Record Number:
Author, Monographic: Ashkar, F.//Bobée, B.
Author Role:
Title, Monographic: Statistical analysis of annual flood series and partial duration series
Translated Title:
Reprint Status:
Edition:
Author, Subsidiary:
Author Role:
Place of Publication: Québec
Publisher Name: INRS-Eau
Date of Publication: 1985
Original Publication Date: Février 1985
Volume Identification:
Extent of Work: 123
Packaging Method: pages
Series Editor:
Series Editor Role:
Series Title: INRS-Eau, Rapport de recherche
Series Volume ID: 177
Location/URL:
ISBN: 2-89146-175-4
Notes: Rapport annuel 1985-1986
Abstract: 20.00$
Call Number: R000177
Keywords: rapport/ ok/ dl

STATISTICAL ANALYSIS OF ANNUAL FLOOD SERIES AND PARTIAL DURATION SERIES

by

Fahim Ashkar

and

Bernard Bobée

Rapport scientifique # 177

INRS-Eau

February, 1985

CONTENTS

1. FUNDAMENTAL CONCEPTS

The analysis of flood probability distributions plays a major role in hydrologic and economic evaluations of water resources projects and in establishing project design criteria. In general, one wishes to determine a statistical distribution or a probability model suitable for representing a sample of run-off flows. With such a distribution or such a model it then becomes possible to estimate events corresponding to a given probability. These estimations are basic for the construction works and help to permit an efficient design.

1.1. Annual Series and Partial Duration Series

Starting with a recorded hydrograph or with the tabulated data abstracted from this hydrograph, two types of flood peak series may be used in a frequency analysis. These are the annual flood series (a.f.s.) which consists of the largest flood in each year, and the partial flood series (p.f.s.) which consists of all "well-defined" flood peaks above a specified magnitude, often called the flood truncation level or base level. In fact one of the drawbacks of partial flood series (also called partial duration series; p.d.s.) is that it is not completely well-defined which of the flood peaks exceeding the base level should be retained for the analysis and which should be excluded. Since in p.d.s. models now in common use, it is required that sucessive

flood peaks should be independent, some flood investigators (Cunnane, 1979; Water Resources Council, 1976, and others) have proposed putting restrictions on the inter-arrival times of flood events (flood peaks) so that these events will not occur close together in bunches. Water Resources Council (1976) arbitrarily defined separate flood events as events separated by at least as many days as five plus the natural logarithm of square miles of drainage area, with the requirement that the intermediate flows must drop below 75 percent of the lower of the two separate maximum daily flows. Todorovic and Zelenhasic (1970) and Todorovic and Rousselle (1971) defined the partial flood series as all flood peaks above the base level, and in the case of a multiple-peaked flood hydrograph only the largest discharge is considered as the flood peak to be retained. This is done with the expectation that when the base level is sufficiently high the independence of flood peaks (also called exceedances) would become physically plausible.

Although annual flood series are more precisely defined than partial flood series they have the disadvantage that they use only one flood per year. In certain cases the second largest flood in a year may outrank many annual floods of other years and yet it is totally neglected in the a.f.s. approach. This disadvantage is remedied in the p.d.s. approach in which the truncation level is generally selected low enough that the average number of exceedances per year is of the order of two or three.

1.2. The Return Period as a Measure of Risk

Langbein (1949) and Chow (1950) investigated the theoretical relationship between the probability of annual flood series and the expectancy of partial flood series. Let P_p be the expectancy of a variate in the partial flood series being equal to or greater than x, and let m be the average number of events per year, or mN be the total number of events in N years of record. Then P_p/m is the annual probability of an event being equal to or greater than x, and $1 - P_p/m$ is the probability of an event being less than x. The probability of an event x being the largest of the m events in a year would then be $(1 - P_p/m)^m$. This probability can be approximated by $\exp(- P_p)$ when P_p is small compared with m. Note also that this is the probability of an annual event of magnitude x and corresponds to the annual series. Hence, the probability P_a of an annual flood series of magnitude x being equaled or exceeded is:

$$P_a = 1 - \exp(- P_p)$$

or

$$P_p = - \ln (1 - P_a) \qquad (1.1)$$

The time elapsing between successive events of magnitude equalling or exceeding a specified value x is a random variable whose mean value is defined as the return period T of x (notation: $T = T(x)$ or $T = T_x$). Alternatively, each flood value x may be considered as a function of its associated value of return period (notation: $x = x(T)$ or $x = x_T$). The return periods of annual series and partial duration series have different meanings. In the first case the return period T_a is the mean recurrence time of an event of a given magnitude as an annual maximum while in the second case the return period T_p carries no implication of annual maximum.

Letting $T_a = 1/P_a$ and $T_p = 1/P_p$ in equation (1.1) yields:

$$T_p = \frac{1}{\ln T_a - \ln (T_a - 1)} \qquad (1.2)$$

from which it can be seen that T_p, the return period in partial flood series is smaller than T_a in annual series but that T_p approaches T_a as both T_p and T_a increase. Langbein (1949) states that the difference between T_p and T_a becomes negligible for

floods greater than about a five-year return period.

Equation (1.2) is an approximate relationship because it is based on approximating $(1 - P_p/m)^m$ by $\exp(- P_p)$. An alternative derivation of relationship (1.2) was presented recently by Takeuchi (1984) who reconfirmed its validity and encouraged its practical use.

The definition we have just given for the return period T_p for partial duration series can be modified to carry more implication of annual maximum. Let $\chi_1, \chi_2, \ldots, \chi_n$ be the sequence of annual maximum values abstracted from a partial flood series. This means that if the year i contains no exceedances or no flood peaks above the base level x_0 then the value of χ_i for that year will be equal to zero, when the discharge x_0 is taken as reference (base level). In other words, if the year j contains m_j exceedances ξ_1, ξ_2, ξ_{mj} then:

$$\chi_j = \begin{cases} \max(\xi_1, \xi_2, \ldots, \xi_{mj}) \text{ for } m_j > 0 \\ 0 \text{ for } m_j = 0 \end{cases} \qquad (1.3)$$

where ξ_i is obtained from the i^{th} peak flood discharge Q_i exceed-

ing x_0 simply by subtracting x_0: $\xi_i = Q_i - x_0$. If the distribution of the ξ_i's and of the m_j's is known, then the distribution of the χ_j's can be deduced and the probability of the variable χ_j equalling or exceeding a specified magnitude x can be calculated. If this probability is now denoted by P_p (the subscript p standing for "p.d.s.") then $T_p = 1/P_p$ would be the new definition of the return period for partial duration series. This definition, which is more convenient for mathematical treatment than the previous one, is the one in common use [see for instance Todorovic and Zelenhasic 1970, Todorovic and Rousselle 1971, Ashkar and Rousselle 1981]. We shall make no further reference to the previous definition of T_p in the rest of our discussion. The new definition of T_p is nearer to the definition of T_a given for the annual series but still differs from it because the term "annual maximum" does not carry the same meaning in a.f.s. as in p.f.s. Only years that produce a peak discharge superior to the base discharge yield the same value of annual maximum for a.f.s. as for p.f.s. In general, therefore, we expect that the higher the average number of exceedances per year, the nearer T_a and T_p are to each other.

Lloyd (1970) considered the exceedance probability $P[X > x_T]$ associated with an arbitrary random variable X and a return period T such that:

$$P\,[X > x_T] = p = \frac{1}{T} \tag{1.4}$$

and showed that T, as a random variable, has a distribution of the form:

$$P\,[T = t] = p\,(1 - p)^{t-1} \tag{1.5}$$

This distribution has for mean and variance:

mean : $E(T) = 1/p$

variance: $\mathrm{Var}\,(T) = (1 - p)/p^2$

If a period of r years is considered, the probability P that the event $X > x_T$ will occur at least once during these r years is $P = 1 - Q$, where Q is the probability of not having a flow $X > x_T$ during r years. This gives

$$P = 1 - (1 - p)^r = 1 - (1 - \frac{1}{T})^r \tag{1.6}$$

which can be written as

$$T = \frac{1}{1 - (1 - P)^{1/r}} \qquad (1.7)$$

This expression can also be written in an approximative form:

$$T = r \left(\frac{1}{P} - \frac{1}{2} \right)$$

The following simple applications of relationships (1.6) and (1.7) are intended to help better understand the concepts of return period and of reciprocal probability.

Applications

(1) Considering a return period of T = 100 years, the probability P that during these 100 years to come, the centenary flow (p = 0.01) will be exceeded at least once, is

$$P = 1 - (1 - 0.01)^{100}$$

$$P = .63$$

There is therefore a 63 % chance that the centenary flow will occur during the 100 years to come.

(2) One wishes to construct a public work having a duration life of r = 50 years and one wishes to determine the return period T such that the flow $X > x_T$ will occur with a probability less than or equal to 20 %. Thus

$$T > 50 \left(\frac{1}{.20} - 0.5 \right)$$

$$T > 225 \text{ years}$$

The work must thus be constructed so that the return period T = 225 years if one wishes that in the 50 years to come, the flow $X > x_T$ will occur with a probability less than or equal to 20 %.

In this last example, the probability 20 % that one or more events will exceed a given flood magnitude (the flood corresponding to a return period of 225 years) within a specified number of years (50 years) is sometimes referred to as the "risk" associated with the given flood magnitude and with the specified number of years. For a one - year period, the probability of exceedance p, which is the reciprocal of the recurrence interval T, expresses this risk. Table 1.1 gives for different return periods T the percent chance (risk) of getting one or more floods of return period T, or greater, within one of a number of different lengths of time.

1.3. A Reliability Criterion for Flood Flow Estimates

The reliability of flood flow estimates obtained from the recorded data by extrapolation is directly related to the length of record. The following criterion is proposed by Hardison (1969):

Number of years of data collection	2 5 10 15 20 25 years
Maximum recurrence interval	2 5 10 15 50 100 years

Hardison (1969) has shown that this criterion gives sufficiently close estimates for x_T.

1.4. Mixed Populations

In areas where high flows are generated by more than one distinct hydrologic process (e.g. snowmelt - and rainfall - generated peaks), peak discharge data should be considered to be drawn from subpopulations with different statistical characteristics. Stoddart and Watt (1970) for example have described how flooding in some watersheds in southern Ontario is created by two different types of events. In these watersheds rain floods occur generally in the summer and floods due to snowmelt, sometimes combined with precipitation, occur in winter and spring. Waylen and Woo (1982) describe also how floods in the Cascade Mountains of southern British Columbia can be due to heavy winter rainfall, or snowmelt in spring.

When it can be shown that floods on record come from two or more distinct populations then it may be more hydrologically reasonable to try to find to what subpopulation each flood belongs and then to analyse each subpopulation separately, rather than separating floods by calendar periods. This, unless of course, the events in the separate periods are clearly caused by different hydrometeorologic conditions.

Consider two independent flood generating processes, a and b say, and suppose that a flood of a given magnitude x would have a return period T_a if it belongs to population a and T_b if it

belongs to population b. The probabilities of not exceeding x are

$$q_a = 1 - 1/T_a$$

and

$$q_b = 1 - 1/T_b$$

The probability of not exceeding x in any year becomes

$$q = q_a \, q_b$$

and the return period for the annual flood associated with the level x becomes

$$T = 1 / (1 - q)$$

or

$$T = T_a T_b / (T_a + T_b - 1) \qquad (1.8)$$

1.5 Plotting Position

In frequency analysis of hydrological data a statistical model may be postulated whose parameters are estimated from the observed data (cf. section 2.1.4), or alternatively, an empirical distribution of the observed magnitudes may be obtained by graphical analysis on a probability plot. In the latter method the ranked data are plotted on probability paper using probability as abscissa values obtained from a plotting position formula. This plotting may help in getting a better interpretation of the data, in detecting any possible errors, or in picking an adequate probability distribution for fitting the data. The scale of the abscissa is frequently arranged in such a way that events distributed according to a given probability distribution will plot as a straight line.

Numerous works have been produced on the subject of plotting position both because of the practical importance of the choice of this empirical probability and because there is no formula which is entirely satisfactory in finding it. Gumbel (1958) states four

postulates which the plotting position P_k of the event of order k of an ordered sample ($x_1 > \ldots > x_k > \ldots > x_N$) must satisfy:

(1) the plotting position should be such that all observations can be plotted;

(2) the plotting position should be between the observed frequencies (k - 1) / N and k/N and should be distribution free;

(3) the return period of a value equal to or larger than the largest observation should approach N, the number of observations;

(4) the observations should be equally spaced on the frequency scale, i.e. the difference between the observations of order (k + 1) and k should be a function of N only and be independent of k.

Among the principal formulas currently found in practical use, the following may be cited:

The Hazen (or Foster) formula:

$$P_k = \frac{k - 0.5}{N}$$

This formula is recommended by Brunet-Moret (1973) for the case where the parameters of the adjusted distribution are estimated from the sample.

The Weibull formula:

$$P_k = \frac{k}{N + 1}$$

This formula is recommended by Chow (1953) for the study of flows. It is the average of the probabilities of all events with rank k in a series of periods, each of N years.

The Chegodayev formula:

$$P_k = \frac{k - 0.3}{N + 0.4}$$

This formula is recommended by Kimball (1960) as well as by Brunet-Moret (1973) for the case where the parameters of the distribution are known a priori. This formula gives the approximate probability of the median of the distribution of the statistic of order k from a sample of size N.

A critical review of Gumbel's postulates has been given by Cunnane (1978) who argued against postulate (3), notably that the return period of a value equal to or greater than the largest observation should converge towards N, the number of observations. His argument was based on statistical properties of the largest value in a sample of size N. He proposed a plotting position that has the property that quantile estimates made from the plot will be unbiased and will have smallest mean square error among all such estimates. This unbiased plotting position is namely $E(y_{(i)})$ the mean of the ith order statistic in samples from the reduced variate population. $E(y_{(i)})$ has the disadvantage, however, that it depends on the form of the distribution being considered, and if the reduced variate depends on a shape parameter then $E(y_{(i)})$ too depends on this parameter.

2. SINGLE SITE ANALYSIS

2.1. DATA BASE

2.1.1. Introduction

In flood frequency analysis, the primary objectives are to determine the return periods of recorded events of known magnitudes and then to estimate the magnitude of events for return periods beyond the recorded range, that can be used in the design of hydraulic structures and the planning and management of water resources systems. In this kind of analysis it is important to try to abstract the maximum information from the available data. Inadequate estimations may come from:

- the use of inadequate data;

- the wrong choice of a representative statistical distribution;

- the inadequate use of a technique for estimating the parameters of the chosen law.

For these reasons we shall be dealing in the remainder of section 2. with:

- conditions that are required of relevant data before the application of a statistical distribution;

- the characteristics of the distributions habitually used to represent run-off flows;

- properties and particularities of the principal methods of estimating parameters.

2.1.2. Conditions Required from the Data

Before one may adjust a statistical law to a given sample it must be demonstrated that the elements of this sample verify three conditions:

A. temporal independence;
B. homogeneity;
C. stationarity.

A. Condition of temporal independence

To estimate the probabilities of hydrologic events, one assumes generally that the observed flows are independently distributed in time. Streamflow sequences, however, tend to be persistent in that high flows tend to follow high flows and low flows tend to follow low flows. This persistence depends on the

lapse of time separating the successive elements of the sequence; dependence among successive daily flow values, for instance, tends to be strong, while dependence among yearly values, is weak.

The independence of the sample elements of flood flows can be checked using either the Wald and Wolfowitz (1943) test or the Anderson (1941) test.

Wald and Wolfowitz test

For a sample of size N (x_1, ..., x_N) we consider the statistic R such that

$$R = \sum_{i=1}^{N-1} x_i \, x_{i+1} + x_1 \, x_N$$

In the case where the elements of the sample are independent, R follows a normal distribution with mean and variance given by

$$\bar{R} = (s_1^2 - s_2) / (N - 1)$$

$$\text{Var}(R) = (s_2^2 - s_4) / (N - 1) - \bar{R}^2 + (s_1^4 - 4s_1^2 s_2$$

$$+ 4 s_1 s_3 + s_2^2 - 2 s_4) / (N - 1)(N - 2)$$

$$\text{with } s_r = N m_r'$$

where m_r' is the r^{th} moment of the sample about the origin (cf. section 2.1.4 A).

The quantity $u = (R - \bar{R}) / (\text{var } R)^{1/2}$ follows a standardized normal distribution (mean 0 and variance 1) and can be used to test the hypothesis of independence.

Anderson test

Let r_1 be the first-order serial correlation of the sample given by

$$r_1 = \left[\sum_{i=1}^{N-1} x_i x_{i+1} + x_1 x_N - \left(\sum_{i=1}^{N} x_i \right)^2 \quad N \right] \Big/ \left[\sum_{i=1}^{N} x_i^2 \right.$$

$$- \left(\sum_{i=1}^{N} x_i \right)^2 / N \Bigg]$$

For a normal random time series of N values, r_1 is nearly normally distributed with

a mean:

$$\bar{r}_1 = - 1 / (N - 1)$$

a variance:

$$\text{Var } r_1 = (N - 2) / (N - 1)^2$$

Considering the quantity $u = (r_1 - \bar{r}_1) / (\text{var } r_1)^{1/2}$ it is possible to test whether r_1 at a given level is significantly different from zero. Although this test is only theoretically valid for samples taken from a normal distribution, it is generally used for other parent populations also.

B. Condition of representativeness of the sample

The condition of representativeness implies that all the elements of the sample originate from the same population. In annual flood series as well as in partial duration series it may happen that the sample is composed of events of different origin belonging to different populations (e.g., snowmelt and rainfall floods). To check whether two samples belong to the same population we can use the Terry (1952) test or the Mann and Whitney (1947) test.

Terry test

Given two samples of size p and q, respectively, the combined set of N = p + q observations is ranked in increasing order. If in the complete series I denotes the ranks of the elements of the first sample and J those of the observations of the second sample, we consider the statistic C given by

$$C = \sum_{J} E\,(X_{J,N}) = - \sum_{I} E\,(X_{I,N})$$

where $E\,(X_{k,N})$ denotes the mathematical expectation of the kth-order statistic in a sample of size N from a standardized normal

population.

For N > 15, under the null hypothesis that the two samples belong to the same population, the quantity

$$t = C\,[(N - 2) / [(N - 1)\ \text{var}\ C - C^2]]^{1/2}$$

with

$$\text{Var}\ C = (pq / N\ (N - 1)) \cdot \sum_{k=1}^{N} E^2\ (X_{k,N})$$

follows approximately a Student distribution with (N - 2) degrees of freedom. In practice, the values of $E\ (X_{k,N})$ may be obtained from the Harter (1961) tables.

Mann-Whitney test

As was done above, we regroup two samples of size p and q (with $p < q$) in a combined set of size $N = p + q$, ranked in increasing order. We consider the quantities

$$V = T - p\,(p + 1) / 2$$

$$W = pq - V$$

where T is the sum of the ranks of the elements of the first sample (of size p) in the combined series and V is the number of times that an item in sample 1 follows in the ranking an item in sample 2; W is computed in a similar way for sample 2 following sample 1.

When $N > 20$, $p, q > 3$, and under the null hypothesis that the two samples come from the same population, V and W are approximately normally distributed with mean $pq / 2$ and variance $pq\,(p + q + 1) / 12$. In practice we consider the quantity

$$u = \frac{V - pq / 2}{[pq\,(p + q + 1) / 12]^{1/2}}$$

and for a test at a level of significance α, u is compared with the standardized normal variate corresponding to a probability of exceedance $\frac{\alpha}{2}$.

In practice, to study the run-off flows at a given station, one can consider the sample formed from the maximum annual flow (or the sample formed from the exceedances, in the p.d.s. approach) during N years and examine the independence of the elements of the sample. If however, the flows are due to two different causes, for example a flow due to snowmelt which occurs in the spring and the autumn flows due to over-abundant precipitation, it is possible to test for heterogeneity between the two samples. In the case where there is heterogeneity, it is reasonable to consider the two types of flow separately.

For an application of tests of independence and of homogeneity refer to the example given in section 2.2.8.

C. Condition of stationarity

The assumption that the natural processes influencing river flow characteristics are stationary with respect to time is difficult to guarantee. Non-stationary behaviour may occur in a number of different forms. There are the slow changes in hydrologic parameters and the rapid changes. An example of the slow changes are evolutionary changes such as gradual movements in climate, involving for instance, increasing or decreasing rainfall. Urbanization and variations in catchment characteristics are another form of slow change. Rapid changes may result, for instance, from earthquakes or from building of dams.

In current hydrologic investigations, the problem of existence of long term variations, conceived as fluctuations of the basic characteristics of hydrologic time series, in function of time, is one of the most controversial problems. The question at stake, is whether or not, trends, periodicity, or other non-stationarity in the probability structure of hydrologic time series beyond the periodicity of the year, do really exist. Existing techniques of time series analysis cannot answer this question. Based on some studies which do not support the concept of non-stationarity in hydrologic series of annual values (see Yevjevich, 1963 for instance) we shall make the conventional assumption that no non-stationarity exists in hydrologic series, beyond the periodicity of the year.

Flood records used in frequency analysis should represent relatively constant watershed conditions. The records should be carefully examined to make sure that no major changes within the watershed have occurred during the period of record since such changes effect record homogeneity. Tests discussed in the previous paragraph can be used to check for any significant nonhomogeneity when it is suspected that such a nonhomogeneity in flood values might be present.

2.1.3. Random Variables and their Statistical Characteristics

From a fairly short record of streamflow, how does one estimate a design flood ? As we mentioned earlier, the general approach is to use the sample data to fit a frequency distribution which in turn is used to extrapolate from the recorded events to the design events. The first step consists of choosing a frequency distribution. Subsequently, the parameters of this distribution are estimated and used to extrapolate beyond the domain of recorded events.

Given a sample of independent hydrologic observations $x_1, x_2, \ldots, x_N$ let X be the random variable (r.v.) representing the population from which this set of observations is drawn. If the population consists of an infinite set of elements distributed over an interval D of finite or infinite length then we have what is called a "continuous population" which may be represented by:

- its continuous probability density function (p.d.f.) $f(x; \Theta_1, \ldots, \Theta_k)$ where $\Theta_1, \ldots, \Theta_k$ are parameters;

- its continuous cumulative distribution function (usually called simply "distribution function") defined as:

$$F(x) = P[X < x] = \int_{-\infty}^{x} f(x)\, dx$$

which means that

$$f(x) = \frac{dF(x)}{dx}$$

While most of the probability distributions used in flood frequency analysis are of the continuous type, some distributions, such as the Poisson distribution used in partial duration series models to represent the number of flood exceedances in an arbitrary but fixed interval of time (section 2.3.1) are not continuous. The Poisson distribution which can assume the discrete values 0, 1, 2, ... is a member of the class of "discrete distributions" which can assume values over a finite or infinite set S of discrete or separate values. If i is an element of the set S then the discrete random variable X defined over S may be represented by:

- its discrete probability density function (also called mass function) $f(i; \Theta_1, \ldots, \Theta_k) = P[X = i]$ where $\Theta_1, \ldots, \Theta_k$ are parameters;

- its (cumulative) distribution function defined as

$$F(i) = P[X < i] = \sum_{j \in S; j<i} f(j)$$

2.1.4 Methods of Estimation

Several methods for the estimation of parameters of a (continuous or discrete) p.d.f. are available and are more or less adequate depending upon the distribution chosen. The two principal methods of estimation used in practice are:

- the method of moments;
- the method of maximum likelihood.

A. Method of moments

For a given distribution, depending on k parameters it is possible to calculate the non-central and central moments about the mean. This gives the non-central moment of order r, μ_r' such that:

$$\mu_r' = \int_D x^r f(x)\, dx$$

and μ_r the central moment of order r around the mean μ_1' such that:

$$\mu_r = \int_D (x - \mu_1')^r f(x)\, dx$$

The variate X defined over the interval D has probability density function f (x). In the case of a discrete random variable the integration over D is replaced by a summation over the set S over which the discrete r.v. is defined.

Since f (x) depends on the parameters $\Theta_1, \ldots, \Theta_k$, the moments μ_r and μ_r' are functions of the parameters. These moments can be estimated numerically by means of the corresponding sample moments. For a sample $x_1, \ldots, x_N$ the non-central sample moment of order r, m_r' is given by:

$$m_r' = \frac{1}{N} \sum_{i=1}^{N} x_i^r$$

and the central sample moment m_r of order r around the mean $m_1' = \bar{x}$

is given by:

$$m_r = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^r$$

where N is the size of the sample.

Thus, for a law of k parameters, the parameters are estimated by setting the k moments of the population equal to the k moments corresponding to the sample. This gives k equations permitting the estimation of $(\Theta_1, \ldots, \Theta_k)$.

In practice, the higher the order of the moment, the more likely it is that one is subject to important sampling errors. It is for this reason that in the method of moments one uses the moments (or functions of the moments) of the lowest possible order. The moments used should be functionally independent, however. For example, the moments μ_1', μ_2' and μ_2 cannot be used together because $\mu_2 = \mu_2' - \mu_1'^2$.

For a law of two parameters, the mean (non-central moment of order 1) and the variance (central moment of order 2) can be used. In the case of a law of three parameters, the skewness coeffi-

cient γ (which is a function of the central moments of order 2 and 3; $\gamma = \mu_3 / \mu_2^{3/2}$) can be considered as well.

The mean of a sample $(x_1 \ldots x_i \ldots x_N)$ of size N is given by $\bar{x} = m_1'$, and the non-biased variance is given by:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \frac{N\, m_2}{N-1}$$

(one considers the non-biased value s^2 such that $E(s^2) = \mu_2 = \sigma^2$; σ^2 being the variance of the population).

The skewness coefficient is given by:

$$C_s = \frac{1}{N} \sum (x_i - \bar{x})^3 \Big/ \left[\frac{1}{N} \sum (x_i - \bar{x})^2\right]^{3/2} = \frac{m_3}{m_2^{3/2}}$$

In fact, it can be shown (Kirby, 1974) that C_s is biased, or in other words that $E(C_s) \neq \gamma$; γ is the skewness of the population.

Different corrections can be used:

- $(C_s)_1 = \alpha_1 C_s$

where

$$\alpha_1 = \frac{\sqrt{N\ (N - 1)}}{N - 2}$$

This classical correction, which is obtained by using the non-biased values of the moments of order 2 and 3, in fact leads to a biased skewness;

- $(C_s)_2 = \alpha_2 C_s$

where

$$\alpha_2 = \left(1 + \frac{8.5}{N}\right) \frac{\sqrt{N\ (N - 1)}}{N - 2}$$

This usual correction is empirical and leads to a non-biased estimation of the skewness for a small interval only of skewness values (Wallis et al., 1974);

- $(C_s)_3 = \alpha_3 C_s$

where α_3 depends on the distribution used.

It can be shown (Bobée and Robitaille, 1975) that:

- for the Pearson type 3 law, one has:

$$\alpha_3 = \left(1 + \frac{6.51}{N} + \frac{20.20}{N^2}\right) + \left(\frac{1.48}{N} + \frac{6.77}{N^2}\right) C_s^2$$

- for the log-normal law of 3 parameters, one has:

$$\alpha_3 = \left(1.01 + \frac{7.01}{N} + \frac{14.66}{N^2}\right) + \left(\frac{1.69}{N} + \frac{74.66}{N^2}\right) C_s^3$$

According to the value of the skewness coefficient of the sample chosen, different estimations of the parameters of the distribution are obtained.

B. Method of maximum likelihood

The method of maximum likelihood is based on the principle that, for a density function f (x) dependent on the parameters $(\Theta_1, \ldots, \Theta_k)$, the probability of obtaining a given sample $(x_1, \ldots, x_N)$ is proportional to the likelihood function L such that:

$$L=f(x_1; \Theta_1, \ldots \Theta_k)\ldots f(x_i; \Theta_1, \ldots \Theta_k)\ldots f(x_N; \Theta_1, \ldots \Theta_k)$$

The method consists in determining the values of the parameters which maximize L, hence which maximize the probability of observing the sample $(x_1, \ldots x_N)$.

In practice, one often maximizes Ln L, which is equivalent to maximizing L since

$$\frac{\partial \mathrm{Ln}L}{\partial \Theta_i} = \frac{1}{L}\frac{\partial L}{\partial \Theta_i}$$

One thus obtains as many equations as one has parameters to determine:

$$\frac{\partial L}{\partial \Theta_i} = 0 \qquad i = 1, \ldots k$$

One must moreover verify that the matrix of general term

$$a_{ij} = \frac{\partial^2 \mathrm{Ln}L}{\partial \Theta_i \partial \Theta_j}$$

is definite negative to assure that a maximum is obtained.

2.1.5 Sampling Variances and Confidence Intervals

All that is available to estimate the parameters ($\Theta_1, \ldots \Theta_k$) of a distribution representing a population is a sample of size N. The estimation $\hat{\Theta}_1, \ldots \hat{\Theta}_k$ are thus distorted due to sampling errors and they therefore have a certain variance (the estimation of the parameters $\hat{\Theta}_1, \ldots \hat{\Theta}_k$ are realizations of a random variable).

An event X_T corresponding to a return period T, thus to an exceedance probability $p = \frac{1}{T}$, is determined by the general relation:

$$X_T = \mu + \chi\sigma \tag{2.1}$$

where:

μ and σ^2 are the mean and the variance of the population respectively;

χ is a frequency factor which depends on the return period T and the moments of the distribution.

In practice, μ, σ, and χ are not exactly known but are estimated by the quantities $\hat{\mu}$, $\hat{\sigma}$ and K, (notation: $K = K(T)$ or $K = K_T$) which have a certain sampling variance. It results that X_T is estimated by the general relation:

$$\hat{X}_T = \hat{\mu} + K_T\hat{\sigma} \tag{2.2}$$

with a sampling variance $\sigma^2_{X_T}$ = var $(\hat{X}_T)$ and a mean X_T = E $(\hat{X}_T)$. It may be shown in the first approximation that $\hat{X}_T$ is distributed asymptotically according to a normal distribution; thus the quantity

$$u = \frac{X_T - \hat{X}_T}{\sigma_{X_T}}$$

follows a standardized normal distribution (of mean 0 and variance 1). It is then possible to determine the confidence intervals of X_T at a given significance level α. This gives:

$$\hat{X}_T - u_{\alpha/2}\ \sigma_{X_T} < X_T < \hat{X}_T + u_{\alpha/2}\ \sigma_{X_T} \qquad (2.3)$$

$u_{\alpha/2}$ is the standardized normal variable of exceedance probability $\alpha/2$ (also called the "$\alpha/2$ - quantile" of u), Var $(\hat{X}_T)$ is frequently written in the form

$$\text{Var}\ (\hat{X}_T) = \frac{\hat{\mu}_2}{N}\ \delta_T \qquad (2.4)$$

where $\hat{\mu}_2 = s^2 = \hat{\sigma}^2$ and δ_T is a function of the return period T and of the estimated parameters.

A. Method of moments

In the method of moments, the moments of the population are estimated by the moments corresponding to the sample. Thus, for a law of 3 parameters, this gives:

$\hat{\mu} = m_1' = \bar{x}$ mean

$\hat{\sigma}^2 = s^2 = m_2$ variance

$\hat{\gamma} = C_s$ skewness coefficient

Since C_s is a function of m_2 and m_3 (central moments of order 2 and 3), $\hat{X}_T$ is a function of m_1', m_2 and m_3: $\hat{X}_T = f(m_1', m_2, m_3)$.

Thus, for a law of 3 parameters:

$$\text{Var } \hat{X}_T = \sigma_{X_T}^2 = \left(\frac{\partial \hat{X}_T}{\partial m_1}\right)^2 \text{var } m_1' + \left(\frac{\partial \hat{X}_T}{\partial m_2}\right)^2 \text{var } m_2 + \left(\frac{\partial \hat{X}_T}{\partial m_3}\right)^2$$

$$\text{var } m_3 + 2 \left(\frac{\partial \hat{X}_T}{\partial m_1}\right) \left(\frac{\partial \hat{X}_T}{\partial m_2}\right) \text{cov } (m_1', m_2)$$

$$+ 2 \left(\frac{\partial \hat{X}_T}{\partial m_1'}\right) \left(\frac{\partial \hat{X}_T}{m_3}\right) \text{Cov } (m_1', m_3) + 2 \left(\frac{\partial \hat{X}_T}{\partial m_2}\right) \left(\frac{\partial \hat{X}_T}{\partial m_3}\right)$$

$$\text{cov } (m_2, m_3) \qquad (2.5)$$

The partial derivatives are deduced from the general relation $X_T = f(m_1', m_2, m_3)$ and the variances and covariances of the moments can be expressed as a function of the moments, thus of the parameters, of the distribution considered.

For a distribution of two parameters, one has $\hat{X}_T = f(m_1', m_2)$ and the expression of var $\hat{X}_T$ does not bring terms relative to m_3 into consideration.

B. Method of maximum likelihood

When considering the method of maximum likelihood, one obtains the estimations $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ for a law of three parameters. The estimates of the moments of the population are function of these parameter estimates. In other words:

$$\hat{\mu} = g\,(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$$

$$\hat{\sigma} = h\,(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$$

$$\hat{\gamma} = k\,(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$$

$\hat{X}_T$ is thus a function of $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$. This gives:

$$\hat{X}_T = \phi\,(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$$

One can deduce:

$$\text{Var } \hat{X}_T = \left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_1}\right)^2 \text{Var } \hat{\Theta}_1 + \left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_2}\right)^2 \text{Var } \hat{\Theta}_2 + \left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_3}\right)^2 \text{Var } \hat{\Theta}_3$$

$$+ 2 \left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_1}\right)\left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_2}\right) \text{Cov } (\hat{\Theta}_1, \hat{\Theta}_2) + 2 \left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_1}\right)\left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_3}\right)$$

$$\text{Cov } (\hat{\Theta}_1, \hat{\Theta}_3) + 2 \left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_2}\right)\left(\frac{\partial \hat{X}_T}{\partial \hat{\Theta}_3}\right) \text{Cov } (\hat{\Theta}_2, \hat{\Theta}_3) \qquad (2.6)$$

The partial derivatives are calculated from the general relation $\hat{X}_T = \phi\ (\hat{\Theta}_1, \hat{\Theta}_2, \hat{\Theta}_3)$.

Given V_{ij}, the general term of the variance-covariance matrix of the parameters, one has:

$$V_{ij} = \text{Cov } (\hat{\Theta}_i, \hat{\Theta}_j) \qquad \text{if } i \neq j$$

$$V_{ii} = \text{Var } \hat{\Theta}_i$$

This matrix is the inverse of the symmetric matrix:

$$a_{ij} = - E \left[\frac{\partial^2 Ln L}{\partial \Theta_i \partial \Theta_j} \right]$$

It is thus possible from the likelihood function L (Θ_1, Θ_2, Θ_3) to determine the variances and covariances of the parameters estimated by the method of maximum likelihood.

2.2 COMMON PROBABILITY DISTRIBUTIONS AND FITTING TECHNIQUES

2.2.1 Introduction

As hydrological processes are bounded by physical limitations, the statistical distributions which are used to represent them must conform, for a flow can neither take a negative value, nor can it exceed an upper bound while keeping its physical meaning in the hydrogeographical context of the watershed. For flood flows, this upper bound has been well studied by Francou and Rodier (1969). With respect to the lower bound of the flood flow, we think like Csoma (1969) that its value should not be inevitably zero, but is dependent on the hydrogeographical system of the watershed. Furthermore, hydrologists generally agree that the statistical distribution of annual floods is positively skewed,

although it has never been proved that a negative skewness is impossible. Klemes (1970) showed moreover that the distribution of mean annual flows could be negatively skewed.

Kite (1976) has effectuated a profound study of several statistical laws. Only the principal distributions used in Canada for the study of extreme values in hydrology (floods in particular) will be considered here. In addition, the principal characteristics of fitting methods related to these distributions will be indicated.

2.2.2. The Normal Distribution

This distribution has a symmetrical p.d.f. given by

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[- \frac{(x - \mu)^2}{2 \sigma^2} \right] \qquad (2.7)$$

The methods of moments and of maximum likelihood yield the same estimates for the parameters μ and σ:

$$\hat{\mu} = m_1'$$

$$\hat{\sigma}^2 = m_2 \tag{2.8}$$

and the value of δ_T involved in the calculation of Var $(\hat{X}_T)$ (equation 2.4) is given by:

$$\delta_T = 1 + U_T^2 / 2 \tag{2.9}$$

where U_T stands for the standard normal variate corresponding to an exceedance probability of $p = 1 / T$.

2.2.3. The Lognormal Distribution

The lognormal distribution is deduced from the normal distribution by a logarithmic transformation. More precisely, when X follows a lognormal distribution with three parameters a, μ_y and σ_y, its p.d.f. is given by

$$f(x) = \frac{1}{(x - a)\ \sigma_y \sqrt{2\pi}} \exp\left\{ - \frac{1}{2}\left[\frac{\text{Ln}\ (x - a) - \mu_y}{\sigma_y}\right]^2 \right\} \tag{2.10}$$

and the random variable $Y = \text{Ln}\ (X - a)$ follows a normal distribution with parameters μ_y and σ_y. When $a = 0$ we obtain the two-parameter lognormal distribution.

A. Method of moments

Kite (1978) gives the estimates of a, μ_y and σ_y by the method of moments in terms of the sample mean m_1, standard deviation $s = \sqrt{m_2}$ and coefficient of skewness C_S:

$$\hat{\sigma}_y = [\text{Ln}\ (Z^2 + 1)]^{1/2}$$

$$\hat{\mu}_y = \text{Ln} \left(\frac{s}{Z}\right) - \frac{1}{2} \text{Ln}\ (Z^2 + 1)$$

$$\hat{a} = m_1' - s/Z \qquad (2.11)$$

where Z is the coefficient of variation of the sample $(x_1 - a)$, ... $(x_n - a)$ which is solution of the equation

$$C_S = 3\ Z + Z^3$$

and is such that

$$Z = \frac{1 - w^{2/3}}{w^{1/3}} \tag{2.12}$$

with

$$w = \frac{- C_S + (C_S^2 + 4)^{1/2}}{2}$$

For a two-parameter lognormal distribution we have a = 0 and Z becomes equal to the coefficient of variation of the observed sample ($Z = s / m_1'$) in which case the first two equations of (2.11) readily yield the solutions for $\hat{\sigma}_y$ and $\hat{\mu}_y$.

The estimate of the design event X_T corresponding to a return period T can be put in the form of equation (2.2):

$$\hat{X}_T = m_1' + K_T \sqrt{m_2}$$

with

$$K_T = \frac{\exp\left[\mathrm{Ln}\ (1 + Z^2)\right]^{1/2} U_T - \frac{1}{2}\,\mathrm{Ln}\left[1 + Z^2\right] - 1}{Z} \qquad (2.13)$$

Z being given by equation (2.12) and U_T being the standard normal variate corresponding to an exceedance probability of p = 1 / T. The calculation of var $(\hat{X}_T)$ can be done as in (Kite, 1978) but for the three-parameter lognormal distribution the expression obtained is not explicit. In the special case of the two-parameter lognormal distribution, using the same notation as in equation (2.4) one obtains (Kite, 1978):

$$\delta_T = \left[1 + (Z^3 + 3Z)\, K_T + (Z^8 + 6Z^6 + 15Z^4 + 16Z^2 + 2)\, K_T^2/4\right] \qquad (2.14)$$

from which var $(\hat{X}_T)$ and confidence limits around X_T may be deduced assuming normality of $\hat{X}_T$ (substitute 2.14 into 2.4, and then 2.4 and 2.13 into 2.3).

B. Method of maximum likelihood

For a random sample $x_1, \ldots, x_N$ of size N from the three-parameter lognormal distribution, the method of maximum likelihood leads to the following system of equations:

$$\mu_y = \frac{1}{N} \sum_i \text{Ln } (x_i - a)$$

$$\sigma_y^2 = \frac{1}{N} \sum_i [\text{Ln } (x_i - a) - \mu_y]^2$$

$$\sum_i \frac{\mu_y - \sigma_y^2}{(x_i - a)} = \sum_i \frac{\text{Ln } (x_i - a)}{(x_i - a)} \qquad (2.15)$$

which may be solved starting with the parameter "a" which may be found numerically by substituting the values of μ_y and σ_y^2 from the first two equations into the last equation, and subsequently determining μ_y and σ_y^2 using the first two equations.

For the two-parameter lognormal distribution we have a = 0 and the first two equations of (2.15) suffice for the

determination of $\hat{\mu}_y$ and $\hat{\sigma}_y^2$. From the form of these two equations it can be seen that applying the method of maximum likelihood to the two-parameter lognormal distribution is equivalent to fitting the normal distribution by maximum likelihood to the logarithms of the data.

It is important to point out that since the properties of the method of maximum likelihood are only asymptotically optimal, this method may not be optimal with small sample sizes found in hydrology.

The calculation of var $(\hat{X}_T)$ is indicated in (Kite, 1978) but for the three-parameter case no explicit expression is obtained. For the two-parameter distribution, we have (notation of equation 2.4):

$$\delta_T = \frac{\{[Ln\ (Z^2 + 1)]\ (1 + K_T Z)^2\ (1 + U_T^2/2)\}}{Z^2} \qquad (2.16)$$

where K_T is given in equation (2.13), $Z = \frac{\sqrt{m_2}}{m_1}$ is the coefficient of variation and U_T is the standard normal variate corresponding to an exceedance probability of 1/T.

2.2.4 The Gumbel Distribution (Type 1 Extremal)

A. Characterization of the distribution

This distribution is based on the theory of extreme values. When one considers N samples of size p and if one takes the largest value from each sample (or the smallest value), one can form a new sample containing the N extreme values.

If each sample of size p is formed of independent values and comes from the same statistical population, it can be shown (Gumbel, 1958) that when p becomes large the sample consisting of N extreme values can be represented by one of three distributions of extreme values. In the Type 1 distribution of extreme values (Gumbel distribution) the population from which the samples originated is of the exponential type.

In the study of maximum annual flood flows, each sample is of the size p = 365 and the maximum annual value is selected to form the sample of N flood flows.

If one poses $y = \alpha(x - \beta)$, the cumulative distribution (non-exceedance probability) is:

$$F(x) = e^{-e^{-y}} \tag{2.17}$$

and the density function is:

$$f(x) = \alpha\, e^{[-y - e^{-y}]} \tag{2.18}$$

One can thus deduce:

the mean:

$$\mu = \beta + \frac{C}{\alpha}$$ (C is the Euler constant and is equal to 0.577)

the variance:

$$\sigma^2 = \mu_2 = \frac{\pi^2}{6\,\alpha^2}$$

the skewness coefficient:

$$\gamma = 1.139$$

B. Method of moments

The estimation of the parameters α and β by the method of moments leads to:

$$\hat{\alpha} = \frac{\pi}{\sqrt{6}} \cdot \frac{1}{s} = \frac{1.2825}{s}$$

$$\hat{\beta} = \bar{x} - 0.4500\ s \qquad (2.19)$$

$\bar{x}$ and s^2 are the mean and the variance of the sample, respectively, which are estimations of the mean μ and the variance σ^2 of the population.

The event X_T of return period T is estimated by:

$$\hat{X}_T = \tilde{\mu} + K \hat{\sigma} \tag{2.20}$$

One has $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = s$.

It can be shown (Kite, 1976) that the frequency function is estimated by:

$$K(T) = - \left[0.45 + .7797 \, Ln \left[- Ln \left(1 - \frac{1}{T} \right) \right] \right] \tag{2.21}$$

As a function of the estimated parameters, the event of return period T is obtained from equations (2.19), (2.20) and (2.21) as:

$$\hat{X}_T = \hat{\beta} - \frac{1}{\hat{\alpha}} Ln \left[- Ln \left(1 - \frac{1}{T} \right) \right] \tag{2.22}$$

The estimation variance of $\hat{X}_T$ is:

$$\text{Var}\ (\hat{X}_T) = \sigma_{X_T}^2 = \frac{\sigma^2}{N}\ \delta_T \quad \text{(equation 2.4)}$$

with

$$\delta_T = (1 + 1.1396\ K_T + 1.1000\ K_T^2) \qquad (2.23)$$

C. Method of maximum likelihood

The method of maximum likelihood leads to the following system of equations (Kite, 1978):

$$\beta = \frac{1}{\alpha} \text{Ln}\ [N\ /\ \sum_i e^{-\alpha x_i}]$$

$$\sum_i x_i\ e^{-\alpha x_i} - (m_1' - 1/\alpha) \sum_i e^{-\alpha x_i} = 0 \qquad (2.24)$$

The second of these equations, in which m_1' is the sample mean may be solved for α by iteration (see Kite, 1976) and subsequently

β may be deduced from the first equation.

As for var $(\hat{X}_T)$, the term δ_T (equation 2.4) is given by:

$$\delta_T = 0.6740 + 0.3125\, y_T + 0.3696\, y_T^2$$

with

$$y_T = - \text{Ln} [- \text{Ln} (1 - 1/T)] \qquad (2.25)$$

D. <u>The interest of the Gumbel distribution in hydrology</u>

The Gumbel distribution has known an increasing popularity because of its apparent theoretical justification (section 2.2.4 A). In fact, however, the hypotheses leading to the law of extreme values are not respected:

- the maximum value is selected in a sample of size p = 365; this value is not very high;

- the daily flows which constitute the sample of 365 values are not independent;

- the probability law of the daily flows is not constant and may vary according to the season.

The Gumbel distribution cannot, therefore, be preferred over other laws for theoretical reasons. Moreover, this law has certain disadvantages:

- the interval of the variate x is not bound (one can have $-\infty < x < +\infty$);

- the skewness coefficient is constant ($\gamma = 1.139$) and there is little reason to think that all the distributions of flood flows have the same skewness.

2.2.5. The Pearson Type 3 Distribution

A. Characterization of the distribution

The density function of the Pearson type 3 distribution is:

$$f(x) = \frac{|\alpha|}{\Gamma(\lambda)} e^{-\alpha (x-m)} [\alpha (x - m)]^{\lambda-1} \qquad (2.26)$$

where $\Gamma(\cdot)$ is the gamma function.

The interval of definition is always such that $\alpha\ (x - m) > 0$.

Therefore:

if $\alpha > 0$, $m \leq x < +\infty$ (form with positive skewness)

if $\alpha < 0$, $-\infty < x \leq m$ (form with negative skewness)

The moments may be expressed as a function of the parameters:

mean:

$$\mu = \mu_1' = m + \frac{\lambda}{\alpha}$$

variance:

$$\sigma^2 = \mu_2 = \frac{\lambda}{\alpha^2}$$

skewness coefficient:

$$\gamma = \frac{2}{\sqrt{\lambda}} \cdot \frac{\alpha}{|\alpha|}$$

In the case where the parameter of origin m is nul, one obtains the Gamma law. This distribution equally includes, as a limiting case when γ tends towards 0 (λ tends towards ∞) the normal law.

B. Method of moments

The method of moments leads to the following parameter estimates:

$$\hat{\lambda} = 4 \frac{m_2^3}{m_3^2} = \frac{4}{C_s^2}$$

$$\tilde{\alpha} = 2 \frac{m_2}{m_3}$$

$$\hat{m} = m'_1 - 2 \frac{m_2^2}{m_3} \qquad (2.27)$$

m'_1 and m_2 are respectively the mean and the (non-biased) variance of the sample. m_3 is the third central moment (around the mean).

C_s is the value of the skewness coefficient of the sample. One can take the values $(C_s)_1$ or $(C_s)_2$ or $(C_s)_3$ defined in section 2.1.4 A) for C_s.

In the case of the Gamma law, by placing $m = 0$ ($\hat{m} = 0$) in the last equation of (2.27) we obtain:

$$\hat{\alpha} = m'_1 / m_2$$

$$\hat{\lambda} = m'^2_1 / m_2 \qquad (2.28)$$

For the Pearson type 3 distribution, the event of return period T is estimated by:

$$\hat{X}_T = \bar{x} + Ks = \left(\hat{m} + \frac{\hat{\lambda}}{\hat{\alpha}}\right) + K \frac{\sqrt{\hat{\lambda}}}{|\hat{\alpha}|} \qquad (2.29)$$

The frequency factor K depends on T and the skewness C_s; the Harter tables (1969) permit the determination of K.

Using the same notation as in equation (2.4), the term δ_T required in the calculation of the sampling variance of X_T is given by (Bobée, 1973):

$$\delta_T = 1 + \frac{K^2}{2}\left(1 + \frac{3}{4}C_s^2\right) + K\,C_s + 6\left(1 + \frac{C_s^2}{4}\right)\left(\frac{\partial K}{\partial C_s}\right)$$

$$\left[\left(\frac{\partial K}{\partial C_s}\right)\left(1 + \frac{5\,C_s^2}{4}\right) + \frac{K}{2}C_s\right] \qquad (2.30)$$

The partial derivative $\left(\frac{\partial K}{\partial C_s}\right)$ can be deduced from the Harter tables (1969) for a given value of T. Kite (1976) equally gives an approximation of K and of $\left(\frac{\partial K}{\partial C_s}\right)$ as a function of the normal standardized variable. In the case when $C_s = 0$ the Pearson type 3 distribution reduces to a normal distribution and equation (2.30) becomes equivalent to equation (2.9).

The method of moments, with the correction of skewness $C_s = (C_s)_3$ defined in section (2.1.4 A) leads to a better adjustment (Bobée and Robitaille, 1977).

In the case of the Gamma distribution (m = 0) it can be shown that $C_s = 2\ Z$ (Z being the coefficient of variation) and δ_T reduces to (Bobée, 1973):

$$\delta_T = (1 + K_T Z)^2 + \tfrac{1}{2}\left[K_T + 2Z\left(\frac{\partial K}{\partial C_s}\right)\right]^2 (1 + Z^2) \qquad (2.31)$$

C. Method of maximum likelihood

The method of maximum likelihood is not preferable to the method of moments in the case of the Pearson type 3 distribution

applied to small samples (Bobée and Robitaille, 1977). The properties of maximum likelihood are only in effect asymptotically optimal, and with the Pearson type 3 distribution the use of this method in practice may involve certain problems which are discussed by Matalas and Wallis (1973). This work may be consulted for details concerning the applicability of the method of maximum likelihood to the Pearson type 3 distribution.

In the case of the Gamma distribution the method of maximum likelihood leads to the following equations:

$$\frac{\lambda}{\alpha} = m_1'$$

$$\text{Ln } \lambda - \frac{d \text{ Ln } \Gamma(\lambda)}{d\lambda} = A$$

with

$$A = - \frac{1}{N} \sum_i \text{Ln} \left(\frac{x_i}{m_1'} \right)$$

for which an approximate solution was obtained by Thom (1958) as

$$\hat{\lambda} = \frac{1 + \sqrt{1 + 4A/3}}{4A}$$

and

$$\hat{\alpha} = \hat{\lambda} / m_1' \qquad (2.32)$$

From the parameter estimates of α and λ, estimates for the population mean μ_1' and variance μ_2 may be deduced and used to calculate $\hat{X}_T$ with the aid of equation (2.2), K_T being given by the Harter tables (1969) for a coefficient of skewness $C_s = 2 / \sqrt{\hat{\lambda}}$ and the desired return period T.

For the calculation of Var $(\hat{X}_T)$, using the notation of equation (2.4) we have (Bobée and Boucher, 1981 a):

$$\delta_T = \frac{1}{\lambda\eta}\left[(\lambda\Psi'-1)\left(1 + \frac{\varepsilon K_T}{\sqrt{\lambda}}\right)^2 + \frac{K_T^2}{4\lambda} + \frac{1}{\lambda^2}\left(\frac{\partial K_T}{\partial C_s}\right)^2 + \frac{\varepsilon K_T}{\lambda\sqrt{\lambda}}\left(\frac{\partial K_T}{\partial C_s}\right)\right]$$

with

$$\Psi' = \frac{d^2 \text{ Log } \Gamma(\lambda)}{d\,\lambda^2} \text{ (trigamma function) [tabulated]}$$

$$\eta = \Psi' - \frac{1}{\lambda}$$

$$\varepsilon = \alpha \,/\, \left|\alpha\right| \qquad (2.33)$$

2.2.6 The Log-Pearson Type 3 Distribution

A. Characterization of the distribution

The log-Pearson type 3 distribution is deduced from the Pearson type 3 distribution by a logarithmic transformation:

If Y = Ln X follows a Pearson type 3 distribution, X follows a log-Pearson type 3 distribution. It can be shown (Bobée, 1975) that the density function of the variate X is:

$$f(x) = \frac{|\alpha|}{\Gamma(\lambda)} \frac{e^{\alpha m}}{x^{1+\alpha}} [\alpha (\text{Ln } x - m)]^{\lambda-1} \qquad (2.34)$$

$\alpha > 0 \qquad e^m \leq x$

$\alpha < 0 \qquad 0 \leq x \leq e^m$

The non-central moment of order r is given by the expression:

$$\mu'_r = \frac{e^{mr}}{\left(1 - \frac{r}{\alpha}\right)^{\lambda}}$$

The log-Pearson type 3 distribution includes the log-normal distribution in the limiting case where λ tends towards infinity.

B. Method of moments

The hydrology committee of the Water Resources Council in the United States recommends the use of the method of moments to logarithmically transformed samples ($y = \text{Ln } x$) of observed values (Benson, 1968). This method may be described as follows: from the observed sample $x_1, \ldots, x_N$, the transformed sample $y_1, \ldots, y_n$ is obtained such that $y_i = \text{Ln } x_i$ (one may also consider a base -10 logarithmic transformation for instance). The mean $(m'_1)_y$, unbiased variance $(s^2)_y$ and corrected coefficient of skewness $(C_s)_y$ are then calculated and the event $\hat{y}_T$ is deduced from

$$\hat{y}_T = (m'_1)_y + K_T s_y \tag{2.35}$$

where K_T is the Pearson type 3 frequency factor corresponding to a return period T and to a skewness coefficient $(C_s)_y$ (equation 2.29). Since we have $\hat{y}_T = \text{Ln } \hat{x}_T$ we can deduce:

$$\hat{x}_T = e^{\hat{y}_T} \tag{2.36}$$

from which we obtain asymptotically:

$$\text{Var}\ (\hat{y}_T) = \left(\frac{d\,\hat{y}_T}{d\,\hat{x}_T}\right)^2 \text{Var}\ (\hat{x}_T)$$

$$= \frac{1}{\hat{x}_T^2} \text{Var}\ (\hat{x}_T) \qquad (2.37)$$

Note that when the base -10 logarithmic transformation is used, we have asymptotically:

$$\hat{x}_T = 10^{\hat{y}_T} \text{ and Var } (\hat{y}_T) = \frac{1}{(\hat{x}_T)^2} \text{ var } (\hat{x}_T) \cdot A^2$$

with

$$A = \text{Log}_{10}\, e = \frac{1}{\text{Ln } 10} = 0.434$$

The method of moments based on the logarithmic transformation of the observed sample, comes to give the same weight to the logarithms of the observed values. Each observed value, however, no longer has the same weight. Moreover, it is the moments of the sample of logarithms which are preserved and not the moments of the sample of observed values. Consequently this method tends to reduce the relative importance of the larger elements of the sample.

C. The method of moments applied to the sample of observed values

This method (Bobée, 1975) conserves the moments of the sample of observed values and gives the same weight to each observation.

We consider the three first non-central moments of the sample, m'_1, m'_2 and m'_3 which are estimations of the moments μ'_1, μ'_2 and μ'_3 of the population. Equating these first three sample moments to the corresponding population moments, the following system of equations is obtained:

$$\frac{\text{Ln } [(1 - 1/\alpha)^3 / (1 - 3/\alpha)]}{\text{Ln } [(1 - 1/\alpha)^2 / (1 - 2/\alpha)]} = \frac{\text{Ln } m'_3 - 3 \text{ Ln } m'_1}{\text{Ln } m'_2 - 2 \text{ Ln } m'_1}$$

$$\lambda = \frac{\text{Ln } m'_2 - 2 \text{ Ln } m'_1}{\text{Ln } [(1 - 1/\alpha)^2 / (1 - 2/\alpha)]}$$

$$m = \text{Ln } m'_1 + \lambda \text{ Ln } (1 - 1/\alpha) \qquad (2.38)$$

The solution of the first of these equations for α may be obtained with the help of available tables (Bobée, 1975) or by approximations given by Bobée and Boucher (1981 a). Kite (1978) also gives approximation formulas for determining α from this equation.

Knowing α we may calculate λ and then m using the last two equations of (2.38). We may hence estimate the moments $(\mu'_1)_P$ and $(\mu_2)_P$ along with the coefficient of skewness γ_P of the corresponding Pearson type 3 distribution. $y_T = \text{Ln } x_T$ is then calculated using relationship (2.2), and finally $x_T = e^{y_T}$ is deduced.

The calculation of Var $(\hat{X}_T)$ for this method of estimation is indicated in (Bobée and Boucher, 1981 b) but the form of Var $(\hat{X}_T)$ (and therefore of δ_T) is not explicit.

From α, λ and m (or from the non-central moments), the mean, variance and coefficient of skewness of the population may be determined.

Hoshi and Burges (1981) describe a method for estimating x_T and Var $(\hat{X}_T)$ for the log-Pearson type 3 distribution based on estimates of the mean, coefficient of variation and skew coefficient obtained from the observed (untransformed) sample. This method leads in practice to the same results as the method of Bobée (1975) that we have just described.

D. Method of maximum likelihood

The equations of maximum likelihood obtained for the variate X, which follows a log-Pearson type 3 law, correspond to the equations of maximum likelihood for the Pearson type 3 distribution with Y = Ln X.

In practice, it suffices to logarithmically transform the originally observed sample $(x_1 \ldots x_N)$, to obtain the sample $(y_1 \ldots y_N)$ with $y_i = \text{Ln } x_i$.

The application of the method of maximum likelihood to the sample y_i, which is supposed to be drawn from a Pearson type 3 law, leads to the solution $(\hat{\alpha}, \hat{\lambda}, \hat{m})$.

One can thus deduce the event $\hat{y}_T$ of return period T and then compute $\hat{x}_T = e^{\hat{y}_T}$. It is equally possible to determine the asymptotic variance Var $(\hat{y}_T)$ and to deduce var $(\hat{x}_T)$ which is equal to $(\hat{x}_T)^2$ var $(\hat{y}_T)$ (to a first order asymptotic approximation).

However, as in the case of the method of maximum likelihood applied to the Pearson type 3 distribution, properties of the method are asymptotically optimal and theoretically the method is only viable for large samples (Bobée, 1979). The restrictions and the particularities of the method of maximum likelihood, for the log-Pearson type 3 distribution, are the same as those described in section (2.2.5 C) for the Pearson type 3 distribution.

2.2.7. Goodness-of-fit Tests and Comparison of Frequency Distributions

Tests of goodness-of-fit, or test of adjustment, are a means of verifying whether a probability density function f (x) represents the observations (or the sample). Among the most commonly used tests of goodness-of-fit are the chi-square test and the Kolmogorov-Smirnov test. Unfortunately, however, with the usual sample sizes of flood data none of the existing tests of adjustment are powerful enough to discriminate between different probability distributions.

A. Chi-square test

The statistic χ^2 is a measure of the deviation between the observed number of events (O_i) and the theoretical number of events (e_i). The deviation

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{O_i^2}{e_i} - N \qquad (2.39)$$

follows approximately a chi-square distribution (χ^2) with γ degrees of freedom, where $\gamma = k - p - 1$, in which:

k: number of class intervals;

p: number of parameters defining the probability density function $f^*(x)$ that are estimated beforehand from the sample, to make $f^*(x)$ a completely specified function.

In practice therefore, to apply the χ^2 test, the N independent observations of the sample are grouped into k classes and the number of observations O_i in each class, is determined. We therefore have $\sum_{i=1}^{k} O_i = N$.

The χ^2 test is then formulated as follows: let F* (x) be some completely specified distribution function and let X represent the random variable from which the observed sample is drawn. Consider the two hypotheses:

H_0: the distribution function of the observed random variable is F* (x);

H_1: the distribution function of the observed random variable is different from F* (x).

Denote by p_j^* the probability of a random observation from the distribution function F* (x) falling in class j. The theoretical number of events (e_j) in class j is therefore given by

$$e_j = p_j^* N \qquad j = 1, 2, \ldots, k \qquad (2.40)$$

Having defined O_J and e_j, the value of the test statistic χ^2 given in equation (2.39) can now be calculated. If this value is greater than $\chi_{1-\alpha}$ the $(1 - \alpha)$ quatile of a chi-square random variable with k - p - 1 degrees of freedom, reject H_0 at the α

level of significance. Otherwise accept H_0. Tables giving the quantiles of the Chi-square distribution may be found in any standard statistics textbook.

The k classes should cover the whole range of definition of the random variable X. If the k cells are equiprobable, i.e. if $p_j^* = 1/k$ so that $e_j = N/k$, $j = 1, \ldots, k$ then equation (2.39) reduces to:

$$\chi^2 = \frac{k}{N} \sum_{i=1}^{k} O_i^2 - N \qquad (2.41)$$

When some of the e_j's are small the chi-square distribution which in fact is theoretically valid only asymptotically, may not be appropriate as a distribution for the test statistic χ^2. Conover (1971) suggests that classes with small e_j be combined with other classes in some meaningful way so that no more than 20 % of the e_j are less than 5.0 and that none are less than 1.0. This rule may be relaxed somewhat if all the e_j are equal (equiprobable cells).

The chi-square test may be used with both continuous as well as discrete random variables. If a discrete random variable is defined over a set X composed of k separate values these values may be chosen to represent the k different classes.

We point out that one may always be quite sure in practice that the true distribution function representing the sample is never totally the same as the hypothesized distribution function. What we are interested in however is a good approximation to the true distribution function. It should be realized that in any test of adjustment the hypothesized distribution will be rejected if the sample size is large enough.

B. Kolmogorov-Smirnov Test

Denote by $F(x)$ the unknown distribution function with which the random sample $x_1, \ldots, x_N$ is associated. Let $F^*(x)$ be a completely specified hypothesized distribution function. We wish to test the following two hypotheses:

H_0: $F(x) = F^*(x)$ for all x from $-\infty$ to $+\infty$

H_1: $F(x) \neq F^*(x)$ for at least one value of x

Let $S(x)$ be the empirical distribution function based on the random sample $x_1, \ldots, x_N$. $S(x)$ is a step function defined as the fraction of the x_i's which are less than or equal to x for each x, $-\infty < x < +\infty$. Comparing $S(x)$ with $F^*(x)$ constitutes a logical way of deciding whether or not it is reasonable to accept $F^*(x)$ as an adequate distribution function to fit the observed

random sample. As a measure of the discripancy between S (x) and F*(x), Kolmogorov (1933) proposed the statistic

$$D_N = \sup_x \left| F^*(x) - S(x) \right| \tag{2.42}$$

which is the largest vertical distance between the two graphs S (x) and F*(x), this maximum being calculated over the whole range of the variable x and not only at the sample values. Reject H_0 at the level of significance α if the statistic D_N exceeds the $1 - \alpha$ quantile $w_{1-\alpha}$ as given by tables of the Kolmogorov-Smirnov test (Haan, 1977).

For the Kolmogorov-Smirnov test to be exact, the hypothesized distribution function F*(x) should be continuous. If this distribution function is discrete, the test is conservative, that is, the true but unknown level of significance is less than the stated one.

We point out that in practice the hypothesized distribution function F*(x) is seldom completely specified because normally its parameters have to be estimated from the sample. The kolmogorov-Smirnov test used in this situation loses some of its power. One way of remedying this problem is to calculate critical values that take account of this estimation of parameters. This bas been done

only for the normal distribution (Lilliefors, 1967) and for the exponential distribution (Lilliefors, 1969). Tables of this modified Kolmogorov-Smirnov test may be found in (Conover, 1971).

2.2.8. Example of Application

A. Description of the example

In this example, the data for the south Saskatchewan River at Saskatoon (station kF62), for which a sample of size N = 59 years is available, will be considered. This data is presented in table 2.1. It can be shown (Bobée and Robitaille, 1977) that:

- the independence condition is verified; in effect, u = 1.07 is found with the Wald-Wolfowitz test and u = 1.04 is found with the Anderson test. These values are within the acceptance zone at a level of 5 %;

- the homogeneity condition is verified; after having separated (1) floods within the period Jan. 1 - Apr. 30 (16 floods) from the remaining floods, and (2) floods within the period Sept. 1 - Dec. 31 (3 floods) from the remaining floods; the application of the Mann-Whitney and Terry tests leads to the acceptance of the homogeneity hypothesis.

It is thus possible to adjust statistical distributions to the sample considered. The different methods for adjusting the Pearson type 3 and log-Pearson type 3 laws, described in sections 2.2.5 and 2.2.6 will be considered.

B. Estimation of the parameters

The characteristics of the sample of observed values are:

mean $\bar{x} = 1484.84$

standard deviation $s = 774.15$

skewness $(C_s)_1 = 1.272$ (section 2.1.4 A)

The characteristics of the sample of logarithms of observed values are:

mean $(\bar{x})_L = 3.119$

standard deviation $(s)_L = .214$

skewness $[(C_{s_1})]_L = .099$

In the case of the Pearson type 3 distribution, one considers:

Method 1 moments with correction (C_{s_1})

Method 2 moments with correction (C_{s_2}) (section (2.1.4 A)

Method 3 moments with correction (C_{s_3})

Method 4 maximum likelihood

In the case of the log-Pearson type 3 distribution, one considers:

Method 5 moments of the series of observed values

Method 6 moments of the sample of logarithms with correction (C_{s_1})

Method 7 moments of the sample of logarithms with correction (C_{s_2})

Method 8 moments of the sample of logarithms with correction (C_{s_3})

Method 9 maximum likelihood

Table 2.2 indicates the values of the parameters α, λ and m in the case of each method. It equally indicates the skewness coefficient of the population of flows (methods 1 to 4) and of the population of logarithms of the flows (methods 5 to 9).

Table 2.2 shows that, for the example considered, there is little deviation between the different adjustment methods of the Pearson type 3 distribution. On the other hand, for the log-Pearson type 3 distribution, method 5 (adjustment with the observed values) leads to results different from other methods. In particular the sign of the skewness is changed. This is not suprinsing since, in methods 6 to 9, the Pearson type 3 law is applied by considering the sample of logarithms of the observed values.

The calculation program used (Bobée and Robitaille, 1976) permits equally to determine the flow Q_T, thus its sampling variance, and confidence intervals, for different return periods T (hence for different exceedence probabilities $P = \frac{1}{T}$).

In the way of an example, one traces, in considering the plotting position of Hazen $P_k = \frac{k - .5}{N}$ (section 1.5):

- the adjustment obtained by considering the Pearson type 3 law

with the skewness correction $(C_s)_3$ (figure 2.1) and with confidence intervals at a level of 80 % and of 95 %;

- the adjustment of the log-Pearson type 3 distribution (figure 2.2) obtained by considering method 5 (adjustment with the series of observed values) and method 6, which is the method suggested by the Water Resources Council (adjustment with the series of logarithms, with the skewness correction $(C_s)_1$).

It is not possible to choose the law and the method of adjustment the most adequate from this example. As we have already mentioned in section (2.2.7) the classical tests of adjustment are just not powerful enough to discriminate between these different laws. A more global comparison (Bobée and Robitaille, 1977), however, has shown that in general, the Pearson type 3 law with the skewness correction $(C_s)_3$ (method 3) and the log-Pearson type 3 law adjusted to the series of observed values (method 5) lead to the best results.

2.2.9. Conclusion

The primary objectives of frequency analysis are to determine the magnitude of events for design return periods. The sample data are used as an estimate of an unknown population to calculate estimates of the parameters of the selected probability distribution. The fitted distribution is then used to estimate event

magnitudes corresponding to return periods greater than or less than those of the recorded events.

There is actually no general agreement among hydrologists as to which of the various theoretical distributions available should be used. The present state of the art is also such that no general agreement has been reached as to preferable techniques. Moreover, statistical tests, such as chi-square or Kolmogorov-Smirnov, do not permit discrimination between the different techniques applied to different laws.

In North America, the log-Pearson type 3 distribution is being used increasingly since its systematic usage by American governmental agencies has been recommended. With respect to the adjustment technique suggested by the hydrology committee of the Water Resources Council, several criticisms can be addressed to it. It is for this reason that a simple technique of adjustment for the log-Pearson type 3 law, which preserves the moments of the sample of observed values and which accords the same weight to each observation, was presented. This technique, which seems more justified on a theoretical basis, leads equally to better practical results. One must hope that in the future global comparisons, by Monte Carlo simulation for example, will lead to definitive conclusions and will permit the recommendation of the systematic use of an adjustment technique across Canada.

The list of distributions that we have presented represents the set of probability laws that are most commonly used in North America. New laws are being introduced and investigated in the literature. We mention in particular the Wakeby distribution which was introduced by Houghton (1978 a). This distribution has aroused a good degree of interest for aesthetic as well as analytical reasons. In addition to being a five-parameter distribution one of its peculiarities is that it is expressable as an inverse distribution function:

$$X = - a (1 - F)^b + c (1 - F)^d + e$$

where F is a uniform random variable over the interval (0, 1) and a, b, c, d, e are parameters.

Houghton introduced the Wakeby distribution as the grand parent of distributions used in hydrology. It gives better fit for flood data than conventional distributions when its parameters are chosen correctly. It has five parameters, however which have to be estimated. This introduces substantial estimation error. In addition, its density function is expressable only in inverse form. This calls for the use of unconventional methods of estimation (such as the "incomplete means" procedure introduced by Houghton (1978 b)). Finally a method of calculation of var $(\hat{X}_T)$ is not yet

available. All these reasons render the Wakeby distribution less attractive for hydrologic applications.

2.3. PARTIAL DURATION SERIES MODELS

2.3.1. Mathematical Presentation

Let us consider the stochastic process described by the streamflow hydrograph and let us select a base level x_0.

If we consider only those flood peaks Q_i in the arbitrary interval of time [0, t] that exceed x_0, we can define

$$\xi_i = Q_i - xo \tag{2.43}$$

where $\xi_i > 0$ is a random variable for all i = 1, 2, ... Associated with each exceedance ξ_i is a random variable $\tau(i)$ which is the time when the corresponding peak occurred (figure 2.3). Only the largest peak is taken into consideration in the case of a multiple-peaked hydrograph (figure 2.3).

Denote by $n(t)$ the number of exceedances in the interval of time [0, t] and let E_ν^t stand for the event $[n(t) = \nu]$ i.e. the

event that there are exactly ν exceedances at or before time t. The event $[n(t_2) - n(t_1) = \nu;\ t_2 > t_1]$ i.e. the event that there are exactly ν exceedances between t_1, and t_2 is denoted by $E_\nu^{t_1,t_2}$.

Let $P(E_1^{t,t+s} | E_k^t)$ be the probability that there is exactly one exceedance in the interval of time (t, t + s] conditional on the event E_k^t that there are k exceedances up to and including time t. Under certain regularity assumptions [discussed in (Todorovic, 1970)] and in the case when

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(E_1^{t,t+\Delta t} \mid E_k^t)}{\Delta t} \tag{2.44}$$

is independent of k; i.e. in the case when $\lambda_k(t) \equiv \lambda(t)$, we have (Todorovic and Zelenhasic, 1970):

$$P(E_k^t) = P\,[\tau(k) \leq t] - P\,[\tau(k+1) \leq t]$$

$$= \exp\left[-\int_0^t \lambda(s)\, ds\right] \left[\int_0^t \lambda(s)ds\right]^k / k!,\ k = 0, 1, 2... \tag{2.45}$$

which is a time-dependent Poisson process. $\lambda(t)$ is called the "intensity function" of this process. This implies that $E[n(t)]$, the mean number of exceedances in $[0, t]$ (which we shall denote by $\Lambda(t)$) is given by:

$$E\,[n(t)] = \Lambda(t) = \int_0^t \lambda(s)\, ds \tag{2.46}$$

It is commonly known that the variance of $n(t)$ (Poissonian) is equal to its mean:

$$\text{Var}\,[n(t)] = \Lambda(t) \tag{2.47}$$

One of the advantages of truncating the hydrograph by the base level x_0 is that it has enabled us (equation 2.45) to attribute a certain stochastic structure to the process $\{n(t);\ t>0\}$. For additional theoretical explanation of the Poisson property of $n(t)$ the reader is referred to (Todorovic, 1978).

Let us consider now the sequence:

$$\chi_1, \chi_2, \cdots$$

of largest exceedance in each year (some of which may be zeros) and define the random variable N_x for an arbitrary $x > 0$, in the following way:

$$N_x = \inf (\nu;\ \chi_\nu > x) \tag{2.48}$$

In other words, N_x is the smallest value of ν for which χ_ν is greater than x. The length of time (in years) that elapses before x is exceeded for the first time determines N_x. Under the assumption that (χ_ν) is a sequence of independent random variables with the common distribution function $F_t^*(x)$:

$$P\ (\chi_\nu \leq x) = F_t^*(x) \tag{2.49}$$

where t* is equal to the one-year period, it can be shown (Todorovic and Zelenhasic, 1970) that the expectation of the random

variable N_x for an arbitrary $x > 0$ is equal to:

$$E(N_x) = \frac{1}{1 - F_t^*(x)} \tag{2.50}$$

This represents the average passage time of the level x, i.e. the average number of years before the first exceedance of magnitude greater than x occurs. It is used to characterize the return period of the discharge x (equation 2.56).

Let $\chi(t)$ denote the magnitude of the largest exceedance within [0, t] i.e.,

$$\chi(t) = \max_{t(i) \leq t} \xi_i \tag{2.51}$$

so $\chi(t)$ is the maximum among a random number of random variables.

Naturally the distribution of $\chi(t)$ is dependent on both the distribution of the number of exceedances (events) in [0, t] which is usually assumed to be Poissonian (equation 2.45) and the distribution of $\{\xi_i\}$.

For a high enough base level x_0, the variables ξ_i may be assumed to be independent. Many authors (Todorovic and Zelenhasic, 1970; Todorovic, 1978; Ashkar and Rousselle, 1981) introduce the further assumption that the ξ_i's are identically distributed, their common distribution being of the exponential type:

$$F_{\xi}(x) = P\,[\xi \leq x] = 1 - e^{-\beta x},\ x > 0 \qquad (2.52)$$

The mean and variance of ξ are:

$$E\,(\xi) = 1 / \beta$$

$$Var\,(\xi) = 1 / \beta^2 \qquad (2.53)$$

From equations (2.45) and (2.52) and the assumption that (ξ_v) and $(\tau(v))$ are mutually independent sequences, the distribution of $\chi(t)$ the largest exceedance in [o, t] may be deduced (Todorovic and Zelenhasic, 1970):

$$F_t(x) = P\,[\chi(t) \leq x] = \exp\,[-\Lambda(t)\ e^{-\beta x}] \qquad (2.54)$$

Taking $t = t^*$ the one-year period and denoting $\Lambda(t^*)$ by λ for simplicity, we obtain (using equations 2.49 and 2.54):

$$F_t^*(x) = P\,[\chi_\nu \leq x] = \exp\,(-\lambda\ e^{-\beta x}) \qquad (2.55)$$

Combining relationships (2.50) and (2.55) we finally obtain:

$$T = E(N_x) = \frac{1}{1 - F_t^*(x)} = \frac{1}{1 - \exp\,\{-\lambda e^{-\beta x}\}} \qquad (2.56)$$

which gives the return period as a function of the exceedance x (or equivalently as a function of the discharge Q which differs from x only by a constant x_0; equation 2.43).

Remark that if one poses $\lambda = e^{\alpha}$ and $y = (\beta x - \alpha)$ in relationship (2.55), one obtains:

$$F_t^*(x) = e^{-e^{-y}} \tag{2.57}$$

which is the Gumbel distribution function already given in relationship (2.17).

The discharge X_T of return period T can be obtained by solving equation (2.56) with respect to x; this yields:

$$X_T = \frac{C_T + \ln \lambda}{\beta} \text{ with } C_T = -\ln \ln \frac{T}{T-1}$$

Note that in this last relationship X_T will be negative if λ is less than e^{-C_T}; so if we want X_T to be positive we should define:

$$X_T = \begin{cases} (C_T + \ln \lambda) / \beta \text{ for } \lambda > e^{-C_T} = \ln [T / (T-1)] \\ 0 \text{ otherwise} \end{cases} \tag{2.58}$$

2.3.2. Estimation of Design Events and Uncertainty of Estimation

In practice, the parameters λ and β in relationship (2.58) are estimated by $\hat{\lambda}$ and $\hat{\beta}$ obtained using a sample of flood data from N years of record. It can be shown (Ashkar and Rousselle, 1981) that the method of moments and the method of maximum likelihood yield the same estimates, given by:

$$\hat{\lambda} = M / N$$

$$\hat{\beta} = M / \left(\sum_{i=1}^{M} \xi_i \right) \tag{2.59}$$

where M is the number of exceedances ξ_i observed in the N years of record.

Replacing λ and β in relationship (2.58) with their estimates $\hat{\lambda}$ and $\hat{\beta}$ makes X_T a random variable, which we shall denote by $\hat{X}_{T,N}$ (to show that it varies with both the return period T, and the length of record N). The p.d.f. of $\hat{X}_{T,N}$ as derived by Ashkar and Rousselle (1981) is given by:

$$f_{T,N}(x) = \begin{cases} \dfrac{e^{-N\lambda}}{x} \sum_m^* [mN\lambda\beta x \exp(-\beta x/u)/u]^m / m!(m-1)!; & x > 0 \\ 1 - \sum_m^* \dfrac{e^{-N\lambda}(N\lambda)^m}{m!}, & x = 0 \end{cases} \qquad (2.60)$$

with

$$u = u_{T,N}(m) = -\ln\left[\ln\left(\frac{T}{T-1}\right)^N\right] + \ln m$$

The summation $\sum^*$ is over positive integers m that are greater than $\ln [T/(T-1)]^N$.

One of the advantages of the density function (2.60) is that it is exact (non-asymptotic), and therefore valid with any sample size N. It can be used to calculate the exact variance of $\hat{X}_{T,N}$ and exact confidence limits around $X_{T,N}$ at any level of confidence.

Cunnane (1973), assuming that Cov $(\hat{\lambda}, \hat{\beta}) = 0$ and using the well-known first-order Taylor approximations of Var (f(x,y)) and Cov (f(x,y), g(x,y)) derived the asymptotic variance of $\hat{X}_{T,N}$ as:

$$\text{Var}\,(\hat{X}_{T,N}) = \lambda^{-1}\,\beta^{-2}\,N^{-1}\,\{1 + [\ln\lambda + C_T]^2\} \qquad (2.61)$$

Ashkar and Rousselle (1981) and Tavares and Da Silva (1983) have shown, however that this assumption of independence of $\hat{\lambda}$ and $\hat{\beta}$ can lead to serious errors. The use of equation (2.61) is therefore not recommended in practice.

It would be very helpful to have tables based on relationship (2.60) that will give confidence limits around X_T for values of T, N, λ, β and confidence coefficient $(1 - \alpha)$, within the range of interest in flood analysis. Unfortunately, such tables are not available at present, but a computer program can easily be constructed to deal with relationship (2.60). Ashkar and Rousselle (1981) used such a program to calculate the density function $f_{T,N}(x)$ for different values of T and N using real data. Note that in order to calculate $f_{T,N}(x)$ the parameters λ and β in relationship (2.60) should be estimated beforehand using relationships (2.59). The cumulative distribution function $F_{T,N}(x)$ of $\hat{X}_{T,N}$ may be obtained from $f_{T,N}(x)$ by numerical integration. This too can easily be done on a computer. Once $F_{T,N}(x)$ is obtained, the calculation of confidence intervals at any level of confidence becomes a very simple task. This kind of calculation has been successfully done by El-Jabi et al. (1982).

2.3.3. Comparison of Annual Series and Partial Duration Series

Cunnane (1973) and also Taesombut and Yevjevich (1978) have compared the (asymptotic) estimation variance of $\hat{X}_{T,N}$ obtained by the partial duration series method with the assumption that Cov $(\hat{\lambda}, \hat{\beta}) = 0$ (relationship 2.61) to that resulting from fitting a Gumbel law to the annual maxima by the method of maximum likelihood (relationship 2.25). Cunnane's study concluded that the estimate of X_T of partial flood series in the case when the average number of exceedances per year is equal to one, has a larger sampling variance than the annual flood series estimate for the return periods greater than 10 years. Both studies (Cunnane, 1973; Taesombut and Yevjevich, 1978) concluded that partial flood series produce smaller sampling variance than annual flood series only if partial flood series contained at least 1.65 N items; N being the number of years of record.

As we have stated earlier, however, the use of equation (2.61) can lead to significant errors; the validity of the above results may therefore be doubtful.

Tavares and Da Silva (1983) carried out a study similar to the studies done by Cunnane (1973) and Taesombut and Yevjevich (1978) but evaded the use of relationship (2.61) by resorting to simulation. They concluded that the partial duration series method has a significantly lower estimation variance than the

annual maxima method if λ the mean number of exceedances per year is greater than 2. This reduction of the estimation variance increases with the return period and with λ.

We believe that more use of equation (2.60) may give a more accurate idea about the relative efficiency of the partial duration series method as compared to the method of annual maxima. Studies along this direction are encouraged.

2.3.4. Treatment of Non Identically Distributed Exceedances

In many cases there is good agreement between Gumbel's distribution and observed annual flood series indicating that the assumptions introduced in the derivation of equation (2.55) which is equivalent to equation (2.57) are basically correct. In this kind of a situation we shall say that "Model A" applies. There are cases, however, when Gumbel's law is clearly found inadequate for fitting the observed sequence of annual maxima. In this case the need for more refined models arises. The natural thing to do is to try and remove the strongest of the hypotheses on which equation (2.55) is based, namely that $\{\xi_i\}$ is a sequence of identically distributed random variables.

It may be found phenomenologically closer to reality to assume that the occurrence of exceedances still follows a time-dependent or "non-homogeneous" Poisson process, but that the

exceedance values are not identically distributed random variables. In fact, it has been remarked by many authors (Todorovic and Rousselle, 1971; North, 1980; Waylen and Woo, 1982) that the distribution of ξ_i is actually dependent on $\tau(i)$. One way of allowing for this time dependence is by retaining the exponential distribution as in equation (2.52) but assuming its parameter β to be time-dependent, i.e.,

$$F_\xi(x) = P\,[\xi_i \leq x \mid \tau(i) = t] = 1 - e^{-\beta(t)x}, \quad x > 0 \qquad (2.62)$$

This is the model suggested by North (1980). In this model (which we shall call Model B) the distribution of the largest exceedance $\chi(t)$ within [0, t] is given by:

$$F_t(x) = P\,[\chi(t) \leq x] = \exp\left[-\int_0^t e^{-\beta(u)x}\,\lambda(u)\,du\right] \qquad (2.63)$$

By taking $t = t^*$ (the one-year period) in this equation the distribution of the largest annual exceedance is obtained.

In areas where high flows are generated by more than one distinct hydrologic process (e.g. snowmelt-and rainfall-generated

peaks) the partial duration series may be modelled by considering ξ_i as the mixture of two or more independent components, each being exponentially distributed (Model C). In the case of two independent components, if the exceedances associated with component i, i = 1,2 occur according to a Poisson process of parameter $\Lambda_i(t)$, as in expression (2.46) then the distribution function of the largest exceedance in a year will be given by (Versace et al., 1981; Waylen and Woo, 1982):

$$F_t^*(x) = P\,[\chi_\nu \leq x] = \exp\,[-\lambda_1\, e^{-\beta_1 x} - \lambda_2\, e^{-\beta_2 x}) \qquad (2.64)$$

where t* stands for the one-year period; $\lambda_i = \Lambda_i(t^*)$ and β_i is the parameter of the exponential distribution associated with component i. Note that with a change of notation like the one introduced in equation (2.57), notably: $y_i = e^{\alpha_i}$ and $\lambda_i = (\beta_i x - \alpha_i)$, i = 1, 2, expression (2.64) reduces to:

$$F_t^*(x) = e^{-e^{-y_1}-e^{-y_2}} = e^{-e^{-y_1}} \cdot e^{-e^{-y_2}} \qquad (2.65)$$

In other words, the overall annual flood distribution (by "flood" we often mean "exceedance"; this should not lead to any confusion)

is the product of the annual flood distributions of the individual components.

On substituting expression (2.63) with $t = t^*$ (Model B) or expression (2.64) (Model C) in equation (2.50), the return period $T_x = E(N_x)$ associated with an exceedance magnitude equal to x can be calculated.

It was said in section 1.4 that when it can be shown that the set of recorded floods come from two or more distinct populations then it may be more hydrologically reasonable to try and separate the subpopulations on the basis of the distinct generating processes that gave rise to them (as was done in Model C) rather than separating floods by calendar periods. It may happen however, and in fact it may not be uncommon, that a separation on the basis of the distinct flood generating processes comes to reveal that the different subpopulations obtained are also separated according to separate calendar periods. These calendar periods may be called "seasons" and the "Seasonal Model" thus obtained (Todorovic and Rousselle, 1971) is nothing but a special case of Model C.

The Seasonal Model is also in effect a special case of Model B. It corresponds to the case of a piecewise constant function $\beta(t)$ in equation (2.62), $\beta(t)$ being constant within each

season, and assuming different values from one season to the other.

Ashkar and Rousselle (1981) derived the p.d.f. $f_{T,N}(x)$ of $\hat{X}_{T,N}$, the estimate of the event of return period T in the case of the Seasonal Model. This derivation is also valid for the more general Model C. The numerical calculation of this probability distribution $f_{T,N}(x)$ requires computer programming and involves a number of summations and numerical integrations that increases with the number of seasons (subpopulations) considered. With two or three seasons (subpopulations) no serious difficulties should occur with this numerical calculation. For the exact expression of $f_{T,N}(x)$ the reader is referred to (Ashkar and Rousselle 1981).

Once $f_{T,N}(x)$ is determined, the calculation of the cumulative distribution function $F_{T,N}(x)$ and of confidence intervals for X_T becomes an easy task.

With regard to Model B, the estimation of its parameters (North, 1980) is complex and requires elaborate computer programming. The distribution of $\hat{X}_{T,N}$ has not yet been studied for this model.

2.3.5. Applications and Additional Comments

It is difficult to construct an example using data from one single station that will apply all the theoretical results we have presented with the partial duration series approach, especially that different models were presented each having its own set of hypotheses. More than one example with more than one set of data are needed in order to clarify all the ideas that were put forward. To simplify our task of providing satisfactory practical applications we shall choose a number of examples already given in scientific journals or readily available publications, and present the reader with a brief summary of each example, adding some comments when we find it necessary.

As for Model B, North (1980) gives a numerical example. Since the kind of calculations involved calls for some quite elaborate numerical techniques and since we are primarily interested in calculations that can be reproduced with reasonable effort by practitioners, we shall not discuss this example here.

For the more practical Models A and C, here are some applications (in referring to equations or paragraphs pertaining to the present study we shall write "AB" (Ashkar and Bobée)):

MODEL A

This model may be widely applicable in parts of Canada where high flows are dominated by a single flood generating process, notably snowmelt.

Todorovic and Zelenhasic (1970, pp. 1645-1648) give a numerical example. A brief summary of this example follows:

(1) The river considered is Susquehanna River at Wilkes-Barre, Pennsylvania; N = 72 years.

(2) The base level x_0, is not chosen by the authors (in the United States it is usually furnished along with p.d.s. data by the U.S. Geological Survey).

(3) The question of which flood peaks above x_0 should be retained and which should be exluded is briefly addressed (cf. AB section 1.1).

(4) The question of independence of the exceedances ξ_i is briefly commented and the assumption of identically distributed exceedances for the whole year period is justified graphically and with a Kolmogorov-Smirnov test (cf. AB section 2.2.7 B).

(5) The observed function $\Lambda'(t)$, the average number of exceedances in [o, t] (AB equation 2.46) is plotted (TZ fig. 7) and a fitting function $\Lambda(t)$ is obtained for $\Lambda'(t)$ using a Fourier series fit procedure. This fitted function which can be computationally burdensome can for all practical purposes be replaced by the observed function $\Lambda'(t)$ in most applications.

(6) Observed and corresponding theoretical (Poisson) distributions of the number of exceedances within periods [o, t] for different values of t ranging from 20 days to 365 days, are plotted and compared (TZ fig. 6). A chi-square test of goodness of fit (AB section 2.2.7 A) could be used in this case to measure the discripancy between the observed and theoretical distributions.

(7) The distribution function $F_t(x)$ of $\chi(t)$ the largest exceedance in [o, t] (AB equation 2.62) is obtained after estimating β by β given in (AB equation 2.54). The result is (TZ equation 30). Observed and theoretical functions $F_t(x)$ are plotted and compared in (TZ fig. 9) for 160 -, 200-and 365-day periods. Observed and theoretical (exponential) distributions $F_\xi(x)$ of the magnitude of exceedances (AB equation 2.52) are plotted in (TZ fig. 8). A Kolmogorov- Smirnov test of goodness of fit could be used in the case of both $F_t(x)$ and $F_\xi(x)$ to measure their adequacy for fitting

the observed data.

(8) The event having a return period of 100 years is calculated using (AB equation 2.58 or equivalently, AB equation 2.55 ~ TZ equation 31).

MODEL C

This model may be applicable when floods are brought about by more than one generating process, snowmelt and rainfall for instance.

The following is a brief summary of a numerical example given in (Versace et al., 1981):

(1) The example considers the flood values obtained from a 36-year record of daily flows at the Amato River in Southern Italy.

(2) The coefficient of skewness C_s of the annual flood series is computed and found to be incompatible with what the Gumbel distribution should produce. This directs the attention to checking the validity of the hypotheses on which Model A is based.

(3) The base level is chosen such that λ, the observed mean number of exceedances per year is equal to 2.06. The hypothesis of a Poisson distribution of the number of events per year is checked graphically and by a statistical test. It is found acceptable. For more details on the test used, which employs a test statistic R equalling the ratio of the observed variance to the observed mean the reader is referred to (Cunnane, 1979).

(4) Observed and theoretical (exponential) distribution functions of exceedances are plotted and compared under both Models A and C. In Model C floods associated with desastrous storms are distinguished from other floods and considered as belonging to a separate population. From the plots obtained, it is shown that considering two subpopulations of flood exceedances (Model C) for the river under investigation improves substantially the fit to the observed data as compared with Model A which considers all exceedances as coming from the same population. A plot of the observed and theoretical distributions of the largest annual exceedance for Model A (AB equation 2.55) and for Model C (AB equation 2.64) demonstrates further the superiority of Model C over Model A.

(5) The same comparisons done between Models A and C are done also between Models B and C. This again points out the good performance of Model C.

Waylen and Woo (1982) applied Model C to the study of floods in southwestern British Columbia. The following is a brief summary of a numerical example they give:

(1) The river considered is Coquihalla River in the Cascade Mountains. The number of years of record is 34 years (daily discharge).

(2) In the region covered by the study, high flows may result from heavy winter rainfall or from snowmelt in spring.

(3) Observed annual floods and the fitted Gumbel distribution are represented graphically and the divergence of the data from the fitted function shows that Model A is not applicable. It is hypothesized that the floods are generated by two distinct processes: rainfall (including rain on snow) and snowmelt.

(4) The authors propose a differentiation between these two flood generating processes on the basis of antecedent precipitation. The available records of precipitation permit them to plot 4-day antecedent precipitation against annual flood discharge for each year. From this plot two subpopulations

are found to be distinguishable. When Gumbel distributions are fitted to the annual maxima series of each of these two subpopulations separately a better fit is obtained (WW fig. 4) as compared to the case where all the annual maxima are considered to belong to only one population (WW fig. 1). A point of interest in this case study is that the two subpopulations are associated with two separate seasons. The subpopulation of snowmelt-generated floods is associated with the period April - July, while the subpopulation of rainfall - generated floods is associated with the months October - March (WW fig. 3). Model C reduces in this case to the Seasonal Model that was mentioned in section 2.3.4.

(5) The two Gumbel ditributions are compounded to yield the distribution function of the annual flood (AB Equation 2.64). A comparison of the fit between observed and theoretical distributions in (WW fig. 5; Model C) and (WW fig. 1; Model A) demonstrates the improvement that Model C introduces as compared to the more restrictive Model A.

The base level

The choice of the base level x_0 is of particular importance in p.d.s. models. This importance stems from the fact that the (time-dependent) Poisson process as a model for flood count and the assumption of stochastic independence of flood exceedances

cannot be expected to be physically plausible if the truncation level or base level is not "relatively high".

The truncation level is usually chosen in such a way that on the average, no more than two or three exceedances occur in a year period (Langbein, 1949; Dalrymple, 1960). This criterion is no more than a "rule of thumb", however. Often in practice interest lies with a precise value of x_0, determined by the nature of the engineering problem at hand.

The natural question that comes to mind is the following: what are the effects of changing the base level upon the distributions employed in Models A, B and C above ?

Ashkar and Rousselle (1983) have proven that with any of the Models A, B or the "Seasonal Model" that we described previously (which is a special case of Model C), if the time-dependent Poisson process is found applicable with a certain truncation level x_0 then it should remain so with any level y_0 higher than x_0. Although this proof as given by Ashkar and Rousselle does not cover Model C in its general form, a simple modification may be introduced in the proof, that allows Model C to be covered too. If we denote by $n'(t)$ the number of exceedances within $[o,t]$ obtained from a truncation at the level $y_0 = x_0 + h$ $(h > 0)$, then the intensity function $\lambda'(t)$ of the Poisson process at the level y_0 is related to the corresponding intensity function $\lambda(t)$ at the

level x_0 by the following equation:

$$\lambda'(t) = p_{t,h} \cdot \lambda(t)$$

where $p_{t,h}$ is the probability of an exceedance occuring at time t being greater than h. Note that in Model C an exceedance ocurring at time t may belong to one of a number of different populations, in which case $p_{t,h}$ should be taken as the overall probability of this exceedance being greater than h.

Another point of practical interest regarding Models, A, B and C is that the exponential distribution used in these three models as the distribution of flood magnitude shares the same property that we have just described in relation to the Poisson process. In fact, if the exponential distribution (with parameter β) is found applicable at a certain truncation level x_0 then as the truncation level is raised, not only does this distribution remain applicable, but its parameter β remains constant too (Ross, 1976).

These interesting characteristics of the Poisson process and of the exponential distribution give a great degree of freedom to the practicing engineer to choose the truncation level that he

finds adequate for the problem at hand without having to worry too much about the sensitivity of the obtained results to the choice of the truncation level. As a general rule, the truncation level should be chosen high enough so as to satisfy the Poisson model and the hypothesis of independence of exceedances, and low enough to get a good number of events that permits reliable estimates of distribution parameters.

REFERENCES

ANDERSON, R.L. 1941. Distribution of the serial correlation coefficient. Annals of Mathematical Statistics, 8(1), pp. 1-13.

ASHKAR, F. and ROUSSELLE, J. 1981. Design discharge as a random variable: A risk study. Water Resources Research, 17(3), pp. 577-591.

ASHKAR, F. and ROUSSELLE, J. 1983. Some remarks on the truncation used in partial flood series models. Water Resources Research, 19(2), pp. 477-480.

BENSON, M.A. 1968. Uniform flood frequency estimating methods for federal agencies. Water Resources Research, 4(5), pp. 891-908.

BOBÉE, B. 1973. Sample error of T-year events computed by fitting a Pearson type 3 distribution. Water Resources Research, 9(5), pp. 1264-1270.

BOBÉE, B. 1975. The log-Pearson type 3 distribution and its application in hydrology. Water Resources Research, 11(5), pp. 681-689.

BOBÉE, B. 1979. Comment on the log-Pearson tye 3 distribution: the T-year event and its asymptotic standard error by maximum likelihood theory by R. Condie. Water Resources Research, 15(1), pp. 189-190.

BOBÉE, B. and BOUCHER, P. 1981a. Ajustement des distributions Pearson type 3, Gamma, Log-Pearson type 3 et Log-Gamma. Rapport scientifique no 105, INRS-Eau.

BOBÉE, B. and BOUCHER, P. 1981b. Calcul de la variance d'un événement de période de retour T. Cas des lois Log-Pearson type 3 et Log-Gamma ajustées par la méthode des moments sur la série des valeurs observées. Rapport scientifique no 135, INRS-Eau.

BOBÉE, B. and ROBITAILLE, R. 1975. Correction of bias in the estimation of the coefficient of skewness. Water Resources Research, 11(6), pp. 851-854.

BOBÉE, B. and ROBITAILLE, R. 1976. Ajustement des distributions Pearson 3, Gamma, Log-Pearson 3, Log-Gamma. Rapport scientifique no 70, INRS-Eau, 76 p.

BOBÉE, B. and ROBITAILLE, R. 1977. The use of the Pearson type 3 and log-Pearson type 3 distributions revisited. Water Resources Research, 13(2), pp. 427-443.

BRUNET-MORET, Y. 1973. Statistique de rangs. Cah. ORSTOM, Ser. Hydrol, X(2), pp. 133-151.

CHOW, V.T. 1950. Discussion of Annual floods and the partial duration flood series. Transactions American Geophysical Union, 31(6), pp. 939-941.

CHOW, V.T. 1953. Frequency analysis of hydrologic data with special application to rainfall intensities. Univ. Illinois, Eng. Exper. Stat. Bull.: 414.

CONOVER, W.J. 1971. Practical nonparametric statistics, Wiley, New York.

CSOMA, J. 1969. An estimate of discharge probabilities floods and their computation. A.I.H.S., Vol. 84.

CUNNANE, C. 1973. A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. Journal of Hydrology, 18, pp. 257-271.

CUNNANE, C. 1978. Unbiased plotting positions - A review. Journal of Hydrology, 37, pp. 205-222.

CUNNANE, C. 1979. A note on the Poisson assumption in partial duration series models. Water Resources Research 15(2), pp. 489-494.

DALRYMPLE, T. 1960. Flood-frequency analysis, U.S. Geological Survey, Water Supply Paper 1543-A, 80 p.

EL-JABI, N. RICHARD, D., ASHKAR, F. and ROUSSELLE, J. 1982. Analyse stochastique des débits de crue pour la province de Québec. Ministère de l'Environnement du Québec, 615 p.

FRANCOU, J. and RODIER, J.A. 1969. Essais de classification des crues annuelles. A.I.H.S., vol. 84.

GUMBEL, E.J. 1958. Statistics of extremes. Columbia University Press, New York, NY, 375 p.

HAAN, C.T. 1977. Statistical methods in hydrology. Iowa State University Press, 378 p.

HARDISON, C.H. 1969. Accuracy of streamflow characteristics in geological survey research. U.S. Geol. Survey Pro. paper 650-D, pp. D210-D214.

HARTER, H.L. 1961. Expected values of normal order statistics. Biometrika, 48(1), pp. 151-165.

HARTER, H.L. 1969. A new table of percentage points of the Pearson type 3 distribution. Technometrics, 2(1), pp. 177-187.

HOSHI, K and BURGES, S.J. 1981. Sampling properties of parameter estimates for the log-Pearson tye 3 distribution, using moments in real space. Journal of Hydrology, 53, pp. 305-316.

HOUGHTON, J.C. 1978a. Birth of a parent: The Wakeby distribution for modeling flood flows. Water Resources Research, 14(6), pp. 1105-1109.

HOUGHTON, J.C. 1978b. The incomplete means estimation procedure applied to flood frequency analysis. Water Resources Research, 14(6), pp. 1111-1115.

KIMBALL, B.F. 1960. On the choice of plotting positions on probability paper. Journal of the American Statistical Association, 55(291), pp. 546-560.

KIRBY, W. 1974. Algebraic boundedness of sample statistics. Water Resources Research, 10(2), pp. 220-222.

KITE, G.W. 1976. Frequency and risk analysis in hydrology. Networkd planning and forecasting section applied hydrology division. Water Resources branch. Department of Environment, Ottawa.

KITE, G.W. 1978. Frequency and risk analyses in hydrology. Water Resources Publications, Fort Collins, Colorado, 224 p.

KLEMES, J. 1970. Negatively skewed distribution of runoff. Symposium on representative and experimental basin. AIHS no 96.

KOLMOGOROV, A. 1933. Sulla determinazione empirica di una legge di distribuzione, Giornata, Inst. Ital. Attuari, 4, pp. 83-91.

LANGBEIN, W.B. 1949. Annual floods and the partial duration series. Transactions American Geophysical Union, 30(6), pp. 879-881.

LILLIEFORS, H.W. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62, pp. 399-402.

LILLIEFORS, H.W. 1969. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association, 64, pp. 387-389.

LLOYD, E.H. 1970. Return period in the presence of persistence. Journal of Hydrology, 10(3), pp. 291-298.

MANN, H.B. and WHITNEY, D.R. 1947. On the test of whether one of two random variables is stochastically larger than the order. Annals of Mathematical statistics, 18, pp. 50-60.

MATALAS, N.C. and WALLIS, J.R. 1973. Eureka! it fits a Pearson type 3 distribution. Water Resources Research, 9(2), pp. 281-289.

NORTH, M. 1980. Time-dependent stochastic model of floods, ASCE Journal of Hydraulics Division, 106 (HY5), pp. 649-665.

ROSS, S.M. 1976. Introduction to probability models. Academic Press, New York, NY, 272 p.

STODDART, R.B.L. and WATT, W.E. 1970. Flood frequency prediction for intermediate drainage basins in southern Ontario. C.E. Research Report 66, Queen's University at Kingston.

TAESOMBUT, V. and YEVJEVICH, V. 1978. Use of partial flood series for estimating distribution of maximum annual flood peak. Colorado state University, Hydrology Paper 97, 71 p.

TAKEUCHI, K. 1984. Annual maximum series and partial-duration series - Evaluation of Langein's formula and Chow's discussion. Journal of Hydrology, 68, pp. 275-284.

TAVARES, L.V. and DA SILVA, J.E. 1983. Partial duration series method revisited. Journal of Hydrology, 64, pp. 1-14.

TERRY, M.E. 1952. Some rank order tests which are most powerful against specific parametric alternatives. Annals of Mathematical Statistics, 23, pp. 346-366.

THOM, H.C.S. 1958. A note on the gamma distribution. Monthly Weather Review, 86(4), pp. 117-122.

TODOROVIC, P. 1970. On some problems involving random number of random variables. Annals of Mathematical Statistics, 41(3), pp. 1059-1063.

TODOROVIC, P. 1978. Stochastic models of floods. Water Resources Research, 14(2), pp. 345-356.

TODOROVIC, P. and ROUSSELLE, J. 1971. Some problems of flood analysis. Water Resources Research, 7(5), pp. 1144-1150.

TODOROVIC, P. and ZELENHASIC, E. 1970. A stochastic model for flood analysis. Water Resources Research 6(6), pp. 1641-1648.

VERSACE, P., FIORENTINO, M. and ROSSI, F. 1981. Analysis of flood series by stochastic models. Proceedings of the International Conference on Time Series Methods in Hydrosciences, October 6-8, 1981, Burlington, Ontario, pp. 315-324.

WALD, A., and WOLFOWITZ, J. 1943. An exact test for randomness in the non-parametric case based on serial correlation. Annals of Mathematical Statistics, 14, pp. 378-388.

WALLIS, J.R., MATALAS, N.C. and SLACK, J.R. 1974. Just a moment! Water Resources Research, 10(2), pp. 211-219.

WATER RESOURCES COUNCIL 1976. Guidelines for determining flood flow frequency. Bull. n° 17, the hydrology committee.

WAYLEN, P. and WOO, M.K. 1982. Prediction of Annual floods generated by mixed processes. Water Resources Research, 18(4), pp. 1283-1286.

YEVJEVICH, V. 1963. Fluctuations of wet and dry years part 1 research data assembly and mathematical models. Colorado State University, Hydrology Paper, n° 1.

TABLE 1.1: Risk of obtaining one or more floods of return period T or greater within the specified length of time

Percent chance of getting one or more such or bigger floods in this many years					Return period (Years)
100 years	50 years	25 years	10 years	1 year	
				50	2
				20	5
	99.9	94	65	10	10
	90.5	71	40	5	20
86	63	40	18	2	50
63	39	22	9.6	1	100
39	22	12	5	0.5	200
18	9.5	5	2	0.2	500
9.5	4.8	2.5	1	0.1	1000
5	2.3	1.2	0.5	.05	2000
2	1.0	0.5	0.2	.02	5000
1	0.5	.25	0.1	.01	10000

TABLE 2.1: Maximum daily discharge values for the South Saskatchewan River at Saskatoon - Station n° KF62

YEAR	MAXIMUM DAILY DISCHARGE	YEAR	MAXIMUM DAILY DISCHARGE
1912	1424 m^3/sec on Jul. 17	1942	1818 m^3/sec on Jun. 13
1913	1209 m^3/sec on Jul. 6	1943	1280 m^3/sec on Apr. 9
1914	994 m^3/sec on Jun. 26	1944	631 m^3/sec on Jun. 25
1915	3143 m^3/sec on Jul. 2	1945	1169 m^3/sec on Jun. 14
1916	2633 m^3/sec on Jul. 4	1946	1053 m^3/sec on Jun. 16
1917	1974 m^3/sec on Jun. 9	1947	1826 m^3/sec on Mar. 28
1918	1184 m^3/sec on Jun. 22	1948	2696 m^3/sec on Apr. 24
1919	980 m^3/sec on Jun. 5	1949	541 m^3/sec on Jun. 11
1920	1761 m^3/sec on May 14	1950	1274 m^3/sec on Jun. 30
1921	1416 m^3/sec on Apr. 15	1951	1540 m^3/sec on Jul. 1
1922	1070 m^3/sec on Jun. 13	1952	2418 m^3/sec on Apr. 10
1923	3370 m^3/sec on Jun. 7	1953	3936 m^3/sec on Jun. 15
1924	900 m^3/sec on Jun. 24	1954	1535 m^3/sec on Sep. 2
1925	1365 m^3/sec on Apr. 5	1955	2183 m^3/sec on Apr. 10
1926	1150 m^3/sec on Sep. 19	1956	1192 m^3/sec on Apr. 15
1927	2330 m^3/sec on Jun. 17	1957	855 m^3/sec on May 30
1928	2486 m^3/sec on Jul. 7	1958	852 m^3/sec on Apr. 17
1929	3058 m^3/sec on Jun. 9	1959	1252 m^3/sec on Jul. 5
1930	855 m^3/sec on Jun. 19	1960	1124 m^3/sec on Apr. 3
1931	583 m^3/sec on Jun. 29	1961	1070 m^3/sec on Jun. 4
1932	3143 m^3/sec on Jun. 8	1962	815 m^3/sec on Apr. 16
1933	1082 m^3/sec on Jun. 25	1963	1540 m^3/sec on Jul. 8
1934	1138 m^3/sec on Jun. 15	1964	1424 m^3/sec on Jun. 18
1935	793 m^3/sec on Jun. 28	1965	1634 m^3/sec on Jun. 27
1936	861 m^3/sec on Apr. 18	1966	1260 m^3/sec on Jun. 13
1937	1107 m^3/sec on Jun. 21	1967	583 m^3/sec on Apr. 30
1938	1365 m^3/sec on Jun. 3	1968	595 m^3/sec on Jan. 9
1939	1422 m^3/sec on Jun. 24	1969	1778 m^3/sec on Jul. 6
1940	1572 m^3/sec on Apr. 22	1970	399 m^3/sec on Nov. 30
1941	926 m^3/sec on Apr. 7		

TABLE 2.2: Estimation of the parameters (station kF 62)

Method / Parameters	PEARSON TYPE 3				LOG-PEARSON TYPE 3				
	1	2	3	4	5	6	7	8	9
α	.0020	.0018	.0018	.0020	-29.42	94.74	82.81	87.10	77.97
λ	2.473	1.889	1.943	2.360	44.85	411.20	314.16	347.50	274.00
m	267.00	420.29	405.15	327.47	4.639	-1.221	-0.674	-0.871	-0.395
(C_s)	1.272	1.455	1.435	1.302	-0.299	.099	0.113	0.107	0.121

FIGURE 2.1: Adjustment of the Pearson type 3 distribution with the skewness correction $(C_s)_3$

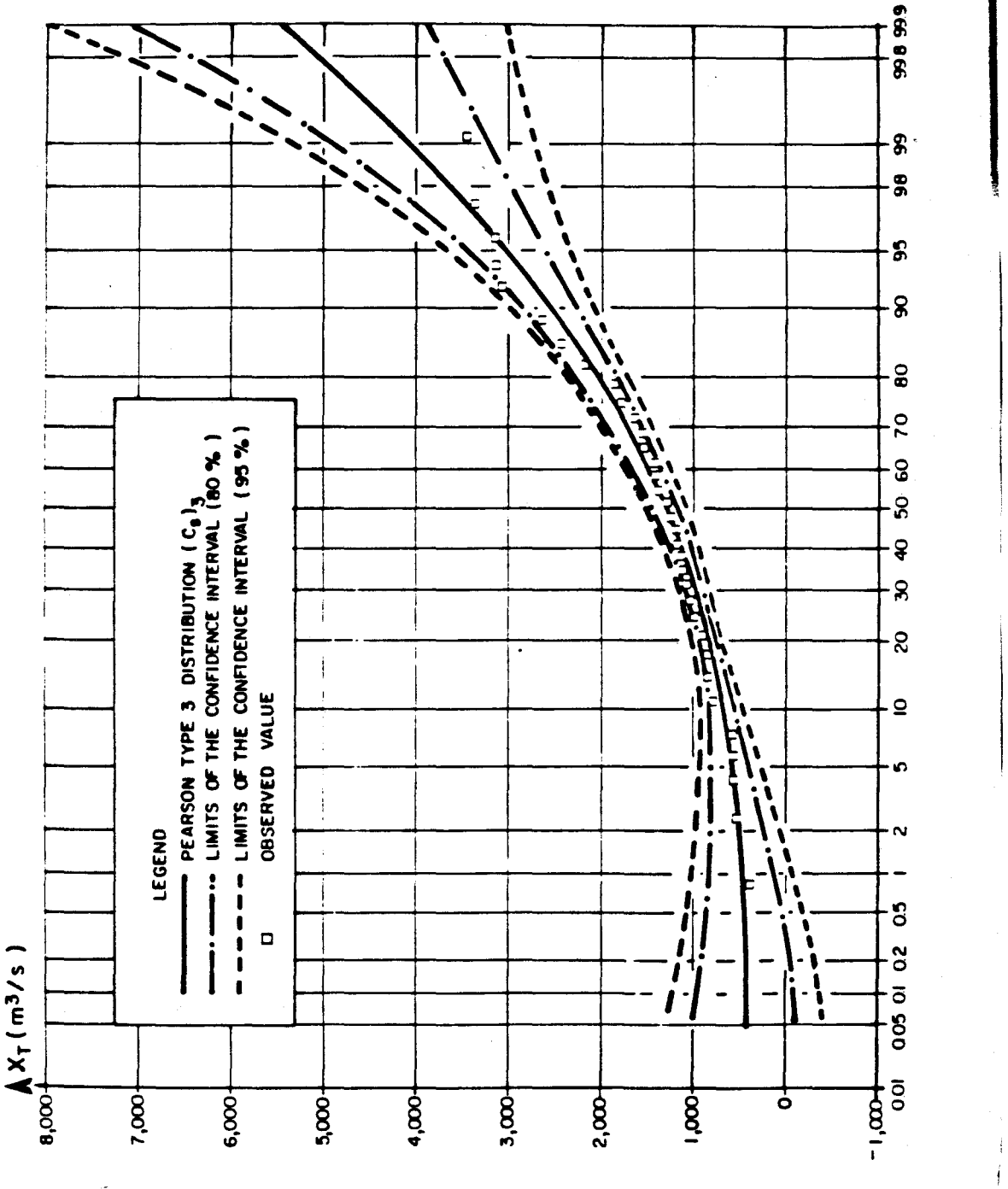

FIGURE 2.2: Adjustment of the log-Pearson type 3 distribution by considering the methods 5 and 6

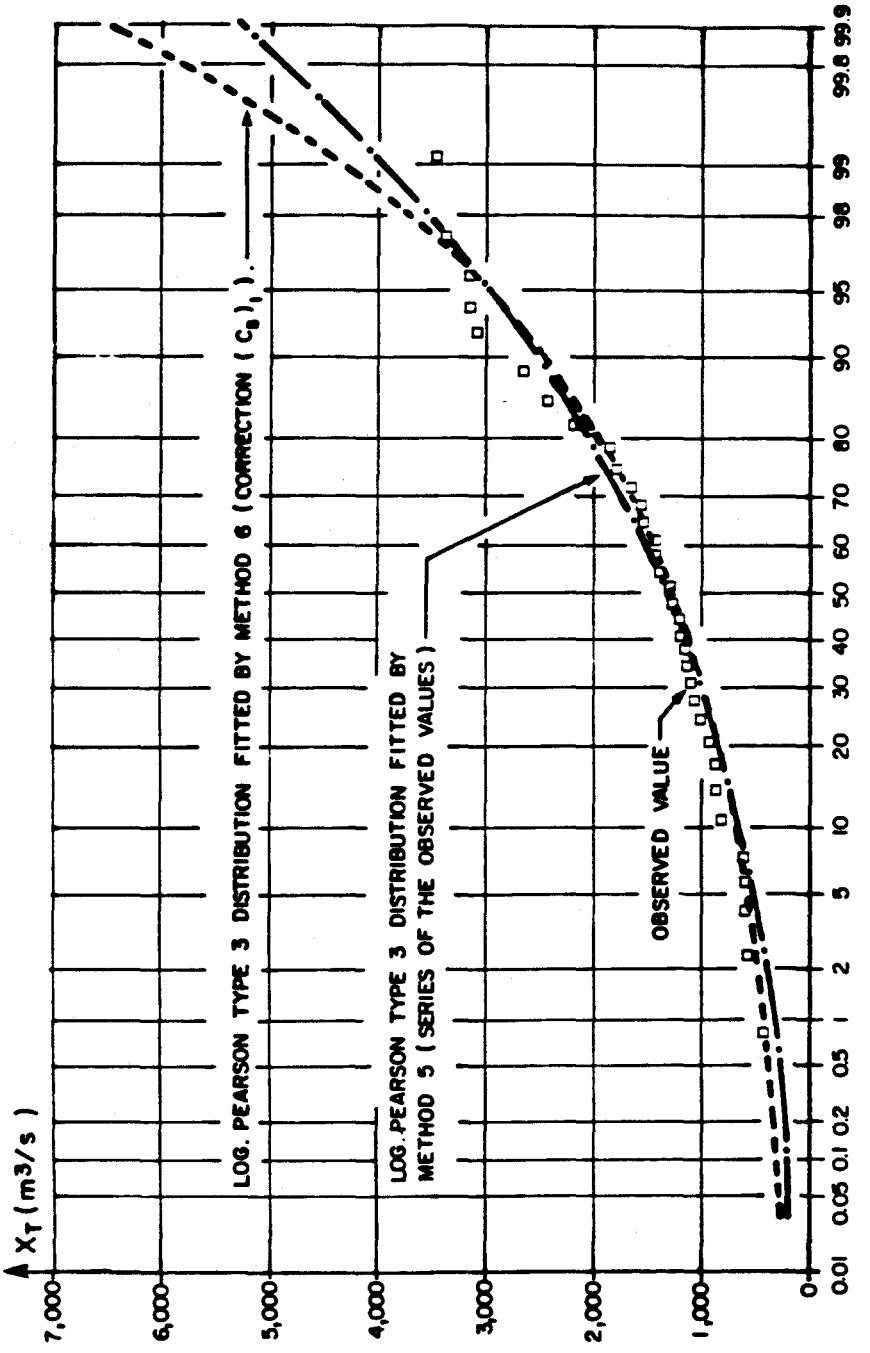

FIGURE 2.3: Stochastic representation of a streamflow hydrograph