Environment Canada Environnement Canada

CFA

CONSOLIDATED FREQUENCY ANALYSIS
VERSION 3.1

REFERENCE MANUAL

SURVEYS AND INFORMATION SYSTEMS BRANCH
RELEVÉS ET SYSTÈMES D'INFORMATION

ECOSYSTEM SCIENCES AND EVALUATION DIRECTORATE
DIRECTION GÉNÉRALE DES SCIENCES ET DE L'ÉVALUATION DES ÉCOSYSTÈMES

CONSOLIDATED FREQUENCY ANALYSIS (CFA)

Version 3.1

REFERENCE MANUAL

by

Paul J. Pilon and K. David Harvey

INTERPRETATION AND APPLICATIONS DIVISION
SURVEYS AND INFORMATION SYSTEMS BRANCH
ECOSYSTEM SCIENCES AND EVALUATION DIRECTORATE
ENVIRONMENT CANADA
OTTAWA, ONTARIO
K1A 0H3

March 1993

DISCLAIMER

This program has been tested, but exhaustive testing is naturally impossible. Hence, the Interpretation and Access Division makes no warranty, expressed or implied, as to the performance of this program. The users of the program are expected to make the final evaluation as to the usefulness and correctness of the program in their own set of circumstances.

It is recognized that computers and other electronic aids are useful and necessary parts of modern engineering practice. However, the use of such tools does not relieve the engineer from the requirement to provide a safe and adequate design. Ultimately, the engineer is responsible for the correct working of the software and hardware used and of the design that results.

1. ABSTRACT

Program Language: FORTRAN 77

Computer: IBM PC or compatible with a minimum 640K ram, math co-processor and graphics card.

Available From: Interpretation and Access Division, Surveys and Information Systems Branch, Environment Canada

Purpose/Technique: To provide a user-friendly interactive package to:

1) enter or import flow data into the program;
2) list contents of stored data sets;
3) edit existing data sets;
4) save entered or edited data by placing it into permanent storage;
5) perform nonparametric tests, identify low and high outliers and plot data;
6) compute the flood frequency regime using one or more frequency distributions and/or nonparametric methods.

Input: The sample series of annual flood data is input to the program via the keyboard or importing of files. Historic information may, as well, be entered via the keyboard.

Output: The monitor is the usual mode of display. Hard copy may be obtained using a variety of printers.

Output includes:

i) the ranked input series of annual floods, with high and low outliers denoted, and empirical probabilities.
ii) Estimates of population statistics and distribution parameters.
iii) Flows for preselected return periods.
iv) Plots of the frequency curves and other displays of data.

Comments: The number of observations in a data set cannot exceed 150.

1. RÉSUMÉ

Language du programme: FORTRAN 77

Ordinateur: IBM PC ou compatible avec 640K RAM minimum, co-processeur mathématique et carte graphique

Fournisseur: Interprétation et Accès
Relevés et Systèmes d'information
Environnement Canada

But/Technique: Fournir un progiciel interactif à la portée de l'utilisateur pour:

1) l'entrée de données relatives à l'écoulement en période de crue.
2) le listage de fichiers de données enregistrés.
3) la révision de fichiers existants
4) la préservation de fichiers entrés ou révisés en les introduisant dans la mémoire permanente.
5) la réalisation de tests non paramétriques, l'identification de données aberrantes en deçà et au delà de la moyenne et le traçage de courbes.
6) la calcul de la fréquence des crues en utilisant une ou plusieurs distribution(s) de fréquences.

Entrée: La série d'événements hydrologiques annuels peut être introduite par clavier ou par transfert des fichiers de données enregistrés dans la mémoire de grande capacité. Les renseignements historiques peuvent aussi être introduits par clavier.

Sortie: L'écran cathodique est le mode usuel d'affichage. Les sorties d'ordinateur peuvent être obtenues en utilisant une des plusieurs imprimantes compatibles.

Les sorties comprennent:

i) les données d'entrée concernant les crues annuelles parmi lesquelles sont indiquées les données aberrantes en deçà et au delà de la moyenne et les probabilités empiriques.
ii) des estimations faites à partir de l'échantillon ainsi que des paramètres de distribution.
iii) les débits pour des périodes de récurrence déterminées en avance.
iv) des tracés facultatifs des courbes de fréquence et d'autres affichages des données.

Observations: Le fichier de données ne peut contenir plus de 150 observations.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

TABLE **Page**

LIST OF TABLES (Continued)

Page

LIST OF TABLES (Continued)

LIST OF FIGURES

Page

LIST OF FIGURES (Continued)

Page

LIST OF FIGURES (Continued)

Page

2. INTRODUCTION

In 1976, the governments of Canada and the Provinces embarked on a program of flood damage reduction emphasizing flood plain mapping. It was envisaged that such an effort would be of benefit to current and future generations of Canadians through the alleviation of unnecessary hardship and sacrifice. To this end, Environment Canada released guidelines entitled "Hydrologic and Hydraulic Procedures for Flood Plain Delineation" and made available documented computer programs for performing single-site flood frequency analyses.

Since 1976, Environment Canada has made several additional computer programs available for: 1) performing nonparametric tests for homogeneity, trend, independence and randomness, (Shiau and Condie, 1980); 2) detecting low and high outliers (Pilon and Condie, 1982); and 3) handling the "problem" sample containing zero flows, low outliers, and/or historic information (Condie and Harvey, 1984).

The technique of flood frequency analysts has progressed and our capabilities in analyzing and interpreting flood data have increased. The culmination of these advances is represented in the user-friendly interactive CFA computer package documented herein. This package allows the user to: easily enter data for analysis from standard Water Survey of Canada flood data files (Inland Waters Directorate, 1980); modify data sets so as to add, delete, or modify all stored information; save modified data sets for future use; perform nonparametric tests for homogeneity, trend, independence and randomness; perform tests for low and high outliers; determine T-year events for the straightforward case, samples with historic information, samples with low outliers, samples with zero flow values, and combinations of historic information, low outliers, and zero flow values. The probability distributions incorporating these enhancements include:

1) the generalized extreme value;
2) the three-parameter lognormal;
3) the Log Pearson Type III; and
4) the Wakeby.

The Weibull distribution also has all the mentioned enhancements, save the capability of handling historic information.

CFA has been designed to handle not only the standard "no problem" sample, but also to handle all combinations of highs, lows, and zeros. Eight combinations are possible:

a) the standard case;
b) historic highs;
c) historic highs and low outliers;
d) historic highs, low outliers, and zeros;
e) historic highs and zeros;
f) low outliers;
g) low outliers and zeros; and
h) zeros.

All cases except (d) and (e) have been found in hydrometric records from Canadian rivers. The package's components have been tested many times using real and simulated data and have proven reliable. Testing of cases (d) and (e) was done by adding false zeros to real records that contained historic highs and identified low outliers, or to real records that contained historic highs only.

This version of CFA now supersedes programs previously released by this office. However, it is emphasized that any previous analyses made using earlier programs are not necessarily invalidated.

The authors would appreciate hearing from users who have any difficulties, find any errors, or have suggestions for improving the package. If you are quite happy with the package, we would be pleased to hear from you too!

This version of CFA was designed to run on an IBM PC or compatible with at least 640K RAM, a math co-processor, and a graphics card. A hard disk is highly recommended. CFA is written in Microsoft Fortran 5.0. Graphics output is generated with HALO Professional Graphics.

3. TECHNICAL BACKGROUND

For clarity, this user manual is principally descriptive with a minimum of mathematical content. All the necessary equations and solution algorithms have been grouped in the appendices.

3.1 CHARACTERISTICS OF THE DISTRIBUTIONS

This section gives a brief description of the distributions incorporated in the package. These are the generalized extreme value, the three-parameter lognormal, the Log Pearson Type III, the Wakeby, and the Weibull distributions.

3.1.1 The Generalized Extreme Value (GEV) Distribution

The extreme value (EV) distributions figure prominently in hydrologic literature. They represent attempts to deduce, on purely theoretical grounds, how annual maximum floods are distributed. The GEV family can be divided into three classes, corresponding to different ranges of the shape parameter, k. The three classes are referred to as Fisher-Tippett Type 1, Type 2, and Type 3 (Fisher and Tippett, 1928). They are also sometimes referred to as EV1, EV2, and EV3. In practice, k values lie in the range of -0.6 to + 0.6. The range is divided among the three classes as follows: a negative k corresponds to Type 2, a k of zero corresponds to Type 1, and a positive k corresponds to Type 3. Figure 1 shows how these variates are related to each other.

The GEV is a three parameter distribution. Parameters μ and α represent location and scale, respectively. If k is equal to zero, then the distribution appears as a two parameter EV1. If k is negative, the lower bound of the EV2 is $(\mu + \alpha/k)$. A positive k value implies an EV3 which is upper bounded at $(\mu + \alpha/k)$.

Figure 2 shows the relationship of the skewness of the sample with the shape parameter k. When the skewness is 1.14, the corresponding shape parameter is zero — the EV1 distribution. The theoretical kurtosis of the EV1 is 5.4. Samples having a skewness greater than 1.14 are EV2, while a skewness less than 1.14 infers an upper bounded EV3.

Figure 3 shows the relationship for the third L-moment ratio (skewness) and the shape parameter k. When the third L-moment ratio is .1699, the corresponding shape parameter is zero. The fourth L-moment ratio (kurtosis) for the EV1 distribution is .1504. L-moment ratios greater than .1699 infer an EV2 distribution, ratios less than .1699 infer an EV3 distribution.

The EV1 distribution is also known as the Gumbel I or the double exponential distribution. If a variate x has an EV3 distribution, then -x is said to have a Weibull distribution.

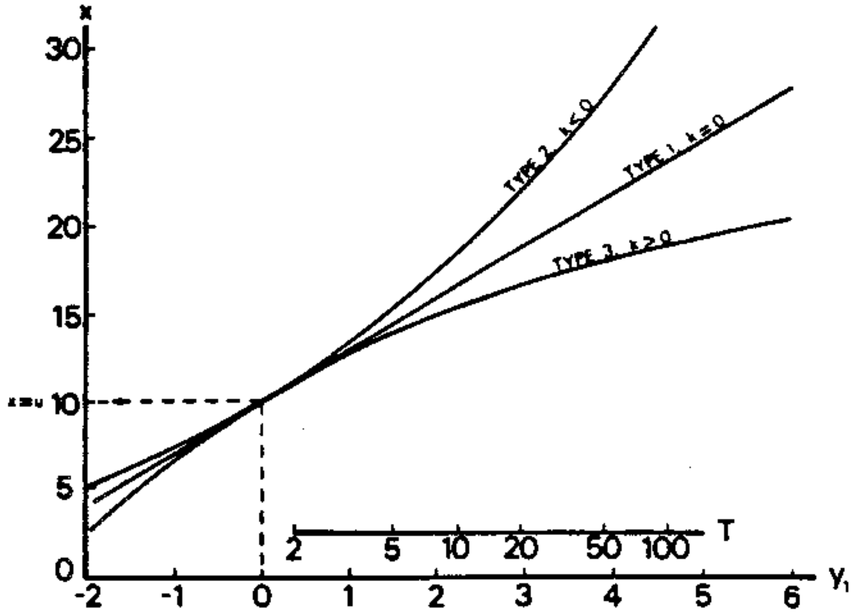

Figure 1: The three types of extreme value variate shown as functions of the Type 1 reduced variate, y, by the relation

$$x = \mu + \alpha(1 - e^{ky_1})/k.$$

(Natural Environment Research Council, 1975)

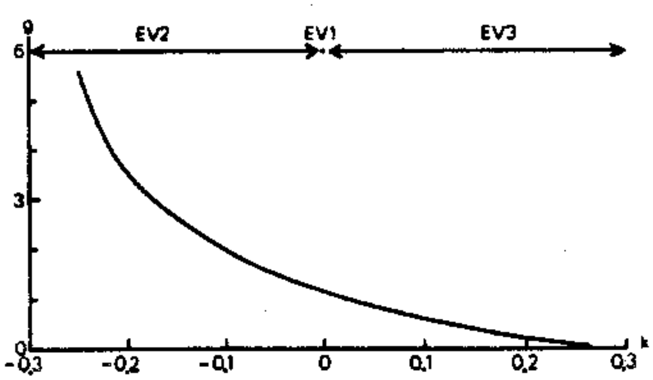

Figure 2: Skewness, g, of extreme value variates as a function of the shape parameter, k (Natural Environment Research Council, 1975).

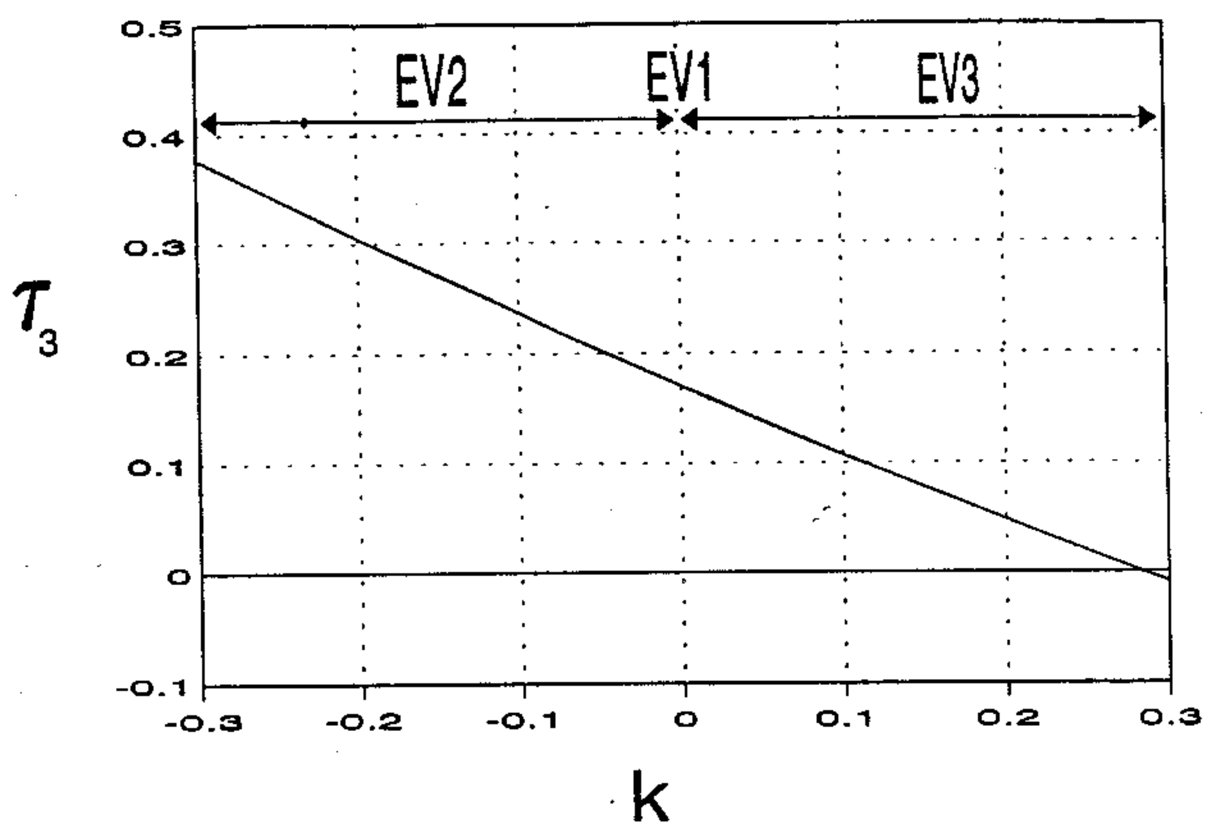

Figure 3: L-Moment Ratio, τ_3, as a function of the shape parameter k for the generalized extreme value distribution.

The EV distributions have an interesting property. Suppose that x is an EV variate, and let Z be the maximum (or extreme) value of a random sample of size N from this distribution. Because the sample is random, Z is also random with the special property that the distribution of Z has the same form as that of x although the mean and possibly the standard deviation of Z differ from those of x. This is the major characteristic of the EV distributions, namely that maxima of random samples drawn from such populations have the same form of distribution as the parent.

The utility of a GEV distribution is demonstrated if an extreme value distribution is wanted, but the type is unknown. When sample data are available, the GEV's distribution parameters μ, α, and k may be obtained by several different methods. If no historic information is present, then the method of L-moments (Hosking 1988, 1989, 1990) is used to estimate the parameters. The value of k indicates to which of the extreme value distributions the sample belongs. If historic information is present, then historically weighted moments are used. Appendix B gives a more detailed account of the fitting procedures and the distribution.

3.1.2 The Three-Parameter Lognormal Distribution

Hydrologic events can seldom be described by the normal distribution since these events are more commonly skewed. Several theoretical transformations have been developed for normalizing a skewed distribution, because it is generally easier to draw statistical conclusions from the normal distribution whose theoretical properties are well known. In flood frequency analysis it is generally necessary to extrapolate far beyond the range of observed events and the idea of a normalizing transformation is very attractive. One such transformation leads to the lognormal distribution. Essentially, use of this distribution implies that the logarithms of the data set are normally distributed. By virtue of the logarithmic transform, the theoretical coefficients of skewness and kurtosis of the transformed data should be 0.0 and 3.0, respectively. These theoretical characteristics give a handy subjective assessment of how well the lognormal distribution fits the data sample of annual foods.

Experience with Canadian rivers shows that the logarithmic transform generally results in overdoing the normalizing process. Basic data that are positively skewed sometimes show a substantially negative skew coefficient of the logarithmic transform.

It has been found that the inclusion of a third parameter will greatly improve the normalizing transformation, hence the name, the three-parameter lognormal distribution. In its simplest terms, the function $y = \ln(x-a)$ is normally distributed. The variate x is bounded below by the parameter a, and in its conventional form the distribution cannot have negative skewness. If the transform $y = \ln(a-x)$ is assumed to be normally distributed, then the distribution of x is negatively skewed, is unbounded below, and is upper bounded by the parameter a.

Estimates of the parameters of the distribution can be made by several different methods and, during development of this program, results obtained by fitting the distribution by moments were compared with those obtained by a maximum likelihood fit. The maximum likelihood fitting method was clearly superior in that the computed coefficients of skew and kurtosis of the transformed data were much closer to the theoretical values of 0.0 and 3.0. Thus only the maximum likelihood fit is used, except on rare occasions where it seems unobtainable and a moment fit is used. Historically weighted moments are used as a backup to the maximum likelihood technique for the historic analysis.

The three-parameter lognormal distribution is exceptionally flexible and very well suited to flood frequency analysis. When the parameter a becomes zero, then the distribution becomes the lognormal. Extensive testing during the development of this program showed that the coefficient of skew of the transformed data averaged 0.055 with a low value of 0.0008, an excellent check on the success of the fit. The mean kurtosis was a reasonable 3.255.

Appendix B further documents this distribution.

3.1.3 The Log Pearson Type III Distribution

A group of frequency distributions can be derived from a generalized differential equation proposed by Karl Pearson. This generalized equation has four constants. By equating some of them to zero or to each other and solving the differential equation, a series of symmetrical or skewed distributions is found. One possible solution leads to Pearson's Type III distribution. If it is assumed that the logarithm of the variate follows a Pearson Type III distribution, the distribution of the variate itself is the Log Pearson Type III.

The Log Pearson Type III distribution has three parameters. Bobée (1975) shows the many different shapes the distribution can take, depending on the relationship between the parameters and their signs. Not all these shapes seem acceptable in flood frequency analysis and the user of this program is advised to study Bobée's 1975 article particularly with regard to the distribution's shape. The range of the variate also depends on the sign of one of the parameters. The variate may have a positive lower boundary and be unbounded above, or, it may have a zero lower boundary with a positive upper bound. Parameters of the distribution are a, b, and m. The boundary parameter is m, while a and b represent scale and shape, respectively. A positive upper bound was found in about 67 percent of the test stations, and, depending on the position of this upper bound relative to the largest observed flood in the data sample, it can be very difficult to judge how well this distribution fits the data. Cases have been found using a moment fit where an upper boundary parameter was found to be less than the largest observed flood.

When combinations of parameters lead to apparently impractical shapes of the distribution, or where an upper bound is found to be less than the greatest observed flood, the program does not give the flood frequency regime. The program is designed to provide assistance to the user by indicating the upper boundary when it occurs. The distribution parameters will always be given in the output and the user may relate them to the combinations shown in Bobée (1975). It is very difficult to give any advice on the interpretation of an upper boundary and its implications. Points to be considered are the proximity of the upper boundary to the largest observed flood and the general appearance of the plot.

For the conventional analysis, the method of maximum likelihood is used as the primary technique to estimate the parameters of the distribution. If a maximum likelihood solution is not obtained, the program reverts to the method of moments. When historical information is present, the program estimates the parameters using historically weighted moments. Appendix B gives further information concerning this distribution.

3.1.4 The Wakeby Distribution

The Wakeby distribution is a very flexible five-parameter distribution proposed by Houghton (1978) for modelling flood flows. An advantage is that the distribution function can assume shapes such as "S-bends" and "hockey sticks", unrealizable by conventional distributions. The Wakeby was introduced as the "parent" of distributions because of its ability to mimic most conventional hydrologic distributions. This is true only if the parameters are chosen correctly, while the converse is not true.

The probability function of the Wakeby in inverse form is:

$$x = -a\,(1-F)^{b} + c\,(1-F)^{-d} + e \qquad \text{B.92}$$

where F is the probability of not exceeding x, e is a location parameter, a and c are scale parameters, and b and d are shape parameters. One of the most attractive features of the Wakeby is that the right and left-hand tails of the distribution can be modelled separately. That is, parameters a and b govern the left-hand (lower floods) tail while parameters c and d govern the right-hand (higher floods) tail. Thus, two parameters are required to define each of the tails and the fifth, e, is a location parameter. Houghton (1978) states that

> "in traditional estimation procedures the smallest observations can have a substantial effect on the right-hand side (large observations) of the distribution. But the left-hand side (small observations) does not necessarily add information to an estimate of quantile on the right-hand side. Indeed, since floods are not known to follow any particular distribution, it seems intuitively better to divorce the left-hand side from the right".

An added feature of this distribution is its capability to have less "separation effect" than that observed by conventional distributions (Matalas, et al., 1975). Greis (1983) explains that the separation effect

> "refers to the differences which appear between samples of synthetic streamflow data and natural streamflow data when the standard deviation of skew is plotted versus the mean of skew for regional data. The natural data consistently display a larger standard deviation of skew than the synthetic data indicating that nature has been given more inherently unstable skews than most statistical distributions".

Greenwood, et al. (1979) introduced the concept of probability weighted moments and demonstrated their use in estimating the Wakeby parameters, giving a solution algorithm later improved by Landwehr, et al. (1979 a,b). Hosking (1990), introducing the concept of L-moments, indicates that probability weighted moments "...can be expressed as linear combinations of L-moments." He states as well that L-moments are "more convenient" as "they are more directly interpretable as measures of the scale and shape of probability distributions." For standard samples and samples with low outliers, the method of L-moments is used herein, and the computer program is an adaptation of one developed by Hosking (1988).

If historic information is available, then Houghton's (1978) algorithm is used with a modification of the rank statistic due to Benson (1950), in conjunction with the Cunnane plotting position formula. If low outliers and historic information are present, then the sample exclusive of low outliers is treated first using the historic algorithm, and then the effect of the low outliers is accounted for using the conditional probability function.

When performing a conventional analysis, the method of probability weighted moments is used to obtain the parameters. If a historical analysis is performed, then parameters are estimated using a least squares algorithm similar to that prepared by Houghton (1978).

The distribution itself, as well as programming, imposes limits on the magnitudes of parameters b and d. Furthermore, certain combinations of both magnitudes and signs of parameters a, b, c, and d lead to improper definitions of the probability function. Since it is arithmetically possible to obtain these combinations, a sub-routine in the program checks the parameters, and if a valid set of parameters is not obtained then the program will inform the user. If the parameter d is negative, then the distribution has an upper boundary at e. Appendix B lists the valid parameter combinations and tests for admissible Wakeby parameter combinations, as well as further documenting the distribution. A special case of the Wakeby occurs when parameters a and b or c and d are zero. In such cases, the Wakeby parameters are obtained by fitting a generalized Pareto distribution to the data by L-moments (Hosking, 1988, p.58).

3.1.5 The Weibull Distribution

The interpretation of the frequency analysis of negatively skewed runoff samples can prove difficult for some of the distributions of this package. The generalized extreme value (EV3), the three-parameter lognormal, and the Log Pearson Type III may be upper bounded for the untransformed sample having a negative skew. The Wakeby may be upper bounded depending on the sign of parameter d. A problem arises if the upper boundary is located very close to the maximum observed sample member. Such a problem may lead to the maximum observed sample member. Such a problem may lead to an underestimation of extreme flood events. When a moment fitting method is used, the upper boundary may even be less than the observed maximum, an impossibility, thus making the analysis useless.

The Weibull distribution offers a useful alternative, since it can be fitted to samples with skewness as low as -1.08. The Weibull distribution is bounded below and remains unbounded above. Parameter estimates are obtained using the method of moments. This distribution is used only for "non-historic" samples having negative skewness in the untransformed data. Appendix B gives further information concerning this distribution.

3.2 NONPARAMETRIC FREQUENCY ANALYSIS

The previous section describes the parametric approaches to flood frequency analysis available in this package. In hydrologic practice, there does not as yet appear to be any theoretical justification for the selection of one distribution over another. Simulation studies, on the other hand, have demonstrated that the generalized extreme value, three parameter lognormal, and Wakeby distributions can yield more accurate estimates of quantile than can the Log Pearson Type III distribution (Wallis and Wood, 1985; Pilon, et al., 1987). This even occurs when the synthetic flood generator is based on the Log Pearson Type III distribution.

It is unfortunate that nature is not as simple as the simulation studies depict it to be. This is particularly so with regards to the form of the density function, e.g. unimodal. In an attempt to overcome certain distributional assumptions and limitations, the nonparametric method has been developed and advocated (Adamowski, 1989). The approach does not require assumptions to be placed regarding the form of the density function of the generating process. And, in application, the method is suitable when the generating mechanism produces floods having multimodal or mixed densities.

The non-parametric kernel density estimation method requires the selection of a kernel function, K (·), and the computation of a smoothing factor, h (Adamowski, 1985). The kernel function is itself a density function and is assumed to be Gaussian or normal in this package. The smoothing factor is estimated using the cross-validation procedure (Rudemo, 1982). These procedures have been adapted for the case when historic information is present.

3.3 OUTLIER ANALYSIS

The presence of outliers in a data sample will cause difficulties in satisfactorily fitting a parametric frequency distribution to the sample. Depending on whether the outliers are high or low, and on the chosen frequency distribution, the estimates of the T-Year event will often be underestimated or overestimated. Techniques are available for appropriately dealing with these outliers; but, these outliers must first be detected.

It is emphasized that if any historic information can be found for the highest member or members of the sample, then this information should be included in the analysis, even though a statistical test may fail to identify these members as high outliers.

The theory of outliers is still incomplete and has only been satisfactorily developed for a normal population. Application of the test is simple, requires only the mean and standard deviation of the sample, and tabulated values of the Grubbs and Beck (1972) statistic for various sample sizes and significance levels. Tabulated values of this statistic at the 10 percent significance level are listed in Table 1. The following polynomial can be used for estimating the tabulated values:

$$Y = -3.62201 + 6.28446\ N^{1/4} - 2.49835\ N^{1/2} + .491436\ N^{3/4} - .037911\ N$$

where N is the number of observations. Truncating the values of Y beyond the third decimal place should give results identical to the tabulated Grubbs and Beck statistic (K_N) for N observations as per Table 1.

The Grubbs and Beck outlier test has been adopted in modified form by the Hydrology Subcommittee (1982) of the United States, and this package follows the sequence herein. Since the test is applicable only to samples from a normal population, the assumption is made that the logarithms of the sample members are normally distributed. Rearranging the Grubbs and Beck test as done by the Hydrology Subcommittee (1982), the two following equations are obtained:

$$X_H = \exp(\bar{x} + K_N s) \qquad (1a)$$

$$X_L = \exp(\bar{x} - K_N s) \qquad (1b)$$

where $\bar{x}$ and s are the mean and standard deviation of the natural logarithms of the sample, respectively, and K_N is as previously defined.

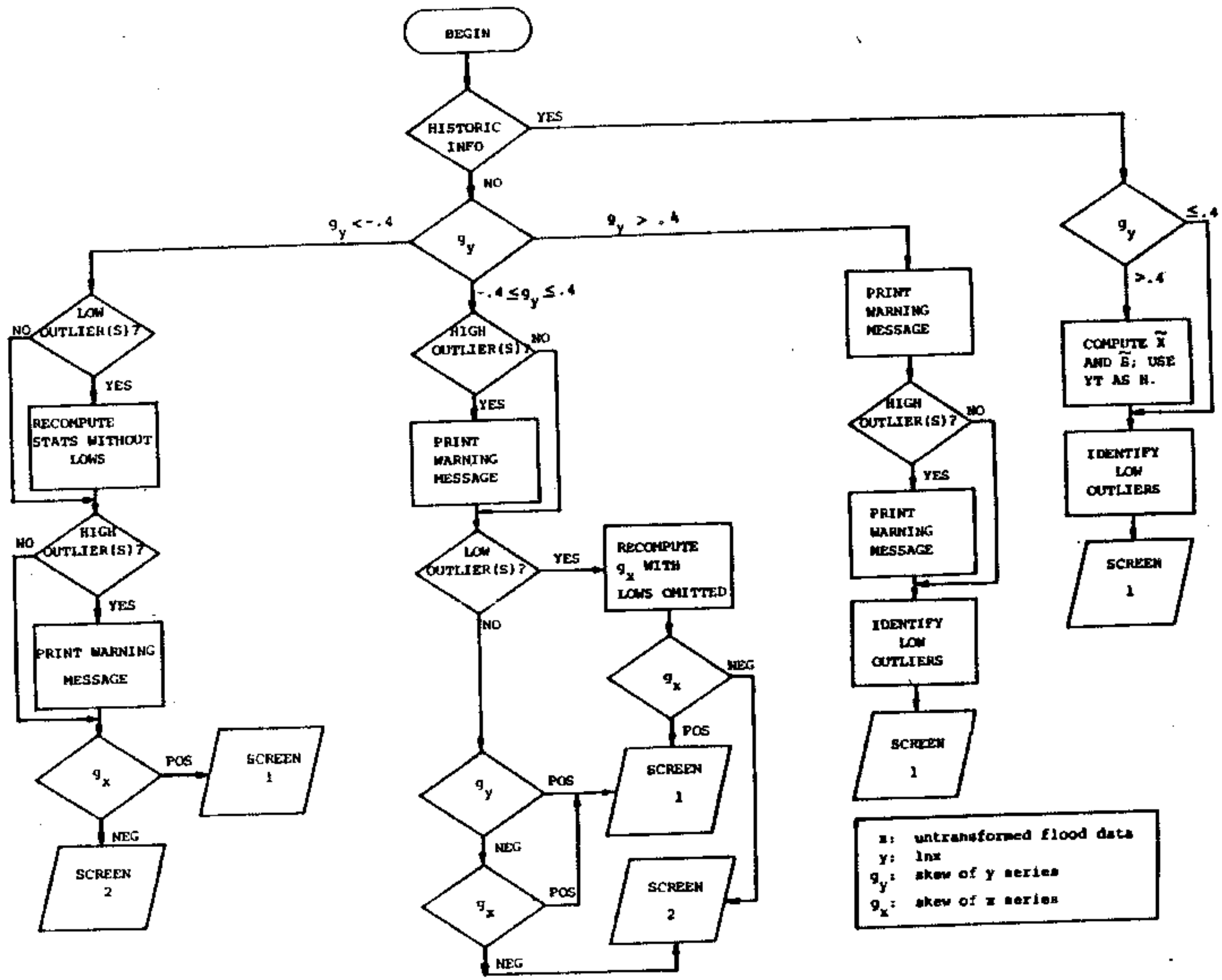

Figure 4: Flow diagram of outlier testing, where SCREEN 1 corresponds to Figure 11a and SCREEN 2 corresponds to Figure 11b.

3.4 ANALYSIS WITH ZEROS

The occurrence of zeros in a sample can cause difficulties when using some logarithmic types of parametric distributions because the logarithm of zero is minus infinity. However, for other distributions such as the generalized extreme value and three-parameter lognormal, the logarithm of the variate is not taken. So, unless (x-a) is exactly zero, there would be no mathematical difficulty in fitting the three-parameter lognormal distribution.

The three-parameter lognormal distribution was used to determine if the procedure proposed for the analysis of samples with zeros gave reasonable results. The records from 16 unregulated seasonal stations, each of which had experienced at lest one season of total dryness, were selected to test the method. The analyses were not very impressive, particularly the plots, and it was decided to rework the analyses using conditional probability.

If the probability of occurrence of zero flow is P(0), then the probability that flow does occur is [1-P(0)]. In standard statistical notation, the probability that a variate does not exceed x is F(x), and hence the probability that flow exceeds x is [1-F(x)]. Then the probability that flows occurs and exceeds x is:

$$P(x) = [1-P(0)]\,[1-F(x)] \qquad (2)$$

If a hydrometric record of length N years contains Z years in each of which no flow was recorded, then the probability of an entirely dry year in the future is Z/N. This follows from the empirical definition that gives the probability of an event as its relative frequency of occurrence in an extended series of trials. It is assumed in this definition that the probability tends to a definite limit as the number of trials tends to infinity. Although it may seem to be labouring a simple point, it is important to be aware of these conditions when estimating P(0) from the short periods of record usually available in hydrology. For the three-parameter lognormal distribution, F(x) in the preceding equation is the integral of the three-parameter lognormal distribution from the boundary parameter a to x. Note that the parameters of F(x) are estimated using only non-zero flow events.

The records of the 16 test stations were reanalysed using conditional probability, and on comparison this method gave subjectively superior analyses in 13 cases. Hence, the conditional probability method has been chosen for all the distributions.

3.5 ANALYSIS WITH LOW OUTLIERS

The presence of low outliers in a sample of annual maximum floods can create problems in producing a satisfactory parametric flood frequency analysis, particularly when the sample is small. Most noticeably, the low outliers affect the skewness, which becomes small and sometimes even negative, with the attendant difficulty in fitting the more common density functions in their conventional forms.

Assume for the moment that in a sample of size N, L of these observations are identified or are to be treated as low outliers. Now let n be the number of observations excluding the L low outliers; thus, n = N-L. The first step of the analysis proceeds exactly as in the With Zeros case. Now rearranging Equation (2):

$$F(x) = 1 - P(x)\,(n+L)/n \qquad (3)$$

then since F(x) cannot be less than zero, and n and L are positive numbers, P(x) cannot exceed n/(n + L), or floods of return periods less than (n + L)/n cannot be computed. Furthermore the probability function corresponding to Equation (3) is not truly that of the theoretical distribution, and curves more sharply downward at the lower return periods. These are not serious drawbacks since they have their largest effects at the shorter return periods that are not usually within the range of interest. At return periods greater than 2 years the conditional probability function can be simulated almost exactly by mathematically retrofitting the generalized extreme value, the three-parameter lognormal, and the Weibull distribution. The Log Pearson Type III uses a method based on synthetic statistics as illustrated by the Hydrology Subcommittee's publication (1982). A retrofitting algorithm for the Wakeby distribution was not found.

Retrofitting leads to a probability function of the distribution in question and simulates exactly the conditional probability function between return periods of 2 and 100 years for the three-parameter lognormal and Weibull distribution and between return periods of 1.582 and 100 years for the generalized extreme value distribution. The retrofitted function is only off by less than one percent at a return period of 500 years. When extrapolated downward to return periods of less than the lower return period, the shape of the curve is much improved and reduces the drawbacks previously mentioned. When retrofitting is used by the program, the statement "PARAMETERS OF THE DISTRIBUTION WHICH DUPLICATES THE CONDITIONAL FUNCTION:" appears prior to the listing of the parameters. Appendix B documents the retrofitting algorithms used in the package.

Note that the retrofitting of the conditional probability function for the generalized extreme value, the three-parameter lognormal, and the Weibull distributions is only performed if no zeros are present in the sample. A conditional probability function is used if lows and/or zeros are present in the sample. The Log Pearson Type III uses the conditional probability adjustment as recommended by the Hydrology Subcommittee (1982) for samples having years of zero flow and/or low outliers.

3.6 TYPES OF HISTORIC INFORMATION

The conventional sample available for flood frequency analysis consists of a series of maximum flows obtained from a continuous record of discharge at a hydrometric station. These maxima may be either instantaneous peaks or daily flows occurring in each calendar or water year, or in some specified season throughout the period of record.

If the period of operation of the hydrometric station is N years, then the sample size is N. N is typically quite small. Obviously, any historic information which effectively enlarges N will improve the frequency analysis. The following is a description of the types of historic information and the notation and terminology used.

3.6.1 Extreme floods prior to the gauged record

On many rivers some large floods may have occurred in the past, often many years prior to the installation of the hydrometric station. If their magnitudes and years of occurrence are known, then that historic information can be incorporated in the frequency analysis using this program. Listed below are the definitions of symbols which appear on the printouts and are further explained using a hypothetical example.

If	NA	=	number of fully specified floods above the threshold
	NB	=	number of fully specified floods below the threshold
	NC	=	number of censored floods below the threshold
	N	=	number of observed or fully specified floods
	YT	=	total time span in years
	NHA	=	number of historic floods above the threshold

then N = NA + NB and YT = NA + NB + NC

Let the example hydrometric record be available for the years 1940 through 1982. A flood of known magnitude occurred in 1920 and is known to be the largest since 1900. At first glance, the data available for analysis are the 1920 flood and the recent series from 1940 onward. There is, however, the additional information that in the 39 missing years - 1900 through 1919 and 1921 through 1939 - the annual flood was less than the 1920 value. These missing years are the censored data and the censoring threshold is the value of the 1920 flood. So NA = 1, NB = 43, NC = 39, N = 44, YT = 83, and NHA =1. When performing an historical frequency analysis, only YT, NHA, and the censoring threshold need be specified supplementary to the conventional data required by CFA.

3.6.2 Extreme floods in the gauged record

The term historic need not apply only to floods which occurred before the installation of a hydrometric station to collect a continuous record. Methods used in CFA are equally adaptable to the case of the occurrence in the gauged record of the largest flood or floods in the history of the area, provided that there is reliable local information that the flood or floods were the largest since some known date. Suppose that at the same hypothetical location described in (a), the 1960 flood is reliably known to be the second highest sinçe 1900. Then NA = 2 and NB = 42. NC, N, YT, and NHA remain at 39, 44, 83, and 1, respectively, but the censoring threshold then becomes the value of the 1960 flood. Note that NHA does not affect the computations and is included only for display purposes.

3.6.3 Out of bank floods

This type of record can be found in the British Isles, the U.S.A. and possibly other countries. Floods that rose above a certain level, generally bank-full, were marked by dated stone markers and discharges have since been estimated, and the assumption is that in the years for which no markers exist, the flood was less than bank-full discharge. In an analysis of this type, the total time span, YT, is obtained from the data of the earliest marker, and the censoring threshold is bank-full discharge. The number of fully defined floods, N, is clearly the number of annual maximum floods in the gauged record plus the number of markers. The program computes NA, NB, and NC. The number of historic floods above the threshold, NHA, is merely the number of markers and is for display purposes only.

3.7 ANALYSIS WITH HISTORIC INFORMATION

There are at least two methods of fitting parametric distributions to samples containing historic information. The oldest and most commonly used method is that of historically weighted moments; and a more recent development is to consider the series of floods to be a censored sample from some postulated distribution, and then fit the distribution to the sample using maximum likelihood theory.

If a sample of x's, size N, is drawn from a postulated distribution but with the parameters as yet unknown, the likelihood function L can be expressed in terms of the sample and the unknown parameters. This likelihood function L is the probability that all the members of the sample were drawn from the postulated distribution, and the principle of maximum likelihood states that the unknown parameters should be chosen to maximize L.

Suppose that the magnitude of a sample member is unknown, but the sample member is known to be less than a certain value x_c, the censoring threshold. Such a member is called a censored member. Then, the probability that the sample member was less than x_c, and came from the postulated distribution, is the probability function evaluated at x_c, $F(x_c)$. By extension the probability that r sample members were less than x_c is $[F(x_c)]^r$. For a censored sample from a postulated distribution with parameters yet to be determined, since the likelihood function is a probability, then L can be expressed in terms of the fully specified sample members, the number of censored values below the censoring threshold, and the distribution parameters. Maximizing L by taking partial derivatives with respect to each parameter in turn and equating them to zero gives a set of simultaneous equations, the maximum likelihood estimators, that when solved give maximum likelihood estimates of the distribution parameters.

When the parameters have been estimated, the floods of the required exeedance probabilities or return periods can easily be computed.

For the three-parameter lognormal distribution, it has been shown by Condie and Lee (1982) that the censored maximum likelihood approach is superior to the traditional historically weighted moments method in terms of providing the least biased estimates of the T-year floods. The censored maximum likelihood method is used primarily in this program and in more than 1000 tests has never failed to give a solution. The historically weighted moment fit is included as a backup method, in case of failure. Appendix B further documents the parameter estimation via historically weighted moments.

Condie and Pilon (1983) have developed the censored maximum likelihood approach for the Log Pearson Type III. The technique requires refinement, thus historically weighted moments are used. The generalized extreme value distribution uses the historically weighted moment approach when dealing with historic information.

If historic information is available, the Wakeby distribution parameters are estimated using a regression method similar to that proposed by Houghton (1978). Rewriting equation B.80 gives:

$$x = -aP^{b} + cP^{-d} + e \qquad (4)$$

where P is estimated using the Cunnane (1978) probability plotting formula, adjusted using the methods of Benson (1950). Appendix B further documents the parameter estimation via least squares regression.

The kernel estimate of the nonparametric method is adjusted to reflect historic information in a fashion similar to historically weighted moments and the expectation theory of the maximum likelihood approach. Appendix C further documents the parameter estimation when historic information is present.

3.8 PLOTTING POSITION

The Cunnane (1978) plotting position is used to plot the data on the probability paper. The plotting formula can be written as:

$$T = \frac{N + .2}{m - .4}$$

where T is the return period, N the sample size, and m the rank, starting with rank 1 for the largest. It is emphasized that the plot should be used as a guide to show how well the distribution has been fitted to the sample series of floods.

The position may be adjusted if historic information exists using Benson's (1950) technique. When historic information is present, the return periods for floods above and equal to the threshold are computed from:

$$T = \frac{(YT + .2)}{m - .4}$$

The return period for each flood below the threshold is computed from:

$$T = \frac{(YT + .2)}{m_a - .4}$$

where

$$m_a = NA + (YT - NA)\ (m - NA)/NB$$

and where YT, NA, NB, and m are as previously defined.

3.9 STATISTICAL TESTS FOR INDEPENDENCE, TREND, HOMOGENEITY, AND RANDOMNESS

Statistical frequency analysis assumes that the sample to be analyzed is a reliable set of measurements of independent random events from a homogenous population. The validity of this assumption can be verified using statistical significance tests. Nonparametric tests are used exclusively in this program. In addition to the statistical tests of this section, various graphical displays can be used to assess the aforementioned assumption. These plots are available and are accessed under the "SCREEN DATA" option of the program, which is documented in Chapter 4.

Brief descriptions of the rationale for each test are given here; fuller descriptions are given by the Natural Environment Research Council (1975) and Siegel (1956). The theory of the tests is not given but the required functions to be evaluated and the determination of their significance are given in Appendix A.

3.9.1 Test for Independence

Two events can be considered independent only if the possibility of occurrence of either is unaffected by the occurrence of the other. This definition can be extended to a sample of size N. Practically, in a time series, independence can be measured by the significance of the correlation coefficient between the N-1 pairs of the (i) and (i+1) members of the series and if the correlation coefficient is not significantly greater than zero, then independence is assumed. It is noted here that, in the strict mathematical sense, this does not necessarily define independence. To avoid the assumptions made in the coefficient, the nonparametric Spearman rank order serial correlation coefficient is used.

3.9.2 Test for Trend

If successive measurements of members of a time series have been made during a period of gradually changing conditions, then there will be a more or less noticeable trend in the magnitude of the members of the series when arranged in chronological order. As an example from hydrology, it would be expected that gradual land use changes in a drainage basin would affect the magnitude of the annual flood. Similarly long term climatic changes will be reflected in the hydrology of a basin, although it is customary to assume climatic time invariance.

3.9.3 Test for Homogeneity

If some more or less abrupt change occurred during the sampling period, then some difference could be expected between the means of the subsamples before and after the change. Examples from hydrology could include the construction of an ungated reservoir in the basin, or a forest fire that denuded a substantial portion of the basin. Assuming a normal distribution and that the two subsamples have the same variance, then the difference in the subsample means can be tested for significance using the distribution of Student's t. These assumptions are not commonly met in hydrology and so the Mann-Whitney nonparametric test is used instead. If two subsamples of approximately the same size are chosen, it would be expected that if there were no changes in conditions, then the sums of the ranks of the two subsamples would not differ by too much. The question to be answered is whether the difference is significant or not. The Mann-Whitney U statistic is a function of the subsample sizes and their sums of ranks. The distribution of U is known and critical values at various levels of significance have been tabulated. Hence a decision can be made on whether the means of the subsamples differ significantly.

The program provides a histogram of the data sample by months of occurrence, and the user may then choose the most sensible seasonal split from the hydrologic point of view. Differences may be expected between the means of floods occurring in different seasons, where again the question is whether the means of the subsamples differ significantly. The Mann-Whitney nonparametric test can be used to check this hypothesis. Computation methods are shown in Appendix A.

3.9.4 Test for General Randomness

This is a very simple nonparametric test. Data are ranked in chronological order, and the median is determined. The number of runs of observations above and equal to or below and equal to the median are counted. Theoretically, the number of runs, RUNAB, could be as high as the total number of observations, indicating an extreme short term cyclic pattern, or as low as 2, indicating an abrupt change half way through the period over which the sample was collected. Notice that the median is used since the probability of exceeding the median is always 0.5, regardless of the probability distribution from which the sample was drawn, thus making the test nonparametric or distribution free. The distribution of RUNAB is known and upper and lower critical values have been tabulated, thus enabling a decision to be made on whether the data are random or not. Appendix A further describes this test and gives tables of critical values.

This page intentionally left blank.

4. PROGRAM DOCUMENTATION

4.1 STARTING THE PROGRAM

NOTE: Installation instructions are included on the software distribution diskette in a file called READ.ME.

A main menu will be displayed on the monitor from which the user can select the various tasks to be performed. Figure 5 displays the main menu of CFA. Once it is displayed, the user should select Data Entry or Data Set Directory before performing any other task. "Data Entry" places a data set into the virtual memory of the computer. Once a data set is in virtual memory, the user can then perform tasks on the data.

CONSOLIDATED FREQUENCY ANALYSIS PACKAGE - MAIN MENU	
STOP THE PROGRAM	
SINGLE SITE DATA MANIPULATIONS	Data Entry Data List Data Edit Data Save
DIRECTORY (H22FLOOD)	List Purge Sort & Compress
STATISTICAL ANALYSIS	Screen Data Parametric Frequency Analysis Nonparametric Frequency Analysis
BATCH MODE	Print Output
MODIFY	Set-up File

Figure 5. Main Menu of CFA

4.2 DATA ENTRY

Once Data Entry has been selected by the user, the "Data Entry Menu" should appear as shown in Figure 6. This menu item allows the user to place a data set of flood record for a site into the computer. The data are primarily the series of year, month, and flows, along with site specific information. Data can be manually entered using the keyboard or data can be read from mass storage provided the data have been previously stored or transferred onto the system. Note that the disk master file (H22FLOOD.DAT) is a data bank created for the storage of data that are to be used by CFA. CFA also has the ability to read files from the HYDAT CD-ROM 4.0 (Ecosystem Sciences and Evaluation, 1993) as well as 3.0 (Inland Waters, 1991) and ASCII format.

4.2.1 CD-ROM Files

CD-ROM implies Compact Disc Read-Only Memory and is a laser-read optical disc containing historical discharge, water level and sediment data files of Environment Canada's HYDAT database. Daily discharges, annual maximum daily discharges, and annual instantaneous peak discharges can be extracted using the software that accompanies the HYDAT CD-ROM. Item numbers 4 and 5 of the Data Entry Menu of Figure 6 correspond with the two forms of data as provided from CD-ROM. If you are a HYDAT CD-ROM user, refer to Appendix B of the HYDAT version 4.0 or Appendix C of version 3.0 USER's MANUAL for an explanation of the export formats.

CONSOLIDATED FREQUENCY ANALYSIS PACKAGE - DATA ENTRY MENU
RETURN TO MAIN MENU OF CFA
Disk Master File (H22FLOOD. DAT)
Keyboard Entry
CD-ROM Annual Maximum Instantaneous Peak Discharges or Levels
CD-ROM Annual Maximum Daily Discharges or Levels
ASCII File (LOTUS Print and CFA Export Formats)

Figure 6: Data Entry Menu of CFA

4.2.2 ASCII Files

ASCII files are a convenient and widely-used standard that can be used to import data and information into CFA. Many commercially available software packages that perform text editing, word processing, spreadsheet manipulations, etc., allow for the storage of material into ASCII files. The arrangement of the contents of the ASCII file is described in Section 4.5 of this report under the title "Data Save".

4.3 DATA LIST - LIST CONTENTS OF DATA SET

This menu item allows the user to view the information contained in virtual memory regarding a certain data set. The user has the choice of obtaining a hardcopy listing or simply viewing the data set via the monitor.

4.4 DATA EDIT

This menu item is designed to allow the user to perform a variety of editing procedures on the data set. Figure 7 shows the "Data Edit Menu". Note that "Data Edit" affects only the data set in virtual memory. If the user wants to save the modifications, then the "Data Save" feature should be used.

If observations of YEAR, MONTH, FLOW are being deleted, added or modified, then the historic information may need alteration. For example, if an additional five years of record is being added to a previously created historical data set, the historic time span should be altered. The censoring threshold and the number of peaks above or equal to it may also need modification.

CONSOLIDATED FREQUENCY ANALYSIS PACKAGE - DATA EDIT MENU	
RETURN TO MAIN MENU OF CFA	
EDIT	Station Number, Name and Area Any Entry of Year, Month, Flow Existing Historic Information
ADD	Historic Information Observation(s) Anywhere in the Sequence
DELETE	All Historic Information Observation(s) of Year, Month, Flow The Entire Record from Mass Storage
LIST	

Figure 7: Data Edit Menu of CFA

4.5 DATA SAVE

This Main Menu option allows the user to store in a permanent file the data set currently residing in virtual memory. Once selected, the user is shown Figure 8. Three avenues exist for the storage of the data and information.

CONSOLIDATED FREQUENCY ANALYSIS PACKAGE - DATA SAVE MENU
RETURN TO MAIN MENU OF CFA
Disk Master File (H22FLOOD.DAT) LOTUS Import File ASCII File

Figure 8: Data Save Menu of CFA

If the first choice of the Data Save Menu is selected, then the data and information are stored in CFA's master file H22FLOOD.DAT. The data set will be stored in the master file under the name of the "WSC STATION NO.". This station number must be less than or equal to 10 alphanumeric characters. If a previous data set has been stored in the same master file under the same WSC STATION NO., then the dat^ set will be over-written. Caution should be exercised when original data sets and modifications are both wanted in permanent memory. Renumbering of the data set's "WSC STATION NO." ($\leq$10 alphanumeric characters) is suggested using the "Data Edit" option of the Main Menu. For example, 01AK001 could be renumbered to 01AK001B.

The second choice of the Data Save Menu permits the storage of the data and information into a LOTUS import file format. The resultant file can be directly imported into LOTUS 1-2-3.

```
050F003
BOYNE RIVER NEAR CARMAN
33      976.000     90      1       187.000
   33        NUMBER OF OBSERVATIONS
  976.00     AREA
   90        HISTORIC TIME SPAN
    1        NUMBER OF FLOODS ABOVE CENSORING THRESHOLD
  187.00     CENSORING THRESHOLD
050F003          1893   4       187.00
050F003          1919   4        13.500
050F003          1922   4        15.300
```

Figure 9: ASCII File Format for CFA

The third choice stores the data in ASCII format. An example of the format is shown in Fig. 9. Note that the first line contains the "WSC STATION NO.". This should be a unique identifier for the data and information that follows. The second line is the station name. The third line is a sequence of five numbers described in lines 4 through 8, inclusive, of Figure 9. Lines 4 to 8 are included in saved ASCII files in order to assist users in decoding the third line. Thus, lines 4 to 8 are not required when creating an ASCII file similar to that of Figure 9 for importing of data to CFA by the "Data Entry Menu". Evidently, the site depicted in Figure 9 contains historic information. If this site had no historic information then lines 3 to 5 would be

```
33      976.000       0     0       0.0
   33        NUMBER OF OBSERVATIONS
  976.00     AREA
```

Lines 6 to 8 would not be contained in the file, and the remainder would be identical. Lines 9 to the last in Figure 9 contain the YEAR, MONTH, FLOW data for the 33 years of existing record. Note the first column of these lines should be the station number. Again, if no historic information was present, then lines 6 to the last would contain the 33 years of flow data.

4.6 DATA SET DIRECTORY

This menu item offers three choices. The first allows the user to view the station number and name of all data sets in a CFA master file. The user may obtain a hardcopy of all station numbers and names on the file or simply view the information on the monitor. The second choice allows the user to delete unwanted staions from the master file. The third choice allows the user to sort the master file by station number or station name and in the process, compress the master file. You should compress the H22FLOOD.DAT file if you delete any stations. This will remove unused space from the indexed file. The original H22FLOOD.DAT file will be saved as H22FLOOD.BAK. Delete this file if it is unnecessary.

4.7 SCREEN DATA - GRAPHICAL DISPLAYS AND NONPARAMETRIC TESTING

This menu item allows the user to test the assumption that the sample is a reliable set of measurements of independent random events from a homogeneous population. The validity of this assumption can be investigated using the nonparametric tests of the "Screen Data" menu. In addition, this menu item allows the user to view the data via graphical displays; thus, permitting qualitative assessments to the suitability and characteristics of the data.

Nonparametric tests include: 1) the Spearman rank order serial correlation coefficient test for independence; 2) the Spearman rank order correlation coefficient test for trend; 3) a general randomness test; and 4) the Mann-Whitney split sample test for homogeneity. Section 3.9 of this report describes these tests in more detail. The graphical displays include: 1) rank-time plot; 2) discharge-rank plot; 3) discharge-time plot; 4) frequency histogram by % of maximum discharge; 5) frequency histogram by month; and 6) frequency histogram by discharge. Examples of the output from these tests are in section 5.2 and 5.3.

Figure 10 shows the "Screen Data Menu" as it should appear on your screen. Note that the second selection on the menu is the nonparametric tests for independence, trend, and randomness. The third selection is the homogeneity test. This test was not included with the other nonparametric tests to allow for a greater freedom when analyzing the data. For example, it would be advantageous to view and assess a rank-time plot, a discharge-rank plot, and a discharge-time plot before performing a homogeneity test based on a split sampling by years. A viewing and an assessment of a frequency histogram by month would be advantageous before performing a homogeneity test based on split-sampling by season.

CONSOLIDATED FREQUENCY ANALYSIS PACKAGE - SCREEN DATA MENU
Return to main menu of CFA
Independence, trend, and randomness
Homogeneity (Mann-Whitney)
Rank-time plot
Discharge-rank plot
Discharge-time plot
Frequency histogram by % of maximum discharge
Frequency histogram by month
Frequency histogram by discharge

Figure 10: Screen Data Menu of CFA

Menu items: discharge-rank plot; frequency histogram by % of maximum discharge; and frequency histogram by discharge permit the assessment of the nature of the distribution (unimodal, bimodal...).

> ***HINT:*** *rather than use <CTRL> <P> to print the pages of output you want, note the desired pages, then print all the pages at once using BATCH MODE - PRINTER OUTPUT. See sections 4.8 and 4.9.*

4.8 FREQUENCY ANALYSIS

This menu item provides a hydrologic frequency analysis of a typical sample, a sample with zeros, a sample with low outliers, a sample with historic information, or combinations of the latter three cases. The user has the choice of performing parametric and/or non-parametric frequency analysis.

4.8.1 Parametric Frequency Analysis

Outlier analysis is performed automatically. The user has the capability of altering the number of low outliers prior to conducting the frequency analysis. The distributions available for conventional and historical analysis include the generalized extreme value, the three-parameter lognormal, the Log Pearson Type III, and the Wakeby. The Weibull distribution is available for the conventional analysis of a sample having negative skewness in the untransformed data. Flows corresponding to selected return periods are computed from the probability function of the given distribution.

The output consists of input data: ranked data and adjusted ranking if historic information is available; high outliers, low outliers, and zeros identified if present; empirical probabilities; and return periods. Then follow the sample statistics, the distribution's parameters and a tabular frequency regime with a flood frequency plot.

The program automatically checks if historic information has been given for the site. If a conventional frequency analysis is to be performed, that is no historic information is present and the coefficient of skewness of the natural logarithms of the flow series exceeds .4, then the program displays:

"YOUR SAMPLE DISPLAYS A HIGH (>.4) SKEW. YOU SHOULD CHECK FOR HISTORIC INFORMATION AND ACCURACY."

If a conventional analysis is being performed, the program automatically performs an outlier analysis on the data sample. If a high outlier is detected, then the program displays:

"...HIGH OUTLIER(S) DETECTED. YOU SHOULD CHECK FOR HISTORIC INFORMATION AND ACCURACY."

These two comments infer that the data set may contain a typing error, or that the station should be reviewed for historic information and/or accuracy of the record before proceeding with a frequency analysis.

If a historic analysis is being performed, the program automatically performs a low outlier analysis.

Figure 11(a) and 11(b) show the two menus of parametric distributions. The characteristics of the data set determines which of the two menus will be displayed on your monitor. The sole difference between these menus is the Weibull distribution in Figure 11(b). The rules for knowing which one of the two will appear are:

(1) if the data set contains historic information, Figure 11(a) will appear.
(2) if the data set contains no historic data and if the coefficient of skewness of the untransformed data sample excluding all low outliers is less than zero, then Figure 11(b) will appear; otherwise, 11(a) will appear.

DISTRIBUTION MENU	
RETURN TO MAIN MENU OF CFA Generalized Extreme Value (GEV) 3 Parameter Lognormal (HILO) LOG Pearson Type III Wakeby All of the above Alter number of low outliers	(a)

DISTRIBUTION MENU	
RETURN TO MAIN MENU OF CFA Generalized Extreme Value (GEV) 3 Parameter Lognormal (HILO) LOG Pearson Type III Wakeby Weibull (Negskew) / Gumbel III All of the above Alter number of low outliers	(b)

Figure 11 (a) and (b): The Two Distribution Menus in CFA

4.8.2 Nonparametric Frequency Analysis

This menu item allows the user to determine flood quantiles based on nonparametric frequency approaches. Section 3.2 and Appendix C should be referred to for more information regarding the formulation of the technique. Note that outlier analysis is not performed when proceeding with a nonparametric frequency analysis due to the inherent nature of the technique.

4.9 BATCH MODE - PRINTER OUTPUT

This allows you to print in one batch job the output for the station(s) you are analyzing. When this option is selected you are first given a list of the stations in your master H22FLOOD.DAT file. You must TAG the stations you are analyzing by pressing <F10>. When you have completed tagging stations press <ENTER>. The menu in Figure 12 then appears.

CONSOLIDATED FREQUENCY ANALYSIS PACKAGE - BATCH MENU
Nonparametrics Tests (independence, trend, randomness, homogeneity)
Rank - time plot
Discharge - rank plot
Discharge - time plot
Frequency histogram by % of maximum discharge
Frequency histogram by month
Frequency histogram by discharge
Generalized Extreme Value (GEV)
3 Parameter Lognormal (HILO)
Log Pearson Type III
Wakeby
Weibull (Negskew) / Gumbel III
Gauss Kernel (nonparametric)

Figure 12: Batch Print Menu for CFA

You then TAG the output you would like by pressing <F10>. When you have completed tagging the output press <ENTER>.

> ***HINT:*** *Printing uses the printer settings in the SET_UP.DAT file. Run your output off in draft (low density) first to ensure you have what you want, then adjust your SET_UP.DAT file to a higher print density for your final version. On an HP LaserJet III expect to wait about 2 minutes per page at low density, about 15 minutes at high density.*

The limit for the number of stations that can be analyzed in one batch job is 1000.

Currently, batch mode allows only default settings for plots, i.e. there is no provision for truncation of plots. If truncation is desired this must be done in interactive mode.

4.10 MODIFY SET-UP FILE

This allows you to modify set-up parameters initially specified at installation. When this is selected the following menu appears:

```
              CFA PACKAGE - MODIFY SET_UP FILE MENU
    RETURN TO MAIN MENU

    CHANGE    Screen Mode

    CHANGE    Number of copies
    CHANGE    Output direction (printer/file)

    CHANGE    Default printer
              Page size
              Printer density
              Printer port
              Graph Orientation

    SAVE Modified file
```

Figure 13: Modify Set-up Menu for CFA

The printer output direction is not saved in the SET_UP.DAT file. This is a temporary redirection of output to file. This option allows you to obtain hardcopy output on another system with a printer without installing CFA on that system. The output file has to be created for the printer that will be used to generate the hard copy. The graphic printer output in different runs of CFA will always number from HALOP**1.GRF, so it will overwrite existing generated graphs. Files must be renamed to prevent this. The tabular output is sent to a file named by the user.

Pressing the <ESC> key erases any modifications made without saving the modified file. If you make some changes and want them to apply only for the current session, then do not SAVE the modified file. Exit the menu by selecting "RETURN TO MAIN MENU". Selecting "SAVE Modified File" saves any modifications in the SET_UP.DAT file. The changes will be stored for use in the current and future sessions.

5. EXAMPLES

5.1 INTRODUCTION

This section contains examples of the output from "Screen Data", "Parametric Frequency Analysis" and "Nonparametric Frequency Analysis". Examples of both a conventional frequency analysis and a historical frequency analysis are given. For each case, nonparametric tests and graphical plots of "Screen Data" are presented and reviewed.

It has been mentioned previously that parametric frequency analysis assumes that the sample to be analyzed is a reliable set of measurements of independent random events from a homogenous population. The validity of this assumption can be investigated using the techniques incorporated in "Screen Data". Statistical testing and qualitative graphical assessment of the sample should be performed prior to attempting a frequency analysis.

The output format of the frequency analysis option will differ somewhat, depending on the presence or absence of zeros, low outliers, and historic information. Note that low outlier information is not included when nonparametric frequency analysis is being performed. In all cases, the output will be in three parts. The first part includes a listing of the station identification, input data, ranked data, empirical probabilities, and return periods. Adjustments are made if historic information is available. The second part includes the estimates of the parameters of the distribution and a tabular frequency regime. The third and final output is a titled plot on lognormal probability scale showing the data points and the plotted function of the distribution in question.

5.2 CONVENTIONAL ANALYSIS - ENGLISH RIVER NEAR SIOUX LOOKOUT - 05QA001

This gauging station was established in 1921 and was discontinued in 1981. A continuous record of annual maximum daily flows is available from 1922 to 1981, giving a sample of size 60 for analysis. Table 2 to 12 and Figures 14 to 24 comprise the output for the conventional analysis. Tables 2 to 6 list the results of the nonparametric tests performed on the flow series. Test results indicate that the data do not display significant serial dependence or trend. In addition, the data appear random and homogeneous. Figures 14 to 19 show the graphical analysis available using the "Screen Data Menu". Analysis of these graphical displays substantiate the nonparametric test results.

Figure 18 shows the histogram for discharge by month. This is a graphical representation of the number of occurrences of annual maximum mean daily flows for each month. Two distinct groupings of observations are apparent. They are May to July and September to November. These correspond with the spring and early summer season and the fall season. Unfortunately, causation factors may not always be seasonally dependent. However, are these two groups homogeneous? Do they appear as though they were drawn from the same population? The Mann-Whitney U-test for homogeneity is used in an attempt to quantitatively answer these two questions. Table 6 lists the results of this test for the groupings previously described and indicates that no significant location difference in the magnitude of observed floods was noted. That is, the floods collectively appear homogeneous and as though they were drawn from the same population; however, this does not negate the possibility that floods from the two seasons are due to different mechanisms.

Figures 14 and 16 show the rank-time plot and the discharge-time plot of the sample, respectively. These two graphical displays are used in the qualitative assessment of the data with respect to trend or jumps or heteroscedasticity (non-constant variance) or dependence structure. They show no evidence of trend or heteroscedasticity or jumps. No dependence structure in the flow series is apparent. These results are in agreement with the findings of the nonparametric tests as listed in Tables 2, 3 and 4. Figure 15 shows no abrupt change in slope, while Figure 17 shows no evidence of an excessively low or high observation. Figure 19 shows the frequency histogram of the sample, indicating a positively skewed unimodal distribution.

From the above exercise, the sample appears to fulfil the prerequisite assumptions of frequency analysis. The sample appears well behaved and there should be no problems in attempting a frequency analysis.

Table 7 represents the first of three sections of output from the "Frequency Analysis" selection of the main menu. The first section, after titling gives in the first three columns the month, year, and magnitude of the flood. The fourth and fifth columns rank the data in descending order of magnitude and a rank number is assigned, starting with 1 for the largest. The sixth and seventh columns give the probability of exceedance as a percentage, and the return period in years based on Cunnane's plotting position.

Tables 8 to 12 represent the output of the second section. This section, after titling, etc., gives the best estimates of population mean, standard deviation (S.D.), coefficient of variation (C.V.), coefficient of skewness (C.S.), and coefficient of kurtosis (C.K.) as estimated from the sample. These statistics are given for both the sample and its transformation in Naperian logarithms. At this point, Table 9 is slightly different than the three other tables - Tables 8, 10 and 12. When the boundary parameter "a" of the three-parameter lognormal distribution has been found, the statistics are repeated for the ln(x-a) series. Table 10 lists the coefficient of skewness of the ln(x-a) series as - .056, which is close to the theoretical value of zero. The coefficient of kurtosis is listed at 2.896, which is close to the theoretical value of 3.0. This is an indication that a good fit has been obtained for this distribution. In addition, Tables 8 and 11 are slightly different than Tables 9, 10, 12. These two tables include the first two L-moments (mean and standard deviation), the ratio of the second to the first (C.V.), the third (C.S.) and the fourth (C.K.). The L-moment statistics are included only in the summary of statistics for the generalized extreme value and Wakeby distributions as they are used to obtain parameter estimates for the conventional sample. The output then includes the sample's maximum and minimum, the lower outlier limit (parametric distributions only), the total sample size, the number of low outliers (parametric distributions only), and the number of zero flows. The solution method (e.g. moments) is given, followed by the estimate of the parameters of the distribution. Table 10 indicates that the Log Pearson Type III distribution is upper-bounded at 105 000. Table 11 shows that the Wakeby distribution is upper bounded at 2 945. Both of these upper boundaries appear large in comparison with the 500 year event. A viewing of the frequency plot may give further information as to the practical limitations of the distribution for this sample. That is, the upper boundary may appear sufficiently large so as to not affect the shape of the frequency curve in our range of interest. Finally, the tabular flood frequency regime is given for various preselected return periods and exceedance probabilities.

The third and final section of the output from the frequency analysis menus is the plot on lognormal probability paper. Figures 20 to 23 show the plot of the floods indicated by an asterisk and the fitted probability function by a continuous line. Figure 24 shows the nonparametric estimate of the frequency of the floods where the lower portion of the density is not drawn for return periods below 1.05 years.

In summary, from a comparison of the output for each frequency distribution, it would appear that no one parametric distribution gives a superior answer than the next for this sample. The nonparametric method gives answers similar to those obtained from the parametric approaches. Figure 24 shows the flexibility of the nonparametric approach, especially in the high return period portion of the graph. One characteristic of this approach is the very small probability of occurrence that exists when extrapolating beyond the highest observed flood in the sample. This point is illustrated in the example, but must be kept in mind when estimating floods of high return periods.

```
        --- SPEARMAN TEST FOR INDEPENDENCE ---

   05QA001            ENGLISH RIVER NEAR SIOUX LOOKOUT
   ANNUAL MAXIMUM DAILY FLOW SERIES  1922 TO 1981  DRAINAGE AREA =  13570.00

   SPEARMAN RANK ORDER SERIAL CORRELATION COEFF =  .103    D.F.= 57
                      CORRESPONDS TO STUDENTS T =  .782
                   CRITICAL T VALUE AT 5% LEVEL = 1.673    NOT SIGNIFICANT
                       -    -    -    -  1%   -  = 2.395    NOT SIGNIFICANT

   Interpretation: The null hypothesis is that the correlation is zero.

   At the 5% level of significance, the correlation is not significantly
   different from zero.  That is, the data do not display significant
   serial dependence.
```

Table 2: Output of CFA for the Spearman rank order serial correlation coefficient as a test of the independence of the annual maximum daily flows of the English River near Sioux Lookout - 05QA001.

```
        --- SPEARMAN TEST FOR TREND ---

   05QA001            ENGLISH RIVER NEAR SIOUX LOOKOUT
   ANNUAL MAXIMUM DAILY FLOW SERIES  1922 TO 1981  DRAINAGE AREA =  13570.00

   SPEARMAN RANK ORDER CORRELATION COEFF = -.117    D.F.= 58
               CORRESPONDS TO STUDENTS T = -.898
            CRITICAL T VALUE AT 5% LEVEL =-2.002    NOT SIGNIFICANT
                -    -    -    -  1%   -   =-2.664    NOT SIGNIFICANT

   Interpretation: The null hypothesis is that the serial(lag-one) correlation
                   is zero.

   At the 5% level of significance, the correlation is not significantly
   different from zero. That is, the data do not display significant
   trend.
```

Table 3: Output of CFA for the Spearman rank order serial correlation coefficient as a test of the trend of the annual maximum daily flows of the English River near Sioux Lookout - 05QA001.

```
        --- RUN TEST FOR GENERAL RANDOMNESS ---

05QA001           ENGLISH RIVER NEAR SIOUX LOOKOUT
ANNUAL MAXIMUM DAILY FLOW SERIES  1922 TO 1981  DRAINAGE AREA =  13570.00

THE NUMBER OF RUNS ABOVE AND BELOW THE MEDIAN (RUNAB) = 29
      THE NUMBER OF OBSERVATIONS ABOVE THE MEDIAN(N1) = 30
      THE NUMBER OF OBSERVATIONS BELOW THE MEDIAN(N2) = 30

(NOTE: Z IS THE STANDARD NORMAL VARIATE.)

For this test, Z =   .521
Critical Z value at the 5% level = 1.960           NOT SIGNIFICANT

Interpretation: The null hypothesis is that the data are random.

At the 5% level of significance, the null hypothesis cannot be
rejected. That is, the sample is significantly random.
```

Table 4: Output of CFA for the run test for general randomness of the annual maximum daily flows of the English River near Sioux Lookout - 05QA001.

```
   --- MANN-WHITNEY SPLIT SAMPLE TEST FOR HOMOGENEITY ---

05QA001           ENGLISH RIVER NEAR SIOUX LOOKOUT
ANNUAL MAXIMUM FLOW SERIES  1922 TO 1981  DRAINAGE AREA=  13570.00

SPLIT BY TIME SPAN, SUBSAMPLE 1 SAMPLE SIZE= 29
                    SUBSAMPLE 2 SAMPLE SIZE= 31

(NOTE: Z IS THE STANDARD NORMAL VARIATE.)
                         For this test, Z =  -.777
CRITICAL Z VALUE AT 5% SIGNIFICANT LEVEL = -1.645    NOT SIGNIFICANT
   -     -  -   -  1%        -          -   = -2.326    NOT SIGNIFICANT

Interpretation: The null hypothesis is that there is no
                location difference between the two samples.

At the 5% level of significance, there is no significant
location difference between the two samples. That is, they
appear to be from the same population.
```

Table 5: Output of CFA for the Mann-Whitney split sample test for homogeneity of the annual maximum daily flows of the English River near Sioux Lookout - 05QA001.

```
   --- MANN-WHITNEY SPLIT SAMPLE TEST FOR HOMOGENEITY ---

05QA001          ENGLISH RIVER NEAR SIOUX LOOKOUT
ANNUAL MAXIMUM FLOW SERIES  1922 TO 1981   DRAINAGE AREA=  13570.00

SEASONAL SPLIT, SUBSAMPLE 1 IS AUG THROUGH APR  SAMPLE SIZE=  9
                SUBSAMPLE 2 IS MAY THROUGH JUL  SAMPLE SIZE= 51

(NOTE: Z IS THE STANDARD NORMAL VARIATE.)
                        For this test, Z =  -.704
CRITICAL Z VALUE AT 5% SIGNIFICANT LEVEL = -1.645    NOT SIGNIFICANT
   -     -   -   -  1%      -          -  = -2.326    NOT SIGNIFICANT

Interpretation: The null hypothesis is that there is no location
                difference between the two samples.

At the 5% level of significance, there is no significant location
difference between the two samples. That is, they appear to be from the
same population.
```

Table 6: Output of CFA for the Mann-Whitney split sample test for the homogeneity of the seasonal occurrences of the annual maximum daily flows of the English River near Sioux Lookout - 05QA001.

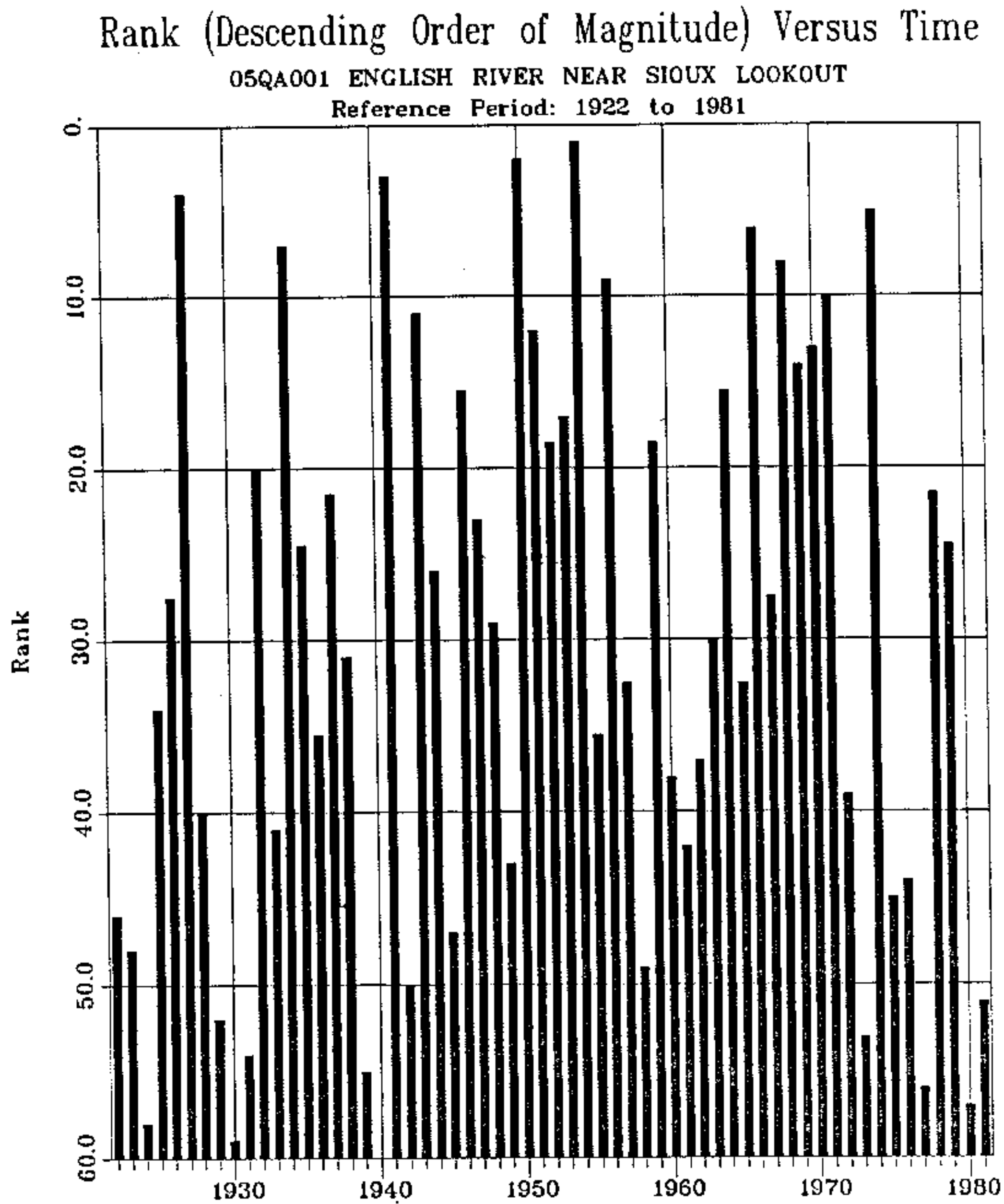

Figure 14: Output of CFA showing the rank-time plot for the English River near Sioux Lookout - 05QA001

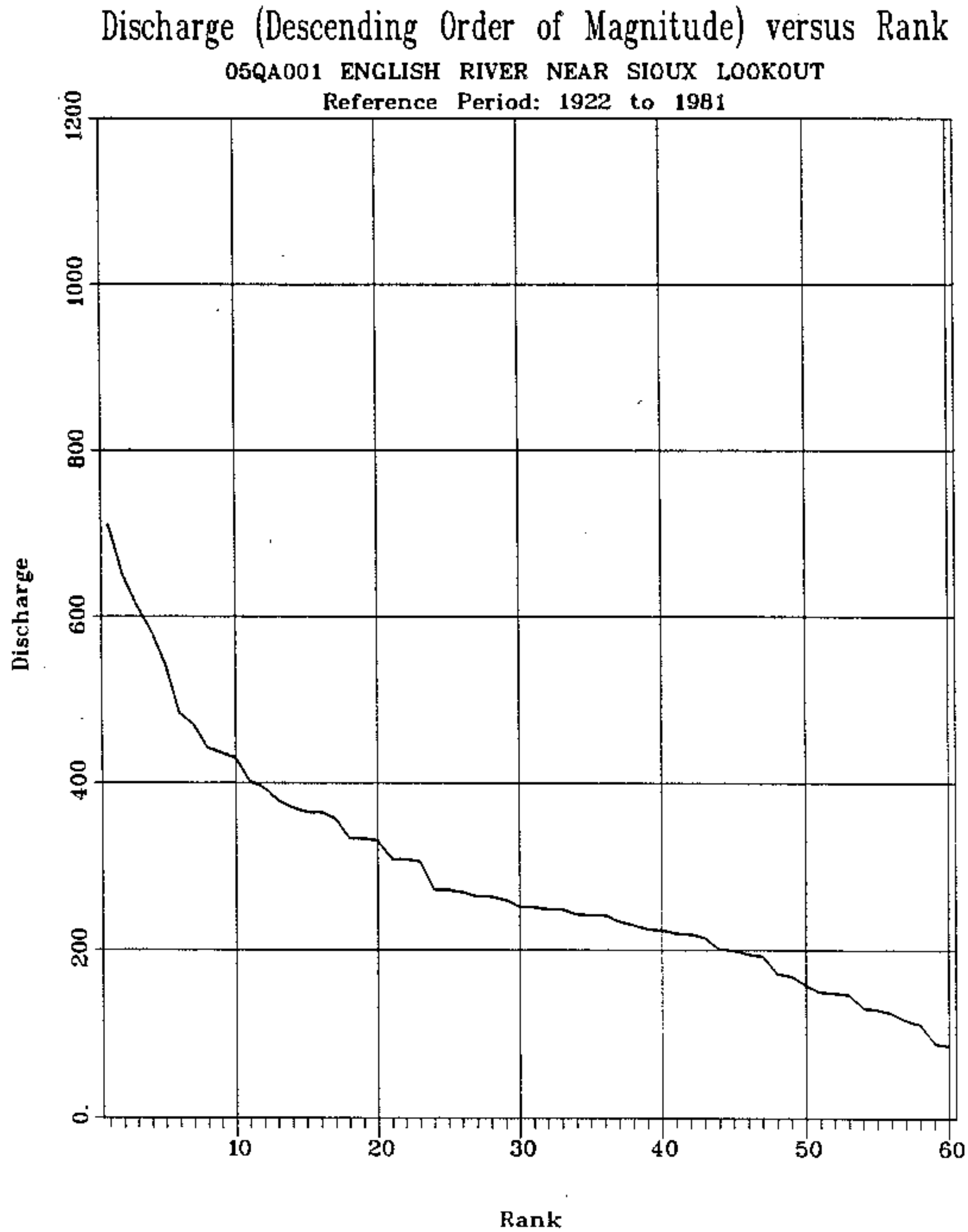

Figure 15: Output of CFA showing the discharge-rank plot for the English River near Sioux Lookout - 05QA001

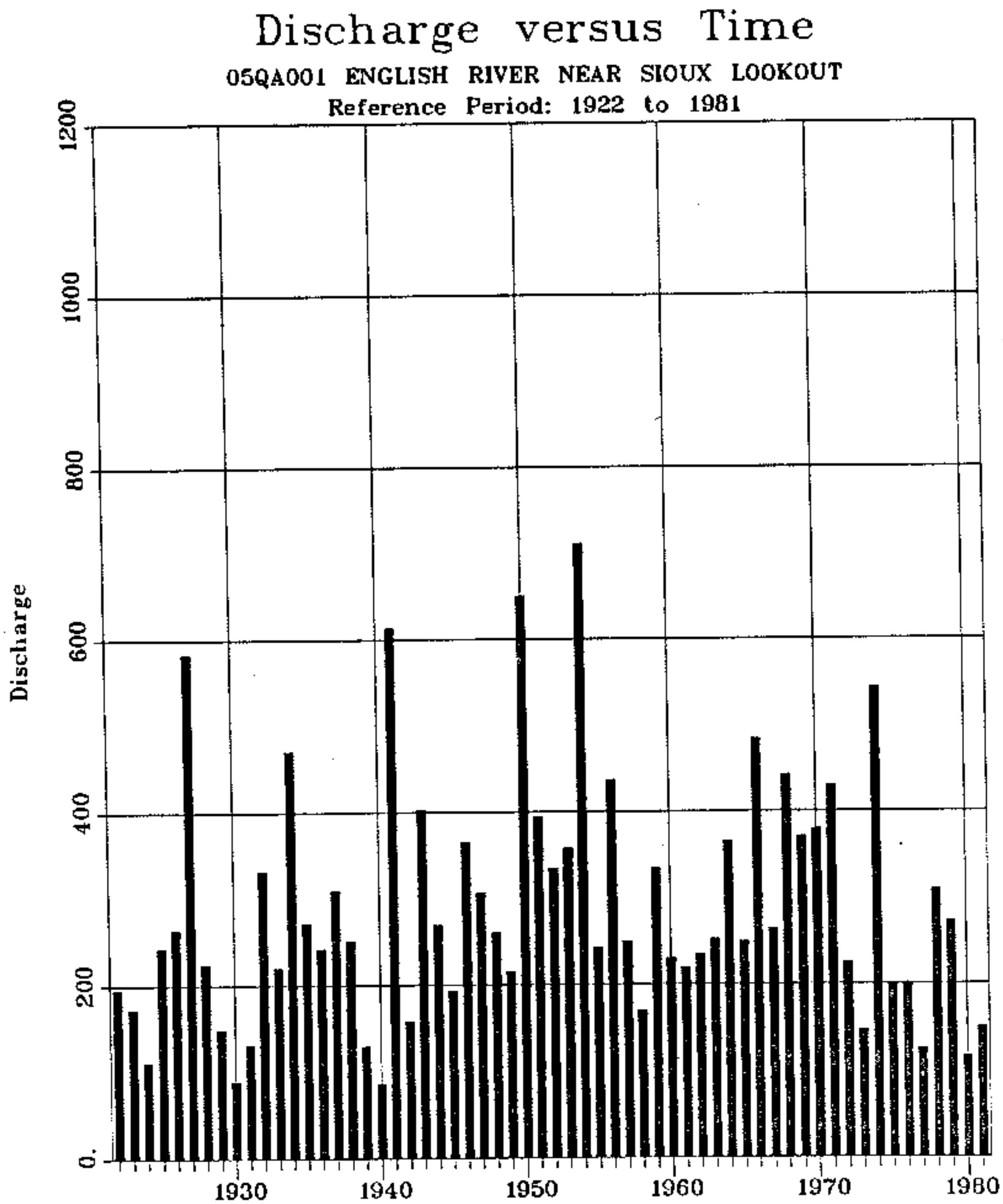

Figure 16: Output of CFA showing the annual maximum daily discharge-time plot for the English River near Sioux Lookout - 05QA001

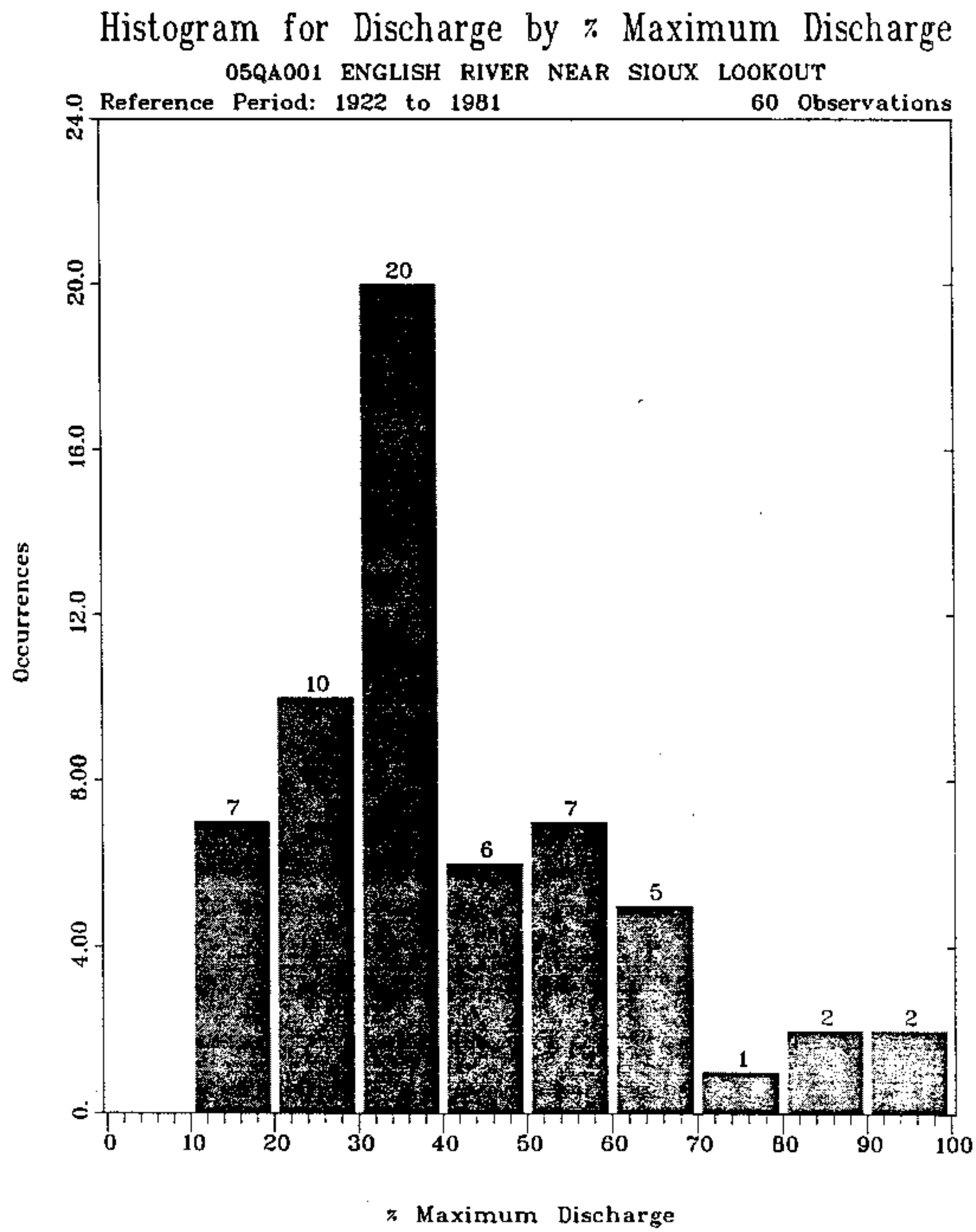

Figure 17: Output of CFA showing the histogram for discharge by % maximum discharge for the English River near Sioux Lookout - 05QA001

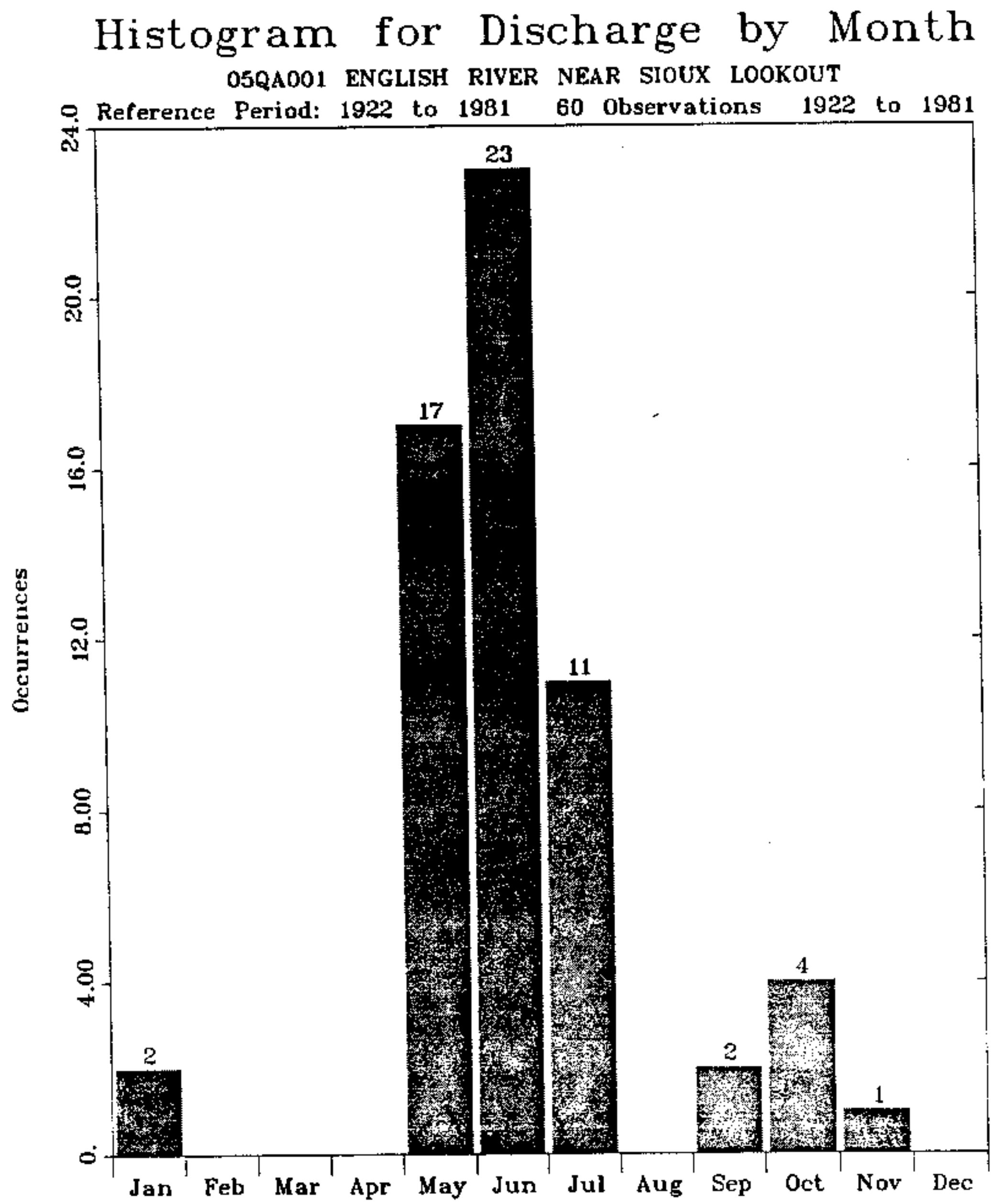

Figure 18: Output of CFA showing the histogram for discharge by month for the English River near Sioux Lookout - 05QA001

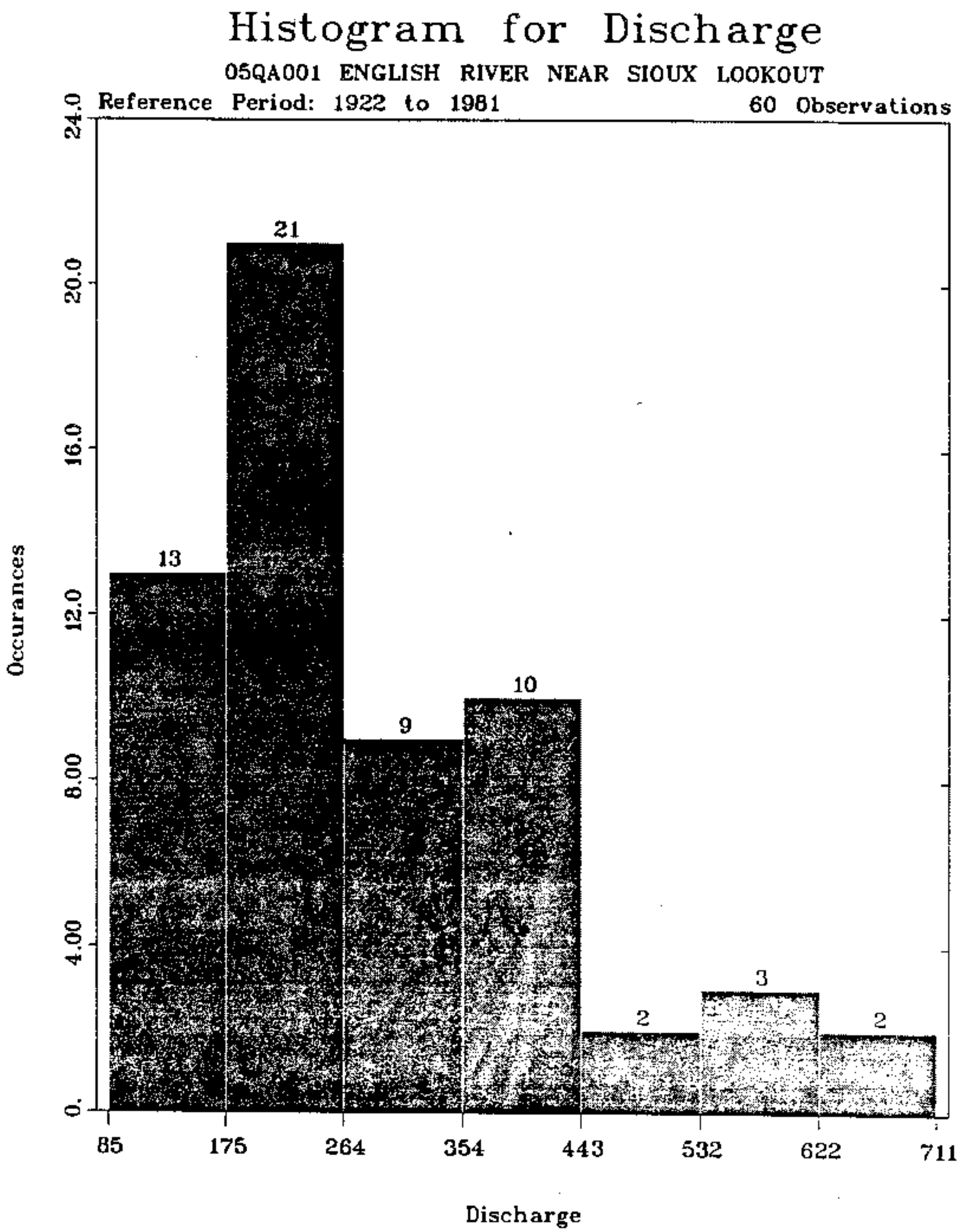

Figure 19: Output of CFA showing the histogram for discharge for the English River near Sioux Lookout - 05QA001

WSC STATION NO=05QA001
WSC STATION NAME=ENGLISH RIVER NEAR SIOUX LOOKOUT

MONTH (1)	YEAR (2)	DATA (3)	ORDERED (4)	RANK (5)	PROB. (6) (%)	RET. PERIOD (7) (YEARS)
5	1922	195.000	711.000	1	1.00	100.333
5	1923	172.000	651.000	2	2.66	37.625
6	1924	111.000	614.000	3	4.32	23.154
6	1925	243.000	583.000	4	5.98	16.722
10	1926	264.000	544.000	5	7.64	13.087
5	1927	583.000	484.000	6	9.30	10.750
9	1928	224.000	470.000	7	10.96	9.121
7	1929	149.000	442.000	8	12.62	7.921
7	1930	88.300	436.000	9	14.29	7.000
7	1931	131.000	430.000	10	15.95	6.271
5	1932	331.000	402.000	11	17.61	5.679
6	1933	220.000	394.000	12	19.27	5.190
6	1934	470.000	379.000	13	20.93	4.778
5	1935	272.000	371.000	14	22.59	4.426
5	1936	242.000	365.000	15	24.25	4.123
6	1937	309.000	365.000	16	25.91	3.859
5	1938	251.000	357.000	17	27.57	3.627
7	1939	129.000	334.000	18	29.24	3.420
7	1940	85.500	334.000	19	30.90	3.237
10	1941	614.000	331.000	20	32.56	3.071
1	1942	158.000	309.000	21	34.22	2.922
7	1943	402.000	309.000	22	35.88	2.787
6	1944	269.000	306.000	23	37.54	2.664
5	1945	193.000	272.000	24	39.20	2.551
5	1946	365.000	272.000	25	40.86	2.447
6	1947	306.000	269.000	26	42.52	2.352
5	1948	260.000	264.000	27	44.19	2.263
5	1949	215.000	264.000	28	45.85	2.181
6	1950	651.000	260.000	29	47.51	2.105
6	1951	394.000	252.000	30	49.17	2.034
7	1952	334.000	251.000	31	50.83	1.967
7	1953	357.000	249.000	32	52.49	1.905
6	1954	711.000	249.000	33	54.15	1.847
6	1955	242.000	243.000	34	55.81	1.792
6	1956	436.000	242.000	35	57.48	1.740
5	1957	249.000	242.000	36	59.14	1.691
10	1958	169.000	234.000	37	60.80	1.645
6	1959	334.000	230.000	38	62.46	1.601
6	1960	230.000	225.000	39	64.12	1.560
6	1961	219.000	224.000	40	65.78	1.520
9	1962	234.000	220.000	41	67.44	1.483
7	1963	252.000	219.000	42	69.10	1.447
5	1964	365.000	215.000	43	70.76	1.413
6	1965	249.000	201.000	44	72.43	1.381
6	1966	484.000	200.000	45	74.09	1.350
6	1967	264.000	195.000	46	75.75	1.320
6	1968	442.000	193.000	47	77.41	1.292
6	1969	371.000	172.000	48	79.07	1.265

Table 7: Output of CFA listing the station's flows, ranked flows, probability, and return period as obtained from the Cunnane formula for the English River near Sioux Lookout - 05QA001.

WSC STATION NO=05QA001
WSC STATION NAME=ENGLISH RIVER NEAR SIOUX LOOKOUT

MONTH (1)	YEAR (2)	DATA (3)	ORDERED (4)	RANK (5)	PROB. (6) (%)	RET. PERIOD (7) (YEARS)
6	1970	379.000	169.000	49	80.73	1.239
11	1971	430.000	158.000	50	82.39	1.214
1	1972	225.000	150.000	51	84.05	1.190
10	1973	147.000	149.000	52	85.71	1.167
6	1974	544.000	147.000	53	87.38	1.144
5	1975	200.000	131.000	54	89.04	1.123
5	1976	201.000	129.000	55	90.70	1.103
7	1977	125.000	125.000	56	92.36	1.083
6	1978	309.000	116.000	57	94.02	1.064
5	1979	272.000	111.000	58	95.68	1.045
5	1980	116.000	88.300	59	97.34	1.027
7	1981	150.000	85.500	60	99.00	1.010

Table 7 continued

```
        FREQUENCY ANALYSIS - GENERALIZED EXTREME VALUE DISTRIBUTION
        05QA001      ENGLISH RIVER NEAR SIOUX LOOKOUT


                              SAMPLE STATISTICS

                      MEAN         S.D.          C.V.        C.S.        C.K.
   X SERIES        288.947      140.700          .487       1.058       4.105
LN X SERIES          5.554         .485          .087       -.124       2.936
L-MOM RATIO        288.947       76.800          .266        .212        .166


X(MIN)=      85.500                           TOTAL SAMPLE SIZE=  60
X(MAX)=     711.000                         NO. OF LOW OUTLIERS=   0
LOWER OUTLIER LIMIT OF X=   65.260            NO. OF ZERO FLOWS=   0


          SOLUTION OBTAINED VIA L - MOMENTS


        GEV PARAMETERS:      U=     221.45   A=  103.646   K=    -.070


                          FLOOD FREQUENCY REGIME

                RETURN        EXCEEDANCE         FLOOD
                PERIOD        PROBABILITY

                 1.003           .997             49.8
                 1.050           .952            110
                 1.250           .800            173
                 2.000           .500            260
                 5.000           .200            385
                10.000           .100            474
                20.000           .050            564
                50.000           .020            686
               100.000           .010            784
               200.000           .005            886
               500.000           .002           1030
```

Table 8: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the English River near Sioux Lookout (05QA001) for the generalized extreme value distribution.

```
        FREQUENCY ANALYSIS - THREE-PARAMETER LOGNORMAL DISTRIBUTION
        05QA001     ENGLISH RIVER NEAR SIOUX LOOKOUT

                              SAMPLE STATISTICS

                    MEAN         S.D.          C.V.        C.S.         C.K.
     X SERIES     288.947     140.700         .487       1.058        4.105
   LN X SERIES      5.554        .485         .087       -.124        2.936
LN(X-A) SERIES      5.611        .458         .082       -.056        2.896

   X(MIN)=      85.500                            TOTAL SAMPLE SIZE=  60
   X(MAX)=     711.000                          NO. OF LOW OUTLIERS=   0
   LOWER OUTLIER LIMIT OF X=   65.260             NO. OF ZERO FLOWS=   0


          SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

          3LN PARAMETERS:      A=    -13.726   M= 5.611    S=  .458


                            FLOOD FREQUENCY REGIME

                  RETURN        EXCEEDANCE          FLOOD
                  PERIOD        PROBABILITY

                   1.003           .997              64.0
                   1.050           .952             114
                   1.250           .800             172
                   2.000           .500             260
                   5.000           .200             388
                  10.000           .100             478
                  20.000           .050             567
                  50.000           .020             686
                 100.000           .010             779
                 200.000           .005             875
                 500.000           .002            1010
```

Table 9: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the English River near Sioux Lookout (05QA001) for the three-parameter lognormal distribution.

```
        FREQUENCY ANALYSIS - LOG PEARSON TYPE III DISTRIBUTION
        05QA001     ENGLISH RIVER NEAR SIOUX LOOKOUT


                              SAMPLE STATISTICS

                    MEAN        S.D.          C.V.        C.S.         C.K.
   X SERIES      288.947     140.700          .487       1.058        4.105
LN X SERIES        5.554        .485          .087       -.124        2.936

X(MIN)=      85.500                          TOTAL SAMPLE SIZE=  60
X(MAX)=     711.000                        NO. OF LOW OUTLIERS=   0
LOWER OUTLIER LIMIT OF X=   65.260           NO. OF ZERO FLOWS=   0


          SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD


          DISTRIBUTION IS UPPER BOUNDED AT M=  .1050E+06
 LP3 PARAMETERS: A= -.3852E-01 B=  156.0      LOG(M)=  11.56
                                                  M =  .1050E+06


                           FLOOD FREQUENCY REGIME

                RETURN         EXCEEDANCE          FLOOD
                PERIOD         PROBABILITY

                 1.003            .997              63.2
                 1.050            .952             113
                 1.250            .800             173
                 2.000            .500             261
                 5.000            .200             388
                10.000            .100             474
                20.000            .050             557
                50.000            .020             665
               100.000            .010             747
               200.000            .005             829
               500.000            .002             940
```

Table 10: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the English River near Sioux Lookout (05QA001) for the Log Pearson Type III distribution.

```
       FREQUENCY ANALYSIS - WAKEBY DISTRIBUTION
       05QA001     ENGLISH RIVER NEAR SIOUX LOOKOUT

                               SAMPLE STATISTICS

                    MEAN        S.D.          C.V.        C.S.        C.K.
   X SERIES       288.947    140.700         .487       1.058       4.105
LN X SERIES         5.554       .485         .087       -.124       2.936
L-MOM RATIO       288.947     76.800         .266        .212        .166


X(MIN)=      85.500                               TOTAL SAMPLE SIZE=  60
X(MAX)=     711.000                             NO. OF LOW OUTLIERS=   0
LOWER OUTLIER LIMIT OF X=   65.260                 NO. OF ZERO FLOWS=   0


THE FOLLOWING WAKEBY PARAMETERS WERE OBTAINED VIA L-MOMENTS

M=    58.043    A=      96.191    B= 10.18   C=  -2790.846   D= -.054
          DISTRIBUTION IS UPPER BOUNDED AT E=  .2945E+04


                              FLOOD FREQUENCY REGIME

                    RETURN        EXCEEDANCE          FLOOD
                    PERIOD        PROBABILITY

                     1.003            .997             61.4
                     1.050            .952            103
                     1.250            .800            178
                     2.000            .500            257
                     5.000            .200            387
                    10.000            .100            481
                    20.000            .050            572
                    50.000            .020            687
                   100.000            .010            770
                   200.000            .005            850
                   500.000            .002            952
```

Table 11: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the English River near Sioux Lookout (05QA001) for the Wakeby distribution.

```
FREQUENCY ANALYSIS - NONPARAMETRIC METHOD
05QA001      ENGLISH RIVER NEAR SIOUX LOOKOUT


                          SAMPLE STATISTICS

                  MEAN        S.D.        C.V.        C.S.        C.K.
   X SERIES     288.947     140.700       .487       1.058       4.105
LN X SERIES       5.554        .485       .087       -.124       2.936

X(MIN)=      85.500                          TOTAL SAMPLE SIZE=  60
X(MAX)=     711.000                          NO. OF ZERO FLOWS=   0

      SMOOTHING PARAMETER H =    52.130

                FLOOD FREQUENCY REGIME

      RETURN        EXCEEDANCE          FLOOD
      PERIOD        PROBABILITY

       1.003            .997              3.25
       1.050            .952             86.1
       1.250            .800            167
       2.000            .500            263
       5.000            .200            399
      10.000            .100            495
      20.000            .050            598
      50.000            .020            677
     100.000            .010            716
     200.000            .005            745
     500.000            .002            775
```

Table 12: Output of CFA listing the summary statistics of the sample, the computed smoothing parameter H, and the tabular flood frequency regime of the English River near Sioux Lookout (05QA001) for the Nonparametric Method.

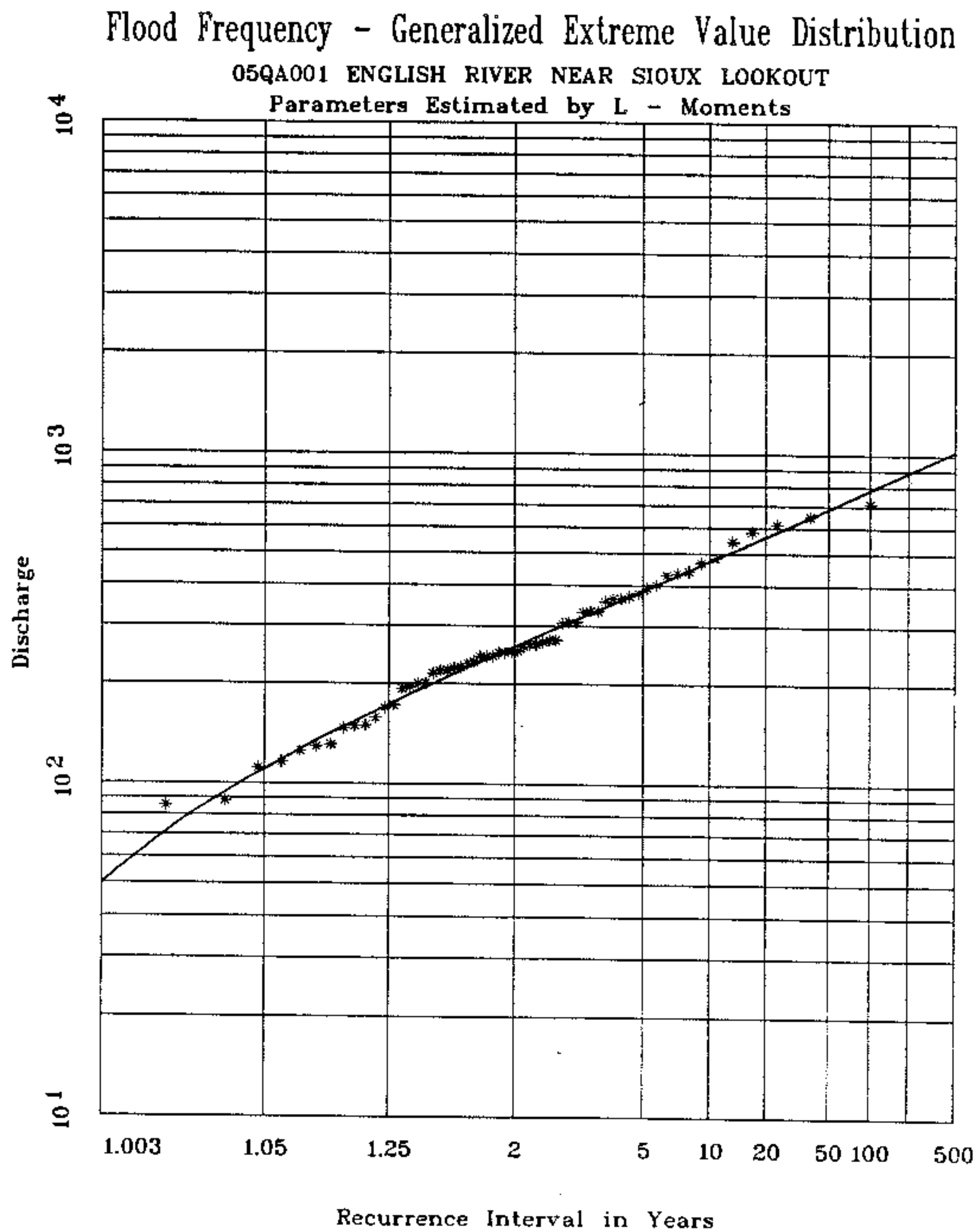

Figure 20: Output of CFA showing the frequency plot based on the generalized extreme value distribution for the English River near Sioux Lookout - 05QA001

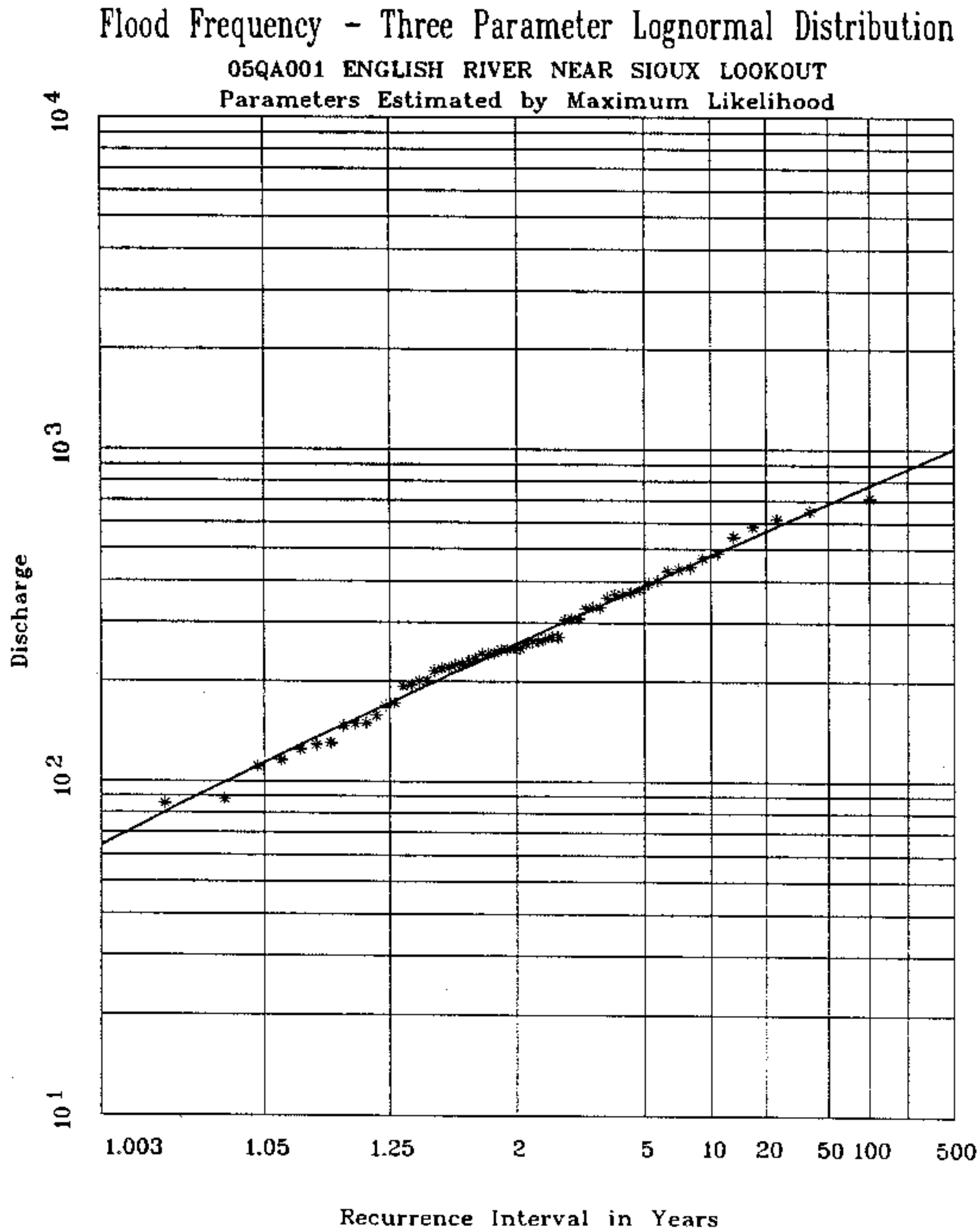

Figure 21: Output of CFA showing the frequency plot based on the three-parameter lognormal distribution for the English River near Sioux Lookout - 05QA001

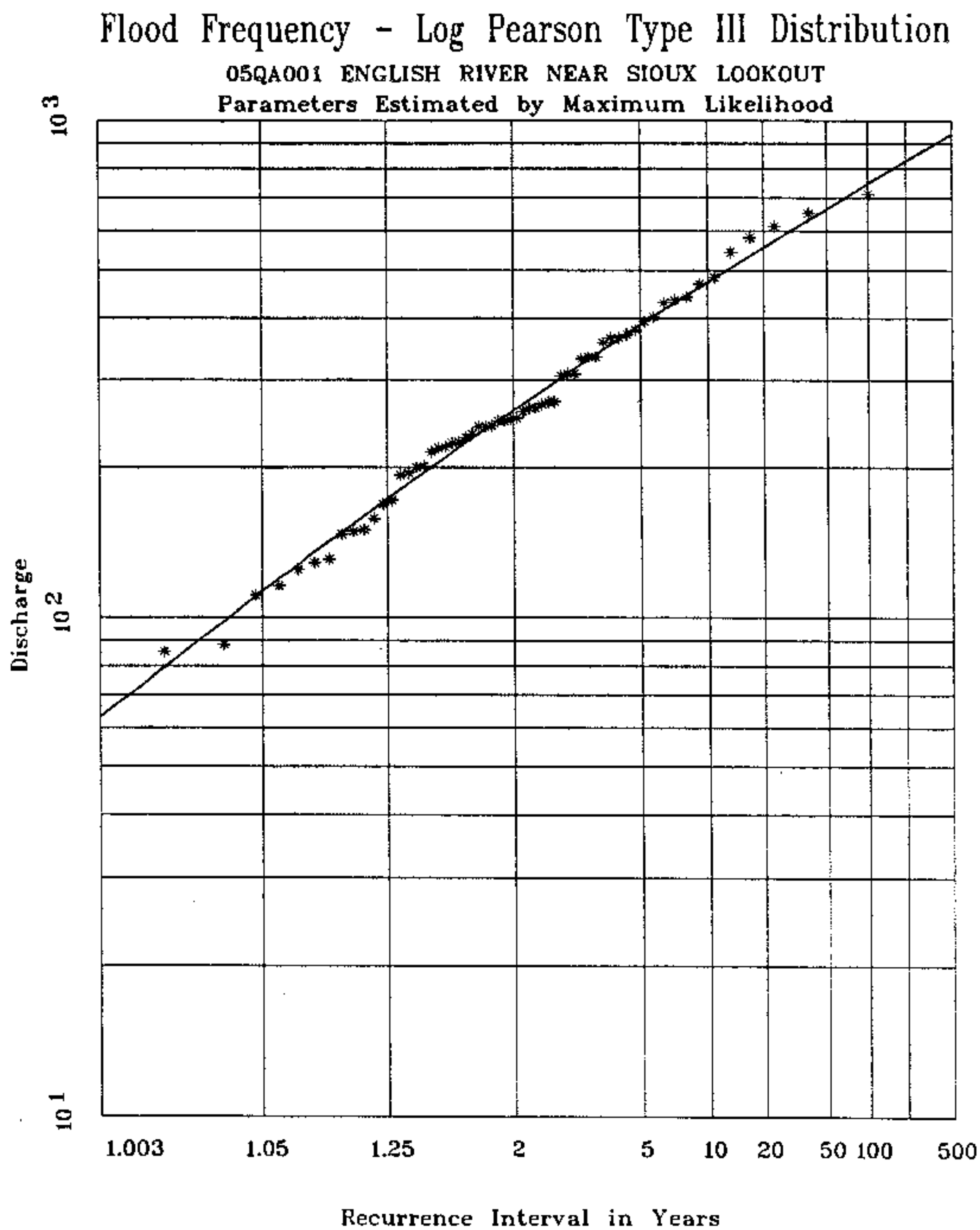

Figure 22: Output of CFA showing frequency plot based on the Log-Pearson Type III distribution for the English River near Sioux Lookout - 05QA001

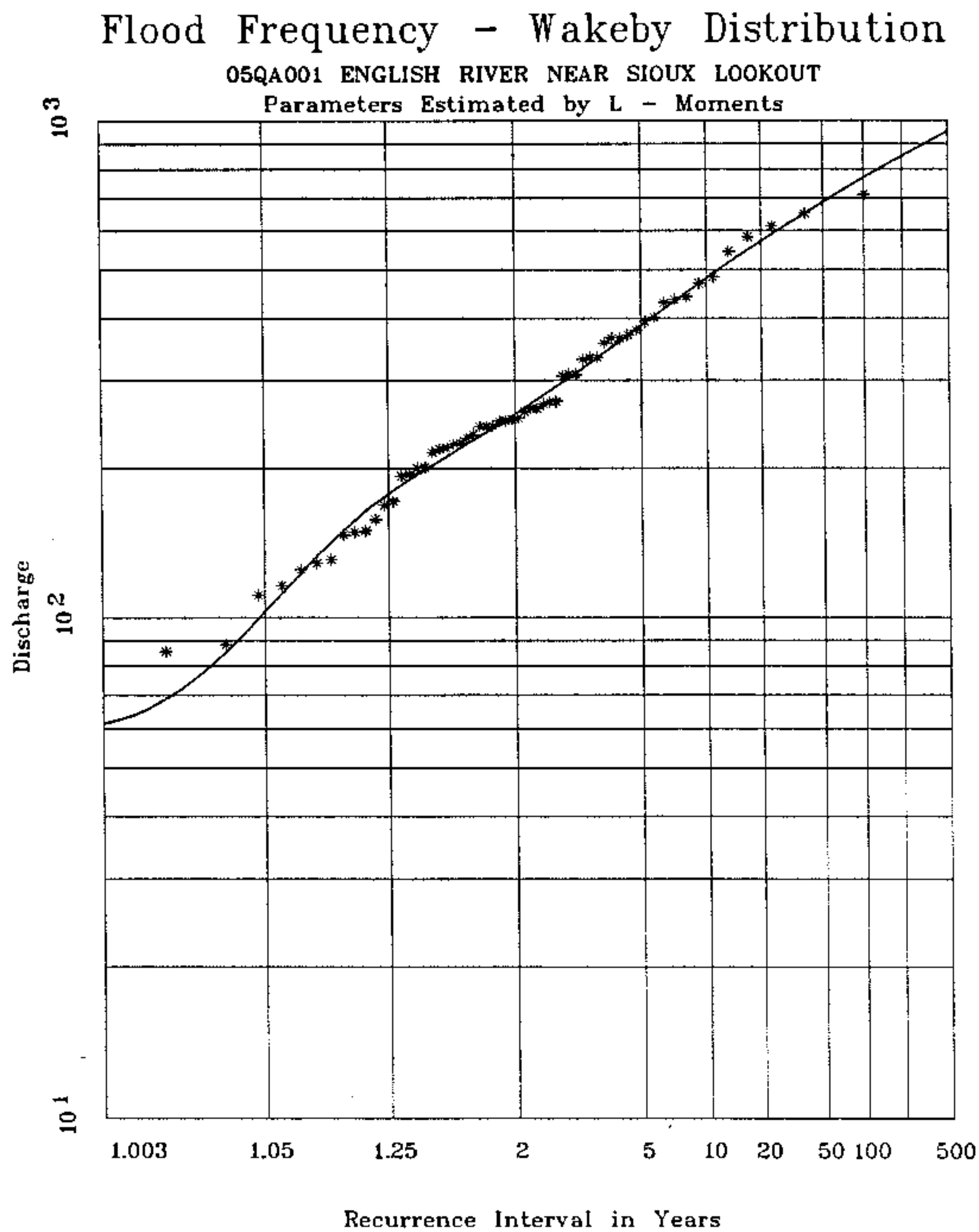

Figure 23: Output of CFA showing the frequency plot based on the Wakeby distribution for the English River near Sioux Lookout - 05QA001

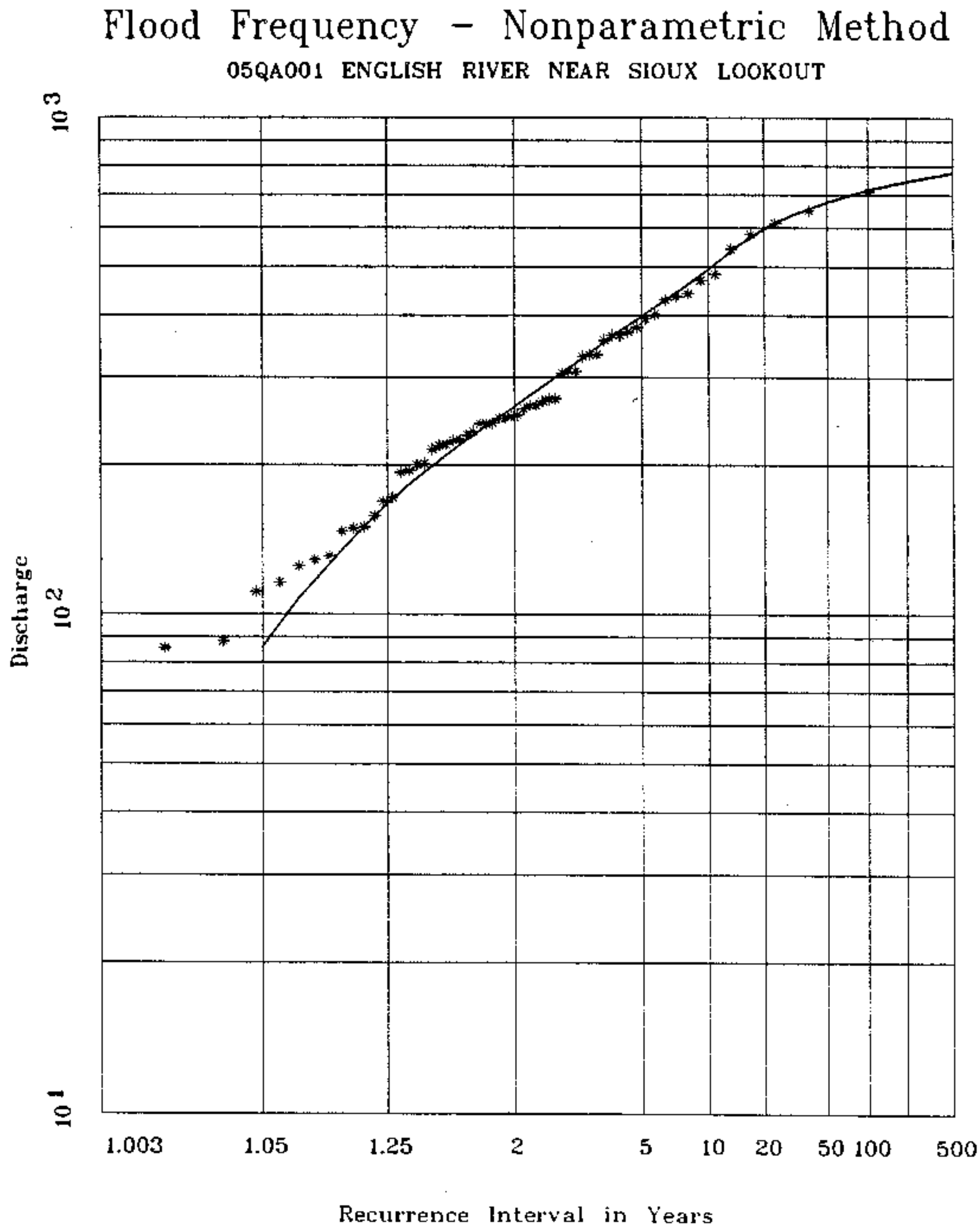

Figure 24: Output of CFA showing the frequency plot based on Nonparametric Method for the English River near Sioux Lookout - 05QA001

5.3 HISTORICAL ANALYSIS - BOYNE RIVER NEAR CARMAN - O50F003

The Boyne River near Carman gauging station was established in 1919 and was still active in 1982, although the annual floods for the years 1920, 1921, 1923 through 1926, 1929, and 1931 through 1955 are missing. There is also historic information that a flood of 1 897 m^3/s occurred in 1893 and is the largest on record.

Tables 13 to 22 and Figures 25 to 35 comprise the output for the historical analysis. Tables 13 to 16 list the results of the nonparametric tests performed on the series. Summaries of the statistical analysis are provided in each of these tables. Figures 25 to 30 show the graphical analysis available using the "Screen Data Menu".

Figure 29 shows the histogram for discharge by month. There appears only one grouping of observations in Figure 29, thus no homogeneity test based on second groupings is performed. Figures 25 and 27 shows little evidence of trend or heteroscedasticity in the series, although figure 27 does show an increasing trend in the annual maximum discharges from 1963 to about 1970. Table 16 summarizes the results of a Mann-Whitney U-test performed on the annual maximum daily flows of 1893 to 1965 and 1966 to 1982. No significant location difference was found. That is, the two samples appear to have been drawn from the same population. In addition, the nonparametric tests summarized in Tables 13 and 14 found no significant dependency or trend in the sample. However, Figure 26 shows an abrupt change in shape between ranks 4 and 5. Such an abrupt change may indicate the existence of mixed populations and/or the presence of historic highs. The causation factors for the floods above and below the break may be different. A distinct break in the slope of the line is an indicator that the frequency distributions of this program may not fit the sample well if historic information is not found for the high flow values. In addition, the lowest ranked observation dips below the trend of the line. This point later proves to be a low outlier based on the Grubbs and Beck test. Nonparametric frequency analysis may yield superior results in such cases. Also, note that the outlier analysis is not performed when this approach is taken.

Note that the ranks of Figures 25 and 26 are not adjusted for historic information. If observations in the series have floods of equal size, they are given a tied adjusted rank value equal to the arithmetic average of their unadjusted ranks:

e.g.

Flood	Rank	Adjusted Rank
.	.	.
.	.	.
.	.	.
1 300	20	20
1 200	21	21.5
1 200	22	21.5
1 100	23	23
1 000	24	25
1 000	25	25
1 000	26	25
800	27	27
.	.	.
.	.	.
.	.	.

Figure 28 shows the histogram for discharge by % of the maximum discharge. This plot shows that the maximum observed flood is much larger than the remainder of the sample. Figure 30 shows the conventional histogram for discharge. The sample appears positively skewed with the maximum flood being located quite distant from the remainder of the sample. This usually is an indication that the sample has a high outlier and that the skewness will be significantly different from zero. In such a case, the data should be checked for accuracy. If the maximum observation is not erroneous, then this flood should be investigated for historic significance. It may be recalled that historic information was found for this event. Unfortunately, Figure 30 does not display a classical unimodal histogram. This is not surprising due to the low number of observations available (33) for histogram construction.

Table 17 represents the first of three sections of output from the "Frequency Analysis" selection of the main menu. The first section, after titling, gives information on the censoring parameters that were described in Section 3.6. Relating them to the example: since the 1893 flood was the largest known from 1893 to 1982, the total time span, YT of the analysis is 90 years, and since no flood during that period exceeded 187 m^3/s, the censoring threshold is 187 m^3/s. The number of observed peaks is N = 33. The three aforementioned censoring parameters should be carefully checked as the computations will be affected if they have been wrongly entered. Since there is only one historic peak above the threshold, NHA = 1. The remaining information about the sample, such as the observed peaks above the threshold, NA, observed peaks below the threshold, NB, and the missing peaks below the threshold, NC, are all calculated by the program.

The first three columns of Table 17 give the month, year, and magnitude of the flood. The fourth and fifth columns rank the data in descending order of magnitude and a rank number is assigned, starting with 1 for the largest. Column (6) assigns an adjusted rank as suggested by Benson (1950). Columns (7) and (8) give the empirical values of exceedance probability and return period according to Cunnane (1978). Note that the lowest observed flood has been identified as a low outlier, and is denoted by an asterisk.

The second section of the output is listed in Tables 18 to 22. This section gives the values of the population moments as estimated from the sample of fully specified floods only. They are given for the untransformed variate and its logarithmic transform. If the output is for the generalized extreme value and Wakeby distributions, then the L-moments and L-moment ratios are given for the sample. If the output is for the three-parameter lognormal distribution, then the logarithmic transform with the lower boundary parameter is also given. Next comes the sample's minimum and maximum values, the lower outlier limit, the total sample size, the number of low outliers (not given for the nonparametric method), and the number of zero flows. The solution method (e.g. moments) is given, followed by the estimate of the parameters of the distribution. Table 21 lists that the Wakeby distribution is upper bounded at 416. This upper boundary appears large in comparison with the 500 year event. The frequency plot, shown in Figure 34, can be used to glean further information as to the practical limitations of the distribution for this sample. The upper boundary does not appear to affect the shape of the frequency curve in our range of interest. Finally, the tabular flood frequency regime is given for various preselected return periods and exceedance probabilities.

The third and final section of the output consists of the plots of the frequency distributions and nonparametric density estimates on lognormal probability scale. Figures 31 to 35 shows these plots for our example. Note that the cumulative density plot of Figure 35 has been suppressed for return periods less than 1.05 years.

The generalized extreme value and the Wakeby distributions and the nonparametric method give similar design floods for the higher return periods. The three-parameter lognormal and the Log Pearson Type III, as well, give similar design floods for the higher return periods. However, both of these groups give distinctively different results. The upper bound on the Wakeby is 416, while the 500 year event for the second group is 409. Thus, a more conservative estimate of the extreme events is obtained using the three-parameter lognormal and the Log Pearson Type III distributions for this sample. (Note that more conservative does not imply more accurate nor better.)

However, at this point it would be beneficial to review Figure 26, which shows the "Screen Data" display of discharge versus rank. Remember, that an abrupt change in the slope of the line was noted between ranks 4 and 5. Table 17 lists the ordered data, and it is evident that the four highest floods are quite different in magnitude than the remainder of the sample.

The fourth highest flood occurred in 1970 and represents the highest recorded flood to that point in time, except for the historic event of 1893. An investigation found that the flood of 1970 was indeed the worst event since that of 1893. This additional historic information permits the censoring threshold, X_c, to be lowered to 105 m^3/s from 187 m^3/s. Four floods are then larger than or equal to the new censoring threshold.

An analysis with the additional historic information was performed. Table 23 lists the station's flows, probability, and return periods. Note how the historic adjustment is different from that of Table 17 due to the lowering of the censoring threshold. Tables 24 through 28 list the estimates of the parameters based on the additional historic information and certain information of the Wakeby distribution. Note that the Wakeby distribution is unbounded above when the censoring threshold is 105 m^3/s, but is bounded above when the censoring threshold is 187 m^3/s.

Figures 36 through 40 show the plots of the frequency distributions and nonparametric methods on lognormal probability scale. Note that the lower portion of the cumulative frequency curve of Figure 36 is suppressed at 1.1 years and Figure 40 at 1.05 years so that the number of cycles in the figures remains constant. From a visual inspection of these figures and the flood frequency regimes of Tables 24 to 28, it is obvious that the T-year flood estimates of the different distributions correspond more closely than when the censoring threshold was set at 187 m^3/s. In addition, the refinement of the historic information has yield distributions that more closely resemble the estimated plotting positions of the data. Figure 36 shows that the generalized extreme value distribution may be underestimating the extreme events. Figure 37 shows that the three-parameter lognormal distribution may be slightly overestimating the extreme T-year floods, while Figure 38 shows that the Log Pearson Type III distribution corresponds very clearly with the plotted data. The Wakeby plot, shown in Figure 39, appears to be slightly under the extreme data. Figure 40 shows the nonparametric frequency plot. This method again appears to give poor estimates beyond the probability associated with the largest flood. In this case, the nonparametric approach appears to underestimate the design floods for return periods of larger than 150 years; however, the flexibility of the nonparametric density is evident. The estimates of the 100-year flood for these four distributions and the nonparametric approach are 138, 185, 171, 162, and 181 m^3/s, respectively.

Statistical frequency analysis assumes that the sample to be analyzed is a reliable set of measurements of independent random events from a homegeous population. The analyst should be cautious that climate or land-use changes may go undetected during the ungauged portion of the historic record. A non-homogenous or time-variant sample may result from such causation factors, possibly invalidating the historic frequency analysis. It could be postulated that for non-homogenous or mixed distributions that the Wakeby and, in particular, the nonparametric method would yield more accurate estimates of flood quantiles than those obtained by use of the three parameter distribution.

```
        --- SPEARMAN TEST FOR INDEPENDENCE ---

  05OF003            BOYNE RIVER NEAR CARMAN
  ANNUAL MAXIMUM DAILY FLOW SERIES  1893 TO 1982  DRAINAGE AREA =  976.0000

  SPEARMAN RANK ORDER SERIAL CORRELATION COEFF =  .100    D.F.= 25
                    CORRESPONDS TO STUDENTS T =  .500
                 CRITICAL T VALUE AT 5% LEVEL = 1.708    NOT SIGNIFICANT
                     -    -    -    - 1%    -   = 2.485    NOT SIGNIFICANT

  Interpretation: The null hypothesis is that the correlation is zero.

  At the 5% level of significance, the correlation is not significantly
  different from zero.  That is, the data do not display significant
  serial dependence.
```

Table 13: Output of CFA for the Spearman rank order serial correlation coefficient as a test of the independence of the annual maximum daily flows of the Boyne River near Carman - 05OF003.

```
        --- SPEARMAN TEST FOR TREND ---

  05OF003            BOYNE RIVER NEAR CARMAN
  ANNUAL MAXIMUM DAILY FLOW SERIES  1893 TO 1982  DRAINAGE AREA =  976.0000

  SPEARMAN RANK ORDER CORRELATION COEFF =  .138    D.F.= 31
             CORRESPONDS TO STUDENTS T =  .778
          CRITICAL T VALUE AT 5% LEVEL = 2.040    NOT SIGNIFICANT
              -    -    -    - 1%    -   = 2.745    NOT SIGNIFICANT

  Interpretation: The null hypothesis is that the serial(lag-one) correlation
                  is zero.

  At the 5% level of significance, the correlation is not significantly
  different from zero. That is, the data do not display significant
  trend.
```

Table 14: Output of CFA for the Spearman rank order serial correlation coefficient as a test of the trend of the annual maximum daily flows of the Boyne River near Carman - 05OF003.

```
        --- RUN TEST FOR GENERAL RANDOMNESS ---

050F003           BOYNE RIVER NEAR CARMAN
ANNUAL MAXIMUM DAILY FLOW SERIES  1893 TO 1982  DRAINAGE AREA =  976.0000

THE NUMBER OF RUNS ABOVE AND BELOW THE MEDIAN (RUNAB) = 18
      THE NUMBER OF OBSERVATIONS ABOVE THE MEDIAN(N1) = 16
      THE NUMBER OF OBSERVATIONS BELOW THE MEDIAN(N2) = 16
Range at 5% level of significance: 12. to 22.    NOT SIGNIFICANT

Interpretation: The null hypothesis is that the data are random.

At the 5% level of significance, the null hypothesis cannot be
rejected. That is, the sample is significantly random.
```

Table 15: Output of CFA for the run test for general randomness of the annual maximum daily flows of the Boyne River near Carman - 05OF003.

```
   --- MANN-WHITNEY SPLIT SAMPLE TEST FOR HOMOGENEITY ---

050F003           BOYNE RIVER NEAR CARMAN
ANNUAL MAXIMUM FLOW SERIES  1893 TO 1982   DRAINAGE AREA=  976.0000

SPLIT BY TIME SPAN, SUBSAMPLE 1 SAMPLE SIZE= 16
                    SUBSAMPLE 2 SAMPLE SIZE= 17

                                 MANN-WHITNEY U = 131.0
     CRITICAL U VALUE AT 5% SIGNIFICANT LEVEL =  89.0    NOT SIGNIFICANT
      -   -   -   -   -  1%      -           -   =  71.0    NOT SIGNIFICANT

Interpretation: The null hypothesis is that there is no
                location difference between the two samples.

At the 5% level of significance, there is no significant
location difference between the two samples. That is, they
appear to be from the same population.
```

Table 16: Output of CFA for the Mann-Whitney split sample test for homogeneity of the annual maximum daily flows of the Boyne River near Carman - 05OF003.

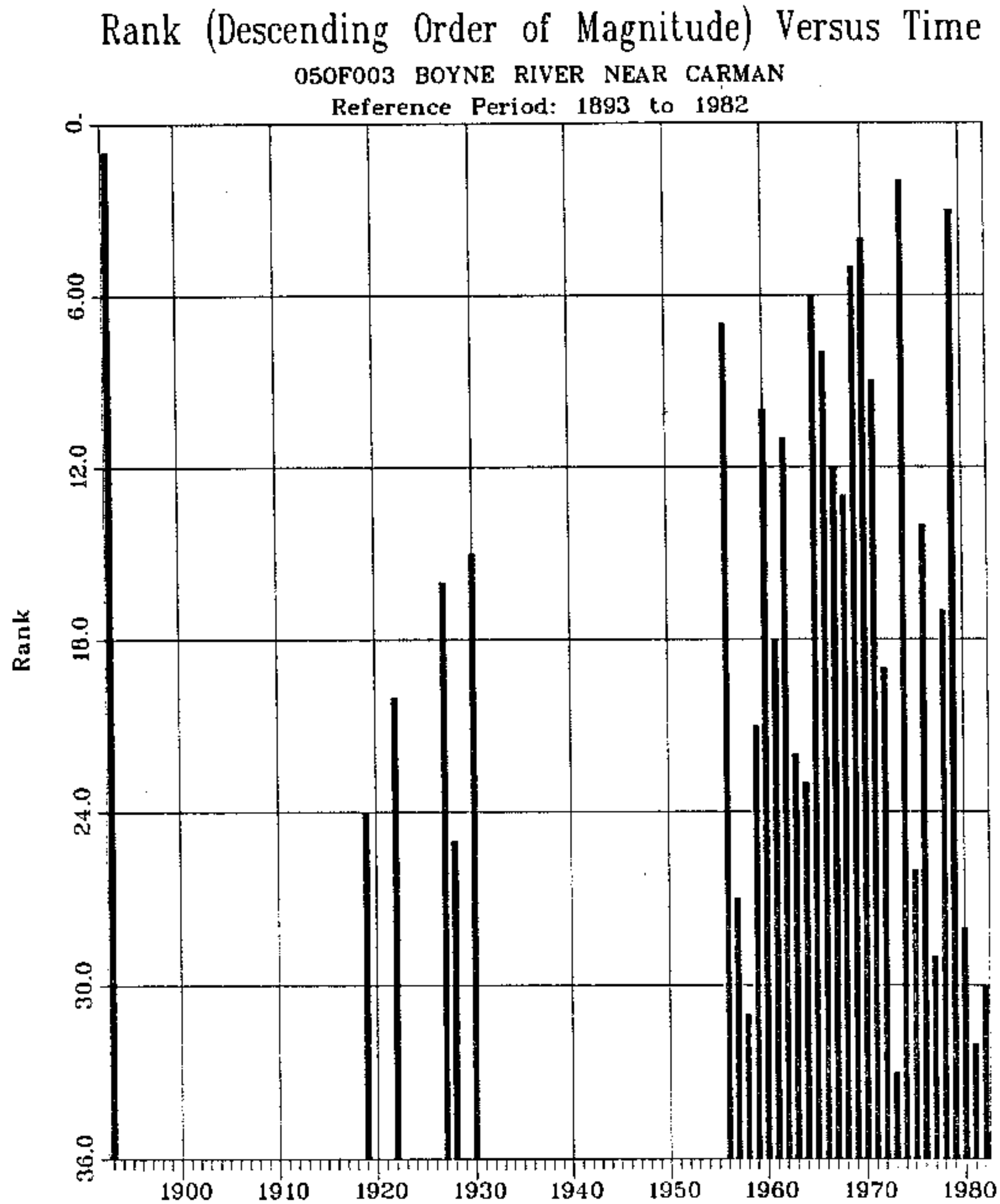

Figure 25: Output of CFA showing the rank-time plot for the Boyne River near Carman - 05OF003

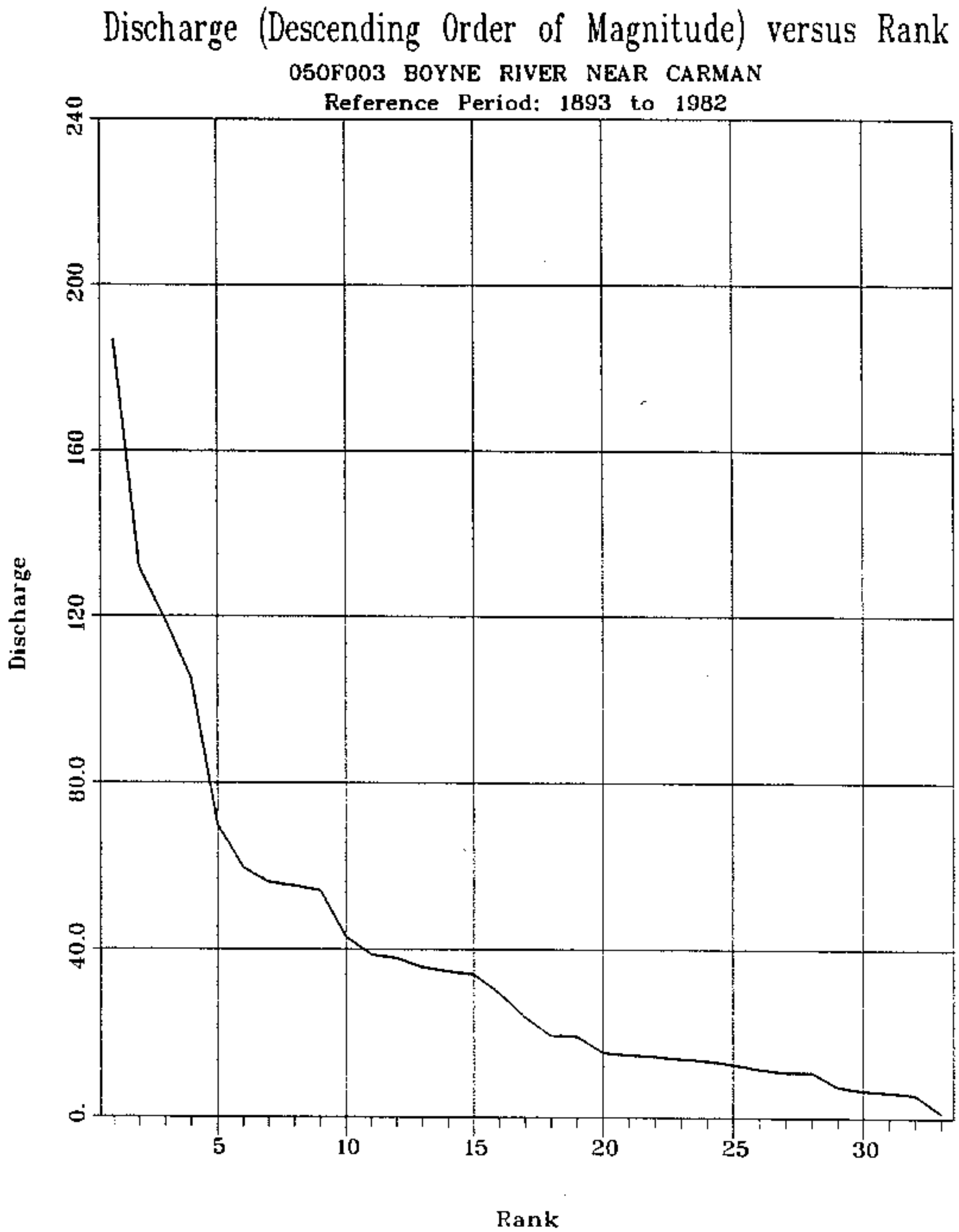

Figure 26: Output of CFA showing the discharge-rank plot for the Boyne River near Carman - 05OF003

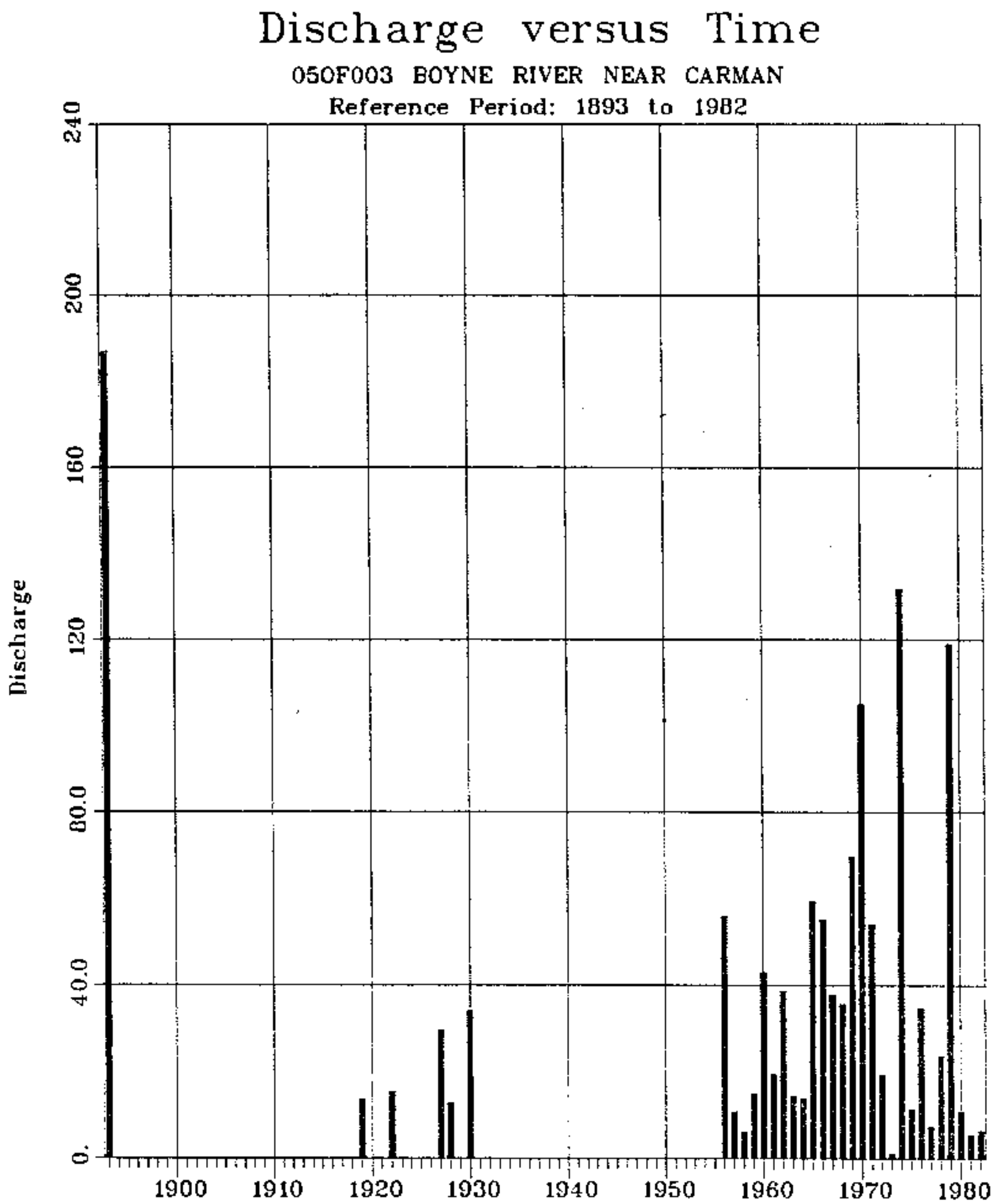

Figure 27: Output of CFA showing the annual maximum daily discharge-time plot for the Boyne River near Carman - 05OF003

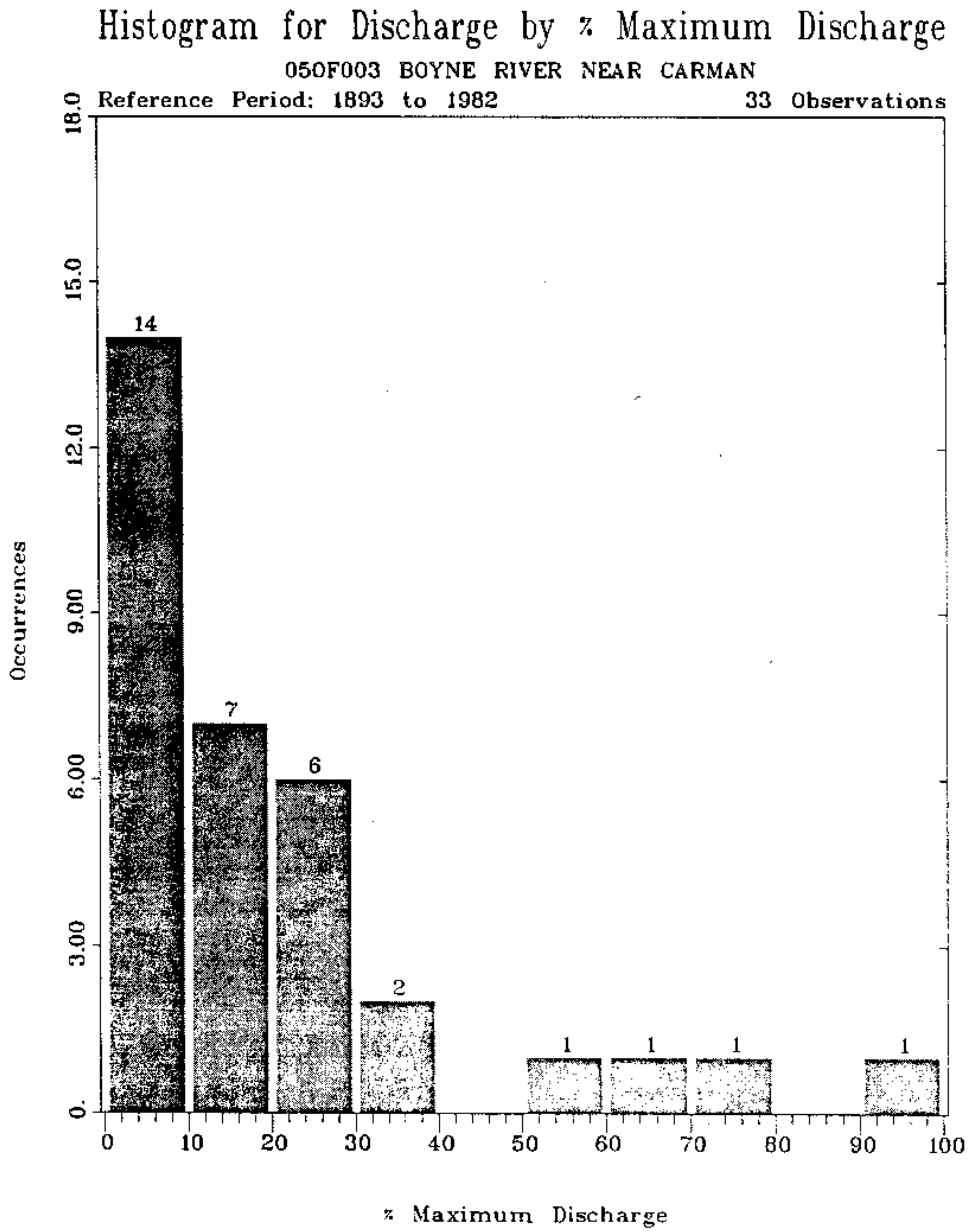

Figure 28: Output of CFA showing the histogram for discharge by % maximum discharge for the Boyne River near Carman - 05OF003

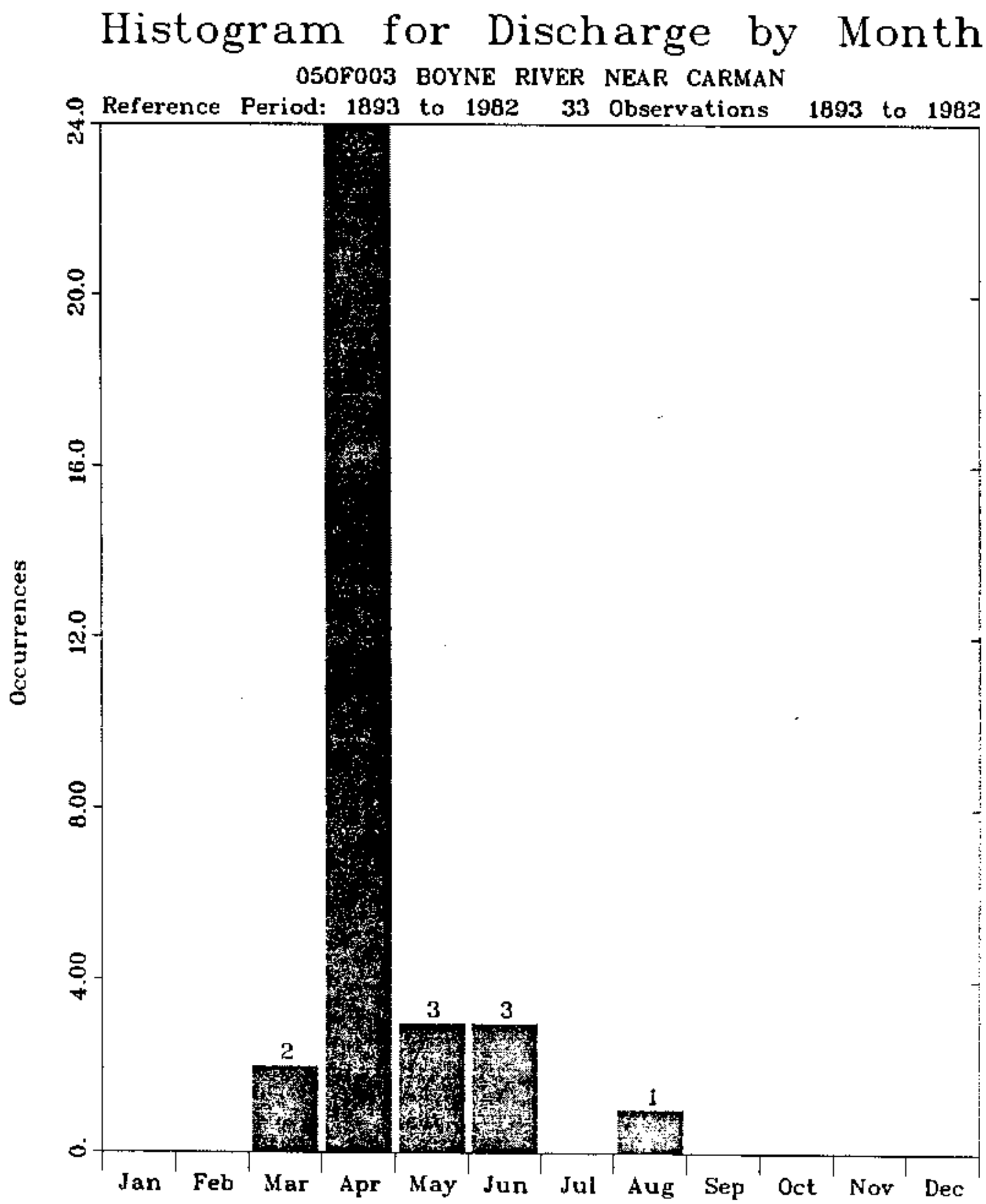

Figure 29: Output of CFA showing the histogram for discharge by month for the Boyne River near Carman - 05OF003

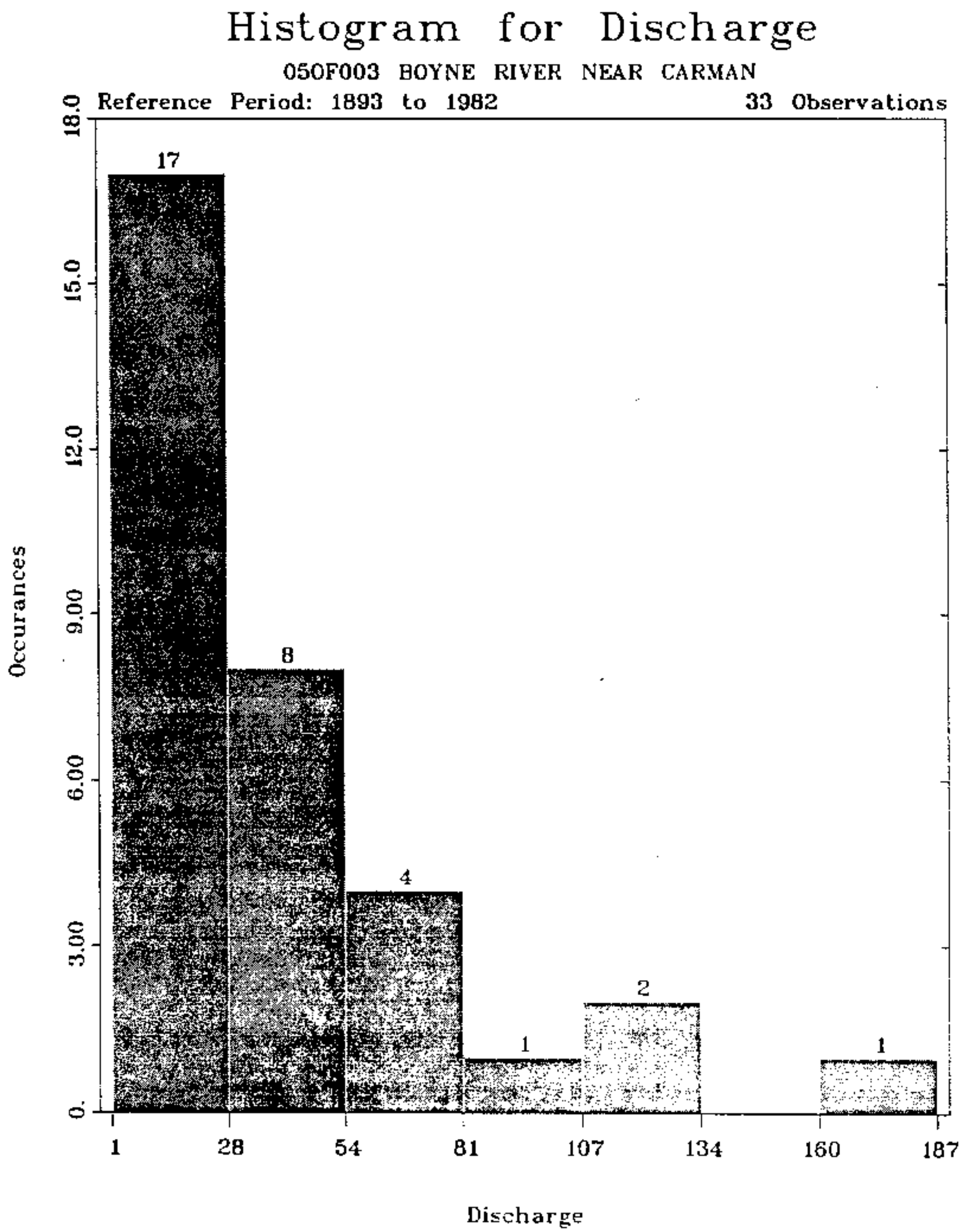

Figure 30: Output of CFA showing the histogram for discharge for the Boyne River near Carman - 05OF003

WSC STATION NO=05OF003
WSC STATION NAME=BOYNE RIVER NEAR CARMAN

TOTAL TIME SPAN, YT= 90 YRS. FLOW THRESHOLD = 187.000
OBSERVED PEAKS, N= 33 HISTORIC PEAKS ABOVE THRESHOLD, NHA= 1

OBSERVED PEAKS ABOVE THRESHOLD, NA= 1
OBSERVED PEAKS BELOW THRESHOLD, NB= 32
MISSING PEAKS BELOW THRESHOLD, NC= 57

MONTH (1)	YEAR (2)	FLOOD (3)	DESCENDING ORDER (4)	RANK M (5)	RANK ADJ. (6)	CUM. PROB. (7)	RET.PERIOD YEARS (8)
4	1893	187.000	187.000	1	1.00	.67	150.33
			THRESHOLD				
4	1919	13.500	132.000	2	3.78	3.75	26.68
4	1922	15.300	119.000	3	6.56	6.83	14.64
5	1927	29.700	105.000	4	9.34	9.92	10.09
5	1928	12.600	69.700	5	12.13	13.00	7.69
4	1930	34.000	59.500	6	14.91	16.08	6.22
4	1956	56.100	56.100	7	17.69	19.17	5.22
3	1957	10.800	55.200	8	20.47	22.25	4.49
4	1958	6.090	54.100	9	23.25	25.33	3.95
4	1959	14.900	43.000	10	26.03	28.42	3.52
4	1960	43.000	38.800	11	28.81	31.50	3.17
3	1961	19.400	37.900	12	31.59	34.58	2.89
4	1962	38.800	35.700	13	34.38	37.67	2.65
6	1963	14.500	34.800	14	37.16	40.75	2.45
4	1964	13.900	34.000	15	39.94	43.83	2.28
4	1965	59.500	29.700	16	42.72	46.92	2.13
4	1966	55.200	23.800	17	45.50	50.00	2.00
4	1967	37.900	19.400	18	48.28	53.08	1.88
8	1968	35.700	19.300	19	51.06	56.17	1.78
4	1969	69.700	15.300	20	53.84	59.25	1.69
4	1970	105.000	14.900	21	56.63	62.33	1.60
4	1971	54.100	14.500	22	59.41	65.42	1.53
4	1972	19.300	13.900	23	62.19	68.50	1.46
4	1973	1.180	13.500	24	64.97	71.58	1.40
4	1974	132.000	12.600	25	67.75	74.67	1.34
6	1975	11.400	11.400	26	70.53	77.75	1.29
4	1976	34.800	10.800	27	73.31	80.83	1.24
5	1977	7.390	10.700	28	76.09	83.92	1.19
4	1978	23.800	7.390	29	78.88	87.00	1.15
4	1979	119.000	6.530	30	81.66	90.08	1.11
4	1980	10.700	6.090	31	84.44	93.17	1.07
6	1981	5.470	5.470	32	87.22	96.25	1.04
4	1982	6.530	1.180*	33	90.00	99.33	1.01

Table 17: Output of CFA listing the station's flows, ranked flows, probability, and return period as obtained from the Cunnane formula for the Boyne River near Carman (05OF003), using a censoring threshold of 187 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - GENERALIZED EXTREME VALUE DISTRIBUTION
05OF003     BOYNE RIVER NEAR CARMAN


                         SAMPLE STATISTICS

                 MEAN        S.D.        C.V.        C.S.        C.K.
    X SERIES    39.341      41.975      1.067       2.009       7.540
 LN X SERIES     3.173       1.076       .339       -.364       3.887
 L-MOM RATIO    39.341      20.549       .522        .430        .235

 X(MIN)=      1.180                         TOTAL SAMPLE SIZE=  33
 X(MAX)=    187.000                       NO. OF LOW OUTLIERS=   1
 LOWER OUTLIER LIMIT OF X=     1.452        NO. OF ZERO FLOWS=   0


              AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                 MEAN        S.D.        C.V.        C.S.        C.K.
    X SERIES    40.534      42.075      1.038       1.998       7.466
 LN X SERIES     3.267        .945       .289        .233       2.582
 L-MOM RATIO    40.534      20.605       .508        .435        .229

               SOLUTION OBTAINED VIA MOMENTS

   PARAMETERS OF THE GEV WHICH DUPLICATES THE CONDITIONAL FUNCTION:
               U=      19.54    A=    24.807    K=     -.091


                       FLOOD FREQUENCY REGIME

              RETURN        EXCEEDANCE        FLOOD
              PERIOD        PROBABILITY

               1.003           .997             -
               1.050           .952             -
               1.250           .800            7.98
               2.000           .500           28.8
               5.000           .200           59.4
              10.000           .100           81.5
              20.000           .050          104
              50.000           .020          136
             100.000           .010          161
             200.000           .005          188
             500.000           .002          227
```

Table 18: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the generalized extreme value distribution, with the censoring threshold at 187 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - THREE-PARAMETER LOGNORMAL DISTRIBUTION
05OF003      BOYNE RIVER NEAR CARMAN


                              SAMPLE STATISTICS

                      MEAN        S.D.        C.V.        C.S.        C.K.
    X SERIES        39.341      41.975      1.067       2.009       7.540
 LN X SERIES         3.173       1.076       .339       -.364       3.887

 X(MIN)=       1.180                              TOTAL SAMPLE SIZE=  33
 X(MAX)=     187.000                            NO. OF LOW OUTLIERS=   1
 LOWER OUTLIER LIMIT OF X=     1.452              NO. OF ZERO FLOWS=   0


                AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                      MEAN        S.D.        C.V.        C.S.        C.K.
    X SERIES        40.534      42.075      1.038       1.998       7.466
 LN X SERIES         3.267        .945       .289        .233       2.582
LN(X-A) SERIES       3.066       1.121       .366       -.058       2.649

             SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

   PARAMETERS OF THE 3LN WHICH DUPLICATES THE CONDITIONAL FUNCTION:
                 A=      2.985   M= 2.927    S= 1.070


                          FLOOD FREQUENCY REGIME

                RETURN         EXCEEDANCE          FLOOD
                PERIOD         PROBABILITY

                 1.003            .997               3.97
                 1.050            .952               6.12
                 1.250            .800              10.6
                 2.000            .500              21.7
                 5.000            .200              48.9
                10.000            .100              76.5
                20.000            .050             111
                50.000            .020             171
               100.000            .010             228
               200.000            .005             297
               500.000            .002             409
```

Table 19: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the three-parameter lognormal distribution, with the censoring threshold at 187 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - LOG PEARSON TYPE III DISTRIBUTION
05OF003     BOYNE RIVER NEAR CARMAN


                              SAMPLE STATISTICS

                 MEAN        S.D.        C.V.        C.S.        C.K.
   X SERIES     39.341      41.975      1.067       2.009       7.540
LN X SERIES      3.173       1.076       .339       -.364       3.887

X(MIN)=      1.180                          TOTAL SAMPLE SIZE=  33
X(MAX)=    187.000                        NO. OF LOW OUTLIERS=   1
LOWER OUTLIER LIMIT OF X=     1.452          NO. OF ZERO FLOWS=   0


             AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                 MEAN        S.D.        C.V.        C.S.        C.K.
   X SERIES     40.534      42.075      1.038       1.998       7.466
LN X SERIES      3.267        .945       .289        .233       2.582

               SOLUTION OBTAINED VIA MOMENTS

  PARAMETERS OF THE LP3 WHICH DUPLICATES THE CONDITIONAL FUNCTION:
   A=  .8610E-01 B=  110.8      LOG(M)= -6.347     M =  .1752E-02
   SYNTHETIC STATISTICS: MEAN=   3.194  S.D.=   .906  C.S.=  .190


                         FLOOD FREQUENCY REGIME

               RETURN        EXCEEDANCE          FLOOD
               PERIOD        PROBABILITY

                1.003           .997               -
                1.050           .952               5.66
                1.250           .800              11.3
                2.000           .500              23.7
                5.000           .200              51.8
               10.000           .100              79.2
               20.000           .050             114
               50.000           .020             172
              100.000           .010             228
              200.000           .005             296
              500.000           .002             409
```

Table 20: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the Log Pearson Type III distribution, with the censoring threshold at 187 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - WAKEBY DISTRIBUTION
05OF003     BOYNE RIVER NEAR CARMAN

                         SAMPLE STATISTICS

                MEAN        S.D.        C.V.        C.S.        C.K.
   X SERIES    39.341      41.975      1.067       2.009       7.540
LN X SERIES     3.173       1.076       .339       -.364       3.887
L-MOM RATIO    39.341      20.549       .522        .430        .235


X(MIN)=      1.180                          TOTAL SAMPLE SIZE=  33
X(MAX)=    187.000                        NO. OF LOW OUTLIERS=   1
LOWER OUTLIER LIMIT OF X=     1.452          NO. OF ZERO FLOWS=   0


             AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                MEAN        S.D.        C.V.        C.S.        C.K.
   X SERIES    40.534      42.075      1.038       1.998       7.466
LN X SERIES     3.267        .945       .289        .233       2.582
L-MOM RATIO    40.534      20.605       .508        .435        .229

THE FOLLOWING WAKEBY PARAMETERS WERE OBTAINED VIA LEAST SQUARES REGRESSION

E=     416.04   A=     -24.046   B=  1.89   C=    -434.077   D= -.128
          DISTRIBUTION IS UPPER BOUNDED AT E=  .4160E+03

                        FLOOD FREQUENCY REGIME

              RETURN        EXCEEDANCE         FLOOD
              PERIOD        PROBABILITY

               1.003           .997              -
               1.050           .952              6.20
               1.250           .800              9.23
               2.000           .500             24.1
               5.000           .200             62.6
              10.000           .100             91.9
              20.000           .050            119
              50.000           .020            152
             100.000           .010            174
             200.000           .005            195
             500.000           .002            219
```

Table 21: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the Wakeby distribution, with the censoring threshold at 187 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - NONPARAMETRIC METHOD
05OF003      BOYNE RIVER NEAR CARMAN


                          SAMPLE STATISTICS

                 MEAN        S.D.        C.V.        C.S.        C.K.
   X SERIES     39.341      41.975      1.067       2.009       7.540
LN X SERIES      3.173       1.076       .339       -.364       3.887

X(MIN)=      1.180                          TOTAL SAMPLE SIZE=  33
X(MAX)=    187.000                          NO. OF ZERO FLOWS=   0

     SMOOTHING PARAMETER H =     5.053

                 FLOOD FREQUENCY REGIME

     RETURN          EXCEEDANCE          FLOOD
     PERIOD          PROBABILITY

      1.003             .997              -
      1.050             .952              2.06
      1.250             .800              9.84
      2.000             .500             24.0
      5.000             .200             55.8
     10.000             .100             99.1
     20.000             .050            122
     50.000             .020            135
    100.000             .010            181
    200.000             .005            188
    500.000             .002            192
```

Table 22: Output of CFA listing the summary statistics of the sample, the computed smoothing parameter H, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the Nonparametric Method, with the censoring threshold at 187 m^3/s.

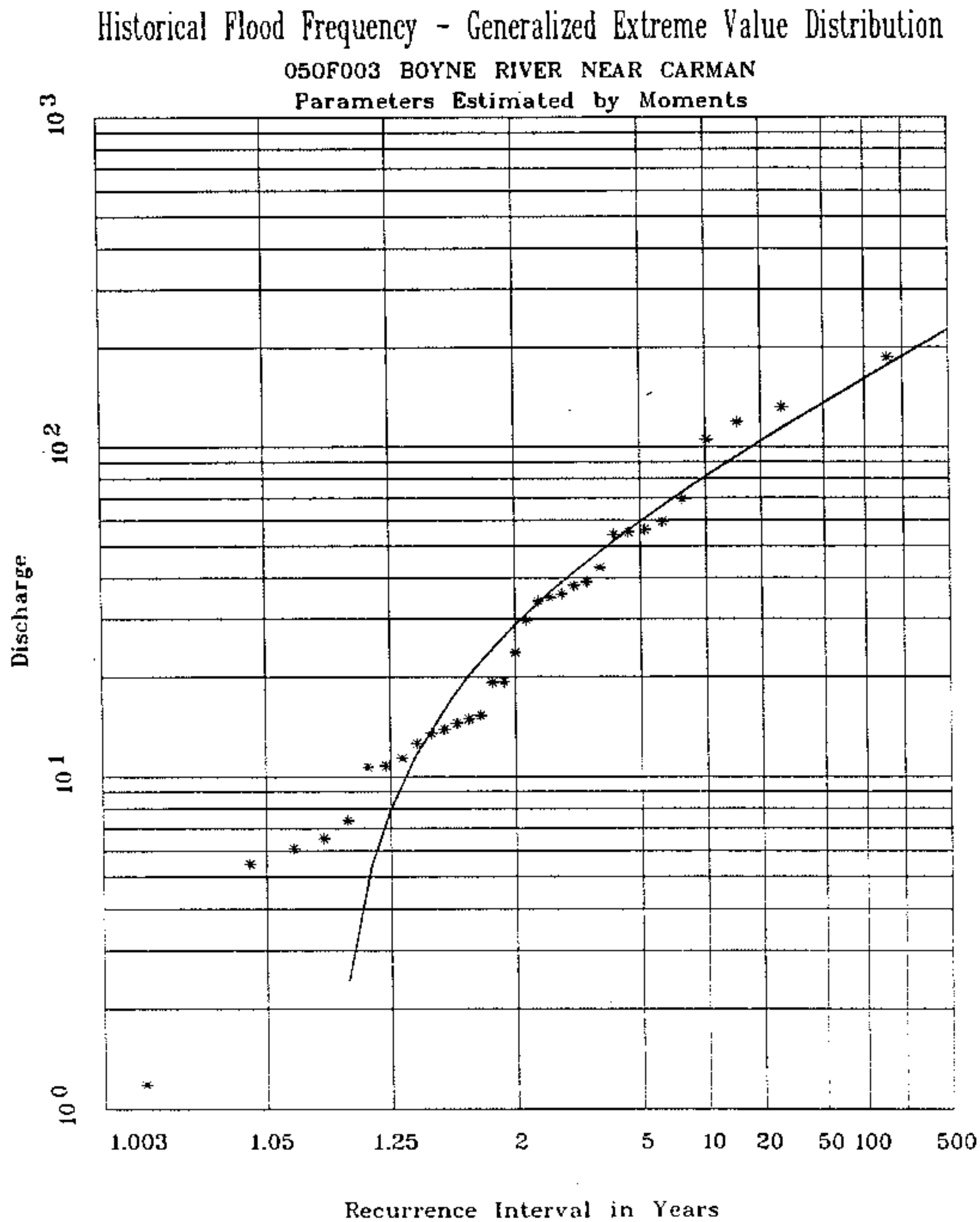

Figure 31: Output of CFA showing the frequency plot based on the generalized extreme value distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 187 m^3/s.

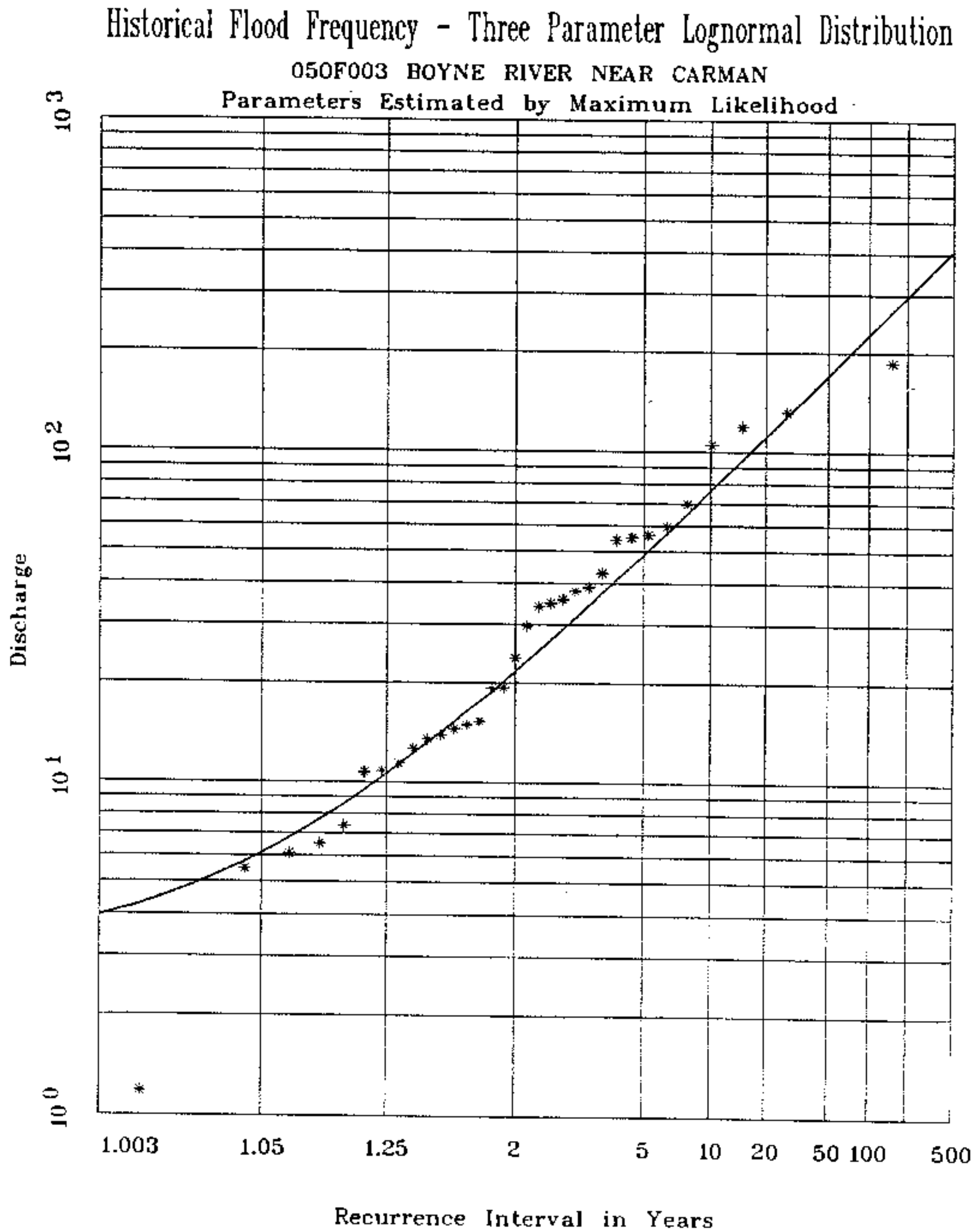

Figure 32: Output of CFA showing the frequency plot based on the three-parameter lognormal distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 187 m^3/s.

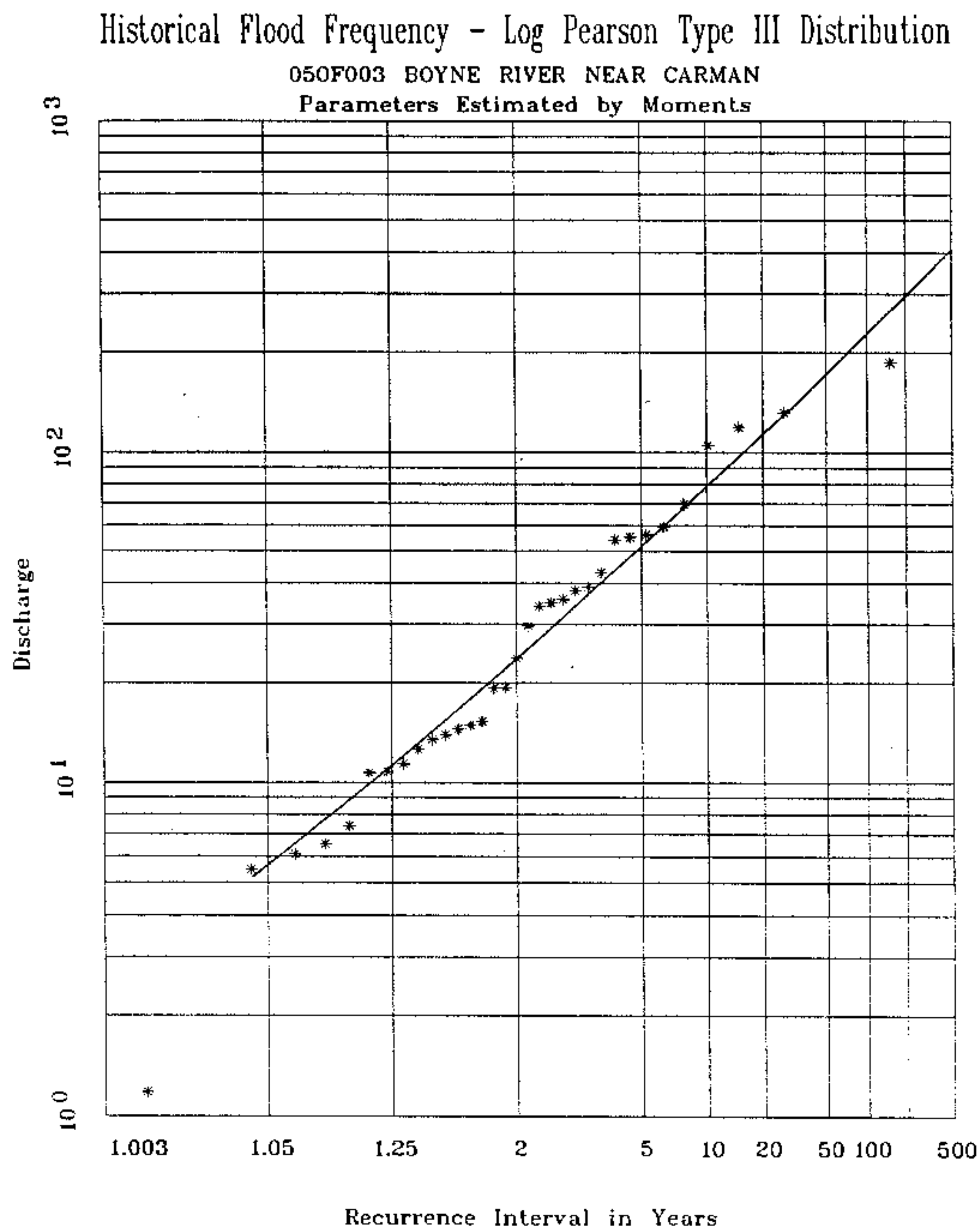

Figure 33: Output of CFA showing the frequency plot based on the Log Pearson Type III distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 187 m^3/s.

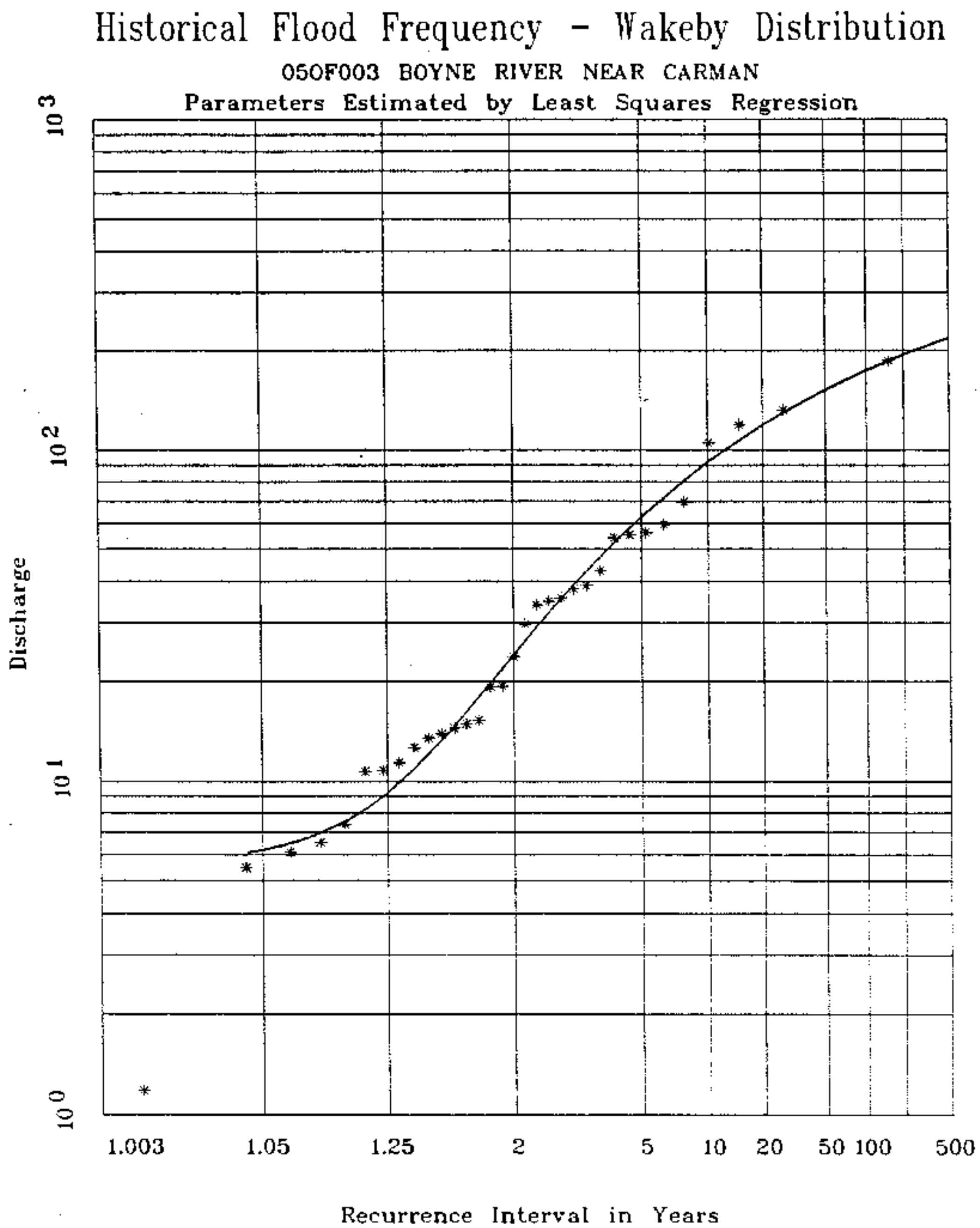

Figure 34: Output of CFA showing the frequency plot based on the Wakeby distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 187 m^3/s.

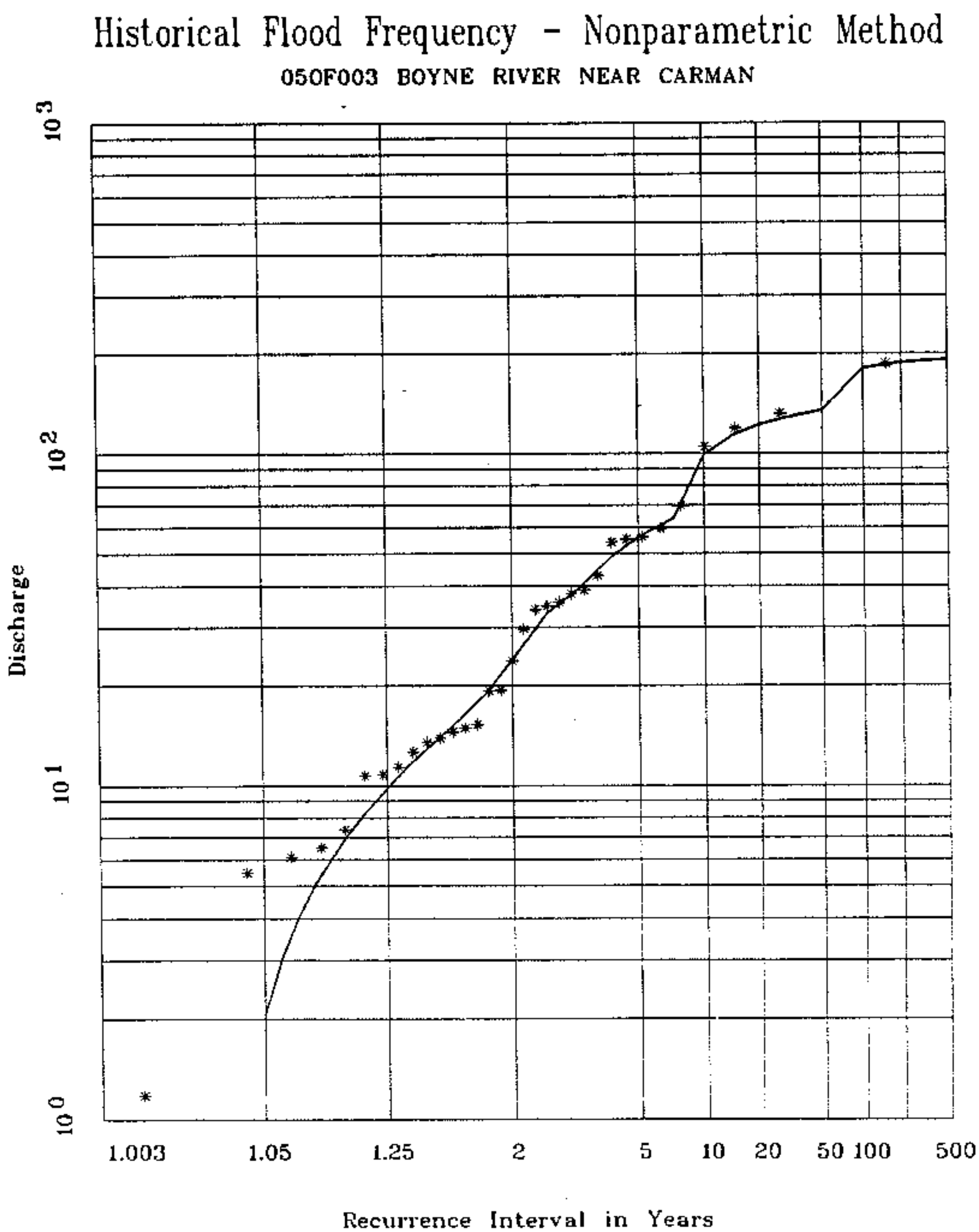

Figure 35: Output of CFA showing the frequency plot based on the Nonparametric Method for the Boyne River near Carman - 05OF003, using a censoring threshold of 187 m^3/s.

```
WSC STATION NO=05OF003
WSC STATION NAME=BOYNE RIVER NEAR CARMAN
--------------------------------------------------------------------------

TOTAL TIME SPAN, YT= 90 YRS.    FLOW THRESHOLD =   105.000
OBSERVED PEAKS, N= 33  HISTORIC PEAKS ABOVE THRESHOLD, NHA= 4

        OBSERVED PEAKS ABOVE THRESHOLD, NA=  4
        OBSERVED PEAKS BELOW THRESHOLD, NB= 29
         MISSING PEAKS BELOW THRESHOLD, NC= 57
```

MONTH (1)	YEAR (2)	FLOOD (3)	DESCENDING ORDER (4)	RANK M (5)	RANK ADJ. (6)	CUM. PROB. (7)	RET.PERIOD YEARS (8)
4	1893	187.000	187.000	1	1.00	.67	150.33
4	1919	13.500	132.000	2	2.00	1.77	56.38
4	1922	15.300	119.000	3	3.00	2.88	34.69
5	1927	29.700	105.000	4	4.00	3.99	25.06
			THRESHOLD				
5	1928	12.600	69.700	5	6.97	7.28	13.74
4	1930	34.000	59.500	6	9.93	10.57	9.46
4	1956	56.100	56.100	7	12.90	13.85	7.22
3	1957	10.800	55.200	8	15.86	17.14	5.83
4	1958	6.090	54.100	9	18.83	20.43	4.89
4	1959	14.900	43.000	10	21.79	23.72	4.22
4	1960	43.000	38.800	11	24.76	27.01	3.70
3	1961	19.400	37.900	12	27.72	30.29	3.30
4	1962	38.800	35.700	13	30.69	33.58	2.98
6	1963	14.500	34.800	14	33.66	36.87	2.71
4	1964	13.900	34.000	15	36.62	40.16	2.49
4	1965	59.500	29.700	16	39.59	43.44	2.30
4	1966	55.200	23.800	17	42.55	46.73	2.14
4	1967	37.900	19.400	18	45.52	50.02	2.00
8	1968	35.700	19.300	19	48.48	53.31	1.88
4	1969	69.700	15.300	20	51.45	56.59	1.77
4	1970	105.000	14.900	21	54.41	59.88	1.67
4	1971	54.100	14.500	22	57.38	63.17	1.58
4	1972	19.300	13.900	23	60.34	66.46	1.50
4	1973	1.180	13.500	24	63.31	69.75	1.43
4	1974	132.000	12.600	25	66.28	73.03	1.37
6	1975	11.400	11.400	26	69.24	76.32	1.31
4	1976	34.800	10.800	27	72.21	79.61	1.26
5	1977	7.390	10.700	28	75.17	82.90	1.21
4	1978	23.800	7.390	29	78.14	86.18	1.16
4	1979	119.000	6.530	30	81.10	89.47	1.12
4	1980	10.700	6.090	31	84.07	92.76	1.08
6	1981	5.470	5.470	32	87.03	96.05	1.04
4	1982	6.530	1.180*	33	90.00	99.33	1.01

Table 23: Output of CFA listing the station's flows, ranked flows, probability, and return period as obtained from the Cunnane formula for the Boyne River near Carman (05OF003), using a censoring threshold of 105 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - GENERALIZED EXTREME VALUE DISTRIBUTION
05OF003     BOYNE RIVER NEAR CARMAN


                           SAMPLE STATISTICS

                  MEAN        S.D.        C.V.        C.S.        C.K.
    X SERIES     39.341      41.975      1.067       2.009       7.540
 LN X SERIES      3.173       1.076       .339       -.364       3.887
 L-MOM RATIO     39.341      20.549       .522        .430        .235


 X(MIN)=      1.180                        TOTAL SAMPLE SIZE=  33
 X(MAX)=    187.000                     NO. OF LOW OUTLIERS=   1
 LOWER OUTLIER LIMIT OF X=     1.452       NO. OF ZERO FLOWS=   0


            AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                  MEAN        S.D.        C.V.        C.S.        C.K.
    X SERIES     40.534      42.075      1.038       1.998       7.466
 LN X SERIES      3.267        .945       .289        .233       2.582
 L-MOM RATIO     40.534      20.605       .508        .435        .229

              SOLUTION OBTAINED VIA MOMENTS

   PARAMETERS OF THE GEV WHICH DUPLICATES THE CONDITIONAL FUNCTION:
              U=      17.23   A=    18.440    K=     -.146


                      FLOOD FREQUENCY REGIME

             RETURN        EXCEEDANCE          FLOOD
             PERIOD        PROBABILITY

              1.003           .997              -
              1.050           .952              -
              1.250           .800              8.75
              2.000           .500             24.2
              5.000           .200             48.2
             10.000           .100             66.4
             20.000           .050             85.8
             50.000           .020            114
            100.000           .010            138
            200.000           .005            165
            500.000           .002            204
```

Table 24: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the generalized extreme value distribution, with the censoring threshold at 105 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - THREE-PARAMETER LOGNORMAL DISTRIBUTION
05OF003     BOYNE RIVER NEAR CARMAN

                              SAMPLE STATISTICS

                       MEAN        S.D.         C.V.        C.S.        C.K.
         X SERIES     39.341      41.975       1.067       2.009       7.540
      LN X SERIES      3.173       1.076        .339       -.364       3.887

      X(MIN)=       1.180                              TOTAL SAMPLE SIZE=  33
      X(MAX)=     187.000                            NO. OF LOW OUTLIERS=   1
      LOWER OUTLIER LIMIT OF X=     1.452              NO. OF ZERO FLOWS=   0


                 AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                       MEAN        S.D.         C.V.        C.S.        C.K.
         X SERIES     40.534      42.075       1.038       1.998       7.466
      LN X SERIES      3.267        .945        .289        .233       2.582
   LN(X-A) SERIES      3.082       1.106        .359       -.032       2.631

              SOLUTION OBTAINED VIA MAXIMUM LIKELIHOOD

     PARAMETERS OF THE 3LN WHICH DUPLICATES THE CONDITIONAL FUNCTION:
                    A=     2.797   M= 2.864   S= 1.006


                             FLOOD FREQUENCY REGIME

                     RETURN        EXCEEDANCE          FLOOD
                     PERIOD        PROBABILITY

                      1.003           .997               3.90
                      1.050           .952               6.07
                      1.250           .800              10.3
                      2.000           .500              20.3
                      5.000           .200              43.6
                     10.000           .100              66.4
                     20.000           .050              94.5
                     50.000           .020             141
                    100.000           .010             185
                    200.000           .005             237
                    500.000           .002             320
```

Table 25: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the three-parameter lognormal distribution, with the censoring threshold at 105 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - LOG PEARSON TYPE III DISTRIBUTION
05OF003      BOYNE RIVER NEAR CARMAN


                        SAMPLE STATISTICS

                MEAN      S.D.       C.V.       C.S.       C.K.
   X SERIES    39.341    41.975     1.067      2.009      7.540
LN X SERIES     3.173     1.076      .339      -.364      3.887

X(MIN)=      1.180                        TOTAL SAMPLE SIZE=  33
X(MAX)=    187.000                      NO. OF LOW OUTLIERS=   1
LOWER OUTLIER LIMIT OF X=    1.452        NO. OF ZERO FLOWS=   0


             AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                MEAN      S.D.       C.V.       C.S.       C.K.
   X SERIES    40.534    42.075     1.038      1.998      7.466
LN X SERIES     3.267      .945      .289       .233      2.582

               SOLUTION OBTAINED VIA MOMENTS

  PARAMETERS OF THE LP3 WHICH DUPLICATES THE CONDITIONAL FUNCTION:
   A=  .7567E-01 B=  121.7     LOG(M)= -6.121     M =  .2196E-02
   SYNTHETIC STATISTICS: MEAN=   3.090  S.D.=   .835  C.S.=  .181


                     FLOOD FREQUENCY REGIME

              RETURN        EXCEEDANCE         FLOOD
              PERIOD        PROBABILITY

               1.003           .997              -
               1.050           .952             5.72
               1.250           .800            10.8
               2.000           .500            21.4
               5.000           .200            44.0
              10.000           .100            65.0
              20.000           .050            90.5
              50.000           .020           132
             100.000           .010           171
             200.000           .005           218
             500.000           .002           293
```

Table 26: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the Log Pearson Type III distribution, with the censoring threshold at 105 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - WAKEBY DISTRIBUTION
05OF003     BOYNE RIVER NEAR CARMAN

                         SAMPLE STATISTICS

                  MEAN       S.D.       C.V.       C.S.       C.K.
   X SERIES     39.341     41.975      1.067      2.009      7.540
LN X SERIES      3.173      1.076       .339      -.364      3.887
L-MOM RATIO     39.341     20.549       .522       .430       .235

X(MIN)=      1.180                        TOTAL SAMPLE SIZE=  33
X(MAX)=    187.000                      NO. OF LOW OUTLIERS=   1
LOWER OUTLIER LIMIT OF X=     1.452       NO. OF ZERO FLOWS=   0


             AFTER REMOVAL OF ZEROES AND/OR LOW OUTLIERS

                  MEAN       S.D.       C.V.       C.S.       C.K.
   X SERIES     40.534     42.075      1.038      1.998      7.466
LN X SERIES      3.267       .945       .289       .233      2.582
L-MOM RATIO     40.534     20.605       .508       .435       .229

  THE FOLLOWING WAKEBY PARAMETERS WERE OBTAINED VIA LEAST SQUARES
REGRESSION

E=    -27.12   A=      16.880   B=    .93   C=      47.125   D=  .304

                       FLOOD FREQUENCY REGIME

              RETURN        EXCEEDANCE          FLOOD
              PERIOD        PROBABILITY

               1.003           .997              -
               1.050           .952              3.66
               1.250           .800              8.73
               2.000           .500             21.4
               5.000           .200             45.1
              10.000           .100             64.8
              20.000           .050             87.9
              50.000           .020            126
             100.000           .010            162
             200.000           .005            206
             500.000           .002            282
```

Table 27: Output of CFA listing the summary statistics of the sample, the solution method, the estimated parameters of the distribution, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the Wakeby distribution, with the censoring threshold at 105 m^3/s.

```
HISTORICAL FREQUENCY ANALYSIS - NONPARAMETRIC METHOD
05OF003      BOYNE RIVER NEAR CARMAN


                    SAMPLE STATISTICS

               MEAN       S.D.       C.V.       C.S.       C.K.
   X SERIES   39.341     41.975     1.067      2.009      7.540
LN X SERIES    3.173      1.076      .339      -.364      3.887

X(MIN)=      1.180                         TOTAL SAMPLE SIZE=  33
X(MAX)=    187.000                         NO. OF ZERO FLOWS=   0

      SMOOTHING PARAMETER H =     4.956

                FLOOD FREQUENCY REGIME

      RETURN         EXCEEDANCE         FLOOD
      PERIOD         PROBABILITY

       1.003            .997              -
       1.050            .952              1.89
       1.250            .800              9.38
       2.000            .500             21.4
       5.000            .200             49.4
      10.000            .100             61.1
      20.000            .050             74.5
      50.000            .020            128
     100.000            .010            181
     200.000            .005            188
     500.000            .002            192
```

Table 28: Output of CFA listing the summary statistics of the sample, the computed smoothing parameter H, and the tabular flood frequency regime of the Boyne River near Carman (05OF003) for the Nonparametric Method, with the censoring threshold at 105 m^3/s.

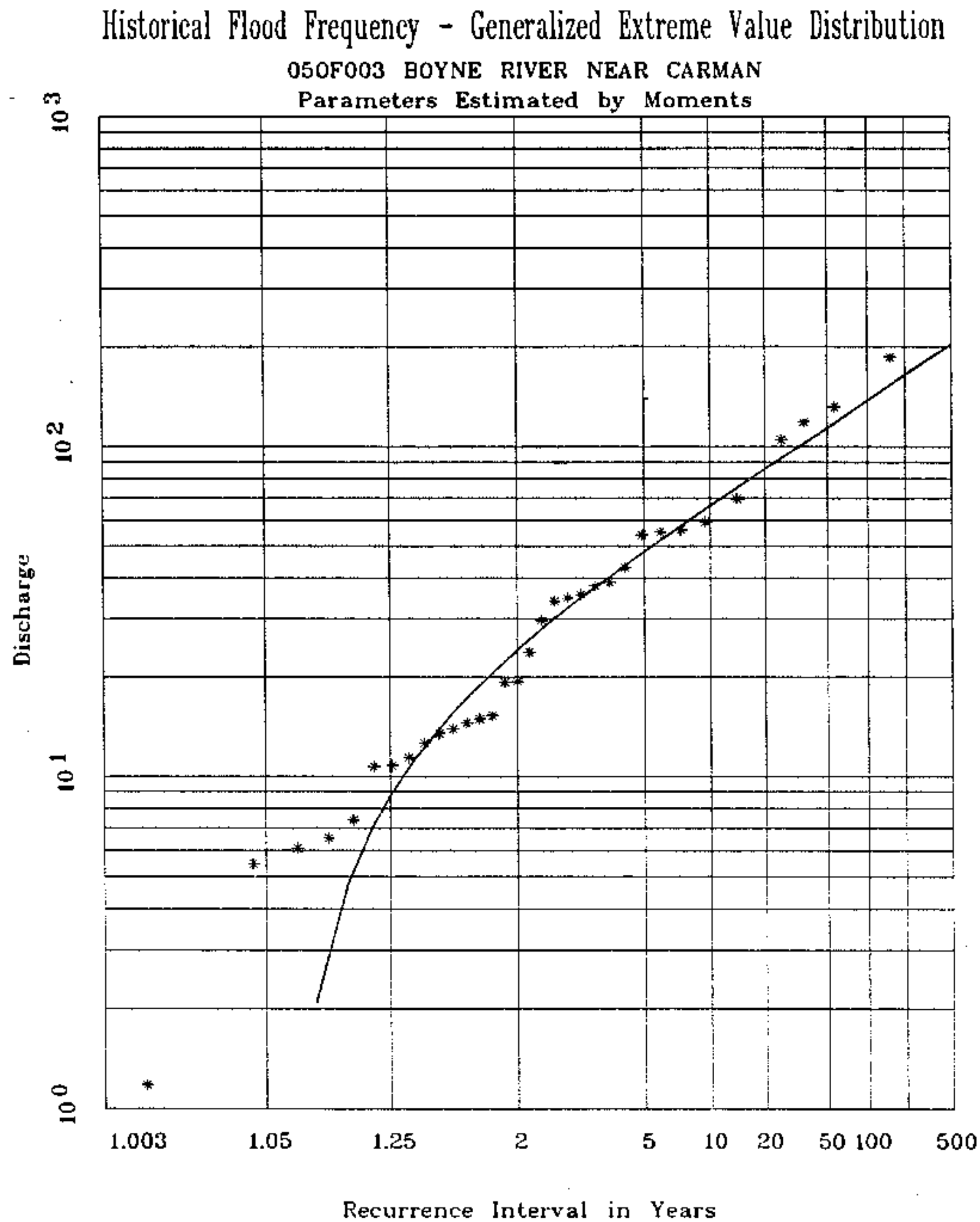

Figure 36: Output of CFA showing the frequency plot based on the generalized extreme value distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 105 m^3/s.

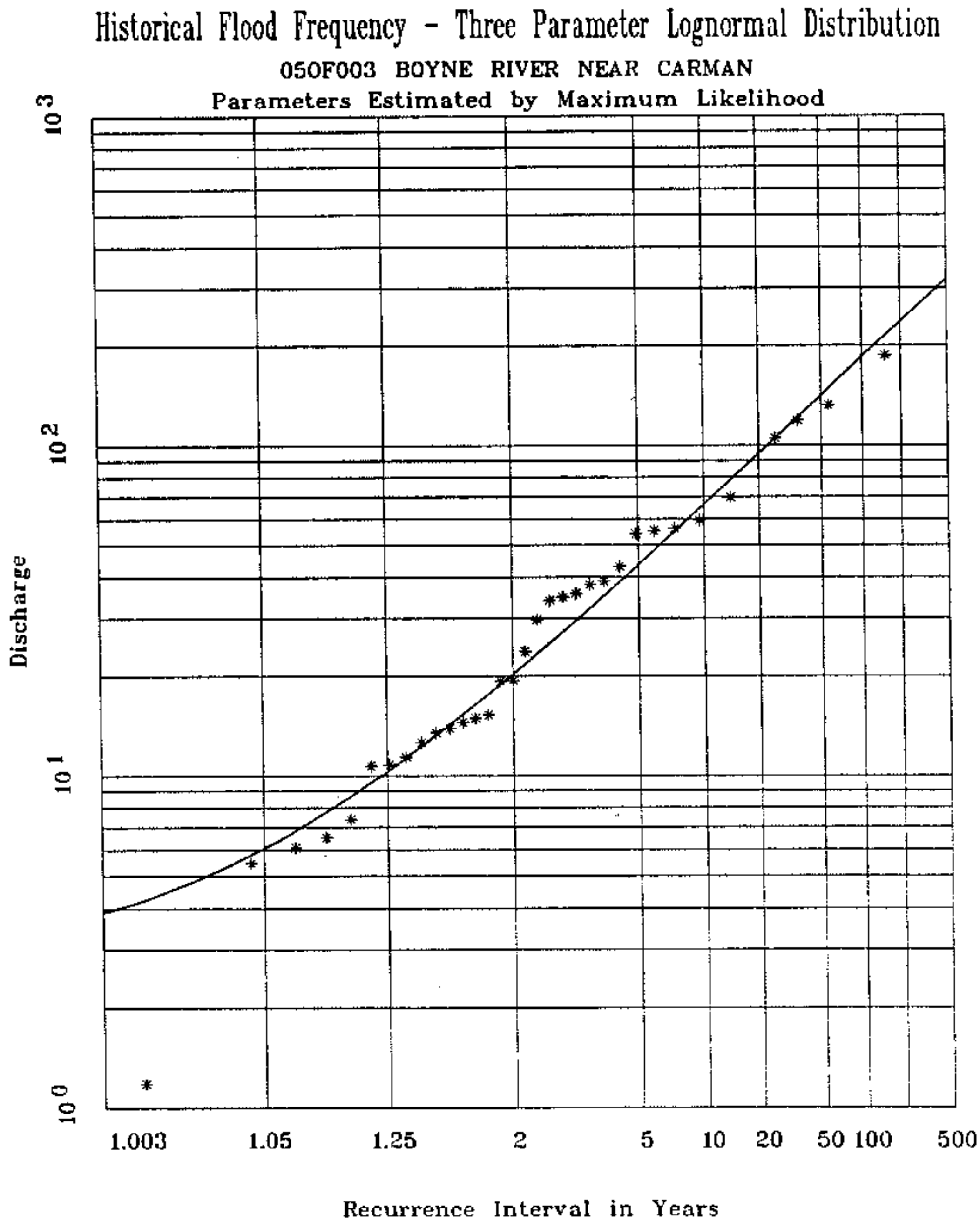

Figure 37: Output of CFA showing the frequency plot based on the three-parameter lognormal distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 105 m^3/s.

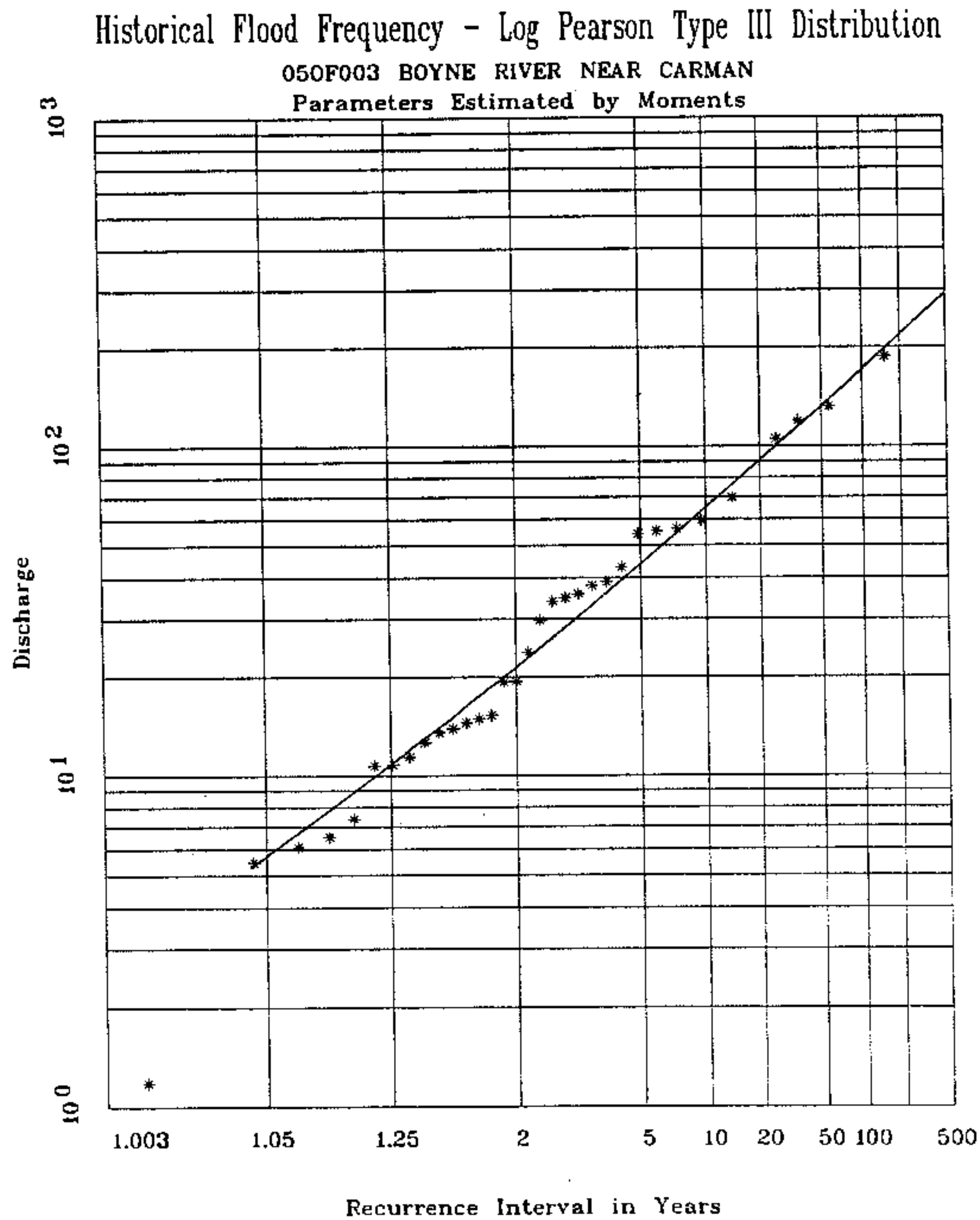

Figure 38: Output of CFA showing the frequency plot based on the Log Pearson Type III distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 105 m^3/s.

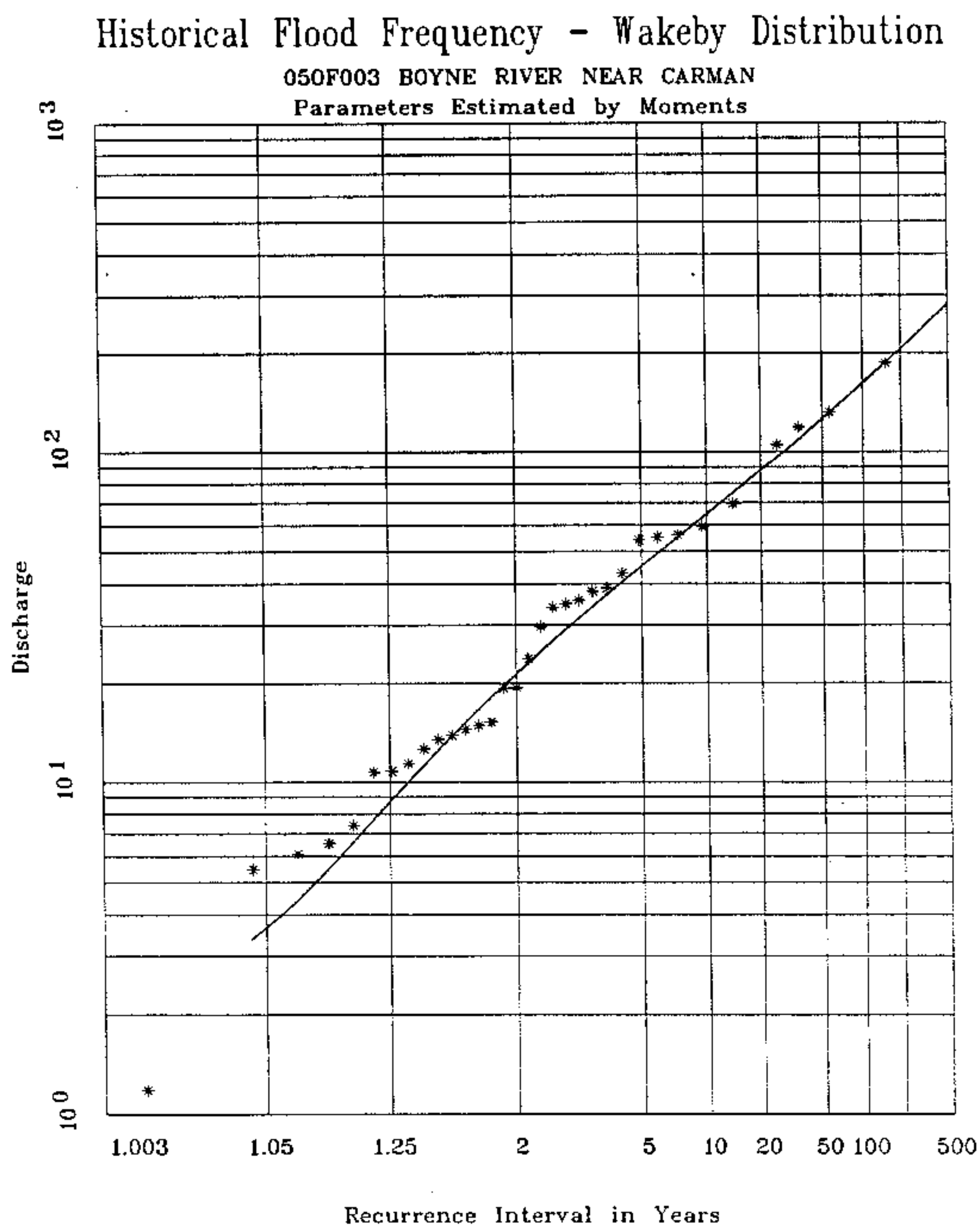

Figure 39: Output of CFA showing the frequency plot based on the Wakeby distribution for the Boyne River near Carman - 05OF003, using a censoring threshold of 105 m^3/s.

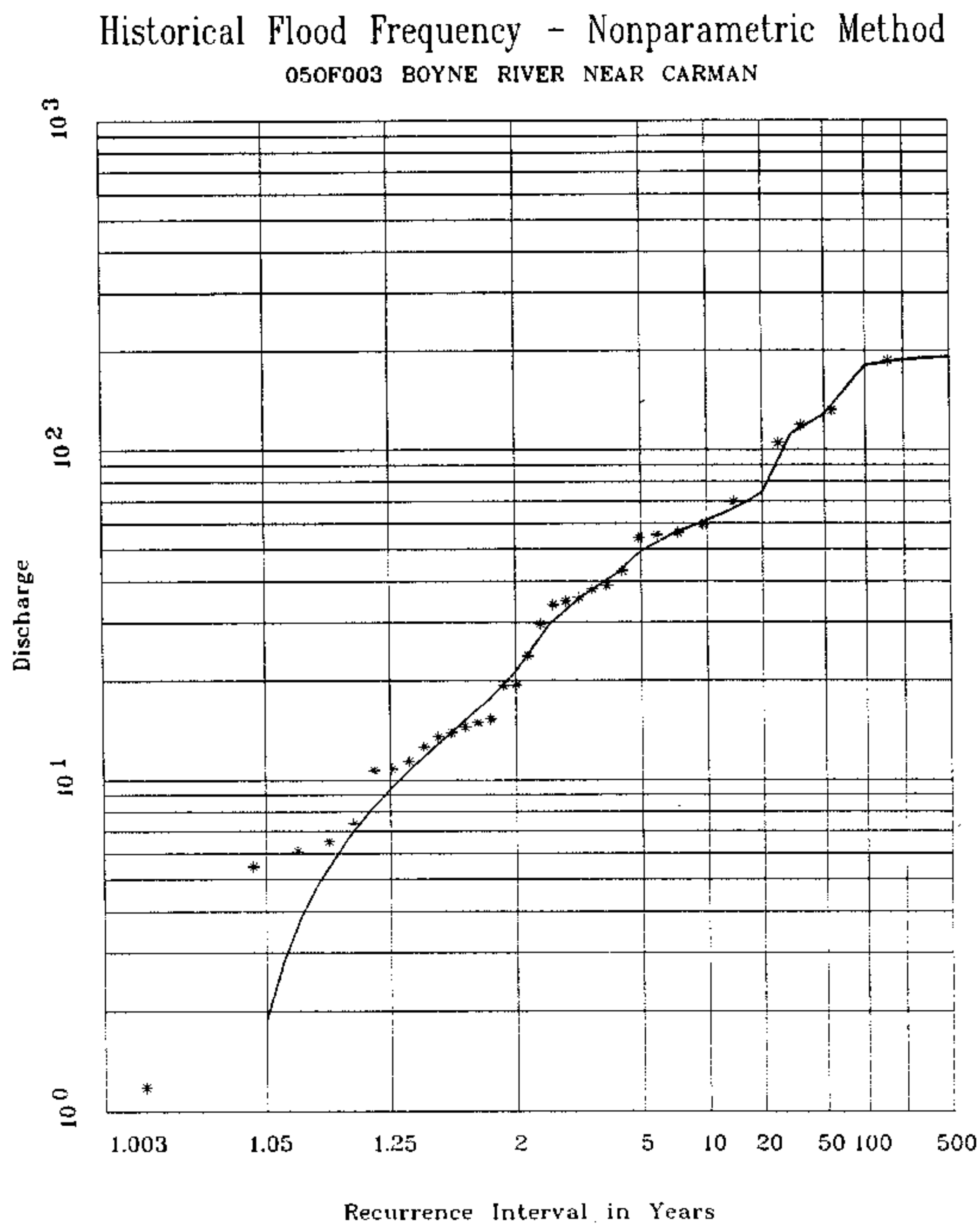

Figure 40: Output of CFA showing the frequency plot based on the Nonparametric Method for the Boyne River near Carman - 05OF003, using a censoring threshold of 105 m^3/s.

6. REFERENCES

Adamowski, K., 1985, Nonparametric kernal estimates of flood frequencies, *Water Resources Res.*, 21(11): 1585-1590.

Adamowski, K., 1988, Reply to a comment on nonparametric kernal estimates of flood frequencies, *Water Resources Res.*, 24(6): 890-891.

Adamowski, K., 1989, A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies, *J. of Hydrology*, 108: 295-308.

Adamowski, K. and W. Feluch, 1990, Nonparametric flood frequency analysis with historic information, *J. Hydraulics Div.*, ASCE, August.

Adamowski, K. and P.J. Pilon, 1989, "Worth of Historic Information in the Nonparametric Flood Frequency Analysis", American Geophysical Union,, Spring Meeting, May 8-12.

Benson, M.A., 1950, Use of historical data in flood-frequency analysis, *Trans. Am. Geophysical Union*, 31(3): 419-424.

Beran, M., Personal Communication, Institute of Hydrology, Wallingford, England.

Bobée, B., 1975, The Log Pearson type 3 distribution and its application in hydrology, *Water Resources Res.*, 11(5): 681-689.

Condie, R. and K.D. Harvey, 1984, *Flood Frequency Analysis with Historic Information, Low Outliers, and Zeros, Program HILO*, Inland Waters Directorate, Environment Canada, Ottawa.

Condie, R. and K.A. Lee, 1982, Flood frequency analysis with historic information, *J. of Hydrology*, 58:47-61.

Condie, R. and P.J. Pilon, 1983, Fitting the Log Pearson type 3 distribution to censored samples: an application to flood frequency analysis with historic information, *Proceedings of the 6th Canadian Hydrotechnical Conference*, pp.11-22, Ottawa, Ontario, June 2-3.

Cunnane, C., 1978, Unbiased plotting positions - a review, *J. of Hydrology*, 37(May): 205-222.

Ecosystem Sciences and Evaluation Directorate, 1993, HYDAT CD-ROM Version 4.0 User's Manual, Surveys and Information Systems Branch, Environment Canada, Ottawa.

Fisher, R.A., and L.H.C. Tippett, 1928, Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample, *Proceedings of the Cambridge Philosophical Society*, 24: 180-190.

Greenwood, J.A., J.M. Landwehr, N.C. Matalas, and J.R. Wallis, 1979, Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, *Water Resources Res.*, 15(5): 1049-1054.

Greis, N.P., 1983, Flood frequency analysis: a review of 1979-1982, *Reviews of Geophysics and Space Physics*, 21(3): 699-706.

Grubbs, F.E. and G. Beck., 1972, Extension of sample sizes and percentage points for significance tests of outlying observations, *Technometrics*, 14(4): 847-854.

Hastings, C. Jr., 1955, *Approximations for Digital Computers*, Princeton University Press, Princeton, N.J.

Hosking, J.R.M., 1988, Fortran routines for use with the method of L-moments, Research Report RC13844, IBM Research, Yorktown Heights, N.Y.

Hosking, J.R.M., 1989, Some theoretical results concerning L-moments, Research Report RC14492, IBM Research, Yorktown Heights, N.Y.

Hosking, J.R.M., 1990, L-moments: analysis and estimation of distributions using linear combinations of order statistics, *J. of the Royal Statistical Soc.*, Series B, 52(1): 105-124.

Hosking, J.R.M., J.R. Wallis, and E.F. Wood, 1985, Estimation of the generalized extreme value distribution by the method of probability weighted moments, *Technometrics*, 27(3): 251-261.

Houghton, J.C., 1978, Birth of a parent: the Wakeby distribution for modelling flood flows, *Water Resources Res.*, 14(6): 1105-1109.

Hydrology Subcommittee, 1982, *Guidelines for Determining Flood Flow Frequency*, Bulletin #17B, Interagency Advisory Committee on Water Data, U.S. Geological Survey, Office of Water Data Coordination, Reston, Virginia 22092. Revised with Editorial Corrections.

Inland Waters Directorate, 1991, *HYDAT CD-ROM User's Manual*, Hydrology Division, Water Resources Branch, Environment Canada, Ottawa.

Inland Waters Directorate, 1980, *Supplying Hydrometric and Sediment Data to Users*, Second Edition, Water Resources Branch, Environment Canada, Ottawa.

Kennedy, J.B. and A.M. Neville., 1976, *Basic Statistical Methods for Engineers and Scientists*, Second Edition, Harper and Row, New York.

Kimball, B.F., 1946, Sufficient statistical estimation functions for the parameters of the distribution of maximum values, *Annals of Mathematical Stat.*, 17: 299.

Landwehr, J.M., N.C. Matalas, and J.R. Wallis, 1979, Estimation of parameters and quantiles of Wakeby distributions 1. Known lower bounds, *Water Resources Res.*, 15(6): 1361-1372.

Landwehr, J.M., N.C. Matalas, and J.R. Wallis, 1979, Estimation of parameters and quantiles of Wakeby distributions 2. Unknown lower bounds, *Water Resources Res.*, 15(6): 1373-1379.

Lotus Development Corporation, *Lotus 1-2-3 Release 3 Reference*, Cambridge, MA, 1989.

Matalas, N.C., J.R. Slack and J.R. Wallis, 1975, Regional skew in search of a parent, *Water Resources Res.* 11(6): 815-826.

Natural Environment Research Council, 1975, *Flood Studies Report, Volume I - Hydrological Studies*, Water Resources Publications, Fort Collins, Colorado.

Pilon, P.J. and R. Condie, 1982, *Identification of High and Low Outliers, Program OUTLIER*, Inland Waters Directorate, Environment Canada, Ottawa.

Pilon, P.J., K.D. Harvey, and R. Condie, 1987, An assessment of contemporary flood frequency distributions, *Proceedings of the 8th Canadian Hydrotechnical Conference*, Montreal, Quebec, May 19-22.

Roscoe, J.T., 1969, *Fundamental Research Statistics for the Behavioural Sciences*, Holt, Rinehart and Winston, Inc., Toronto.

Rudemo, M., 1982, Empirical choice of histograms and kernel density estimators, *Scandinavian J. of Stat.*, 9: 65-78.

Shiau, S.Y. and R. Condie, 1980, *Statistical Tests for Independence, Trend, Homogeneity and Randomness*, Inland Waters Directorate, Environment Canada, Ottawa.

Siegel, S., 1956, *Nonparametric Statistics for the Behavioural Sciences*, McGraw-Hill, Toronto.

Wallis, R.W., 1980, Risk and uncertainties in the evaluation of flood events for the design of hydrologic structures, paper presented at Seminar on Extreme Hydrological Events: Floods and Droughts: Erice, Italy, March 22-25.

Wilson, E.B. and M.M. Hilferty, 1931, The distribution of Chi-square, *Proceedings of the National Academy of Science*, U.S.A., 17: 684.

APPENDIX A: Nonparametric Tests for Independence, Trend, Homogeneity, and Randomness

A.1 Introduction

This appendix briefly summarizes the functions evaluated in the package and gives the methods used to determine their statistical significance. Statistical tables are provided for ease of reference.

Any statistical test of significance will generally be made using the following steps:

a) State the null hypothesis, H_o. For instance in split sample tests, the null hypothesis may be that there is no difference between the samples means.

b) Choose a significance level, α.

c) Choose an appropriate statistical test. In this program all tests are nonparametric.

d) Compute the test statistic.

e) The sampling distribution of the test statistic is known and has been tabulated, and the chosen significance level then defines the region of rejection.

f) If the computed test statistic lies in the region of rejection, then the null hypothesis is rejected.

A.2 The Spearman Rank Order Serial Correlation Coefficient Test for Independence

The discharge series is put in chronological order. The series is then analyzed to determine the longest sequence of consecutive observations. The longest consecutive series is then denoted as Q_i with i ranging from 1 to N. Two sequences are created and ranks assigned to the series.

Q_1, Q_2 .... Q_{N-1} where x_i is the rank of the series Q_i, i = 1 to N-1

and Q_2, Q_3 .... Q_N where y_i is the rank of the series Q_i, i = 2 to N

Then the Spearman rank order serial correlation coefficient is:

$$S_1 = \frac{1}{2}(\Sigma x_i^2 + \Sigma y_i^2 - \Sigma d_i^2)(\Sigma x_i^2 \; \Sigma y_i^2)^{-1/2} \qquad \text{A.1}$$

where

$$\Sigma x_i^2 = (m^3 - m) / 12 - \Sigma T_x$$

$$\Sigma y_i^2 = (m^3 - m) / 12 - \Sigma T_y$$

$$m = N - 1$$

d_i is the difference in rank between x_i and y_i, and the summations are over the m pairs of x_i, y_i. Ignoring for the moment the terms in T and putting them at zero, equation A.1 becomes:

$$S_1 = 1 - (6\Sigma d_i^2) / (m^3 - m) \qquad \text{A.2}$$

the more familiar form of the Spearman rank correlation coefficient.

The terms in T adjust for tied ranks and are computed as follows. If for instance three observations in the x series were tied for ranks 17, 18, and 19 then each observation is given the rank 18; if two were tied for ranks 24 and 25, then each is ranked 24.5.

For each tied set, T is computed from:

$$T_x = (J^3 - J) / 12$$

where J is the number of observations tied at a given rank. $\sum T_x$ and $\sum T_y$ are defined by the extension of the foregoing.

For N less than 10, Table A.1 can be used for defining the region of rejection for a computed S_1 at given significance level α. When N is 10 or greater, then the function

$$t = S_1[(m - 2) / (1 - S_1^2)]^{1/2} \qquad \text{A.3}$$

is distributed like Student's t with m-2 degrees of freedom. A one-tail test must be used. Table A.2 can be used to obtain the critical values of t for various levels of significance α.

A.3 The Spearman Rank Order Correlation Coefficient Test for Trend

If the series Q_1 with i = 1 to N is put in chronological order and ranks are assigned to the series

Q_1, Q_2 .... Q_N by y_i, the rank of Q_i
and 1,2, ... N by x_i, the sequential order of Q_i

then the Spearman rank order correlation coefficient r_s is calculated as in equation A.1, except that m = N, $T_x = 0$, and the summations are taken over the N pairs of x_i, y_i.

For N less than 10, Table A.1 can be used to obtain the region of rejection for a computed value of r_s at a given significance level α.

For N = 10 or greater, then the function

$$t = r_s[(N - 2) / (1 - r_s^2)]^{1/2} \qquad \text{A.4}$$

is distributed like Student's t with N-2 degrees of freedom. The null hypothesis is that there is no trend, either upward or downward with time, and so a two-tail test is used. Table A.2 can be used to obtain the critical values.

A.4 The Mann-Whitney Split Sample Test for Homogeneity

The sample is split into two subsamples, and ranks assigned. Then the Mann-Whitney U statistic is defined by the smaller of:

$$U_1 = n_1 n_2 + n_1(n_1 + 1)/2 - R_1 \qquad \text{A.5}$$

$$\text{or } U_2 = n_1 n_2 - U_1 \qquad \text{A.6}$$

where n_1 is the size of the smaller subsample
n_2 is the size of the larger subsample
R_1 is the sum of the ranks in subsample n_1

For both n_1 and n_2 less than 21, the critical values of U have been tabulated which define the region of rejection. These values can be found in Table A.3. For n_1 greater than 4 and n_2 greater than 20, the sampling distribution of U rapidly tends to normality with:

$$z = \frac{U - n_1 n_2/2}{\left\{ \left[\frac{n_1 n_2}{N(N-1)} \right] \left[\frac{N^3 - N}{12} - \Sigma T \right] \right\}^{1/2}} \qquad \text{A.7}$$

$T = (J^3 - J)/12$, where J is the number of observations tied at a given rank. The summation of T is over all groups of tied observations in both subsamples.

z is an N (0,1) variate and in the applications of the Mann-Whitney test used in this program, the region of rejection is:

z less than - 1.645 for $\alpha = 0.05$
z less than - 2.326 for $\alpha = 0.01$

Table A.4 lists value of area for the standard normal variate, z.

A.5 RUNS ABOVE AND BELOW THE MEDIAN FOR GENERAL RANDOMNESS

This randomness test is based on the order or sequence in which the individual scores or observations where obtained. Actually, the test is based on the number of runs that a sample exhibits. A run is defined as a succession of identical symbols that are followed and preceded by different symbols or by no symbols at all.

The total number of runs in a sample of any given size gives an indication of whether or not the sample is random. If very few runs occur, a time trend or some bunching due to lack of independence is suggested. If a great many runs occur, systematic short-period cyclical fluctuations seem to be influencing the sample.

For example, once the median of the sample has been determined, each observation can be labelled as being above and equal to or below and equal to the median. If "A" represents above and equal to the median and "B" represents below and equal to the median, then a sample may be ordered as:

AABBBABBBBAABA

(A run represents a succession of identical symbols.) For our example, each run is underscored and numbered consecutively:

$$\underset{1}{\underline{AA}}\ \underset{2}{\underline{BBB}}\ \underset{3}{\underline{A}}\ \underset{4}{\underline{BBBB}}\ \underset{5}{\underline{AA}}\ \underset{6}{\underline{B}}\ \underset{7}{\underline{A}}$$

This sample begins with 2 observations above or equal to the median, followed by a run of 3 observations below or equal to the median, etc.

Seven runs are observed in all: that is, the total number of runs above and below the median, RUNAB, is 7. If n_1 represents the number of events of type A, then $n_1 = 6$. If n_2 denotes the number below the median, type B, then $n_2 = 8$. Thus, the number of observations is equal to $(n_1 + n_2)$.

In order to apply this run test, one must determine n_1, n_2, and RUNAB. If both n_1 and n_2 are equal to or less than 20, then Table A.5 gives the critical values of RUNAB under H_o for $\alpha = .05$. These are critical values from the sampling distribution of RUNAB under H_o. If the observed value of RUNAB falls between the critical values, H_o is accepted. If the observed value of RUNAB is equal to or more extreme than one of the critical values, then H_o is rejected.

Table A.5 consists of two tables. Table A.5(a) gives values of RUNAB which are so small that the probability associated with their occurrence under H_o is .025. Table A.5(b) gives values of RUNAB that are so large that the probability associated with their occurrence under H_o is .025. Any observed value of RUNAB that is equal to or less than the value shown in Table A.5(a) or is equal to or larger than the value shown in Table A.5(b) is in the region of rejection for $\alpha = .05$.

The null hypothesis, H_o, is that the A's and B's occur in random order. The alternate hypothesis, H_1, is that the order of the A's and B's deviates from randomness.

When either n_1 or n_2 is greater than 20, the sampling distribution of RUNAB tends to normality with:

$$z = \frac{|\,\text{RUNAB} - [(2n_1n_2)/(n_1 + n_2) + 1]\,|}{\{2n_1n_2(2n_1n_2 - n_1 - n_2)/[(n_1 + n_2)^2(n_1 + n_2 - 1)]\}^{1/2}} \qquad \text{A.8}$$

where z is an N(0,1) variate as described in Table A.4. This package uses a region of rejection defined by

z greater than 1.96 for $\alpha = .05$.
z greater than 2.575 for $\alpha = .01$.

TABLE A.1

Table of critical values of r_s,
the Spearman rank correlation coefficient (Siegel, 1956).

N	Significance level (one-tailed test)	
	.05	.01
4	1.000	
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	.359	.508
24	.343	.485
26	.329	.465
28	.317	.448
30	.306	.432

df	Level of significance for one-tailed test					
	.10	.05	.025	.01	.005	.0005
	Level of significance for two-tailed test					
	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Table A.2: Table of critical values of Student's t (Siegel, 1956)

Mann-Whitney *U*

One-tailed test at .05 level; two-tailed test at .10 level

m \ n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	0																	
4	—	—	0	1																
5	—	0	1	2	4															
6	—	0	2	3	5	7														
7	—	0	2	4	6	8	11													
8	—	1	3	5	8	10	13	15												
9	—	1	4	6	9	12	15	18	21											
10	—	1	4	7	11	14	17	20	24	27										
11	—	1	5	8	12	16	19	23	27	31	34									
12	—	2	5	9	13	17	21	26	30	34	38	42								
13	—	2	6	10	15	19	24	28	33	37	42	47	51							
14	—	3	7	11	16	21	26	31	36	41	46	51	56	61						
15	—	3	7	12	18	23	28	33	39	44	50	55	61	66	72					
16	—	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83				
17	—	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96			
18	—	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109		
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
21	0	5	11	19	26	34	41	49	57	65	73	81	89	97	105	113	121	130	138	146
22	0	5	12	20	28	36	44	52	60	68	77	85	94	102	111	119	128	136	145	154
23	0	5	13	21	29	37	46	54	63	72	81	90	98	107	116	125	134	143	152	161
24	0	6	13	22	30	39	48	57	66	75	85	94	103	112	122	131	141	150	160	162
25	0	6	14	23	32	41	50	60	69	79	89	98	108	118	128	137	147	157	167	177
26	0	6	15	24	33	43	53	62	72	82	92	103	113	123	133	143	154	164	174	185
27	0	7	15	25	35	45	55	65	75	86	96	107	117	128	139	149	160	171	182	192
28	0	7	16	26	36	46	57	68	78	89	100	111	122	133	144	156	167	178	189	200
29	0	7	17	27	38	48	59	70	82	93	104	116	127	138	150	162	173	185	196	208
30	0	7	17	28	39	50	61	73	85	96	108	120	132	144	156	168	180	192	204	216
31	0	8	18	29	40	52	64	76	88	100	112	124	136	149	161	174	186	199	211	224
32	0	8	19	30	42	54	66	78	91	103	116	128	141	154	167	180	193	206	218	231
33	0	8	19	31	43	56	68	81	94	107	120	133	146	159	172	186	199	212	226	239
34	0	9	20	32	45	57	70	84	97	110	124	137	151	164	178	192	206	219	233	247
35	0	9	21	33	46	59	73	86	100	114	128	141	156	170	184	198	212	226	241	255
36	0	9	21	34	48	61	75	89	103	117	131	146	160	175	189	204	219	233	248	263
37	0	10	22	35	49	63	77	91	106	121	135	150	165	180	195	210	225	240	255	271
38	0	10	23	36	50	65	79	94	109	124	139	154	170	185	201	216	232	247	263	278
39	1	10	23	38	52	67	82	97	112	128	143	159	175	190	206	222	238	254	270	286
40	1	11	24	39	53	68	84	99	115	131	147	163	179	196	212	228	245	261	278	294

This table is reprinted from *American Statistical Association Journal* (September 1964), pp. 927-932.

Mann-Whitney *U* (*Continued*)

One-tailed test at .025 level; two-tailed test at .05 level

m \ n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	0																
5	—	—	0	1	2															
6	—	—	1	2	3	5														
7	—	—	1	3	5	6	8													
8	—	0	2	4	6	8	10	13												
9	—	0	2	4	7	10	12	15	17											
10	—	0	3	5	8	11	14	17	20	23										
11	—	0	3	6	9	13	16	19	23	26	30									
12	—	1	4	7	11	14	18	22	26	29	33	37								
13	—	1	4	8	12	16	20	24	28	33	37	41	45							
14	—	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	—	1	5	10	14	19	24	29	34	39	44	49	54	59	64					
16	—	1	6	11	16	21	26	31	37	42	47	53	59	64	70	75				
17	—	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	—	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		
19	—	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	—	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	—	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	—	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	—	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	—	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	—	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	163
26	—	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	—	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	—	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	—	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	—	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200
31	—	5	14	24	34	45	56	67	78	90	101	113	125	136	148	160	172	184	196	208
32	—	5	14	24	35	46	58	69	81	93	105	117	129	141	153	166	178	190	203	215
33	—	5	15	25	37	48	60	72	84	96	108	121	133	146	159	171	184	197	210	222
34	—	5	15	26	38	50	62	74	87	99	112	125	138	151	164	177	190	203	217	230
35	—	6	16	27	39	51	64	77	89	103	116	129	142	156	169	183	196	210	224	237
36	—	6	16	28	40	53	66	79	92	106	119	133	147	161	174	188	202	216	231	245
37	—	6	17	29	41	55	68	81	95	109	123	137	151	165	180	194	209	223	238	252
38	—	6	17	30	43	56	70	84	98	112	127	141	156	170	185	200	215	230	245	259
39	0	7	18	31	44	58	72	86	101	115	130	145	160	175	190	206	321	236	252	267
40	0	7	18	31	45	59	74	89	103	119	134	149	165	180	196	211	227	243	258	274

Table A.3: Table of critical values of Mann-Whitney's U (Roscoe, 1969)

Mann-Whitney *U* (*Continued*)
One-tailed test at .01 level; two-tailed test at .02 level

m	n = 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	—																
5	—	—	—	0	1															
6	—	—	—	1	2	3														
7	—	—	0	1	3	4	6													
8	—	—	0	2	4	6	7	9												
9	—	—	1	3	5	7	9	11	14											
10	—	—	1	3	6	8	11	13	16	19										
11	—	—	1	4	7	9	12	15	18	22	25									
12	—	—	2	5	8	11	14	17	21	24	28	31								
13	—	0	2	5	9	12	16	20	23	27	31	35	39							
14	—	0	2	6	10	13	17	22	26	30	34	38	43	47						
15	—	0	3	7	11	15	19	24	28	33	37	42	47	51	56					
16	—	0	3	7	12	16	21	26	31	36	41	46	51	56	61	66				
17	—	0	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77			
18	—	0	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88		
19	—	1	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	
20	—	1	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114
21	—	1	5	11	17	23	30	36	43	50	57	64	71	78	85	92	99	106	113	121
22	—	1	6	11	18	24	31	38	45	53	60	67	75	82	90	97	105	112	120	127
23	—	1	6	12	19	26	33	40	48	55	63	71	79	87	94	102	110	118	126	134
24	—	1	6	13	20	27	35	42	50	58	66	75	83	91	99	108	116	124	133	141
25	—	1	7	13	21	29	36	45	53	61	70	78	87	95	104	113	122	130	139	148
26	—	1	7	14	22	30	38	47	55	64	73	82	91	100	109	118	127	136	146	155
27	—	2	7	15	23	31	40	49	58	67	76	85	95	104	114	123	133	142	152	162
28	—	2	8	16	24	33	42	51	60	70	79	89	99	109	119	129	139	149	159	169
29	—	2	8	16	25	34	43	53	63	73	83	93	103	113	123	134	144	155	165	176
30	—	2	9	17	26	35	45	55	65	76	86	96	107	118	128	139	150	161	172	182
31	—	2	9	18	27	37	47	57	68	78	89	100	111	122	133	144	156	167	178	189
32	—	2	9	18	28	38	49	59	70	81	92	104	115	127	138	150	161	173	185	196
33	—	2	10	19	29	40	50	61	73	84	96	107	119	131	143	155	167	179	191	203
34	—	3	10	20	30	41	52	64	75	87	99	111	123	135	148	160	173	185	198	210
35	—	3	11	20	31	42	54	66	78	90	102	115	127	140	153	165	178	191	204	217
36	—	3	11	21	32	44	56	68	80	93	106	118	131	144	158	171	184	197	211	224
37	—	3	11	22	33	45	57	70	83	96	109	122	135	149	162	176	190	203	217	231
38	—	3	12	22	34	46	59	72	85	99	112	126	139	153	167	181	195	209	224	238
39	—	3	12	23	35	48	61	74	88	101	115	129	144	158	172	187	201	216	230	245
40	—	3	13	24	36	49	63	76	90	104	119	133	148	162	177	192	207	222	237	252

Mann-Whitney *U* (*Continued*)
One-tailed test at .005 level; two-tailed test at .01 level

m	n = 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—																			
2	—	—																		
3	—	—	—																	
4	—	—	—	—																
5	—	—	—	—	0															
6	—	—	—	0	1	2														
7	—	—	—	0	1	3	4													
8	—	—	—	1	2	4	6	7												
9	—	—	0	1	3	5	7	9	11											
10	—	—	0	2	4	6	9	11	13	16										
11	—	—	0	2	5	7	10	13	16	18	21									
12	—	—	1	3	6	9	12	15	18	21	24	27								
13	—	—	1	3	7	10	13	17	20	24	27	31	34							
14	—	—	1	4	7	11	15	18	22	26	30	34	38	42						
15	—	—	2	5	8	12	16	20	24	29	33	37	42	46	51					
16	—	—	2	5	9	13	18	22	27	31	36	41	45	50	55	60				
17	—	—	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70			
18	—	—	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81		
19	—	0	3	7	12	17	22	28	33	39	45	51	57	63	69	74	81	87	93	
20	—	0	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105
21	—	0	3	8	14	19	25	32	38	44	51	58	64	71	78	84	91	98	105	112
22	—	0	4	9	14	21	27	34	40	47	54	61	68	75	82	89	96	104	111	118
23	—	0	4	9	15	22	29	35	43	50	57	64	72	79	87	94	102	109	117	125
24	—	0	4	10	16	23	30	37	45	52	60	68	75	83	91	99	107	115	123	131
25	—	0	5	10	17	24	32	39	47	55	63	71	79	87	96	104	112	121	129	138
26	—	0	5	11	18	25	33	41	49	58	66	74	83	92	100	109	118	127	135	144
27	—	1	5	12	19	27	35	43	52	60	69	78	87	96	105	114	123	132	142	151
28	—	1	5	12	20	28	36	45	54	63	72	81	91	100	109	119	128	138	148	157
29	—	1	6	13	21	29	38	47	56	66	75	85	94	104	114	124	134	144	154	164
30	—	1	6	13	22	30	40	49	58	68	78	88	98	108	119	129	139	150	160	170
31	—	1	6	14	22	32	41	51	61	71	81	92	102	113	123	134	145	155	166	177
32	—	1	7	14	23	33	43	53	63	74	84	95	106	117	128	139	150	161	172	184
33	—	1	7	15	24	34	44	55	65	76	87	98	110	121	132	144	155	167	179	190
34	—	1	7	16	25	35	46	57	68	79	90	102	113	125	137	149	161	173	185	197
35	—	1	8	16	26	37	47	59	70	82	93	105	117	129	142	154	166	179	191	203
36	—	1	8	17	27	38	49	60	72	84	96	109	121	134	146	159	172	184	197	210
37	—	1	8	17	28	39	51	62	75	87	99	112	125	138	151	164	177	190	203	217
38	—	1	9	18	29	40	52	64	77	90	102	116	129	142	155	169	182	196	210	223
39	—	2	9	19	30	41	54	66	79	92	106	119	133	146	160	174	188	202	216	230
40	—	2	9	19	31	43	55	68	81	95	109	122	136	150	165	179	193	208	222	237

Table A.3 (continued)

$$F(z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4430	.4441
1.6	.4452	.4463	.4474	.4485	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4762	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4727	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4980	.4980	.4981
2.9	.4981	.4982	.4983	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998	.4998
.	.									
.	.									
.	.									
4.0	.499968									
5.0	.4999997									

This table gives the probability of a random value of a normal variate falling in the range $z = 0$ to $z = z$ (in the *shaded area in figure*). The probability of the same variate having a deviation greater than z is given by 0.5 − probability from the table for the given z. The table refers to a single tail of the normal distribution; therefore the probability of a normal variate falling in the range $\pm z$ = 2 × probability from the table for the given z. The probability of a variate falling outside the range $\pm z$ is 1 − 2 × probability from the table for given z.

Table A.4: Areas under the normal curve (Kennedy and Neville, 1976)

n_1 \ n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

(a)

n_1 \ n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3																			
4				9	9														
5			9	10	10	11	11												
6			9	10	11	12	12	13	13	13	13								
7				11	12	13	13	14	14	14	14	15	15	15					
8				11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9					13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10					13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11					13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12					13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13						15	16	17	18	19	19	20	20	21	21	22	22	23	23
14						15	16	17	18	19	20	20	21	22	22	23	23	23	24
15						15	16	18	18	19	20	21	22	22	23	23	24	24	25
16							17	18	19	20	21	21	22	23	23	24	25	25	25
17							17	18	19	20	21	22	23	23	24	25	25	26	26
18							17	18	19	20	21	22	23	24	25	25	26	26	27
19							17	18	20	21	22	23	23	24	25	26	26	27	27
20							17	18	20	21	22	23	24	25	25	26	27	27	28

(b)

Table A.5: Table of critical values of RUNAB in the runs test (Siegel, 1956). Given in the bodies of Table A.5(a) and Table A.5(b) are various critical values of RUNAB for various values of n_1 and n_2 . For the one-sample runs test, any value of RUNAB which is equal to or smaller than that shown in Table A.5(a) or equal to or larger than that shown in Table A.5(b) is significant at the .05 level.

APPENDIX B: The Distributions

The following appendix is intended to summarize the mathematical relationships upon which the computations are based and to describe the algorithms used in their solution.

B.1 Population Statistics

B.1.1 Moments

Given a sample $x_1, \ldots, x_N$, the best estimates of the population mean, standard deviation, coefficient of skew, and coefficient of kurtosis, as computed herein, are given by:

$$\text{Mean, } \bar{x} = (1/N) \sum x \qquad \text{B.1}$$

$$\text{Standard deviation, } s = \{ [1/(N-1)] \sum (x - \bar{x})^2 \}^{1/2} \qquad \text{B.2}$$

$$\text{Coefficient of skew, } g_1 = \{ [N^2/[(N-1)(N-2)]\} (m_3/s^3) \qquad \text{B.3}$$

$$\text{Coefficient of kurtosis, } g_2 = \{ N^2(N+1)/[(N-1)(N-2)(N-3)]\} (m_4/s^4) \qquad \text{B.4}$$

where m_3 and m_4, the third and fourth central moments, respectively, are defined by

$$m_3 = (1/N) \sum (x-\bar{x})^3 \qquad \text{B.5}$$

$$m_4 = (1/N) \sum (x-\bar{x})^4 \qquad \text{B.6}$$

The summations are carried out over the N terms of the data series. When working in logarithmic units, the same definitions apply with x replaced by its natural logarithm, lnx.

B.1.2 L-Moments

Given an ordered sampled $x_1 \le x_2 \le \ldots \le x_N$, sample L-moments may be estimated and can, in turn, be used to estimate the parameter for various distributions. L-moments can be obtained from the relations (Hosking, 1989):

$$b_r = N^{-1} \sum_{j=1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(N-1)(N-2)\ldots(N-r)} x_j \qquad \text{B.7}$$

$$P^*_{r-1,k} = (-1)^{r-k-1} \binom{r-1}{k} \binom{r+k-1}{k} \qquad \text{B.8}$$

$$l_r = \sum_{k=0}^{r-1} P^*_{r-1,k}\, b_k \qquad \text{B.9}$$

where l_r is an unbiased estimator of the population L-moment, λ_r.

L-moment ratios are defined as:

$$\tau_r = \lambda_r/\lambda_2 \qquad ; r \geq 3 \qquad \text{B.10}$$

and can be estimated by $t_r = l_r/l_2$. Hosking (1989) gives the bounds on λ_1, λ_2, τ_3, and τ_4 to be

$$0 < \lambda_2, \quad -1 < \tau_3 < 1, \quad 1/4\ (5\tau_3^2 - 1) \leq \tau_4 < 1 \qquad \text{B.11}$$

λ_1 is the arithmetic average, and λ_2 is a measure of the dispersion of the sample similar to the standard deviation. τ_3 is considered to be a measure of skewness, while τ_4 reflects the peakedness of the sample — kurtosis.

L-moments can also be estimated based on plotting position formulae. Greenwood, et al. (1979) introduced the concept of probability weighted moments, which can be used for the estimation of the parameters of various distributions. Hosking (1989) showed that probability weighted moments are linear combinations of L-moments. He noted, however, that L-moments "are more convenient ... because they are more directly interpretable as measures of the scale and shape of probability distributions."

Given a plotting position formulae of the form

$$p_{i:N} = (i + A)/(N + B) \text{ for } B > A > -1 \qquad \text{B.12}$$

then the L-moments can be estimated from (Hosking, 1989)

$$\hat{\lambda}_r = \frac{1}{N} \sum_{i=1}^{n} P^*_{r-1}(p_{i:N}) x_i \qquad \text{B.13}$$

where $P^*_{r-1}(p_{i:N})$ is the (r-1)th shifted Legendre polynomial. That is,

$$\hat{\lambda}_2 = \frac{1}{N} \sum_{i=1}^{N} [2\ (p_{i:N}) - 1] x_i \qquad \text{B.14}$$

$$\hat{\lambda}_3 = \frac{1}{N} \sum_{i=1}^{N} [6\ (p_{i:N})^2 - 6\ (p_{i:N}) + 1] x_i \qquad \text{B.15}$$

$$\hat{\lambda}_4 = \frac{1}{N} \sum_{i=1}^{N} [20 (p_{i:N})^2 - 30 (p_{i:N})^2 + 12 (p_{i:N}) - 1] x_i \qquad \text{B.16}$$

while $\hat{\lambda}_1$ is the arithmetic average. The plotting position based method yield consistent but biased estimates of λ_r. Hosking (1989) writes that

> "There is no theoretical reason for preferring plotting-position estimators to the unbiased estimators, but practical experience shows that plotting-

position estimators sometimes yield better estimates of parameters and quantiles when a distribution is fitted to data. In particular the choice $P_{i:n}$ = (i-0.35)/n gives good results...".

The plotting-position based approach is adopted for use in CFA. Programming was adapted from work provided by Hosking (1988).

B.1.3 Historically-Weighted Population Statistics

Given that

NA = the number of floods above and equal to the threshold
NB = the number of floods below the threshold
NC = the number of censored floods

and YT = NA + NB + NC, the total time span,

$$\text{then} \quad W = (YT - NA)/NB \tag{B.17}$$

$$\text{and} \quad \tilde{x} = (W \sum x_b + \sum x_a)/(YT - WL) \tag{B.18}$$

$$\tilde{s} = [(W \sum d_b^2 + \sum d_a^2) \,/\, (YT - WL - 1)]^{1/2} \tag{B.19}$$

$$\tilde{g}_1 = [(YT - WL)(W \sum d_b^3 + \sum d_a^3 \,/\, \tilde{s}^3] \,/\, [(YT - WL - 1)(YT - WL - 2)] \tag{B.20}$$

where x_a is a flood above or equal to the threshold, x_b is a flood below the threshold, d_a and d_b are the deviations of x_a and x_b from $\tilde{x}$ and L is the number of low outliers and may include zero flow events. $\tilde{x}$, $\tilde{s}$, and $\tilde{g}_1$ represent the historically-weighted mean, standard deviation, and coefficient of skew, respectively, of the historic sample. Note that the portion of the record where sampling did not occur but historic information exists is weighted in the guise of the recorded floods below the threshold. The weighted statistics can be used to obtain moment estimates of the parameters of the distribution.

B.2 The Generalized Extreme Value (GEV) Distribution

The concept of a GEV distribution is beneficial if an extreme value distribution is to be used but the type is unknown. When sample data are available, the GEV distribution may be fitted to it by various methods such as moments, maximum likelihood, and L-moments. The fitting procedure used herein is the method of L-moments as adopted from Hosking (1988). Hosking, et al. (1985) and Hosking (1990) have demonstrated that the probability weighted moments (L-moments) approach yielded more accurate quantile estimates than the maximum likelihood approach for sample sizes common in flood hydrology.

The exact extreme value distribution is determined by the value of the parameter k. In the event of an estimated k value being close to zero ($< 10^{-5}$), the data are fitted directly to the EV1 distribution. This is because the parameters μ and α of the EV1 distribution can be more efficiently estimated than the μ and α values estimated for the GEV distribution.

B.2.1 The Density Function of the EV1

The probability density function of the EV1 distribution can be expressed as

$$f(x) = \frac{1}{\alpha} \exp\,[-(x - \mu)/\alpha - e^{-(x-\mu)/\alpha}] \qquad \text{B.21}$$

The distribution function is

$$F(x) = \exp\,(-e^{-(x - \mu)/\alpha}) \qquad \text{B.22}$$

where α is a scale parameter ($\alpha > 0$), μ is a location parameter, and F(x) is the probability of a value of the variate being less than x.

The k referred to previously is zero and does not appear in these expressions. This distribution is unbounded below and above.

B.2.1.1 L-Moment Estimates of the EV1 - No Historic Information

On rare occasions, the shape parameter, k, of the generalized extreme value distribution will be theoretically zero or will have been set to zero as it was found to be less than 1×10^{-5}. In such cases, parameters of the EV1 distribution are estimated via L-moments (Hosking, 1988; 1990) using the following equation:

$$\hat{\alpha} = \hat{\lambda}_2 \,/\, \ln 2 = 1.442695\ \hat{\lambda}_2 \qquad \text{B.23}$$

$$\hat{\mu} = \lambda_1 - (\gamma\lambda^2)/\ln 2 = \lambda_1 - .8327462\ \hat{\lambda}_2 \qquad \text{B.24}$$

where λ_1 is the arithmetic average or first L-moment, $\hat{\lambda}_2$ is the second L-moment found from equation B.14, and γ is Euler's constant. The estimated parameter $\hat{\alpha}$ is then substituted into equation B.24 to give the estimate to $\hat{\mu}$.

B.2.1.2 Moment Estimates of the EV1 - with Historic Information

When historic information is present, parameters are estimated from the moment relations:

$$\hat{\alpha} = \frac{\tilde{s}\sqrt{6}}{\pi} = .779697\ \tilde{s} \qquad \text{B.25}$$

$$\hat{\mu} = \tilde{x} - \gamma\hat{\alpha} = \tilde{x} - .57721566\ \hat{\alpha} \qquad \text{B.26}$$

where γ is Euler's constant, and $\tilde{x}$, and $\tilde{s}$ are the historically-weighted mean and standard deviation as obtained using equations B.18 and B.19. The estimated value of $\tilde{x}$ is substituted into B.26 to obtain $\hat{\mu}$.

B.2.2 *The Density Function of the EV2 and EV3*

The probability density function of the EV2 and EV3 distribution can be expressed as

$$f(x) = \frac{1}{\alpha}\left(1 - \frac{x-\mu}{\alpha}k\right)^{1/k-1} e^{-[1-k(x-\mu)/\alpha]^{1/k}} \qquad \text{B.27}$$

The distribution function is

$$F(x) = e^{-[1-k(x-\mu)/\alpha]^{1/k}} \qquad \text{B.28}$$

where α, μ, and F(x) are as previously defined for the EV1 and k is a shape parameter. For the EV2 distribution, k is less than zero and the distribution is lower bounded at $(\mu + \alpha/k)$ and unbounded above. For the EV3 distribution, k is greater than zero and the distribution is upper bounded at $(\mu + \alpha/k)$ and unbounded below.

B.2.2.1 *L-moment Estimates of the EV2 and EV3 - No Historic Information*

The method of L-moments as documented by Hosking (1988) is adopted for the estimation of the parameters of the EV2 and EV3 distributors. The shape parameter, k, is estimated from the relations (Hosking, 1988)

$$k = 7.817740z + 2.930462z^2 + 13.641492z^3 + 17.206675z^4 \qquad \text{B.29}$$

$$\text{and} \qquad z = 2/(\hat{\tau}_3 + 3) - \ln 2/\ln 3 \qquad \text{B.30}$$

where $\hat{\tau}_3$ is the L-moment ratio of $\hat{\lambda}_3/\hat{\lambda}_2$. Equations B.14 and B.15 are used to estimate $\hat{\lambda}_2$ and $\hat{\lambda}_3$, respectively, when $-0.1 \le \hat{\tau}_3 \le 0.5$. When $\hat{\tau}_3$ lies outside this range, a Newton-Raphson iteration procedure is used to increase the accuracy of the polynomial relationship. Once k has been estimated, α and μ are found from (Hosking, 1990):

$$\alpha = \hat{\lambda}_2 k / \{(1-2^{-k})\Gamma(1+k)\} \qquad \text{B.31}$$

$$\text{and} \qquad \mu = \lambda_1 + \alpha\{\Gamma(1+k) - 1\}/k \qquad \text{B.32}$$

where λ_1 and $\hat{\lambda}_2$ are as previously defined. The estimated value of α from equation B.31 is substituted into equation B.32 to obtain the estimate of the location parameter μ.

B.2.2.2 *Moment Estimates of the EV2 and EV3 - with Historic Information*

When historic information is present, parameters are estimated using historically weighted moments. For the EV2 distribution, the coefficient of skew is related to the shape parameter k by:

$$g_1 = \mu_3 / (\mu_2^{3/2}) \qquad \text{B.33(a)}$$

$$\mu_3 = \Gamma(1+3k) - 3\Gamma(1+2k)\Gamma(1+k) + 2\Gamma^3(1+k) \qquad \text{B.34}$$

$$\mu_2 = \Gamma(1+2k) - \Gamma^2(1+k) \qquad \text{B.35}$$

where Γ implies the gamma function such that

$$\Gamma(k) = (k-1)!$$

For the EV3 distribution, the coefficient of skew is related to the shape parameter k by:

$$g_1 = -\mu_3/(\mu_2^{3/2}) \qquad \text{B.33(b)}$$

The historically-weighted coefficient of the skew, computed using equation B.20, is substituted for g_1 in equation B.33. Parameter k can be obtained via iteration of the above equations or directly from the polynomials:

$$k = .0330113 - .106601\tilde{g}_1 + .0145418\tilde{g}_1^2 - .000940242\tilde{g}_1^3 + .00002316523\tilde{g}_1^4 \quad (3.1 \leq \tilde{g}_1 \leq 11.8) \qquad \text{B.36(a)}$$

$$k = .28732449 - .35732233\tilde{g}_1 + .11195823\tilde{g}_1^2 - .0187768\tilde{g}_1^3 + .001322037\tilde{g}_1^4 \quad (1.20 \leq \tilde{g}_1 < 3.1) \qquad \text{B.36(b)}$$

$$k = .27759 - .32188\tilde{g}_1 + .06203\tilde{g}_1^2 + .01383\tilde{g}_1^3 - .007012\tilde{g}_1^4 \quad (-.21665 < \tilde{g}_1 \leq 1.08) \qquad \text{B.36(c)}$$

$$k = .27752 - .320791\tilde{g}_1 + .068494\tilde{g}_1^2 + .0315096\tilde{g}_1^3 - .00359341\tilde{g}_1^4 \quad (-3.18508 < \tilde{g}_1 \leq -.21665) \qquad \text{B.36(d)}$$

$$k = .2179163 - .470657\tilde{g}_1 - .041434557\tilde{g}_1^2 - .001662769\tilde{g}_1^3 \quad (-8.2 \leq \tilde{g}_1 \leq -3.18508) \qquad \text{B.36(e)}$$

The polynomials accurately cover the range of skew from -8.2 to +11.8. Note that k is assumed equal to zero for the range of skew from 1.08 to 1.20. These polynomials should be sufficient for hydrologic applications. Once k has been estimated from the appropriate polynomial, the scale parameter, α, can be estimated from

$$\alpha = | k | \frac{\tilde{s}}{\sqrt{\mu_2}} \quad \text{B.37}$$

where $\tilde{s}$ is the historically-weighted standard deviation of the untransformed floods as per equation B.19 and μ_2 is as previously defined. The location parameter μ for an EV2 ($k<0$) distribution, can be obtained from

$$\mu = \tilde{x} + [1-\Gamma(1 + k)]\frac{\tilde{s}}{\sqrt{\mu_2}} \quad \text{B.38(a)}$$

μ can be estimated for an EV3 distribution ($k>0$) from

$$\mu = \tilde{x} - [1-\Gamma(1 + k)]\frac{\tilde{s}}{\sqrt{\mu_2}} \quad \text{B.38(b)}$$

where k, $\tilde{s}$, and μ_2 are as previously defined and $\tilde{x}$ represents the historically weighted arithmetic average of the untransformed floods as per equation B.1.

B.2.3 T-Year Flood Estimation - Generalized Extreme Value Distribution

B.2.3.1 Flood Estimation with No Low Outliers Present

If no low outliers are to be treated, the T-year floods are computed from

$$Q_T = \mu + \alpha \{-\ln[-\ln(1-1/T)]\} \quad \text{B.39(a)}$$

or

$$Q_T = \mu - (\alpha/k) \{[-\ln(1-1/T)]^k - 1\} \quad \text{B.39(b)}$$

Equation B.39(a) applies for the EV1 distribution, while equation B.39(b) is for the EV1 and EV2 distributions. The program automatically selects the appropriate equation and, with the appropriate return period, T, computes and displays flood estimates for return periods of 1.003, 1.05, 1.25, 2, 5, 10, 20, 50, 100, 200, and 500 years.

B.2.3.2 The Conditional Probability Function

Consider a sample of N observations, where L observations have been identified as low outliers. Now let n be the number of observations excluding the L low outliers; thus, n=N-L. For the moment, consider these L members as zeros, such as the probability of exceeding x is given by

$$P(x) = [1-F(x)] (N-L)/N \quad \text{B.40}$$

From the sample, N and L are known. F(x) is defined from the estimated parameters -- equations B.22 and B.28. However,

$$F(x) = 1-1/T \qquad \text{B.41}$$

and substituting B.41 into B.40 gives

$$P(x) = [1/T] \quad (N-L)/N \qquad \text{B.42}$$

Thus, equation B.42 can be substituted into equation B.39 in order to obtain the magnitude of the T-year event where 1/T of equation B.39 is equal to P(x). In the case of the EV1:

$$Q_T = \mu + \alpha \{ -\ln[-\ln(1-(1/T)N/(N-L))] \} \qquad \text{B.43(a)}$$

and in the case of the EV2 and EV3:

$$Q_T = \mu - (\alpha/k) \quad \{[-\ln(1-(1/T)N/(N-L))]^k - 1\} \qquad \text{B.43(b)}$$

B.2.3.3 Flood Estimation with Low Outliers Present

In the event that low outliers - zero or non-zero - have been identified in the sample, the L outliers are removed from the sample of size N and the distribution parameters are estimated from the n remaining observations, either by maximum likelihood or by moments as previously outlined. This procedure is followed in the case of a historic analysis as well as a conventional analysis. Floods for any desired return period or probability of exceedance can then be estimated using these parameters, known values of N and L, and equation B.43.

However, this procedure will give a probability function which is not truly that of the distribution. Furthermore, it is not possible to estimate floods of return periods less than N/n because, in the conditional probability function, their exceedance probabilities then become greater than one.

To avoid these difficulties, "synthetic" EV1, EV2, and EV3 distributions are fitted through portions of the "conditional" frequency curve. Since the EV1 has two parameters, two points on the curve are necessary. Parameters for the EV1 follow from

$$\alpha = (Q_{100} - Q_2)/4.23364 \qquad \text{B.44}$$

$$\mu = Q_{100} - 4.60015\alpha \qquad \text{B.45}$$

The EV2 and EV3 have three parameters, thus three points on the curve are necessary. The mathematics are considerably simplified by choosing the floods of return periods 1.582, 10.483, and 100 years. Parameters follow from

$$\mu = Q_{1.582} \qquad \text{B.46}$$

$$k = -.21738 \ln [\, 1 + (Q_{1.582} - Q_{100}) / Z_1 \,] \qquad \text{B.47}$$

$$\alpha = k\, Z_1 \qquad \text{B.48}$$

where

$$Z_1 = [(Q_{1.582} Q_{100} - Q^2_{10.483}) / (Q_{1.582} + Q_{100} - 2Q_{1.582})] - Q_{1.582}$$

The "retro-fitted" parameters of the EV1 and of the EV2 and EV3 distributions can be used in equation B.39(a) or B.39(b) to produce revised flood estimates for the required return periods.

Note that the procedure of retro-fitting the parameters to produce a synthetic EV distribution is used only when all of the low outliers are non-zero. The procedure is not used when zeros are present in the sample because it can result in an unrealistic situation whereby the exceedance probability of a non-zero flow is larger than the exceedance probability of zero flow as indicated by the sample.

B.3 The Three-Parameter Lognormal Distribution

B.3.1 The Density Function

$$f(x) = \frac{1}{\sigma (x-a) \sqrt{2\pi}} \exp \left\{ \frac{\{-[\ln(x-a) - m]^2\}}{2\sigma^2} \right\} \qquad \text{B.49}$$

where m and σ^2 are, respectively, location, and scale parameters for the transformed variate ln(x-a). For the untransformed variate, the skewness is a function of σ^2, and "a" is the lower boundary of the variate x.

B.3.2 Maximum Likelihood Estimates - No Historic Information

Applying maximum likelihood theory to the density function, equation B.49 gives:

$$\partial \ln L / \partial a = (1 - m/\sigma^2) \sum (x-a)^{-1} + (1/\sigma^2) \sum \{[\ln(x-a)](x-a)^{-1}\} = 0 \qquad \text{B.50}$$

$$\partial \ln L / \partial m = (1/\sigma^2) \sum \{\ln(x-a) - m\} = 0 \qquad \text{B.51}$$

$$\partial \ln L / \partial \sigma^2 = -n/2\sigma^2 + [1/(2\sigma^4)] \sum \{\ln(x-a) - m\}^2 = 0 \qquad \text{B.52}$$

Equations B.50, B.51, and B.52 can be re-arranged to give an equation in "a" only:

$$f(a) = \sum \frac{(\frac{1}{n}) \sum \ln^2(x-a) - [(\frac{1}{n}) \sum \ln(x-a)]^2 - (\frac{1}{n}) \sum \ln(x-a)}{(x-a)} + \sum \frac{\ln(x-a)}{(x-a)} = 0 \qquad \text{B.53}$$

Note that, in this application, n is the number of data items after the removal of any low outliers or zeros: n=N-L.

To solve Equation B.53, f(a) is evaluated at an initial value of $a = 0.975x_{min}$. and the sign of the imbalance is noted. At an arbitrary extreme value of a, $a = -80\bar{x}$, f(a) is again evaluated. If a sign change has occurred, indicating that a root lies between $0.975x_{min.}$ and $-80\bar{x}$, then a binary search is made in this interval until the equation is balanced. If no sign change has occurred, then the root is assumed to be greater than $x_{max.}$.

Having found a, then

$$m = (1/n) \sum \ln(x-a) \qquad \text{B.54}$$

and, correcting for bias,

$$\sigma^2 = \{1/(n-1)\} \sum \{\ln(x-a)-m\}^2 \qquad \text{B.55}$$

Hence, the estimated parameters a, m, and σ^2.

Experience has shown that a root, if it exists, will always be greater than -80 times the mean of the data sample. For a positively skewed sample, it must also, of course, be less than the minimum member of the data sample.

If the skewness of the sample is negative, then equations B.49 through B.55 still hold, except that (x-a) is replaced by (a-x) and the distribution becomes bounded above. The root in this case lies to the right of $x_{max.}$ and the solution procedure starts at $1.025x_{max.}$ and searches to the right to $+80\bar{x}$.

B.3.3 Moment Estimates - No Historic Information

On rare occasions, the maximum likelihood method will fail to find a root and, in these cases only, moment estimates are made from the following set of equations. First solve the following function for c:

$$c^3 + 3c - g_1 = 0 \qquad \text{B.56}$$

then

$$a = \bar{x} - s/c \qquad \text{B.57}$$

$$m = \ln(|s/c|) - \tfrac{1}{2}\ln(c^2+1) \qquad \text{B.58}$$

$$\sigma^2 = \ln(c^2+1) \qquad \text{B.59}$$

where $\bar{x}$, s, and g_1 are computed from equations B.1, B.2, and B.3.

In equation B.56, for a positive value of g_1, there is only one real root for c, which is positive. In the event that g_1 is negative, then there is again only one real root, but this time c is negative.

It should be noted, that for negative skew, the three-parameter lognormal distribution becomes unbounded below and has an upper bound at the parameter "a".

B.3.4 *Maximum Likelihood Estimates - With Historic Information*

Given N observed sample members, plus a further r members which are known only to be less than the censoring threshold x_c, maximum likelihood estimates of parameters are obtained by solving the transcendental system:

$$[\Sigma t_i - r f(t_c) / F(t_c)]\, \sigma = 0 \qquad \text{B.60}$$

$$[-N + \Sigma t_i^2 - r t_c f(t_c) / F(t_c)] / \sigma = 0 \qquad \text{B.61}$$

$$\sigma \Sigma [1 / (x_i - a)] + \Sigma t_i / (x_i - a) - r f(t_c) / [(x_c - a) F(t_c)] = 0 \qquad \text{B.62}$$

where $t_i = [\ln(x_i\text{-}a)\text{-}m]/\sigma$, and $f(t_c)$ and $F(t_c)$ are the ordinates of the density and probability functions respectively, at t_c, the standard normal variate corresponding to the censoring threshold.

Although far too complex to be shown here, equations B.60, B.61 and B.62 are reduced in the program to a single transcendental equation in "a" only to give

$$f(a) = 0 \qquad \text{B.63}$$

which can be solved by any numerical analysis method.

Experience has shown that, if a solution for "a" exists, it will always be greater than -80 times the mean of the data sample. Since a is the theoretical lower bound of the distribution, it must also be less than the minimum member of the data sample. Therefore, the search range is limited to $-80\bar{x} < a < x_{min.}$.

To find the root of equation B.63, the function is first evaluated with a trial value of $a = 0.975x_{min.}$. A check is made to see that, in the course of the computations, σ^2 is a positive value. If not, a is successively decremented by 0.01 $x_{min.}$ until σ^2 is a positive value. The equation is then checked for balance and the sign of the imbalance is noted. Next, the equation is evaluated assuming $a = -80\bar{x}$. If the sign of the imbalance is opposite to that of the previous evaluation, then a binary search is made between $.975x_{min.}$ and $-80\bar{x}$ until a value of "a" is found that balances the equation.

If no solution is found within the search range, that is, if the imbalances resulting from the two evaluations are of like sign, then the program switches to the historically-weighted moments technique for estimating the parameters.

In the course of reducing equations B.60, B.61 and B.62 to a single equation in "a", the other parameters, m and σ^2, are isolated. Thus, when equation B.63 is balanced, estimates have been obtained for all three parameters of the distribution.

B.3.5 Moment Estimates - with Historic Information

Section B.2 describes how to obtain the historically-weighted mean, standard deviation, and coefficient of skew. For a historic sample, these estimates can be substituted directly for the moment statistics in equations B.56, B.57, B.58 and B.59. These equations yield estimates of the parameters of the distribution based on the recorded data and the historic information.

B.3.6 T-Year Food Estimation - Three-Parameter Lognormal Distribution

B.3.6.1 Flood Estimation with No Low Outliers Present

If no low outliers are to be treated, the T-year floods are computed from:

$$Q_T = a + \exp(m + t\sigma) \qquad \text{B.64}$$

or

$$Q_T = a - \exp(m - t\sigma) \qquad \text{B.65}$$

Equation B.64 applies in the lower-bounded case, and equation B.65 applies in the upper-bounded case. The program automatically selects the appropriate equation and, with the appropriate values of the standard normal deviate, t, computes and displays flood estimates for return periods of 1.003, 1.05, 1.25, 2, 5, 10, 20, 50, 100, 200, and 500 years.

B.3.6.2 The Conditional Probability Function

Consider a sample of N observations, where L observations have been identified as low outliers Now let n be the number of observations excluding the L low outliers; thus, $n = N-L$. For the moment, consider these L members as zeros, such that the probability of exceeding x is given by

$$P(x) = [1-F(x)]\,(N-L)/N \qquad \text{B.66}$$

Thus, from the cumulative probability function, the probability of exceeding x can be found. The more usual requirement is to find x with a given probability of exceedance, but equation B.66 cannot be solved in this fashion. However, by making the substitution $[\ln(x-a)-m]/\sigma = t$ and rearranging.

$$P(x)N/(N-L) = \int_t^{\infty} \Phi(t)dt \qquad \text{B.67}$$

where t is a standard normal variate, distributed as N(0,1).

For instance, suppose we wish to find the conditional value of t, t^1, that makes P(x) = 1% when N = 20, L=1.

From tables, we have to find t^1 that makes 1-F(t)=0.01 x 20/19 or 0.0105 and $t^1 = 2.037$.

In this program, equation B.67 is solved for $t = t^1$ using the following numerical procedure. The normal probability integral with zero mean and unit standard deviation can be represented as:

$$P(t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} \exp - \left[\frac{t^2}{2}\right] dt \qquad \text{B.68}$$

where P(t) is the probability of exceeding t. Hence, it is then a question of finding the number of standard deviations corresponding to a given probability of exceedance, or solving equation B.68 for t, given P(t). The solution yields t, the number of standard deviations from the mean.

Rather then using equation B.68, the method of development by Hastings (1955) is used. Hastings (1955) gives the following solution for the interval:

$$0 < P(t) < 0.5$$

$$t = V - \frac{c_1 + c_2 V + c_3 V^2}{1 + d_1 V + d_2 V^2 + d_3 V^3} \qquad \text{B.69}$$

where $V = \{\ln[1/P(t)]^2\}^{1/2}$

and $c_1 = 2.515517$ $d_1 = 1.432788$
$c_2 = .802853$ $d_2 = .189269$
$c_3 = .010328$ $d_3 = .001308$

The error is less than 4.5 parts in 10 000.

The restriction on P(t) given is easily overcome since the number of standard deviations to P(t) = 0.95, is the same as to P(t) = 0.05, with the sign reversed.

B.3.6.3 Flood Estimation with Low Outliers Present

In the event that low outliers — zero or non-zero — have been identified in the sample, the L outliers are removed from the sample of size N and the distribution parameters are estimated from the n remaining observations, either by maximum likelihood or by moments as previously outlined. This procedure is followed in the case of a historic analysis as well as a conventional analysis. Floods for any desired return period or probability of exceedance can then be estimated using these parameters, conditional standard normal deviate (t) values derived from equation B.67, and either equation B.64 or B.65.

However, this procedure will give a probability function that is not truly three-parameter lognormal. Furthermore, it is not possible to estimate floods of return periods less than N/n because, in the conditional probability function, their exceedance probabilities then become greater than one.

To avoid these difficulties, a "synthetic" three-parameter lognormal distribution is fitted through that portion of the "conditional" frequency curve between T=2 and T=100. Since three parameters have to be estimated. three points on the curve are necessary. The mathematics are simplified by choosing the floods of return periods 2, 8.17, and 100 years. Parameters follow from

$$a = \frac{Q_{100}Q_2 - Q_{8.17}^2}{Q_2 + Q_{100} - 2Q_{8.17}} \qquad \text{B.70}$$

$$m = \ln[\text{abs}(Q_2 - a)] \qquad \text{B.71}$$

$$\sigma = \text{abs}\{\ln[\text{abs}(Q_{100} - a)] - m\}/2.326 \qquad \text{B.72}$$

Equations B.70, B.71, and B.72 hold for both positively skewed and negatively skewed samples.

These "retro-fitted" parameters are then used with unmodified values of standard normal deviate (equation B.69) and either equation B.64 or B.65 to produce revised flood estimates for the required return periods.

Note that the procedure of retro-fitting the parameters to produce a synthetic three-parameter lognormal distribution is used only when all of the low outliers are non-zero. The procedure is not used when zeros are present in the sample because it can result in an unrealistic situation whereby the exceedance probability of a non-zero flow is larger than the exceedance probability of zero flow as indicated by the sample.

B.4 The Log Pearson Type III Distribution

B.4.1 The Density Function

The probability density function (p.d.f.) of the Log Pearson Type III distribution can be expressed as:

$$f(x) = \frac{\exp\{-(\ln x - m)/a\}}{|a| x \, \Gamma(b)} (\ln x - m)/a^{b-1} \qquad \text{B.73}$$

where a, b, and m are respectively scale, shape, and location parameters and Γ is the gamma function of the argument within parenthesis.

Bobée (1975) shows the many different shapes which this p.d.f. can take. The various shapes depend on the relationship between the parameters, and not all these shapes are credible in flood frequency hydrology. This is particularly so in cases where "a" is negative. Negative values of "a" are common for flood series on Canadian rivers.

B.4.2 Maximum Likelihood Estimates - No Historic Information

Applying maximum likelihood theory to the p.d.f., equation B.73, gives

$$\frac{\partial \ln L}{\partial a} = \frac{1}{a^2} \Sigma (\ln x - m) - \frac{1}{a} N b = 0 \qquad \text{B.74}$$

$$\frac{\partial \ln L}{\partial b} = -N \psi(b) + \Sigma \ln \frac{\ln x - m}{a} = 0 \qquad \text{B.75}$$

$$\frac{\partial \ln L}{\partial m} = -\frac{N}{a} - (b-1) \Sigma \frac{1}{\ln x - m} = 0 \qquad \text{B.76}$$

where $\psi(b)$ is the Digamma or Psi function, $\partial \ln\Gamma(b)/\partial b$.

Equations B.74, B.75 and B.76 are three simultaneous transcendental equations in a, b, and m. Equation B.75 can be reduced to one transcendental equation in m, rather than in three parameters, by rearranging equations B.74 and B.76. This gives:

$$a = \frac{1}{Nb} \Sigma \ln(x-m) \qquad \text{B.77}$$

$$b = \Sigma (\ln x - m)^{-1} / \{ \Sigma (\ln x - m)^{-1} - N^2 / \Sigma (\ln x - m) \} \qquad \text{B.78}$$

Substitution of equation B.77 in equation B.75 eliminates a, then substitution of equation B.78 in equation B.75 eliminates b. The Psi function can be replaced by a combined formula and asymptotic expansion where $\psi(b)$ tends to

$$\ln(b+2) - \frac{1}{2(b+2)} - \frac{1}{12(b+2)^2} + \frac{1}{120(b+2)^4} - \frac{1}{252(b+2)^6} - \frac{1}{(1+b)} - \frac{1}{b} \qquad B.79$$

The distribution of equation B.73 can be either positively or negatively skewed. It can be shown, by considering the moments of the distribution, that if the skew coefficient of the logarithms is positive, "a" must be positive. Therefore, from equation B.75, (lnx-m) must always be positive for all values of lnx. It follows then that the distribution is bounded below at exp. m. It is more common in flood series from Canadian rivers to find that the skew coefficient of the logarithms is negative, and (lnx-m) must also be negative. Thus the distribution is bounded above at exp. m.

The solution of equation B.75, with necessary substitutions, can therefore have roots either less than the minimum of the data series or greater than the maximum of the data series. The curve corresponding to equation B.75 has several inflexions and, for reliability, a Bolzano method is used to isolate the root clear of the effect of any inflexions before the faster Newton-Raphson iteration method takes over. This method of solution has so far proved reliable.

The solution then gives the maximum likelihood estimates $\hat{a}$, $\hat{b}$, and $\hat{m}$ for the distribution parameters.

B.4.3 Moment Estimates - No Historic Information

Taking moments of the p.d.f. given by equation B.73, re-arranging and replacing them with their sample estimates gives

$$\hat{a} = \frac{s g_1}{2} \qquad B.80$$

$$\hat{b} = (2/g_1)^2 \qquad B.81$$

$$\hat{m} = \bar{x} - \frac{2s}{g_1} \qquad B.82$$

where $\bar{x}$, s, and g_1, are computed from equations B.1, B.2, and B.3 with x replaced by its natural logarithm, lnx.

B.4.4 Moment Estimates - with Historic Information

Section B.1.3 describes how to obtain the historically-weighted mean, standard deviation, and coefficient of skew for a historic sample. These estimates can be substituted directly for the moment statistics in equations B.80, B.81, and B.82. These equations yield estimates of the parameters of the distribution based on the recorded data and historic information.

B.4.5 T-Year Flood Estimation - Log Pearson Type III Distribution

B.4.5.1 Flood Estimation with No Low Outliers Present

The cumulative probability function cannot be expressed in closed form. Substituting y = (lnx -m)/a in the p.d.f. (equation B.73) and using the Wilson-Hilferty (1931) approximation of chi-square results in:

$$\ln x_T = m + a\left[\frac{t}{3b^{1/6}} - \frac{1}{9b^{2/3}} + b^{1/3}\right]^3 \qquad \text{B.83}$$

where $\ln x_T$ is the logarithm of the T-year event and t is the standard normal deviate for the required return period. Thus, $x_T = \exp(\ln x_T)$.

B.4.5.2 The Conditional Probability Function

The conditional value of the standard normal variate t^1, is obtained using the methodology outlined in section B.3.6.2. Once t^1 has been determined, it can be substituted into equation B.83 to acquire the natural logarithm of the T-year design flood.

B.4.5.3 Flood Estimation with Low Outliers Present

In the event that low outliers - zero or non-zero - have been identified in the sample, a "synthetic" Log Pearson Type III distribution is fitted using synthetic statistics. Parameters of the distribution are estimated by the moment relationships to equations B.80, B.81, and B.82.

The first step is to determine the Q_2, Q_{10}, and Q_{100} floods using the technique outlined in section B.4.5.2. The synthetic statistics (Hydrology Subcommittee, 1982) are determined based on the values from these three design floods. The relationships are:

$$g_s = -2.5 + 3.12\left[\frac{\ln(Q_{100}/Q_{10})}{\ln(Q_{10}/Q_2)}\right] \qquad \text{B.84}$$

$$s_s = \frac{\ln(Q_{100}/Q_2)}{K_{.01} - K_{.50}} \qquad \text{B.85}$$

$$\bar{x}_s = \ln(Q_{.5}) - K_{.50}(s_s) \qquad \text{B.86}$$

where g_s, s_s, and $\bar{x}_s$ are the synthetic logarithmic coefficient of skew, standard deviation, and mean respectively; and $K_{.01}$ and $K_{.50}$ are Pearson Type III deviates for exceedance probabilities of .01 and .50, respectively, and coefficient of skew. The Hydrology Subcommittee (1982) indicates that equation B.84 "is an approximation appropriate for use between skew values of +2.5 and -2.0". The relationships for determining $K_{.01}$ and $K_{.50}$ are:

$$K_{.01} = \frac{2}{g_s}\left\{\left[\left[2.326 - \frac{g_s}{6}\right]\left[\frac{g_s}{6}\right] + 1\right]^3 - 1\right\} \qquad \text{B.87}$$

$$K_{.50} = \frac{2}{g_s}\left\{1 - \left[\frac{g_s}{6}\right]^2\right\}^3 - 1 \qquad \text{B.88}$$

B.5 The Wakeby Distribution

B.5.1 The Density Function

This distribution cannot be expressed as a probability density function, nor can the probability function be expressed explicitly. In the formulation of Landwehr, et al. (1979, a,b), it is defined as an inverse function by:

$$x = m + a[1-(1-F)^b] - c[1-(1-F)^{-d}] \qquad \text{B.89}$$

where F is the probability of not exceeding x
m is a location parameter
a and c are scale parameters
b and d are shape parameters

B.5.2 L-Moment Estimates - No Historic Information

B.5.2.1 The Standard Case

The method of L-moments is used to obtain parameter estimates. The distribution itself, as well as programming, imposes limits on the magnitudes of parameters b and d, and furthermore certain combinations of both magnitudes and signs of parameters a, b, c, and d lead to improper definitions of the probability function. Since it is arithmetically possible to obtain these combinations, the program checks the parameters, and if a valid set of parameters is not obtained then the program output will inform the user.

The solution method contains three potential scenarios. The first consists of an attempt to obtain a valid set of the five parameters. If the parameters are invalid, then m is set to zero and the remaining four parameters are determined. If this combination is unacceptable, then the sample is fitted to a generalized Pareto distribution by the method of L-moments (Hosking, 1988). When the generalized Pareto distribution is involved, two of the parameters will be zero, depending on the sign of the parameter d.

When d is positive, parameters a and b are set to zero. Otherwise, parameters c and d will be set to zero.

The algorithms used herein for the estimation of parameters of the Wakeby distribution for the standard case are from the work of Hosking (1988).

Having found a valid set of parameters, then event magnitudes for any desired probability of exceedance or return period can be obtained directly from equation (B.89), since

$$P = 1/T = (1-F) \qquad \text{B.90}$$

where P is the probability of exceedance
T is the return period
F is the probability of non-exceedance

B.5.2.2 With Low Outliers and/or Zeros

If in a sample size N such that n = N-L, and L of these sample members are identified as low outliers and/or zeros, then from equations B.89 and B.90 and using the conditional function (i.e. equation B.42)

$$x = m + a\{1-[nT/(n+L)]^{-b} - c\{1-[nT/(n+L)]^{d}\} \qquad \text{B.91}$$

where m, a, b, c, and d are estimated using the n sample members only.

B.5.3 Parameter Estimation via Least Squares Regression - with Historic Information

As given by Houghton (1978), the probability function of the Wakeby in inverse form is

$$x = e - a(1-F)^{b} + c(1-F)^{-d} \qquad \text{B.92}$$

Comparing equations B.89 and B.92, they differ only in the parameters e and m, and m + a -c = e. Houghton's version of the distribution only permits positive values of a, b, c, and d, but e can be either positive or negative. The solution is not given for cases where F does not include historic information. The fitting method used in this program allows for additional combinations of the parameters a, b, c, and d, provided that certain conditions are met. At all stages a check is made to see that a valid set of parameters has been obtained. Furthermore, an estimate of F can be made which will include historic information.

Rewriting equation B.92 gives

$$x = -aP^{b} + cP^{-d} + e \qquad \text{B.93}$$

Houghton estimated F in equation B.92 using the median plotting position and herein P is estimated using the Cunnane (1978) formula, and if historic information is available, then the rank required in the Cunnane formula can be adjusted using the method of Benson (1950). The necessary equations are given in section 3.8. Thus, for every member of the data sample containing historic information, a corresponding estimate of the probability P can be made, and parameter estimates can not be obtained by regression on these probabilities. The regression method is similar to that proposed by Houghton.

Choose a cutoff point $P_c = 0.65$ which will divide the sample into two parts; an upper part such as P_k is less than P_c. Consider the upper portion first and rewrite equation B.93 to give

$$\log(x_k - e + aP_k^b) = \log(c) - d \log P_k \qquad \text{B.94}$$

which is a linear equation of the form $y = c - dx$. Start with $a = 0$, $b = 1$, $P_c = 0.65$ and

$$e = (x_{min} - 0.1) - (x_{min} - 0.1)(I - 1)/5 \qquad \text{B.95}$$

where x_{min} is the minimum member of the upper sample, and I takes on values of 1 to 60 by increments of 1. Successively decrementing e as shown in equation B.95, equation B.94 gives, by linear regression at each step, estimates of c, d, and the coefficient of determination, r^2. When a turning point in r^2 has been reached at the penultimate trial of e, then a quadratic equation is fitted to the last three values of e and r^2 and the value of e which gives a maximum r^2 is selected. Thus, we have first estimates of e, c, and d governing the upper portion of the curve. In this case, c and d are both positive and e may be positive or negative.

Assuming these values are found within the search range, then a and b follow from a further rearrangement of equation B.93:

$$\log[-x_k + e + cP_k^{-d}] = \log(a) + b \log P_k \qquad \text{B.96}$$

for all x_k such that P_k is greater than P_c.

Again, this equation is linear and parameters a and b are obtained by regression, noting that the term inside the square brackets of equation B.96 is the residual of x_k from the downward extrapolation of the preliminary fit of the upper curve. This extrapolation may be above or below the data points and, since logarithms of negative numbers do not exist, the programming has been arranged to accept the two possibilities and the appropriate sign of a parameter allocated after the regression. If the downward extrapolation is above the points, then from equation B.93, a must be positive, and vice versa. Thus, first estimates of a, b, c, and d are obtained and tested for validity. Assuming a valid set has been obtained, the entire procedure is repeated, starting at equation B.94 but using the first estimates of e, a, and b. The second set of parameters are accepted as final estimates, provided they pass the test for validity. If the second set of parameters is invalid, then the first set of parameters is accepted. This is an extremely rare occurrence. However, several problems may arise here depending on the nature of the sample.

(i) In some instances, not all residuals have the same sign, which makes the regression for a and b impossible, since logarithms cannot be taken of negative numbers. In this case, the highest member of the lower part is added to the upper part and the entire procedure is repeated until at least ninety percent of the residuals have the same sign. At this stage, if the residuals are positive, then a and b are estimated from the regression using the positive values only. At this point, a will be positive and b may be positive or negative. If b is negative, then the parameter set is illegal.

If ninety percent of the residuals are negative, then the positive values are discarded. In this case, "a" is negative and equation B.93 is rewritten:

$$\log[x_k - e - cP_k^{-d}] = \log(a) + b \log P_k \qquad \text{B.97}$$

This indicates that the backward projection of the upper curve lies below the points, residuals are positive, and therefore a is negative, but b can still be positive or negative, although b negative is still inadmissible.

(ii) When a negative b is found, either after the initial choice of $P_c = 0.65$, or during the downward accumulation of points to get residuals of the same sign, then P_c is reset to 0.43 and the entire procedure is repeated until a valid set of parameters has been obtained.

(iii) Sometimes, when adding points from the lower portion of the curve, all lower points can be absorbed. When only three lower half points remain, the absorption of another point would make a regression for a and b impossible. Therefore, all three remaining points are absorbed, a and b are assumed zero, and the regression is performed using:

$$\log(x_k - e) = \log(c) - d \log P_k \qquad \text{B.98}$$

Only one flood sample has been found where this has occurred, and a perfectly acceptable fit was obtained.

(iv) When the limit of I is reached in equation B.95, the regression of equation B.94 may not have shown any turning point in the coefficient of determination, r^2. It is then possible that either I is not large enough or, more likely, the Wakeby distribution is bounded above. In the upper bounded case, c is negative and equation B.93 is rewritten:

$$\log(-x_k + e - aP_k^b) = \log(c) - d \log P_k \qquad \text{B.99}$$

is a linear equation of the form y = c - dx. Start with a = 0, b = 1, $P_c = 0.65$ and

$$e = (x_{max} + 10)(I + 3)/4 \qquad \text{B.100}$$

where x_{max} is the largest member of the upper sample and I takes on values of 1 to 200 by increments of 1. Successively incrementing e as shown in equation B.100, equation B.99 gives, by linear regression at each step, estimates of c, d, and the coefficient of determination, r^2. When a turning point has been reached at the penultimate trial of e, then a quadratic equation is fitted to the last three pairs of e and r^2, and the value of e which gives a maximum r^2 is chosen. Thus, again we have first estimates of e, c, and d governing the upper part of the curve. In this case, both c and d are negative and e can be only positive.

Parameters a and b are then estimated by the aforementioned methods. The parameter set is always checked for validity.

Thirty-four sets of data with historic information were tested, and in all cases valid sets of parameters were obtained. Floods of any desired probability of exceedance are then obtained from equation B.93, and if return periods T are preferable, then equation B.93 gives

$$x = -aT^{-b} + cT^{d} + e \qquad \text{B.101}$$

B.5.3.1 *With Low Outliers and/or Zeros and Historic Information*

If, in a sample size n+L, with a historic time span of YT years, L sample members have been identified as low outliers and/or zeros, then estimates of P are made still using the L low outliers and/or zeros. First, estimates of parameters e, c, and d are made as previously described, but estimates of a and b are made omitting the L low outliers and/or zeros. The process repeats, still omitting the outliers and/or zeros, and ensuring as before that a valid set of parameters has been found. To find the floods of various return periods, conditional probability is used and equation B.101 becomes

$$x = -a[nT/(n+L)]^{-b} + c[nT/(n+L)]^{d} + e \qquad \text{B.102}$$

B.6 The Weibull Distribution

B.6.1 *The Density Function*

The probability density function of the Weibull distribution can be expressed as

$$f(x) = \left[\frac{a}{u-e}\right]\left[\frac{x-e}{u-e}\right]^{a-1} \exp\left[-\left[\frac{x-e}{u-e}\right]^{a}\right] \qquad \text{B.103}$$

The distribution function is

$$F(x) = 1-\left[\exp-\left[\frac{x-e}{u-e}\right]^{a}\right] \qquad \text{B.104}$$

where F(x) is the probability of a value of the variate being less than x. The parameters of the distribution are e, u, and a. The distribution is bounded below at e and unbounded above.

B.6.2 Moment Estimates - No Historic Information

Moments of the distribution can be expressed as functions of the parameters, and if the moments are replaced by their best estimates from the sample, a re-arrangement gives

$$g_1 = \frac{\Gamma(1+3/a) - 3\Gamma(1+2/a)\Gamma(1+1/a) + 2\Gamma^3(1+1/a)}{[\Gamma(1+2/a) - \Gamma^2(1+1/a)]^{3/2}} \qquad \text{B.105}$$

$$e = \bar{x} - \frac{s[\Gamma(1+1/a)]}{[\Gamma(1+2/a) - \Gamma^2(1+1/a)]^{1/2}} \qquad \text{B.106}$$

$$u = \bar{x} + \frac{s[1 - \Gamma(1+1/a)]}{[\Gamma(1+2/a) - \Gamma^2(1+1/a)]^{1/2}} \qquad \text{B.107}$$

where $\bar{x}$, s, and g_1 are respectively the best estimates of the population mean, standard deviation, and coefficient of skew as per equations B.1, B.2, and B.3. Γ() is the gamma function of the argument within parenthesis.

When g_1 has been estimated, equation B.104 can be solved by some numerical method to give 1/a. As a simpler alternative, the following three polynomials will give 1/a to a high degree of accuracy:

$$1/a = 0.27730 + 0.3252g_1 + 0.07632g_1^2 + 0.00388g_1^3 \qquad \text{B.108(a)}$$
$$(-1.08 < g_1 < +0.158308)$$

$$1/a = 0.27586 + 0.33071g_1 + 0.05004g_1^2 - 0.01737g_1^3 \qquad \text{B.108(b)}$$
$$(+0.158308 < g_1 < +1.910765)$$

$$1/a = 1.27421 \ln (0.58452g_1 + 1.02291) \qquad \text{B.108(c)}$$
$$(+1.910765 < g_1 < +5)$$

Using this value of 1/a and best estimates of $\bar{x}$ and s, and a subroutine for the gamma function in equations B.94 and B.95 gives estimates of e and u, thus defining the parameters of the distribution.

B.6.3 T-Year Flood Estimation - Weibull Distribution

B.6.3.1 Flood Estimation with No Low Outliers Present

If no low outliers are to be treated, the T-year floods are computed from

$$Q_T = e + (u-e)\ (\ln T)^{1/a} \qquad B.109$$

where T is the return period of the event and e, u, and a are the parameters of the distribution. The program computes and displays flood estimates for return periods of 1.003, 1.05, 1.25, 2, 5, 10, 20, 50, 100, 200, and 500 years.

B.6.3.2 The Conditional Probability Function

Consider a sample of N observations, where L observations have been identified as low outliers. From equations B.40, B.41, and B.42, the magnitude of the T-year event for the Weibull distribution when low outliers are present is:

$$Q_T = e + (u-e)\ \{\ \ln[T\ (N-L)/N]\ \}^{1/a} \qquad B.110$$

B.6.3.3 Flood Estimation with Low Outliers Present

In the event that low outliers — zero or non-zero — have been identified in the sample, the L outliers are removed from the sample of size N and the distribution parameters are estimated from the N-L remaining observations, either by maximum likelihood or by moments as previously outlined. This procedure is followed in the case of a historic analysis as well as a conventional analysis. Floods for any desired return period or probability of exceedance can then be estimated using these parameters, known values of N and L, and equation B.110.

However, this procedure will give a probability function which is not truly that of the distribution. Furthermore, it is not possible to estimate floods of return periods less than N/(N-L) because, in the conditional probability function, their exceedance probabilities then become greater than one.

To avoid these difficulties, a "synthetic" Weibull distribution is fitted through that portion of the "conditional" frequency curve between T=2 and T=100. Since three parameters have to be estimated, three points on the curve are necessary. The mathematics are considerably simplified by choosing the floods of return periods 2, 5.969, and 100 years. Parameters follow from

$$e = \frac{Q_2 Q_{100} - Q_{5.969}^2}{Q_{100} + Q_2 - 2Q_{5.969}} \qquad B.111$$

$$a = \frac{-0.94682}{\ln\frac{(Q_2 - e)}{Q_{5.969} - e}} \quad \text{B.112}$$

$$u = \frac{Q_2 - e}{.69315^{1/a}} + e \quad \text{B.113}$$

The "retro-fitted" parameters of the Weibull distribution can then be used in equation B.109 to produce revised flood estimates for the required return periods.

Note that the procedure of retro-fitting the parameters to produce a synthetic Weilbull distribution is used only when all of the low outliers are non-zero. The procedure is not used when zeros are present in the sample because it can result in an unrealistic situation whereby the exceedance probability of a non-zero flow is larger than the exceedance probability of zero flow as indicated by the sample.

APPENDIX C: The Nonparametric Density Estimation Method

The nonparametric method does not require either the assumption of any functional form of density function, nor estimation of parameters based on the mean, variance and skew. The nonparametric kernel density estimation requires (Adamowski, 1985; Adamowski, 1989) the selection of a kernel function, k(•), which is itself a probability density function, a positive smoothing factor H, and a sample of N observations $x_1, x_2, \ldots x_N$. The kernel estimate of density function f(x) at each fixed point x is

$$f(x) = \frac{1}{NH} \sum_{i=1}^{N} K \left[\frac{x - x_i}{H} \right] \qquad \text{C.1}$$

The principle of kernel estimator as expressed by equation C.1 is that with each observation a probability density function (kernel) of prescribed form (i.e. rectangular, normal, Gumbel etc.) is associated over a specified range (expressed by H) on either side of the observation. The set of such functions constitutes the nonparametric estimate of the density function. The normal form of the kernel is used herein. The adaptation of the nonparametric method is due primarily to the efforts of Prof. K. Adamowski, University of Ottawa.

C.1 An Optimal Kernel

It has been established (Adamowski, 1988) that the optimal (in the mean integrated square error sense) and the most efficient kernel is the Epanechnikov kernel; however, this kernel is bounded and as such, is not particularly desirable in flood frequency analysis when often an extrapolation of density function is required. Therefore, with a very small loss of accuracy, a Gaussian kernel is recommended in flood frequency analysis, that is

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)x^2} \qquad \text{C.2}$$

C.2 Optimal Value of H

The optimal value of H is determined based on the cross-validation procedure which requires numerical solution by successive approximations of the following equation (Rudemo, 1982).

$$\hat{Q}(\hat{f}) = A + B \sum_{i<j} \left[\exp(-\Delta_{ij}^2/4) - C \exp(-\Delta_{ij}^2/2) \right] \qquad \text{C.3}$$

where $\hat{f}$ is an estimate of f(x), $A = (2NH\sqrt{2}\pi)^{-1}$, $B = (N^2H\sqrt{\pi})^{-1}$, $C = 2\sqrt{2}\ N/N\text{-}1)$, and $\Delta_{ij} = (x_i - x_j)/H$. When H is evaluated to be less than 0.01, it is re-evaluated by the expression:

$$H = N^{\cdot 2}(4/3)^{\cdot 2}s \qquad \text{C.4}$$

where N is the sample size and s is the standard deviation of the N observations as obtained from equation B.2

Using an integrated mean square error criterion, it was found (Rudemo, 1982) that the asymptotically optimal value of H for a Gaussian kernal is given by

$$H_{asy} = (24\sqrt{\pi})^{1/3}\, N^{-1/3}\, s \qquad \text{C.5}$$

C.3 Estimation with Historic Information

Assume that there are n_o systematic observations, denoted by $\{x_1, x_2, x_{n_o}\}$ and v historical flows, denoted by $\{x_{n_o+1}, x_{n_o+2}, \ldots x_{n_o+v}\}$ that happen to exceed the censoring threshold Y during the $m = (N-n_o)$ year historical record. It is therefore assumed that there are two sets of data whose probability values are estimated by (Adamowski and Feluch, 1990; Adamowski and Pilon, 1989)

$$\{x_i \le Y\}_{i=1}^{n_o} \text{ with probability } p(x \le Y) = \frac{n-m}{n} \qquad \text{C.6}$$

and

$$\{x_i > Y\}_{i=1}^{m=N-n_o} \text{ with probability } p(x > Y) = \frac{m}{n} \qquad \text{C.7}$$

The unknown density function is estimated by a two component mixture model given by:

$$f(x) = f(x \le y)\, p(x \le y) + f(x > y)\, p(x > y)$$

where the density f(•) is estimated nonparametrically based on equation C.1

Assuming normal kernel (equation C.2), the value of smoothing parameter H is derived based on the cross-validation procedure which requires the solution of the following equation (Adamowski and Feluch, 1990).

$$-n_o + \sum_{\substack{i,j=1 \\ i<j}}^{n_o} d_{ij}\,[(1-2cd_{ij})(\Delta_{ij}^2-1)-1] = 0 \qquad \text{C.9}$$

where $$d_{ij} = \exp\left[\Delta_{ij}^2 / 4\right], \Delta_{ij} = (x_i - x_j)/4, \text{ and } C = 2\sqrt{2}\, N/(N-1) \qquad \text{C.10}$$

Table 1: Grubbs and Beck Outlier Test - 10 percent Significance Level K_N Values

The body of the table below contains one-tail values of the 10 percent significance values of K_N for samples size N from a normal population (Grubbs and Beck, 1972).

Sample size	K_N value	Sample size	K_N value	Sample size	K_N value	Sample size	K_N value
10	2.036	45	2.727	80	2.940	115	3.064
11	2.088	46	2.736	81	2.945	116	3.067
12	2.134	47	2.744	82	2.949	117	3.070
13	2.175	48	2.753	83	2.953	118	3.073
14	2.213	49	2.760	84	2.957	119	3.075
15	2.247	50	2.768	85	2.961	120	3.078
16	2.279	51	2.775	86	2.966	121	3.081
17	2.309	52	2.783	87	2.970	122	3.083
18	2.335	53	2.790	88	2.973	123	3.086
19	2.361	54	2.798	89	2.977	124	3.080
20	2.385	55	2.804	90	2.981	125	3.092
21	2.408	56	2.811	91	2.984	126	3.095
22	2.429	57	2.818	92	2.989	127	3.097
23	2.448	58	2.824	93	2.993	128	3.100
24	2.467	59	2.831	94	2.996	129	3.102
25	2.486	60	2.837	95	3.000	130	3.104
26	2.502	61	2.842	96	3.003	131	3.107
27	2.519	62	2.849	97	3.006	132	3.109
28	2.534	63	2.854	98	3.011	133	3.112
29	2.549	64	2.860	99	3.014	134	3.114
30	2.563	65	2.866	100	3.017	135	3.116
31	2.577	66	2.871	101	3.021	136	3.119
32	2.591	67	2.877	102	3.024	137	3.122
33	2.604	68	2.883	103	3.027	138	3.124
34	2.616	69	2.888	104	3.030	139	3.126
35	2.628	70	2.893	105	3.033	140	3.129
36	2.639	71	2.897	106	3.037	141	3.131
37	2.650	72	2.903	107	3.040	142	3.133
38	2.661	73	2.908	108	3.043	143	3.135
39	2.671	74	2.912	109	3.046	144	3.138
40	2.682	75	2.917	110	3.049	145	3.140
41	2.692	76	2.922	111	3.052	146	3.142
42	2.700	77	2.927	112	3.055	147	3.144
43	2.710	78	2.931	113	3.058	148	3.146
44	2.719	79	2.935	114	3.061	149	3.148

Then, any sample members greater than X_H are considered high outliers and those less than X_L are considered to be low outliers.

If historic information has been found and the skewness of the transformed variate is greater than .4, then the low outlier test is modified to:

$$X_L = \exp\,(\tilde{x} - K_N \tilde{s}) \qquad (1b)$$

where $\tilde{x}$ and $\tilde{s}$ are the historically weighted mean and standard deviation of the natural logarithms of the sample, respectively. K_N is evaluated where N is set to YT, the total historic time span. YT is further described in section 3.5. The methodology for obtaining historically weighted moments is given in Appendix B.2.

The sequence of testing depends on the skewness of the logarithms of the sample g_y. The sequence can be followed using the flow diagram of Figure 4. When a frequency analysis is being performed on a conventional sample, the outlier test determines if low and/or high outliers exit. All outliers for the conventional sample are indicated as being so in the standard tabular output of the flood data. A warning is issued to the user if a high outlier is detected. If historic information is present, the program will only check for low outliers. If a high outlier is detected and no historic information can be associated with the event, then a conventional analysis should proceed with the high outlier included in the sample. The user has the opportunity to change the number of low outliers detected by the program before proceeding with the parametric frequency analysis. Note that outlier analysis is not performed when proceeding with a nonparametric frequency analysis.