Table 5.4 Frequency factors for common distributions

Distribution	*Skewness* (g)	*Value of K_T* *Return Period (years)* 2 *Exceedance probability* .50	5 .20	10 .10	25 .04	50 .02	100 .01
N, LN, 3LN[a]		0	.842	1.282	1.751	2.054	2.326
EV1[b]		−.164	.719	1.305	2.044	2.592	3.137
P3[c],LP3[c]	1.0	−.164	.758	1.340	2.043	2.542	3.023
	0.5	−.083	.808	1.323	1.910	2.311	2.686
	0	0	.842	1.282	1.751	2.054	2.326
	−0.5	.083	.857	1.216	1.567	1.777	1.955
	−1.0	.164	.852	1.128	1.366	1.492	1.588

a. $K_T = z_T$ where z is the standard normal variate. For LN, use mean and standard deviation of $y = \ln x$ to give y_T. For 3LN, use mean and standard deviation of $y = \ln(x - \epsilon)$ to give y_T.

b. For EV1, $K_T = \frac{-\sqrt{6}}{\pi}\{0.577 + \ln \ln[T/(T-1)]\}$

c. For LP3, detailed tables are given in Bulletin 17B (U.S. Interagency Advisory Committee on Water Data 1982). For $-1 \leq g \leq 1$, $K_T = \frac{2}{g}\{[(z - g/6)\frac{g}{6} + 1]^3 - 1\}$

Table 5.5 Coefficients of skewness for two-parameter distributions

Distribution	*Coefficient of Skewness*, γ_1
N	$\gamma_{1,x} = 0$
LN	$\gamma_{1,\ln x} = 0$
EV1	$\gamma_{1,x} = 1.14$

Standard error of quantiles

The standard error of a quantile x_T depends on the form of the distribution and the variance and covariance of parameter estimates. For ML estimates, the solution procedure required for x_T also yields the standard error, $se(x_T)$. For MM estimates, $se(x_T)$ can be expressed in terms of the sample standard deviation, sample size, and a factor δ_T that depends on T and the form of the distribution

$$se(x_T) = \delta_T\, s/\sqrt{N} \qquad [5.10]$$

Expressions for and tables of δ_T for common distributions are summarized by Kite (1977). Table 5.7 gives values of δ_T for the N, LN, and EV1 distributions. For the three-parameter distributions (3LN, EV2, EV3, P3, and LP3), estimates of both x_T and $se(x_T)$ are sometimes provided by available computer programs.

One way to illustrate the relative reliability of estimates for x_T) over the desired range of T is to plot curves at a specified distance, for instance 2 $se(x_T)$, above and below the fitted frequency curve (see Figure 5.5). (If the sampling distribution of x_T) were normal, this would be approximately equivalent to a 95 per cent confidence band.) These curves are sometimes loosely described as defining confidence or reliability bands. Hydrologists differ about the practical utility of these bands.

A slightly different approach to reliability bands on frequency curves was presented by Beard (1962) and reproduced by Viessman *et al.* (1977) in the form of a table of so-called 'error limits' as a function of return period and years of record. For a 90 per cent 'reliability band' (range between 5 per cent and 95 per cent limits) the values are as shown in Table 5.8. The derivation of the table assumes a normal or lognormal distribution.

5.2.4 Peaks Over Threshold (POT) Approach

The POT (or PDS, partial duration series) approach to flood frequency analysis, referred to briefly in Section 5.2.1, is quite different from the more common approach using annual maximum series, described above. For relatively long flood series, however, the practical results are not significantly different (as discussed in subsequent sections). Yet the POT method can be particularly useful when the period of record is short – for instance, less than 15 years – because POT series can usually be selected so as to contain larger numbers of flood peaks than annual maximum series. It should be stressed that the particular POT method discussed below implies a Gumbel distribution of annual maxima, and hence is applicable only when the Gumbel distribution is appropriate. The AM and POT approaches differ, however, in the estimation of parameters and quantiles. In this section only a simple POT model (hereafter referred to as the POT model) is considered. Work on this model by Todorovic and Zelenhasic (1970), Todorovic and Rousselle (1971), El-Jabi *et al* (1982), and others has demonstrated its practical interest. A fairly comprehensive discussion of the POT method is given in NERC (1975).

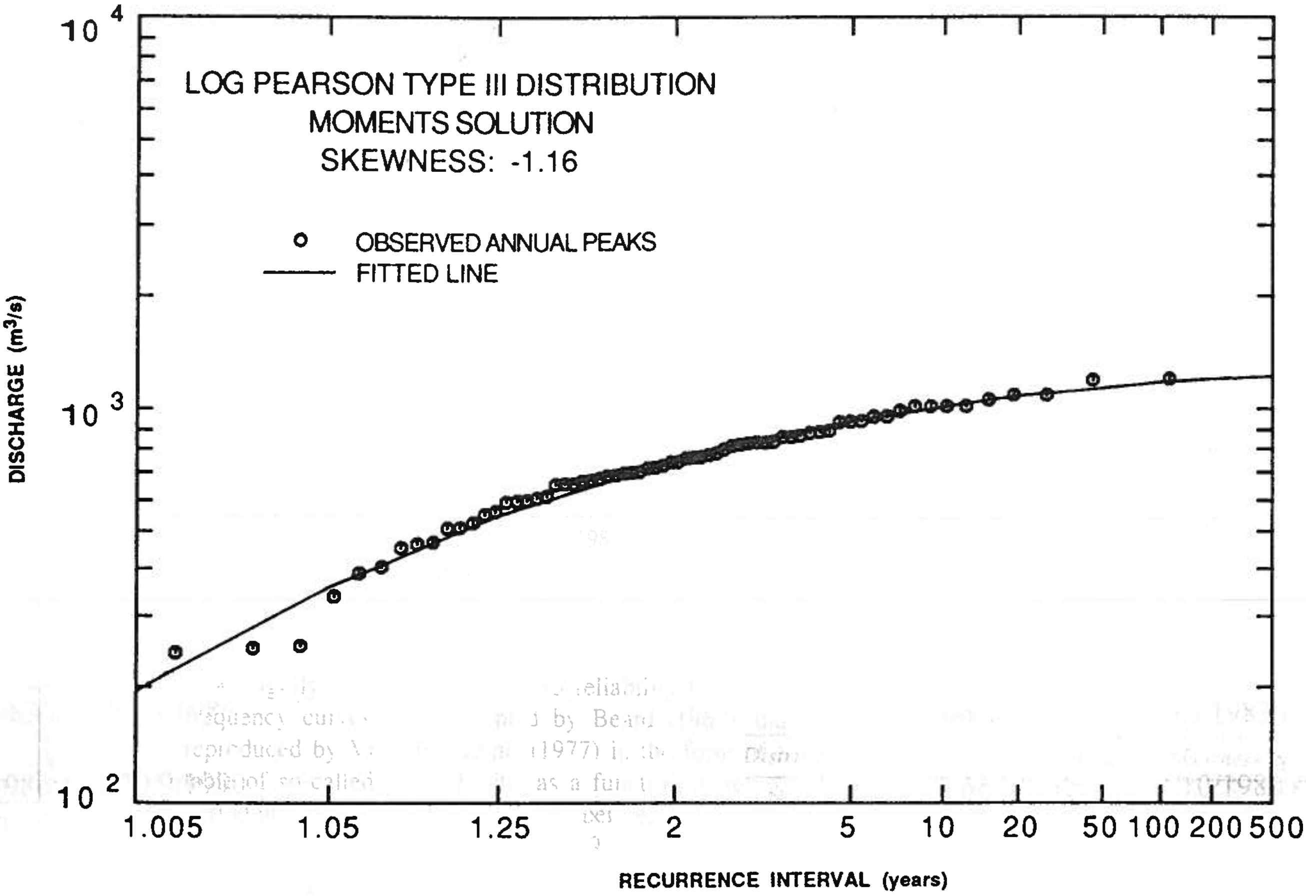

Figure 5.4 Logarithmically transformed flood series with negative skew

Table 5.6 Plotting position formulas

Method	*Equation*	*Reference*	*Comments*
California	$p_m = m/N$	Chow (1964)	Used in parts of Canada.
Hazen	$p_m = (m-0.5)/N$	Hazen (1914)	Not in general use.
Weibull	$p_m = m/(N+1)$	Weibull (1939)	Recommended in Bulletin 17B (U.S. Interagency Advisory Committee on Water Data 1982).
Chegodayev	$p_m = \dfrac{m - 0.3}{N + 0.4}$	Chow (1964)	Not in general use.
Cunnane	$p_m = \dfrac{m - 0.4}{N + 0.2}$	Cunnane (1978)	Used by the Water Resources Branch of Environment Canada (Pilon *et al.* 1985).

Description of the POT procedure

A base level x_0 is selected such that the exceedances can be assumed to be independent, and only those flood peaks that exceed x_0 are considered. In the case of a multiple-peaked event, only the largest peak is considered (Figure 5.6) because the exceedances must be independent.

The number of exceedances in the selected time interval (e.g. one year) is assumed to follow a Poisson distribution whose parameter k is estimated by

$$k = M/N \qquad [5.11]$$

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

It has been shown (Todorovic and Zelenhasic 1970; Todorovic 1978; Ashkar and Rousselle 1981) that in many cases exceedances within a season or within a year can be taken as belonging to a common distribution of the exponential type

$$F(x) = 1 - \exp[-\beta(x-x_0)] \quad x > x_0 \qquad [5.12]$$

where the parameter β is estimated by

$$\beta = M / \sum_{i=1}^{M} (x_i - x_0) \qquad [5.13]$$

The quantile estimate x_T is

$$x_T = x_0 + (\ln T + \ln k) / \beta \qquad [5.14]$$

Standard error of quantiles

The asymptotic standard error of x_T (Cunnane 1973) is given by

$$se(x_T) = \{[1 + (\ln k + \ln T)^2]/(\beta^2 k N)\}^{1/2} \qquad [5.15]$$

For the small sample sizes usually encountered in hydrologic frequency analysis this equation is useful only for indicating the order of magnitude of the standard error.

Choice of base level

Ashkar and Rousselle (1983) discuss the choice of base level x_0, of particular importance because the assumption of stochastic independence of flood exceedances and also the Poisson and exponential hypotheses depend on a base level that is 'relatively high'.

When the Poisson and exponential distributions are found adequate, they permit the engineer to choose a suitable base level for the problem at hand. As a general rule, the base level should be chosen high enough to satisfy the assumptions of Poisson and exponential distributions as well as independent exceedances, but not too high to significantly reduce the sample size.

The Poisson and exponential distributions are relatively simple models to handle, but if either is found inapplicable it should be replaced. The goodness-of-fit of the exponential distribution can be tested by means of a Kolmogorov-Smirnov test, while for the Poisson distribution a chi-square test can be used.

Choice between POT and annual maximum series

Cunnane (1973) compared the two methods on the basis of the standard error of x_T. In the case of the average number of exceedances per year being equal to one, Cunnane found that the estimate of x_T obtained by the POT approach has a larger sampling variance that the AMS estimate for return periods greater than 10 years. He concluded that the POT method produces smaller standard errors of x_T than annual series only if the POT series contains a number of data equal to at least 1.65 N.

Example: POT approach for Saint-Francois River.

For the Saint-Francois River (station 030219, Environment Quebec), a period of record of 39 years is available. The first step is to choose the base level x_0. To do this, the ratio (mean/variance) of the number of exceedances per year is plotted against the mean number of exceedances per year for different base levels. As the base level is raised, the Poisson distribution normally will give a better fit. In other words, as x_0 is increased k will be reduced and the ratio will tend to stabilize around unity as shown in Figure 5.7a. (For the Poisson distribution, the ratio of the mean to the variance of the number of exceedances is one.) With further increases in x_0 (i.e. reduction in k), the ratio will start to fluctuate. This is the value of x_0 that should be selected. In this particular case, the base level chosen corresponds to 133 m^3/s which in turn corresponds to a value of 1.95 exceedances/year.

After estimating x_0 and k, the second step is to plot the Poisson distribution along with the observed distribution of the number of exceedances per year (see Figure 5.7b) and to test the assumption that the Poisson distribution applies using a chi-square test, for example. In Figure 5.7b it is seen that, at the 5 per cent level, the assumption is not rejected.

The third step is to compute β using [5.13]. In this example, $\beta = 0.016$. The cumulative distribution function given by [5.12] is plotted in Fig. 5.7c, along with a plot of the observed distribution function. The hypothesis that an exponential distribution applies cannot be rejected at the 5 per cent level using the Kolmogorov-Smirnov test (see Figure 5.7c).

To obtain some assurance that the exceedances can be taken to be independent, exceedances separated from each other by a period of less than 10 days are plotted against each other, a procedure which does not indicate statistical dependence among the successive exceedances. The last step, in which x_T is plotted against T (Figure 5.7d) is then carried out; the corresponding 95 per cent confidence limits are also plotted.

5.2.5 Multiple Populations

The annual flood series at a station is normally treated as a sample from an implied single population for purposes of frequency analysis. In fact, however, most annual flood series in Canada contain floods of two or more physical types; in particular, snowmelt and rainfall floods may occur at different times of year and have quite different characteristics. In theory then an argument exists for treating most Canadian series as composed of two or more samples from different populations. Practically, however, such treatment considerably complicates the preparation of a frequency analysis, and there is little reason to do so unless treatment as a single population produces a pecul-

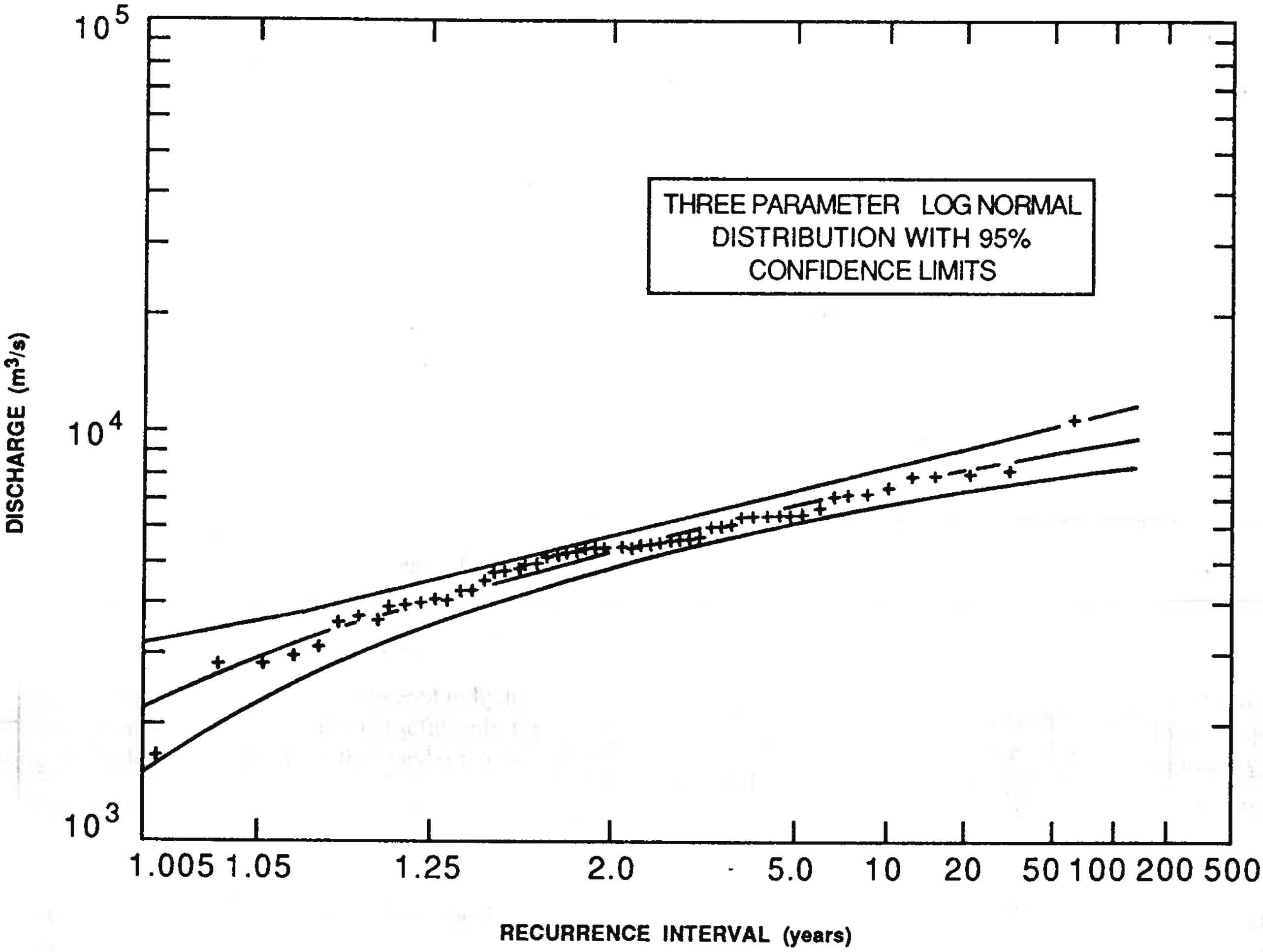

Figure 5.5 Confidence bands on quantile estimates, Saint John River, WSC no. 01AK002 and WSC no. 01AK004 (after Keefe 1979)

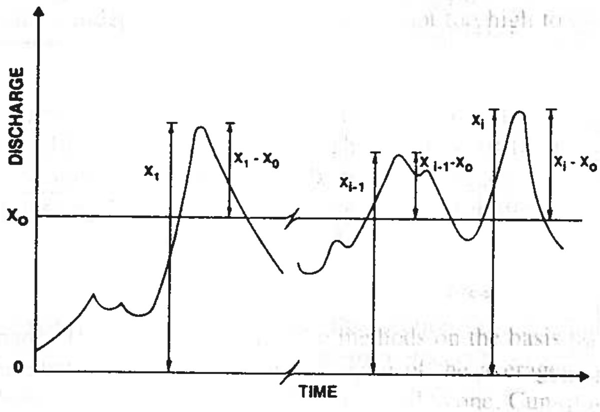

Figure 5.6 Stochastic representation of a streamflow hydrograph

Table 5.7 Standard errors of quantiles

Distri-bution	Value of δ_T Return Period (years) 2	5	10	25	50	100
	Exceedance probability .50	.20	.10	.04	.02	.01
N[a], LN[a,b]	1.000	1.164	1.350	1.592	1.763	1.925
EV1[c]	.918	1.546	1.951	2.815	3.368	4.025

a. $\delta_T = (1 + z_T^2/2)^{0.5}$
b. For LN, application of [5.10] gives *se* (ln x_T)
c. $\delta_T = (1 + 1.1396\, K_T + 1.1000\, K_T^2)^{0.5}$

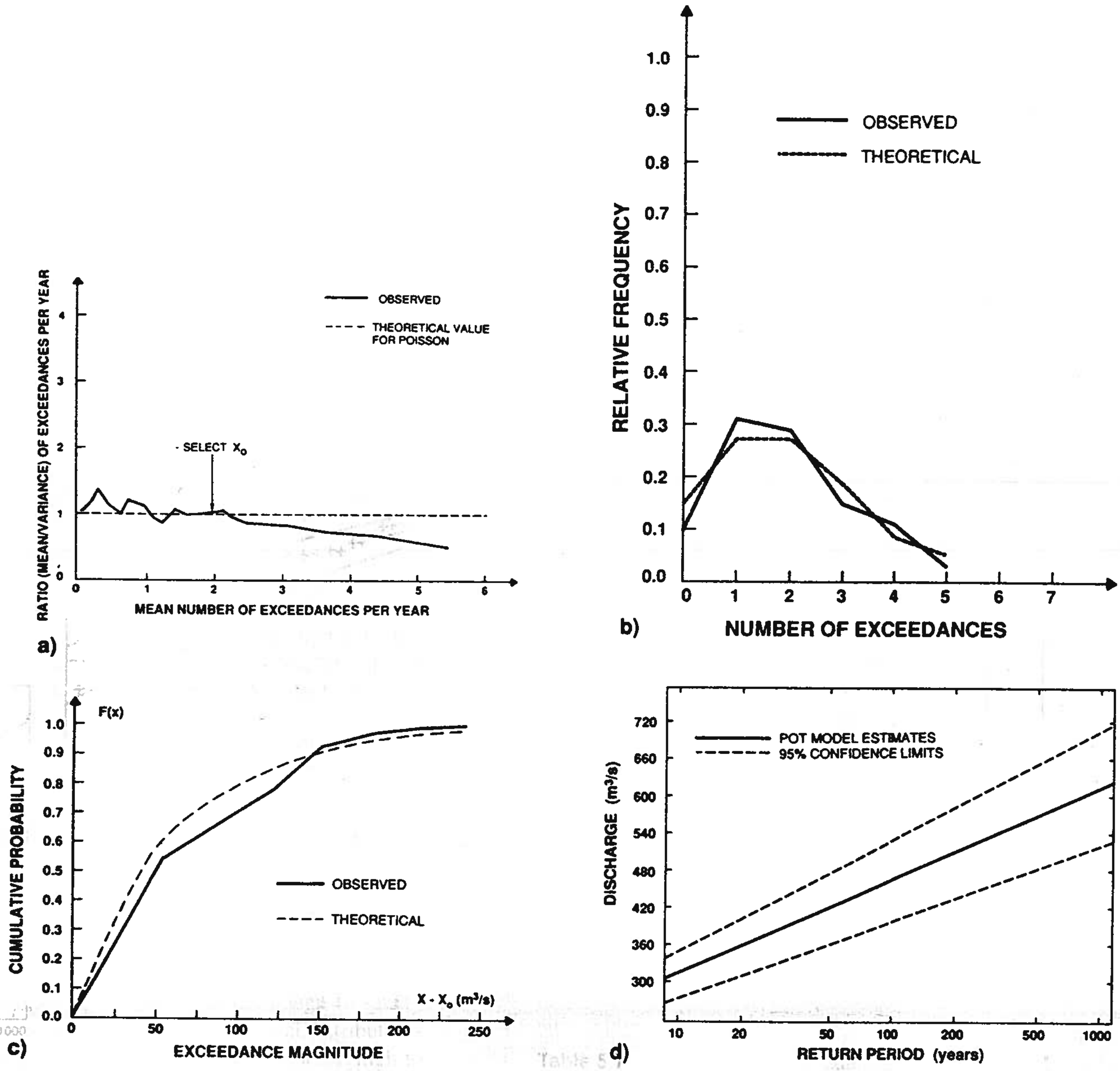

Figure 5.7 Application of the POT model to station no. 030219 (Environment Quebec) on the Saint-Francois River (after El-Jabi *et al.* 1982)

iar shape of frequency curve or there are reasons for determining separate design floods of the two types. If the series is to be divided, some clear basis for division should be apparent. A wide separation in timing, combined with a significant difference in statistical parameters between the two subseries as demonstrated by a homogeneity test, is an example.

When two or more subseries of annual floods are to be analyzed separately, it is necessary to compile the two series from station daily or hourly flow data: the usual historical summaries of annual maxima cannot be used, as they only list one maximum each year. Difficulties may arise in clearly associating certain peaks with either subseries – for example, if the division is on the basis of snow versus rain, some events can involve both.

Examples of dual-population treatment are given by Stoddart and Watt (1970) for southern Ontario and Waylen and Woo (1982) for southern British Columbia. In both studies, two annual maximum series were extracted and fitted to EV1 distributions. Assuming the two underlying populations to be independent, a combined distribution curve was then derived and compared to a plot of the original single annual maximum series. Figure 5.8 shows a set of plots from the British Columbia study.

5.3 *Regional Flood Frequency Analysis*

5.3.1 General Comments

As indicated in Section 5.1.3, the results of a regional analysis may be required for reliable estimation of x_T at a specific site when the records of appropriate gauging stations are short or nonexistent. Although regional analysis can take different forms, there are certain common elements in the procedure.

i) The regional boundaries are defined and gauging stations are screened for acceptable records.
ii) A single-station frequency analysis is carried out for acceptable stations within the region.
iii) Relations are postulated between certain flood statistics and the physiographic and climatic characteristics of the drainage basins.
iv) Values of required physiographic and climatic characteristics are determined for each drainage basin.
v) The relations postulated in (iii) are cast in the form of prediction equations, the coefficients of which are determined using multiple regression or other statistical techniques.

The flood statistics (e.g. quantiles, parameters, moments) are treated as dependent variables and the basin physiographic and climatic characteristics as independent or predictor variables.

Regional flood frequency analyses have been carried out for many regions of Canada. Some of these are referred to in Chapter 3 or in Section 5.3.6. The following sections, 5.3.2 to 5.3.5, provide guidance for one undertaking a regional frequency analysis where required, and indicate the limitations to recognize when existing regional analyses are used.

5.3.2 Data Base

Hydrometric data

The first step in a regional flood frequency analysis is to abstract an annual maximum series for those hydrometric stations (both active and discontinued) with records of 10 years or more. The number of stations affected by storage or other artificial factors which tend to modify flood flows significantly should be kept to a minimum. Canals, ditches, and drains in which flows are subject to substantial control by man should be excluded. If in doubt regarding the impact of storage on floods, one should seek advice from the data collection agency.

Gauged stations on the same stream which differ in area by less than 25 per cent cannot normally be considered as independent. If the records cover more or less the same period, the longer or better one should be used. If the records are for different periods they may be combined into one longer record, with adjustments as appropriate for differences in area (see Section 5.2.6).

Physiographic and climatic data

Geophysical factors such as physiography, geology, topography, and climate have a major influence on flood flows. A study of these characteristics leads to the identification of homogeneous zones: for example, zones can be identified by a flat or mountainous relief, or by similar geology.

A necessary condition for including a particular physiographic or climatic measure in the set of predictor variables for a regression model is that it be justifiable on physical grounds. For example, large basins receive a larger volume of precipitation than small basins and, all other things equal, would be expected to have larger flood flows. Therefore, the inclusion of basin area A can be justified on physical grounds. Similarly, lakes and swamps, as measured by basin surface storage St, tend to attenuate flood peaks. Other physically justifiable physiographic characteristics may include main channel slope S, main channel length L, mean basin elevation E, drainage density DD, basin shape factor SF, soil, and cover type. However, if a strong correlation subsequently emerges between two physiographic characteristics (e.g. A and L) both of these should not be used as independent predictor variables in the final regression model.

The climate of an area is a major consideration. For climatic data, the mean annual precipitation P is often used as a predictor variable.

Not all physically justifiable predictor variables will necessarily be statistically significant in a particular regression model. In the comparison of large basins, surface area is responsible for 90 to 95 per cent of the variance in flood flows and it is often difficult to find a second predictor variable that leads to a significant further reduction in variance.

Definition of region

The definition of a region depends on the method of regional analysis used. For example, if the sample size is large enough and predictor variables could be introduced for all those factors which are significant in affecting the T-year flood, then all of Canada would be one region if flood quantiles were regressed directly on those predictor variables. On the other hand, if the sample size is small, the number of independent variables is necessarily small, and if a method like the index flood method (see Section 5.3.3) is used, then a region must be selected that is relatively homogeneous. The test of homogeneity is dictated by the method and is based on summary statistics for gauged stations in the region. Although physically appealing, a consideration of geographic and meteorological characteristics is not essential to the definition of a region. Nevertheless, the usual first step is to review the data from hydrometric stations in an area with relatively homogeneous geographic, hydrological, and climatic characteristics.

In past Canadian studies, at least one boundary of the adopted region has coincided with a provincial or national boundary. This practice is seldom physically defensible; in most cases, neither the moisture generation process nor the basin response function undergoes drastic change at political boundaries.

Three methods of regional frequency analysis are discussed in Sections 5.3.3, 5.3.4, and 5.3.5.

5.3.3 Index Flood Method (IF)

The index flood method involves development of a regression equation expressing the 'index' flood – usually either mean annual or 2-year – in terms of independent physiographic and/or climatic variables, and a dimensionless frequency curve that relates flood quantiles at any point to the index flood. It therefore implies that within a region or subregion, all frequency curves can be approximated with the same shape.

In the original method as developed by the United States Geological Survey (Dalrymple 1960), a supposedly homogeneous region is first identified on the basis of physiography and climate, and station frequency curves are then prepared for all stations with acceptable lengths and quality of record. A special type of homogeneity test (Gumbel 1958) is then applied to eliminate nonconforming stations and, if they are geographically clustered, to adjust regional boundaries accordingly. The homogeneity test is based on the assumption of an EV1 (Gumbel) distribution, and involves the following procedural steps.

i) The ratio of 10-year to mean annual flood is computed for all stations and averaged over the region.
ii) Each station mean annual flood is multiplied by the regional average to yield a regionally based estimate of each station's 10-year flood. The station-based return period T_{10} of this discharge estimate is then read from the station frequency curve.
iii) All the station-based return periods are plotted against station length of record on a chart like Figure 5.14 on which upper and lower limit curves are shown. Stations falling outside those limits are rejected.

The basis for the upper and lower limit curves is that they represent plus and minus two standard errors (approximately 95 per cent confidence band) of the expected value, supposing the differences between stations to be due to sampling error. Numerical data for plotting the limit curves, assuming the EV1 distribution, are shown in Table 5.13. Corresponding data based on the 3LN distribution (Condie 1975) are shown in Table 5.14.

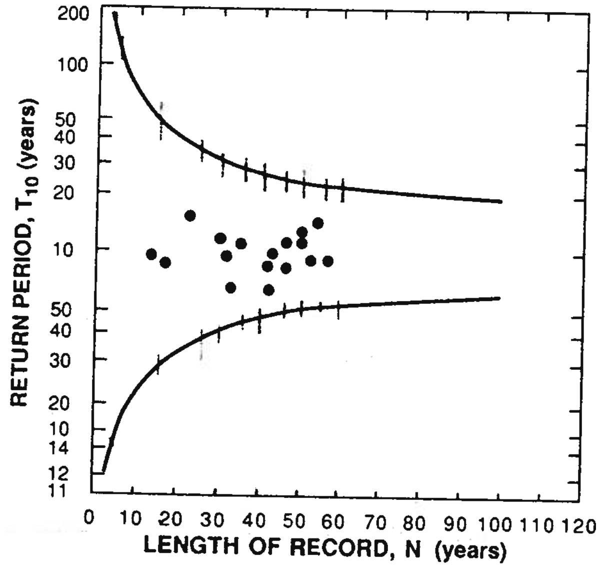

Figure 5.14 Homogeneity test for region D of Quebec (after Desforges and Tremblay 1974)

An example of the EV1 homogeneity test applied to all stations in region D of Quebec (south shore of the St. Lawrence River) is presented in Figure 5.14.

Data from all stations that pass the regional homogeneity test are used to develop a dimensionless regional frequency curve giving an average (median) value of the ratio $x_T/\bar{x}$. Such a curve for region D of Quebec is given in Figure 5.15. Then a regression equation relating $\bar{x}$ to basin characteristics such as A, S, S_t, and so on is developed. In many cases, the drainage area A is found to account for 90 to 95 per cent of the variance in $\bar{x}$. If no other variable is found to yield a statistically significant improvement, the final regression equation takes the form

Table 5.13 Limits on T_{10} for EV1 homogeneity test

Record Length N (years)	*Lower limit* T_L (years)	*Upper limit* T_U (years)
10	1.8	70
20	2.8	40
50	4.4	24
100	5.6	18

Table 5.14 Limits on T_{10} for 3LN homogeneity test

Record Length N (years)	*Lower limit* T_L (years)	*Upper limit* T_U (years)
10	3.0	60
15	3.6	42
20	4.0	33
25	4.4	29
30	4.7	26
40	5.1	23
50	5.4	21
60	5.7	19
100	6.4	17

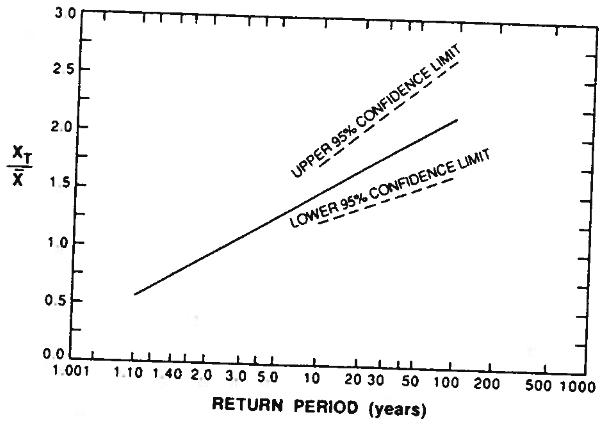

Figure 5.15 Dimensionless frequency curve for region D of Quebec (after Desforges and Tremblay 1974)

$$\bar{x} = KA^n \qquad [5.18]$$

where $\bar{x}$ is the mean annual flood in m³/s; A is drainage area in km², and K and n are the regional coefficient and exponent. For example, for region D of Quebec, $K = 0.446$ and $n = 0.904$.

Although drainage area is normally the dominant predictor variable, other physiographic variables (e.g. river slope, length of main stream, lake and swamp area) should be considered for smaller basins.

Example

Determine 20-year and 100-year floods at an ungauged site in region D of Quebec with a drainage area of 700 km².

i) The mean annual flow can be determined using [5.18] and the values given above for K and n,

$$\bar{x} = 0.446\ (700)^{0.904} = 166 \text{ m}^3/\text{s}$$

ii) The 20-year flood (using Figure 5.15) is

$$x_{20} = 1.69 \times 166 = 281 \text{ m}^3/\text{s}$$

iii) The 100-year flood (using Figure 5.15) is

$$x_{100} = 2.17 \times 166 = 361 \text{ m}^3/\text{s}$$

The index flood method suffers from several limitations, particularly the assumption of a single shape or slope of frequency curve within the region, which is particularly untenable when there is a wide range of drainage areas and where storage effects vary widely throughout the region. Further, the method is not applicable where there is significant flow control (see also comments under Section 5.3.4).

Notwithstanding these limitations, the majority of past Canadian regional flood frequency analyses listed in Table 5.15 are of the index flood type. The method of direct regression for quantiles, described in Section 5.3.4, appears to be conceptually an improvement on the index flood method.

5.3.4 Method of Direct Regression for Quantiles (DRQ)

Within a given region, the flood flow quantiles can be assumed to depend on basin physiographic, geomorphologic, and climatic characteristics. The direct regression method employs multiple regression equations (normally of power form) to relate any quantile (not just $\bar{x}$ or x_2 as in the index flood method) to basin characteristics, that is

$$x_T = K\,A^a\,S^b\,L^c\,E^dSt^eF^f \ldots P^g \ldots \qquad [5.19]$$

where $K, a, b, c, d, e, f \ldots g$ are determined by multiple regression analysis of x_T on A, S, and so on for gauged basins in the region. The multiple regression analysis is normally conducted as a linear analysis using log-transformed data.

Limitations

The regression equations should be based on data representing natural flood conditions. The method is not appropriate where manmade storage or structures have modified the flow appreciably, such as sites downstream from large storage reservoirs. In general, the derived equations should not be applied at any site where flow from 10 per cent or more of the drainage basin is controlled, or where the basin characteristics are out of the range used for developing the regional equations.

5.3.5 Method of Regression for Distribution Parameters (RDP)

It is often possible to adopt a standardized distribution that fits reasonably the flood series at all gauging stations

Table 5.15 Regional flood frequency analyses

Region	Method IF	DRQ	RDP	Reference
Newfoundland	X			Poulin (1971)
Newfoundland		X		Environment Canada and Newfoundland Environment (1984)
New Brunswick – Gaspe	X			Collier and Nix (1967)
New Brunswick	X	X		Acres Consulting Services (1977)
New Brunswick		X		Environment Canada/New Brunswick Municipal Affairs and Environment (1987)
Nova Scotia	X			Coulson (1967)
Nova Scotia		X		MacLaren Atlantic (1980)
Quebec	X			Desforges and Tremblay (1974) Hoang (1977)
Ontario		X		Karuks (1964)
Ontario			X	Stoddart and Watt (1970) Watt and Stoddart (1971)
Ontario	X			Sangal and Kallio (1977)
Ontario		X		Cumming-Cockburn (1985)
Ontario	X			Moin and Shaw (1985)
Ontario		X		Moin and Shaw (1986)
Ontario	X			Ontario MTC (1986)
Manitoba	X			Peterdy (1963)
Prairies	X			Durrant and Blackwell (1959,1961)
Prairies	X			Aaston (1986)
British Columbia	X			Reksten (1987)
British Columbia		X		Leith (1975)
Yukon		X		Kreuder and Leith (1978)
Yukon		X		Janowicz (1986)

within a region. It can then be hypothesized that such a distribution applies not only to gauged rivers but to all sites in the region, and that only the values of the distribution parameters will change. For each gauging station, the mean, standard deviation, and skew coefficient are calculated from sample data; regression equations in the form of [5.19] are then developed between these statistics and basin characteristics. For ungauged sites the magnitude of any specified quantile x_T can be estimated using

$$x_T = \bar{x} + K_T s \qquad [5.8]$$

where $\bar{x}$ and s are estimated using the regression equations and K_T is selected from Table 5.4.

This method is attractive in theory. From a practical point of view, however, links made by multiple regression between the standard deviation s or skew coefficient g and the physiographic characteristics are not always statistically significant. It may therefore be difficult to predict those variables for ungauged sites with an acceptable degree of reliability.

5.3.6 Availability of Regional Analyses

References are given in Table 5.15 for a number of regional flood frequency analyses in various parts of Canada. Some of these are also referred to in Chapter 3. For specific projects, reference should be made to the appropriate federal and provincial agencies for the latest information.

5.4 *Combination of Single Site and Regional Data*

5.4.1 Statement of the Problem

As indicated in Section 5.1, design flood estimates for streams that have no data whatsoever should be based on information obtained from the regional analyses outlined in Section 5.3. When there are some flow data available on the stream in question, they should ideally be used at least as partial information. The problem is that, by itself, a short period of record provides an extremely unreliable base for estimating floods of longer return periods. To illustrate the point, Table 5.16 shows means, standard deviations, and 50-year floods, estimated for the Similkameen River in British Columbia from five separate 10-year periods and also from the entire period of record. For both the Gumbel and lognormal distributions, estimates of the 50-year flood based on the 10-year periods range from about 70 to 130 per cent of the values based on the entire 50 years of record.

It is therefore clear that when there are some flow data available for a stream, but not enough for a reliable single-station frequency analysis, design flood estimates should ideally be based on a combination of these data with information from a regional analysis.

Unfortunately, it is often difficult to integrate these two types of information. The usual tendency is to place most weight on the actual data, however short the period of record, and to check the extrapolated predictions against