X^2)$ mean that the cumulative frequency distributions of $P(X^2)$ are close to parabolas with various parameters.

This maximum deviation method of testing the distributions gives a relatively good and reliable result. Nevertheless, another test is used in order to confirm the above conclusions. The ensemble statistics are used for this purpose. Since the probability of sample chi-square is selected as the measure of deviation of a theoretical function from an observed one, it seems that the ensemble mean of these deviations is a good measure of goodness of fit. The smaller the value of the ensemble mean of probabilities of station sample chi-squares, the smaller is the total deviation between the two distributions and the better is the fitting. Additional statistics in the form of standard deviation, variance, coefficient of variation, skewness coefficient and excess, are used to describe the distribution of these deviations. The computation of statistics is done on a digital computer. The results are as follows:

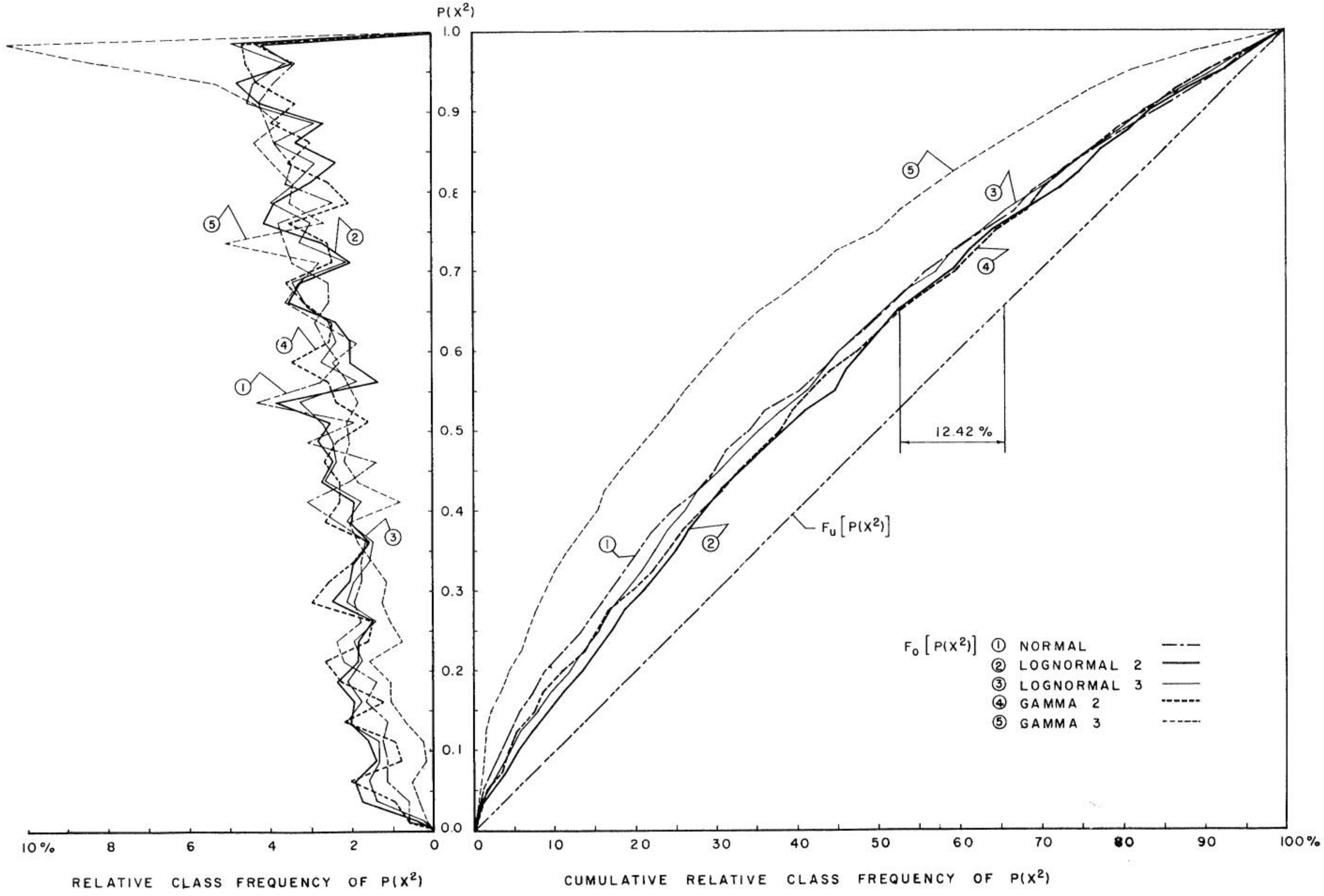

Fig. 7 P_1 - ensemble: frequency and cumulative distribution curves of probabilities of station sample chi-squares

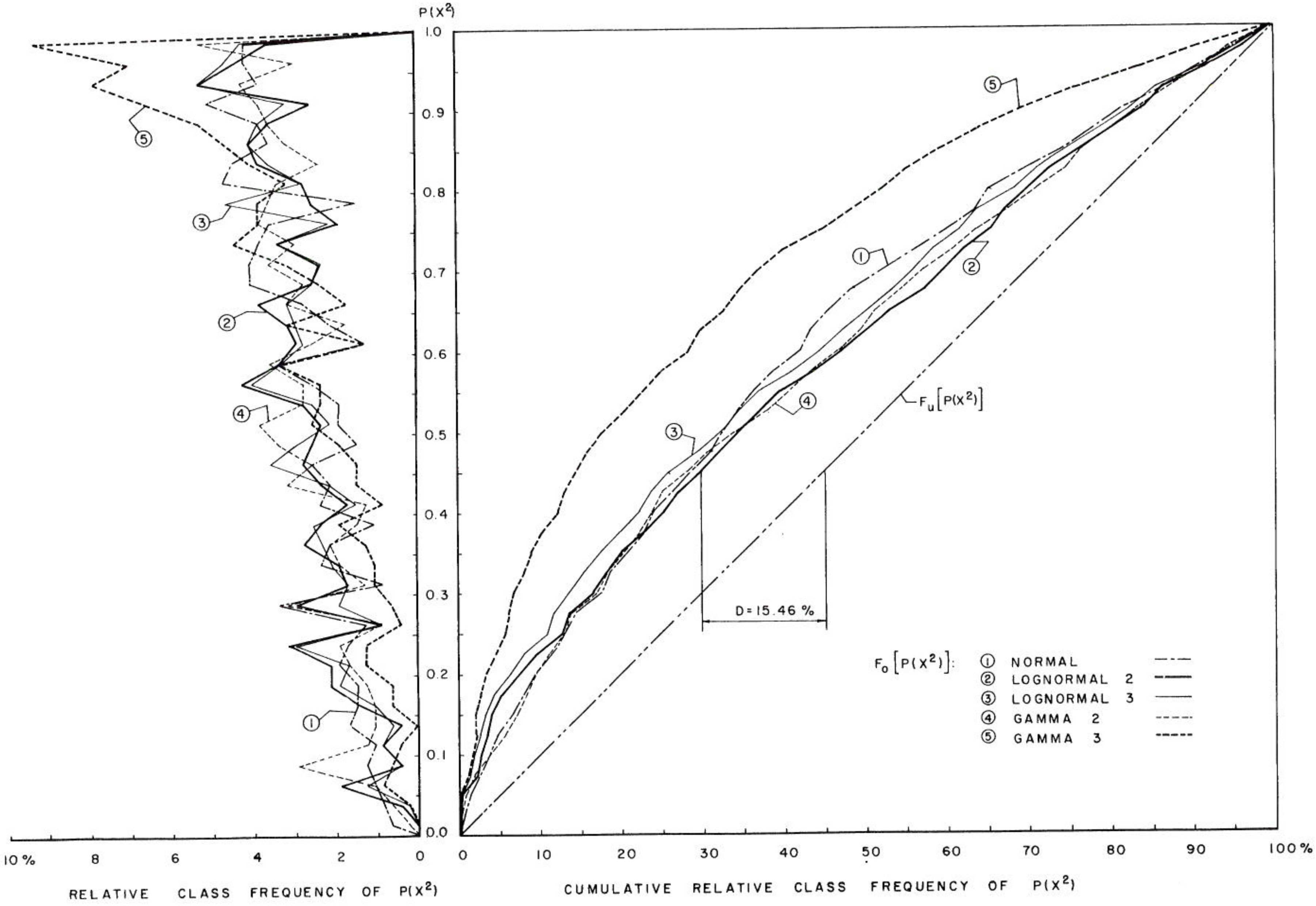

Fig. 8 P_2 - ensemble: frequency and cumulative distribution curves of probabilities of station sample chi-squares

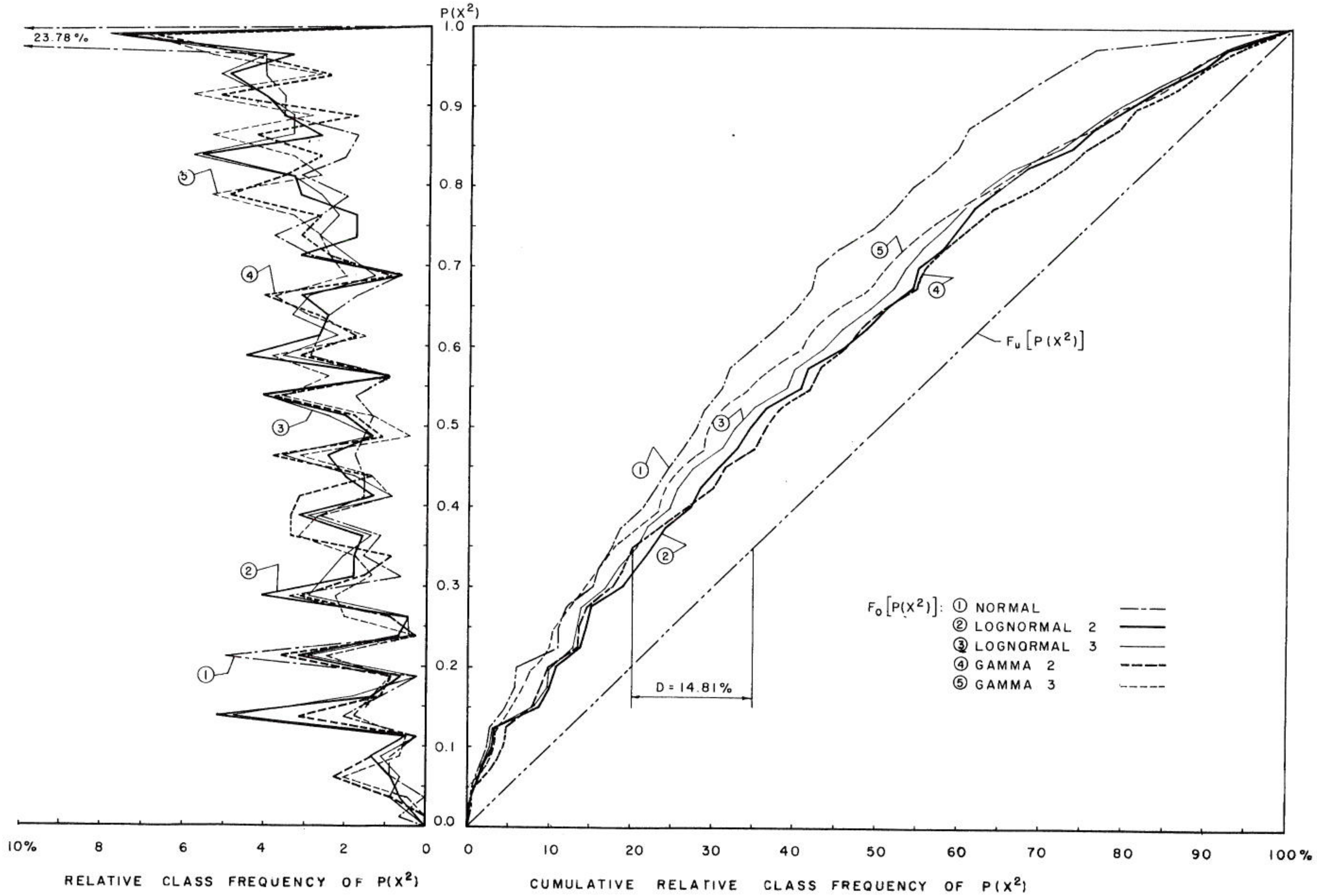

Fig. 9 Q_1 - ensemble: frequency and cumulative distribution curves of probabilities of station sample chi-squares

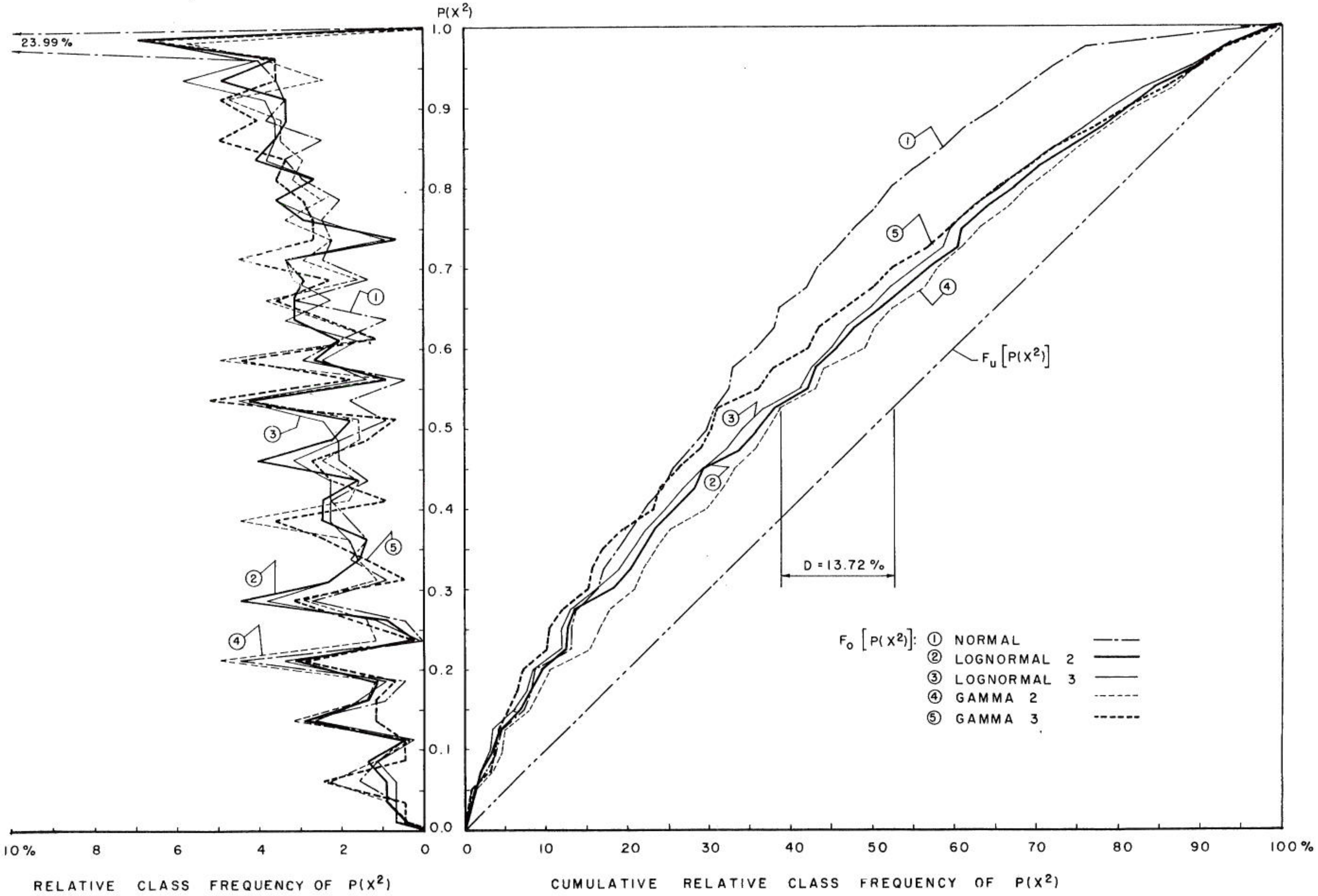

Fig. 10 Q_2 - ensemble: frequency and cumulative distribution curves of probabilities of station sample chi-squares

ENSEMBLE	FUNCTION	$\overline{P(X^2)}$	s	s^2	C_v	C_s	E
P_1	Normal	0.606	0.268	0.072	0.443	-0.370	-0.972
	Log-normal 2	0.582	0.282	0.080	0.485	-0.306	-1.094
	Log-normal 3	0.600	0.278	0.077	0.464	-0.376	-1.023
	Gamma 2	0.590	0.279	0.078	0.472	-0.292	-1.087
	Gamma 3	0.702	0.247	0.061	0.352	-0.737	-0.431
P_2	Normal	0.621	0.274	0.075	0.441	-0.487	-0.942
	Log-normal 2	0.604	0.261	0.068	0.432	-0.288	-1.001
	Log-normal 3	0.625	0.255	0.065	0.408	-0.363	-0.942
	Gamma 2	0.605	0.267	0.071	0.442	-0.373	-0.903
	Gamma 3	0.724	0.236	0.056	0.326	-0.912	-0.009
Q_1	Normal	0.688	0.290	0.084	0.421	-0.587	-0.944
	Log-normal 2	0.608	0.282	0.080	0.464	-0.334	-1.128
	Log-normal 3	0.624	0.279	0.078	0.447	-0.419	-1.018
	Gamma 2	0.605	0.277	0.077	0.458	-0.334	-1.043
	Gamma 3	0.641	0.267	0.071	0.417	-0.518	-0.844
Q_2	Normal	0.681	0.302	0.091	0.443	-0.623	-0.932
	Log-normal 2	0.608	0.276	0.076	0.453	-0.326	-1.050
	Log-normal 3	0.618	0.275	0.076	0.445	-0.353	-1.044
	Gamma 2	0.592	0.283	0.080	0.479	-0.279	-1.130
	Gamma 3	0.633	0.265	0.070	0.419	-0.496	-0.788

These results confirm the statements of previous tests. Namely, P_1 and P_2 - ensembles are best fitted by Log-normal 2, while Q_1 and Q_2 - ensembles are best fitted by the Gamma 2 function. These two functions have the smallest ensemble mean of probabilities of station sample chi-squares for the corresponding ensembles. It is evident from the above results that the differences in fitting observed distributions by Log-normal 2 and Gamma 2 in all four ensembles are very small. This difference could be neglected in most cases. Hence, all four ensembles of observed station sample distributions can be equally approximated either by Log-normal 2 or by Gamma 2 functions.

Considering the distribution of probabilities of station sample chi-squares, they are negatively skewed (negative skew coefficient) and generally flat (negative excess) for all four ensembles and all five functions studied.

3. Effect of various factors on probabilities of chi-square. The statistic - the probability of station sample chi-square - as used in this study is the exclusive measure of goodness of fit of a theoretical function to an observed distribution of annual precipitation or annual river flow. One may wish to know if there is any significant relationship between this statistic and some other factors of station sample characteristics and physiographic parameters. If there is such a relationship one might, a priori, infer some conclusions about observed distributions and give some indications about probability distribution functions of best fit. For this purpose, the P_1 and Q_1 - ensembles of station samples are studied. The eventual relationship can be expected to equal these for the P_2 and Q_2 - ensembles, and for the sake of brevity the P_2 and Q_2 - ensembles are omitted in this analysis.

The characteristic factors to be related to the probability of station sample chi-square, $P(X^2)$, of each of the five probability functions and each of the two considered ensembles, are as follows:

P_1 - ensemble (1141 station samples):

Average annual precipitation, or sample mean, $\overline{P}$ (in./yr.); Standard deviation, s (in./yr.); Coefficient of variation, C_v; and Skewness coefficient, C_s.

Q_1 - ensemble (446 station samples):

Drainage area, A (sq. mi.); Average annual river flow, or sample mean, $\overline{Q}$ (cfs); Average specific yield of river basin, $\overline{q}$ (cfs/sq. mi.); Standard deviation of annual flows, s (cfs); Coefficient of variation of annual flows C_v; and Skewness coefficient of annual flows, C_s.

The coefficient of correlation, r, was chosen as a measure of the linear association between $P(X^2)$, and any of the above factors. Linear correlation coefficients were computed with a digital computer. The results are as follows:

FACTORS	r for $P(X^2)$ of function				
	NORMAL	LOG-NORMAL 2	LOG-NORMAL 3	GAMMA 2	GAMMA 3
P_1 - ensemble					
$\overline{P}$	-0.052	-0.052	-0.039	0.032	0.037
s	-0.027	-0.070	-0.058	-0.048	-0.059
C_v	0.080	0.029	0.011	-0.151	-0.176
C_s	0.193	-0.058	-0.012	0.019	-0.068
Q_1 - ensemble					
A	0.059	0.075	0.062	0.045	0.031
$\overline{Q}$	-0.183	-0.055	0.060	-0.040	-0.057
$\overline{q}$	-0.003	-0.128	0.009	0.030	0.019
s	0.046	-0.031	0.049	0.006	0.000
C_v	0.474	0.032	0.042	0.159	0.108
C_s	0.502	-0.078	-0.029	0.132	0.110

According to these results, since all values of the linear correlation coefficients are small, slightly positive or negative, there is no significant relationship between the statistic $P(X^2)$ and any of the above factors. A somewhat higher value of r is expected for the normal function and the skewness coefficient in the case of the Q_1 - ensemble. Generally, a high value of C_s indicates a more skewed distribution, or further deviance from Normal. Hence, the higher the probability of chi-square the higher is the difference between Normal and observed distributions. Although the values of r are the highest for C_v and C_s of the Q_1 - ensemble and Normal function, all factors considered for all functions and both the P_1 and Q_1 - ensembles, there is no significant indication of any strong relationship between the statistic $P(X^2)$ and the various factors investigated.

CHAPTER VI

CONCLUSIONS

Five probability functions - Normal, Log-normal 2, Log-normal 3, Gamma 2 and Gamma 3 parameter functions - have been fitted to distributions of annual precipitation and annual runoff in the Western United States and the Southwestern Canada. The Chi-square test has been used to measure the goodness of fit of each function to each individually observed distribution of 2506 station samples involved in this investigation. These five functions have been then tested on all station samples which were grouped into four large ensembles: homogeneous precipitation (P_1), nonhomogeneous precipitation (P_2), river flows (Q_1), and river flows corrected for the change in carryover (Q_2). From the results of this study, obtained under criteria and conditions stated earlier in this report, the following conclusions can be drawn.

1. All five probability functions studied are applicable and none is more suitable than the other in fitting an observed individual station sample of annual precipitation or annual river flow distributions.

2. Probability functions described by two parameters have computational advantages in estimating parameters, and less time consuming in their use than those described by three parameters. Furthermore, they are more suitable for ensemble analysis than three parameter functions. This is due to the gain achieved by introducing a third parameter which is less than the loss caused by loosing one degree of freedom in the Chi-square test. When dealing with small sample sizes, small numbers of degrees of freedom, and large ensembles of station samples of annual precipitation and annual runoff, the three parameter functions can be omitted from consideration. This is particularly true in cases of large scale analysis over large regions or continents.

3. Distributions of homogeneous annual precipitation for the ensemble of 1141 station samples are best fitted by Log-normal 2 parameter function. This indicates that, on the average, the annual precipitations are positively skewed.

4. Distributions of nonhomogeneous annual precipitation for the ensemble of 473 station samples are also best fitted by Log-normal 2 parameter function. The nonhomogeneity in data introduces a decrease in the goodness of fit.

5. Distributions of annual river flows for the ensemble of 446 station samples are best fitted by the Gamma 2 parameter function. This indicates that, on the average, the distribution of annual river flows is positively skewed, but somewhat more than the annual precipitation.

6. Distributions of annual river flows corrected for the change in carryover, for the ensemble of 446 station samples, are also best fitted by the Gamma 2 parameter function. The correction for the change in carryover acts in the direction of smoothing the distribution of annual runoff and hence, resulting in slightly better goodness of fit in general than for the annual river flows.

7. Differences in goodness of fit in ensemble analyzes between Log-normal 2 and Gamma 2 functions are very small. For practical purposes they are negligible. Hence, in larger scale distribution analysis, these two functions are interchangeable.

8. No regional characteristic especially favors the use of one of these five probability distribution functions in fitting the observed distributions of either annual precipitation or annual river flows.

9. The use of probabilities of sample chi-squares as measures of goodness of fit of a probability distribution function to an observed distribution is more suitable than the sample chi-squares themselves. These probabilities provide for direct comparison of fitting functions with different degrees of freedom, when the Chi-square distribution is involved.

10. There is no significant linear association between the probability of station sample chi-squares of any function and any ensemble and the station sample means, standard deviations, coefficients of variation, and skewness coefficients. In addition to these drainage areas and average specific yields for river flow ensembles do not show any significant linear correlation with the probabilities of station sample chi-squares for the various functions investigated.

Acknowledgements: The writer acknowledges the support of the U. S. National Science Foundation in the research leading to this paper. Special appreciation is given to Dr. V. M. Yevdjevich, Professor-in-Charge of the Hydrology Program, for valuable suggestions and for reviewing this paper. Appreciation is also extended to Dr. M. M. Siddiqui, Professor of Statistics, for reviewing the statistical part of this paper.

BIBLIOGRAPHY

1. Aitchison, J., and Brown, J.˙A. C., The Lognormal Distribution. Cambridge, Cambridge University Press, 1963, 176 p.

2. Brownlee, K. A., Statistical Theory and Methodology in Science and Engineering, New York, John Wiley and Sons, Inc., 1960, 570 p.

3. Fisher, R. A., Contributions to Mathematical Statistics. New York, John Wiley and Sons, Inc., 1950.

4. Hald, A., Statistical Theory with Engineering Applications. New York, John Wiley and Sons, Inc., 1952, 783 p.

5. Kendall, M. G., The Advanced Theory of Statistics. London, Charles Griffin and Company, Limited, 1943, Vol. 1, 457 p.

6. Kendall, M. G., and Stuart, A., The Advanced Theory of Statistics. New York, Hafner Publishing Company, 1961, Vol. 2, 676 p.

7. Mann, H. B., and Wald, A., On the Choice of the Number of Class Intervals in The Application of the Chi Square Test. The Annals of Mathematical Statistics, 13:306-317, 1942

8. Markovic, R. D., Theoretical Frequency Functions of Best Fit to Distributions of Annual Precipitations and Mean Annual River Flows. Unpublished Master's Thesis, Colorado State University, Fort Collins, Colorado, 1964, 92 p.

9. Parzen, E., Modern Probability Theory and Its Applications. New York, John Wiley and Sons, Inc., 1960, 464 p.

10. Pearson, K., Tables of the Incomplete Gamma Function. Cambridge, The University Press, 1957, 164 p.

11. Thom, H. C. S., A Note on the Gamma Distribution. Weather Bureau, Monthly Weather Review, 86:117-122, 1958.

12. Yevdjevich, V. M., Fluctuations of Wet and Dry Years. Hydrology Papers, No. 1, Colorado State University, 1963.

APPENDIX

NUMERICAL EXAMPLE
WELDON RIVER AT MILL GROVE, MISSOURI, U. S. A.

1. Transformation of observed data into dimensionless form. The observed data for Weldon River at Mill Grove, Missouri, U. S. A., are tabulated at the end of this Appendix, Table 5. By using the sum in Column 3 the sample mean of the actual observed data is

$$\overline{Q} = \frac{1}{n}\sum_{i=1}^{n} Q_i = \frac{1}{31} \times 7957.7 = 256.7 \text{ cfs.}$$

With this, the observed annual flows are transformed into dimensionless form, in terms of the sample mean by eq. (2)

$$K_i = \frac{Q_i}{256.7}$$

and given in Column 4. In order to facilitate the further computation, the modular coefficients are arranged in an array, Column 5.

2. Maximum likelihood estimates.

(1) Normal function. Equation (11) with numerical data in Column 4, leads to

$$\hat{\mu} = \frac{1}{31} \times 31.000 = 1.000$$

and eq. (12) with Column 7

$$\hat{\sigma} = \sqrt{\frac{1}{31} \times 17.018} = 0.741.$$

(2) Log-normal 2. Applying eq. (13) and Column 8

$$\ln\hat{\mu} = \frac{1}{31}(-9.806) = -0.317$$

then eq. (14) and Column 10

$$\hat{\sigma} = \sqrt{\frac{1}{31} \times 21.837} = 0.840$$

(3) Log-normal 3. First the lower boundary parameter is estimated by iteration procedure according to eq. (17). For $K_o = 0.050$ columns 11, 12, 13, 14 and 15 are set up and the above equation checked as follows:

$$73.867\left\{\frac{1}{31} \times 32.492 - \left[\frac{1}{31}(-13.130)\right]^2 - \right.$$
$$\left. - \frac{1}{31}(-13.130)\right\} + (-95.407) = 0$$

$$95.510 - 95.407 \approx 0$$

Hence,

$$\hat{K}_o = 0.050.$$

According to eq. (15) and Column 13,

$$\ln\hat{\mu} = \frac{1}{31}(-13.130) = -0.424$$

then eq. (16) and Column 14

$$\hat{\sigma} = \sqrt{\frac{1}{31} \times 32.492 - (-0.424)^2} = 0.933$$

(4) Gamma 2. Equation (19) with Column 8 gives

$$\hat{\alpha} = \frac{1 + \sqrt{1 + \frac{4}{3}\left[0 - \frac{1}{31}(-9.810)\right]}}{4\left[0 - \frac{1}{31}(-9.810)\right]} - \Delta\hat{\alpha}$$

$$= 1.731 - 0.004 = 1.727$$

the correction factor $\Delta\hat{\alpha}$ being 0.004 for $\hat{\alpha} = 1.731$ according to Table 1. Then eq. (18) and Column 4 yield

$$\hat{\beta} = \frac{1}{1.727} \times \frac{1}{31} \times 31.000 = 0.579$$

(5) Gamma 3. First lower boundary or location parameter is computed by the iteration procedure in accordance with eq. (22) and Columns 16, 17 and 18:

$$\frac{1+\sqrt{1+\frac{4}{3}\,\{\ln\,[1.000-(-1.500)]-\frac{1}{31}\;27.124\}}}{1+\sqrt{1+\frac{4}{3}\{\ln[1.000-(-1.500)]-\frac{1}{n}\;27.124-4\{\ln[1.000-}}\quad *$$

$$\frac{*}{-(-1.500)]-\frac{1}{31}\;27.124\}}-[1.000-(-1.500)] \times \frac{1}{31} \times$$

$$\times 13.418 = 0$$

$$1.087 - 1.083 \approx 0$$

Hence,

$$\hat{K}_o = -1.500$$

Since the lower boundary is negative it should be replaced by zero, and this function reduces to Gamma 2. However, in this particular example the obtained negative value is carried throughout in order to show the computational procedure.

By eq. (20) and Column 18

$$\hat{\alpha} = \frac{1+\sqrt{1+\frac{4}{3}\,\{\ln\,[1.000-(-1.500)]\;-\frac{1}{31}\;27.124\}}}{4\{\ln[1.000-(-1.500)]-\frac{1}{31}\;27.124\}} - \Delta\hat{\alpha}$$

$$= 12.360 - 0.000 = 12.360$$

the correction factor $\Delta \hat{\alpha}$ being zero according to Table 1.

Using eq. (21) and the numerical values from Column 16

$$\hat{\beta} = \frac{1}{12.360} \times \frac{1}{31} \times 77.500 = 0.202$$

3. Class interval limits and observed class frequencies.

(1) Normal. For seven class intervals six class interval limits are computed by eq. (28) and Table 2, and observed class frequencies, O_j, determined and squared as follows:

	O_j	O_j^2
	4	16
$K_1 = 1.000 - 1.068 \times 0.741 = 0.209$		
	10	100
$K_2 = 1.000 - 0.566 \times 0.741 = 0.581$		
	3	9
$K_3 = 1.000 - 0.180 \times 0.741 = 0.867$		
	3	9
$K_4 = 1.000 + 0.180 \times 0.741 = 1.133$		
	1	1
$K_5 = 1.000 + 0.566 \times 0.741 = 1.419$		
	4	16
$K_6 = 1.000 + 1.068 \times 0.741 = 1.791$		
	6	36
	31	187

(2) Log-normal 2. According to eq. (29) and parameter estimates previously computed, the class interval limits are:

	O_j	O_j^2
	5	25
$K_1 = \exp[-0.317 - 1.068 \times 0.840] = 0.297$		
	5	25
$K_2 = \exp[-0.317 - 0.566 \times 0.840] = 0.453$		
	5	25
$K_3 = \exp[-0.317 - 0.180 \times 0.840] = 0.624$		
	2	4
$K_4 = \exp[-0.317 + 0.180 \times 0.840] = 0.847$		
	4	16
$K_5 = \exp[-0.317 + 0.566 \times 0.840] = 1.171$		
	4	16
$K_6 = \exp[-0.317 + 1.068 \times 0.840] = 1.786$		
	6	36
	31	147

(3) Log-normal 3. By using eq. (30) and parameters estimated earlier, the class interval limits are:

	O_j	O_j^2
	5	25
$K_1 = 0.050 + \exp[-0.424 - 1.068 \times 0.933] = 0.291$		
	4	16
$K_2 = 0.050 + \exp[-0.424 - 0.566 \times 0.933] = 0.436$		
	6	36
$K_3 = 0.050 + \exp[-0.424 - 0.180 \times 0.933] = 0.605$		
	1	1
$K_4 = 0.050 + \exp[-0.424 + 0.180 \times 0.933] = 0.824$		
$K_5 = 0.050 + \exp[-0.424 + 0.566 \times 0.933] = 1.159$	4	16
	5	25
$K_6 = 0.050 + \exp[-0.424 + 1.068 \times 0.933] = 1.819$		
	6	36
	31	155

(4) Gamma 2. Equation (35) with the corresponding values of u_j from Table 3, gives:

	O_j	O_j^2
	5	25
$K_1 = \frac{1}{\sqrt{1.727}} \times 4.999 = 0.761 \times 0.384 = 0.292$		
	6	36
$K_2 = 0.761 \times 0.644 = 0.490$		
	4	16
$K_3 = 0.761 \times 0.916 = 0.697$		
	2	4
$K_4 = 0.761 \times 1.234 = 0.939$		
	4	16
$K_5 = 0.761 \times 1.649 = 1.255$		
	4	16
$K_6 = 0.761 \times 2.314 = 1.761$		
	6	36
	31	149

(5) Gamma 3. Solving eq. (36) with the corresponding values of u_j selected for given value of $\hat{\alpha}$ from Table 3, the class interval limits are obtained, the observed class frequencies are determined and squared:

	O_j	O_j^2
	4	16
$K_1 = -1.500 + \frac{1.000 - (-1.500)}{\sqrt{12.360}} \times 2.470 = -1.500 + 0.711 \times 2.470 = 0.256$		
	9	81
$K_2 = -1.500 + 0.711 \times 2.888 = 0.553$		
	2	4
$K_3 = -1.500 + 0.711 \times 3.241 = 0.804$		
	5	25
$K_4 = -1.500 + 0.711 \times 3.612 = 1.068$		
	1	1
$K_5 = -1.500 + 0.711 \times 4.003 = 1.348$		
	4	16
$K_6 = -1.500 + 0.711 \times 4.578 = 1.755$		
	6	36
	31	179

4. Computation of sample chi-squares. The sample chi-squares are computed by eq. (25) for each selected function separately and then converted into corresponding probability by means of Table 4, and fig. 4.

(1) Normal, f = 4 degrees of freedom

$$X^2 = \frac{7}{31} \times 187 - 31 = 11.267 \qquad P(X^2) = 0.976$$

(2) Log-normal 2, f = 4 d.f.

$$X^2 = \frac{7}{31} \times 147 - 31 = 2.222 \qquad P(X^2) = 0.307$$

(3) Log-normal 3, f = 3 d.f.

$$X^2 = \frac{7}{31} \times 155 - 31 = 4.030 \qquad P(X^2) = 0.742$$

(4) Gamma 2, f = 4 d. f.

$$X^2 = \frac{7}{31} \times 149 - 31 = 2.674 \qquad P(X^2) = 0.388$$

(5) Gamma 3, f = 3 d. f.

$$X^2 = \frac{7}{31} \times 179 - 31 = 9.454 \qquad P(X^2) = 0.976$$

5. Analysis of results. Considering this station separately, only Log-normal 2, Log-normal 3, and Gamma 2 are applicable, since each of them has the probability of chi-square less than commonly used level of significance 0.95. Hence, the statistical tests for these three functions are nonsignificant, and for Normal and Gamma 3 they are significant. However, since the smaller probability of chi-square means the better fitting to observed data, it turns out that the Log-normal 2 with the smallest probability of chi-square is of best fit to annual observations at Weldon River at Mill Grove, Missouri, USA. The characteristic histograms of annual river flows at this station, including the discrete time series, the cumulative frequency, the observed and the expected frequency histograms for Log-normal 2, are graphed in fig. 11. It is interesting to note that in the frequency analysis dealing with the class intervals of equal probabilities, the observed frequency curve is transformed to a histogram, and the expected frequency curve to a rectangular. Therefore, the comparison of an observed distribution with a continuous theoretical distribution reduces to the comparison of an observed histogram with a theoretical uniform distribution. This is well illustrated in fig. 11, histograms 3 and 4.

In this particular numerical example, attention is called upon Gamma 3, where the lower case was presented here only to show the computational procedure, otherwise the lower boundary would be zero, and Gamma 3 would reduce to Gamma 2.

Further analysis of this station is done in grouped form for all samples together in the Q_1 - ensemble as shown in Chapter V.

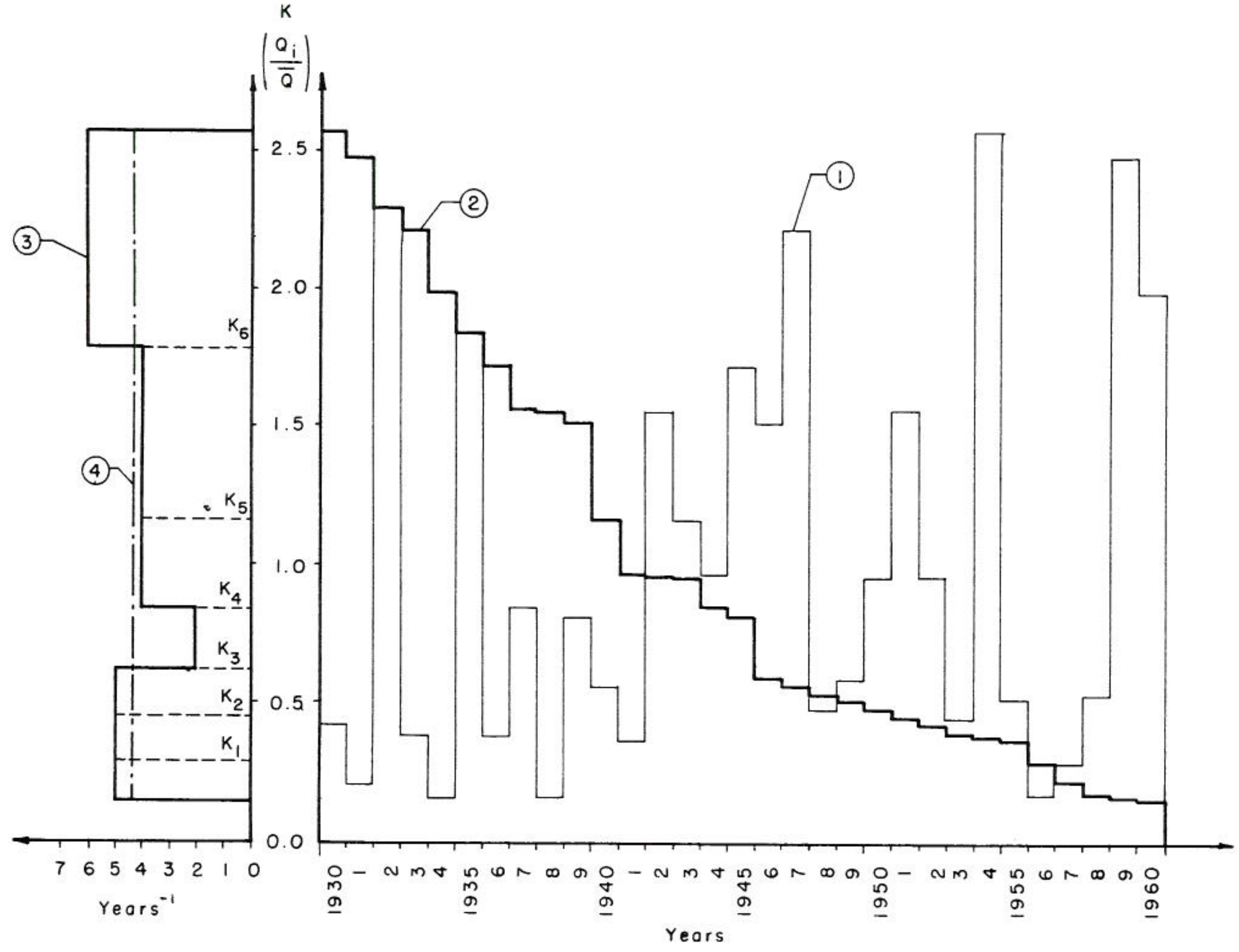

Fig. 11 Characteristic histograms of annual river flows at Weldon River at Mill Grove, Missouri, USA: (1) Discrete time series; (2) Duration or cumulative frequency; (3) Observed frequency; and (4) Expected frequency.

TABLE 5
DATA FOR WELDON RIVER AT MILL GROVE, MISSOURI
Station Sample of Q_1-Ensemble

Order No.	Year of Obser-vation	Annual River Flow Q_i (cfs)	$K_i = \frac{Q_i}{\bar{Q}}$	K_i (in array)	$K_i - \hat{\mu}$ $\hat{\mu} = 1.000$	$(K_i - \hat{\mu})^2$	$\ln K_i$	$\ln K_i - \ln\hat{\mu}$ $\ln\hat{\mu} = -0.317$	$(\ln K_i - \ln\hat{\mu})^2$	$K_i - \hat{K}_o$ $\hat{K}_o = 0.050$	$\frac{1}{K_i - \hat{K}_o}$	$\ln(K_i - \hat{K}_o)$	$\ln^2(K_i - \hat{K}_o)$	$\frac{\ln(K_i - \hat{K}_o)}{K_i - \hat{K}_o}$	$K_i - \hat{K}_o$ $\hat{K}_o = -1.500$	$\frac{1}{K_i - \hat{K}_o}$	$\ln(K_i - \hat{K}_o)$
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1930	108.0	0.421	2.567	1.567	2.455	0.942	1.259	1.585	2.517	0.397	0.923	0.852	0.367	4.067	0.246	1.402
2	1	53.6	0.209	2.474	1.474	2.173	0.905	1.222	1.493	2.424	0.413	0.885	.783	.365	3.974	.252	1.379
3	2	585.0	2.279	2.279	1.279	1.636	0.823	1.140	1.300	2.229	.449	.801	.642	.359	3.779	.265	1.329
4	3	98.1	0.382	2.209	1.209	1.462	0.792	1.109	1.230	2.159	.463	.769	.591	.356	3.709	.270	1.311
5	4	40.6	0.158	1.979	0.979	0.958	0.683	1.000	1.000	1.929	.518	.657	.432	.341	3.479	.287	1.247
6	1935	472.0	1.839	1.839	.839	.704	0.609	0.926	0.857	1.789	.559	.582	.339	.325	3.339	.299	1.205
7	6	96.5	0.376	1.718	.718	.516	0.542	.859	.738	1.668	.600	.511	.261	.306	3.218	.311	1.169
8	7	217.0	0.845	1.558	.558	.311	0.444	.761	.579	1.508	.663	.411	.169	.273	3.058	.327	1.118
9	8	42.7	0.166	1.551	.551	.304	0.438	.755	.570	1.501	.666	.406	.165	.270	3.051	.328	1.115
10	9	208.0	0.810	1.504	.504	.254	0.407	.724	.524	1.454	.688	.374	.140	.188	3.004	.333	1.100
11	1940	143.0	0.557	1.161	.161	.026	0.149	.466	.217	1.111	.900	.105	.011	.095	2.661	.376	0.978
12	1	93.7	0.365	0.966	- .034	.001	-0.035	.282	.080	0.916	1.092	- .088	.008	- .096	2.466	.406	.903
13	2	398.0	1.551	.955	- .045	.002	-0.046	.271	.073	.905	1.105	- .100	.010	- .110	2.455	.407	.898
14	3	298.0	1.161	.951	- .049	.002	-0.051	.266	.071	.901	1.110	- .104	.011	- .115	2.451	.408	.896
15	4	248.0	0.966	.845	- .155	.024	- .168	.149	.022	.795	1.258	- .229	.052	- .288	2.345	.426	.852
16	1945	441.0	1.718	.810	- .190	.036	- .211	.106	.011	.760	1.316	- .274	.075	- .361	2.310	.433	.837
17	6	386.0	1.504	.588	- .412	.170	- .531	- .214	.046	.538	1.859	- .620	.384	- 1.152	2.088	.479	.736
18	7	567.0	2.209	.557	- .443	.196	- .585	- .268	.072	.507	1.972	- .679	.461	- 1.339	2.057	.486	.721
19	8	122.0	0.475	.526	- .474	.224	- .643	- .326	.106	.476	2.101	- .742	.551	- 1.559	2.026	.494	.706
20	9	151.0	0.588	.514	- .486	.236	- .666	- .349	.122	.464	2.155	- .768	.590	- 1.655	2.014	.497	.700
21	1950	244.0	0.951	.475	- .525	.276	- .744	- .427	.182	.425	2.353	- .856	.733	- 2.014	1.975	.506	.680
22	1	400.0	1.558	.444	- .556	.309	- .812	- .495	.245	.394	2.538	- .931	.867	- 2.363	1.944	.514	.665
23	2	245.0	0.955	.421	- .579	.335	- .865	- .548	.300	.371	2.695	- .992	.984	- 2.674	1.921	.521	.652
24	3	114.0	0.444	.382	- .618	.382	- .962	- .645	.416	.332	3.012	-1.103	1.217	- 3.322	1.882	.531	.632
25	4	659.0	2.567	.376	- .624	.389	- .978	- .661	.437	.326	3.067	-1.121	1.257	- 3.439	1.876	.533	.628
26	1955	132.0	0.514	.365	- .635	.403	-1.008	- .691	.477	.315	3.175	-1.155	1.334	- 3.667	1.865	.536	.623
27	6	44.0	0.171	.282	- .718	.516	-1.266	- .949	.901	.232	4.310	-1.461	2.135	- 6.297	1.782	.561	.578
28	7	72.5	0.282	.209	- .791	.626	-1.565	-1.248	1.558	.159	6.289	-1.839	3.382	-11.566	1.709	.585	.536
29	8	135.0	0.526	.171	- .829	.687	-1.766	-1.449	2.100	.121	8.264	-2.112	4.461	-17.455	1.671	.598	.513
30	9	635.0	2.474	.166	- .834	.696	-1.796	-1.479	2.187	.116	8.621	-2.154	4.640	-18.569	1.666	.600	.510
31	1960	508.0	1.979	.158	- .842	.709	-1.846	-1.529	2.338	.108	9.259	-2.226	4.955	-20.611	1.658	.603	.505
31	$\sum_{1}^{31}$	7957.7	31.000	31.000		17.018	-9.810		21.837	29.450	73.867	-13.130	32.492	-95.407	77.500	13.418	27.124

Key Words: Hydrology, Annual Precipitation, Annual River Flows, Frequency, Probability Functions, Parameter Estimation, Fitting, Chi-square Test.

Abstract: Five probability functions - Normal, Log-normal with 2, Log-normal with 3, Gamma with 2 and Gamma with 3 parameters - are fitted to the distributions of annual precipitation and annual river flows on 2506 station samples in the Western United States and Southwestern Canada. The Maximum likelihood method has been used for the estimation of functions parameters from observed data and the probability of chi-square as a measure of goodness of fit of each function to every observed sample distribution. It has been found that all five functions studied are applicable in individually fitting an observed distribution. However, if station samples are grouped in four large ensembles - homogeneous and nonhomogeneous precipitation, river flow, and river flow corrected for carryover - the Log-normal with 2 and the Gamma function with 2 parameters best fit the observed distributions, both precipitation and river flow. For practical purposes these two functions can be used interchangeably for all four ensembles.

Reference: Markovic, Radmilo D., Colorado State University, Hydrology Papers No. 8 (August 1965) "Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff."

Key Words: Hydrology, Annual Precipitation, Annual River Flows, Frequency, Probability Functions, Parameter Estimation, Fitting, Chi-square Test.

Abstract: Five probability functions - Normal, Log-normal with 2, Log-normal with 3, Gamma with 2 and Gamma with 3 parameters - are fitted to the distributions of annual precipitation and annual river flows on 2506 station samples in the Western United States and Southwestern Canada. The Maximum likelihood method has been used for the estimation of functions parameters from observed data and the probability of chi-square as a measure of goodness of fit of each function to every observed sample distribution. It has been found that all five functions studied are applicable in individually fitting an observed distribution. However, if station samples are grouped in four large ensembles - homogeneous and nonhomogeneous precipitation, river flow, and river flow corrected for carryover - the Log-normal with 2 and the Gamma function with 2 parameters best fit the observed distributions, both precipitation and river flow. For practical purposes these two functions can be used interchangeably for all four ensembles.

Reference: Markovic, Radmilo D., Colorado State University, Hydrology Papers No. 8 (August 1965) "Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff."

Key Words: Hydrology, Annual Precipitation, Annual River Flows, Frequency, Probability Functions, Parameter Estimation, Fitting, Chi-square Test.

Abstract: Five probability functions - Normal, Log-normal with 2, Log-normal with 3, Gamma with 2 and Gamma with 3 parameters - are fitted to the distributions of annual precipitation and annual river flows on 2506 station samples in the Western United States and Southwestern Canada. The Maximum likelihood method has been used for the estimation of functions parameters from observed data and the probability of chi-square as a measure of goodness of fit of each function to every observed sample distribution. It has been found that all five functions studied are applicable in individually fitting an observed distribution. However, if station samples are grouped in four large ensembles - homogeneous and nonhomogeneous precipitation, river flow, and river flow corrected for carryover - the Log-normal with 2 and the Gamma function with 2 parameters best fit the observed distributions, both precipitation and river flow. For practical purposes these two functions can be used interchangeably for all four ensembles.

Reference: Markovic, Radmilo D., Colorado State University, Hydrology Papers No. 8 (August 1965) "Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff."

Key Words: Hydrology, Annual Precipitation, Annual River Flows, Frequency, Probability Functions, Parameter Estimation, Fitting, Chi-square Test.

Abstract: Five probability functions - Normal, Log-normal with 2, Log-normal with 3, Gamma with 2 and Gamma with 3 parameters - are fitted to the distributions of annual precipitation and annual river flows on 2506 station samples in the Western United States and Southwestern Canada. The Maximum likelihood method has been used for the estimation of functions parameters from observed data and the probability of chi-square as a measure of goodness of fit of each function to every observed sample distribution. It has been found that all five functions studied are applicable in individually fitting an observed distribution. However, if station samples are grouped in four large ensembles - homogeneous and nonhomogeneous precipitation, river flow, and river flow corrected for carryover - the Log-normal with 2 and the Gamma function with 2 parameters best fit the observed distributions, both precipitation and river flow. For practical purposes these two functions can be used interchangeably for all four ensembles.

Reference: Markovic, Radmilo D., Colorado State University, Hydrology Papers No. 8 (August 1965) "Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff."