6 Statistical treatment of floods

A flood is an unusually high stage in a river that can cause damage to adjacent areas. Floods vary spatially and temporally in magnitude and are often measured through their peak discharges. The structural and hydraulic designs of dams and bridges are based on such extreme flows in water courses. Furthermore, the frequency of occurrence, the maximum stage reached, the volume of flood water, the area inundated and the duration of floods are of importance to the civil engineer when planning and designing roads, buildings and structures. In addition, there are dependent economic problems such as flood-plain zoning and flood insurance.

The peak flow and hydrograph of a flood are controlled by many complex and interrelated factors. In the first place, the amount, intensity and areal extent of the causative storm, antecedent precipitation, accumulated snow, temperature and vegetation are significant climatic (or climatically affected) factors. Secondly, physiographical properties such as the size, shape, slopes and orientation of the catchment, especially in relation to storm movements and isohyetal lines, channel and flood-plain storage and soil composition exert a large influence. In addition, man-made or natural changes in catchment characteristics and hydraulic parameters of flow cause further complications.

On account of these complexities, hydrologists have had to resort to statistical methods. The main objective in this approach is to estimate the magnitudes that are exceeded with specific probabilities. This chapter is for the purpose of describing and critically examining the methods of flood estimation. These include the use of type I (Gumbel), II and III extreme value, lognormal, Pearson type III, log Pearson type III, binomial, Poisson and multinomial distributions. In addition, the peaks-over-threshold method is explained. Empirical methods of regional flood frequency analysis and fundamentals of the probable maximum flood technique are also described[1].

[1] Auxiliary treatment through unit hydrograph theory for calculating flood volumes, durations and the like are outside the scope of this text and reference may be made, for example, to Wilson (1974) and to the Natural Environmental Research Council (1975, vol. 1, chapters 5, 6).

N. T. Kottegoda, *Stochastic Water Resources Technology*

6.1 Annual maximum series and return periods

Although floods are high flows which occur at varying intervals of time, it facilitates the analysis to study flood events within constant intervals of time. Also, it is preferable to choose an interval of 1 year rather than, say, a period of 3 months which brings in additional complications because of the seasonal effect.

Let the random variable X_i denote the maximum instantaneous flow in year i, $i = 1, 2, 3, \ldots$, at a gauging station. These X_i values are said to constitute an annual maximum series. In practice, an observed sequence x_i, $i = 1, 2, 3, \ldots N$, is used to make probabilistic estimates of flood flows. Quite often the annual maxima are taken from discrete daily mean flows, and except for flashy rivers (which rise and fall rapidly at times of flood) these have an approximate linear relationship with the instantaneous peaks[2].

Initially, the treatment will be confined to annual maximum series. The alternative approach is to analyse a so-called partial duration series. This pertains to peak flows that exceed a given threshold value and is the subject of section 6.9.

In all statistical flood studies, a particularly important concept is that of the return period T. This is associated with a fixed magnitude of flood discharge called the T-year flood and is, in fact, the average time interval between exceedances of that magnitude[3]. As will be defined in subsection 6.3.3, T is the reciprocal of probability with which a variate exceeds the given magnitude. This can also be explained in terms of percentiles, the method of describing distributions by identifying particular points on the distribution function; for example, the 10-year flood is the ninetieth percentile of the distribution of annual floods. Note that there will be some T-year periods in which the T-year flood will be exceeded more than once and other such periods in which the highest flood is less than the T-year flood[4].

6.2 Distribution of extreme values

Extreme value theory, which is used in flood estimation, dates back to Fréchet (1927) and to Fisher and Tippett (1928). To explain the fundamentals, consider a set of independent random variables $W_j, j = 1, 2, 3, \ldots, n$, with a common cumulative distribution function $G(x)$, where x is an observed value and n is the

[2] See, for example, Gumbel (1958a).

[3] The term return period was originally used by Fuller (1914) who was also the first to apply frequency methods in flood estimation.

[4] Specifically, if 10 000 years of data are available, there will be, on average, no flood in excess of the 100-year flood in about 37 of the 100 centuries. Also, the expectation is that in each of about 37 other centuries there will be one such flood, and in the remaining period two or more such floods would occur in each century. It is assumed here that the flood peaks are mutually independent; the calculations are based on the Poisson distribution explained in section 6.9.

number of equispaced data points within a fixed period of, say, 1 year. Also, let $\{W_{(1)}, W_{(2)}, W_{(3)}, \ldots, W_{(n)}\}$ represent the ordered set of the same variables in which $W_{(1)}$ is the smallest and $W_{(n)}$ is the largest. The distribution of $W_{(n)}$ is given by

$$\begin{aligned}\Pr\{W_{(n)} \leqslant x\} &= Q_n(x) \\ &= \Pr(W_1 \leqslant x)\Pr(W_2 \leqslant x)\Pr(W_3 \leqslant x)\ldots\Pr(W_n \leqslant x) \\ &= \{G(x)\}^n \end{aligned} \tag{6.1}$$

in which Pr denotes probability.

As n increases indefinitely, $Q_n(x)$ approaches one of three asymptotic types known as the types I, II and III extreme value distributions. In the first type, X is an unbounded variable; the second and third types deal with variables with lower and upper limits respectively. Because the type I distribution was extensively developed and applied to flood events by Gumbel (1958a), it is often referred to as the Gumbel distribution[5]. A simplified derivation of this now follows.

6.3 Gumbel distribution

The Gumbel distribution of extreme values results from any initial distribution of the exponential type. Examples of these are the normal and gamma distributions; the right tails of their density functions converge to the exponential form for large values of the variable. Accordingly, $g(x)$, which is the derivative of $G(x)$ in equation 6.1, can be approximated to the form $\lambda e^{-\lambda x}$. This leads to a probability of non-exceedance given by $G(x) = 1 - e^{-\lambda x}$. Therefore, from equation 6.1

$$\begin{aligned} Q_n(x) &= \Pr\{W_{(n)} \leqslant x\} \\ &= (1 - e^{-\lambda x})^n \end{aligned} \tag{6.2}$$

By changing the location and scale, equation 6.2 can be written as $Q_n(x) = [1 - \exp\{-\alpha(x-u)\}/n]^n$, where u and α are the location and dispersion parameters respectively. Hence,

$$\begin{aligned} \lim_{n\to\infty}\{Q_n(x)\} &= F(x) \\ &= \exp\{-e^{-\alpha(x-u)}\} \end{aligned} \tag{6.3}$$

which is the Gumbel (double-exponential) distribution function[6]. Originally, Fisher and Tippett (1928) using a functional relation derived the general form of

[5] 'It seems that the rivers know the theory,' said Gumbel (1967) in what was to be his last address to engineering hydrologists, 'It remains to convince the engineers—not only in underdeveloped countries—of the validity of this analysis.' Court (1952) gives a simple explanation of the Gumbel procedure as originally formulated.

[6] For a rigorous mathematical proof, see Gumbel (1958a, pp. 156–9), Kendall and Stuart (1977, pp. 352–6) or Bury (1975, pp. 369–71). The particular limit theorem is proved in subsection 6.8.2.

equation 6.3 and the other two types of extreme value distributions included in section 6.4.

If we follow the notation of Gumbel (1958a) and substitute the value y of a reduced (that is, dimensionless) random variate Y where

$$y = \alpha(x - u) \tag{6.4}$$

the basic form of the Gumbel distribution becomes

$$F(y) = \exp(-e^{-y}) \tag{6.5}$$

the density function of which is

$$\begin{aligned} f(y) &= dF(y)/dy \\ &= e^{-y}\exp(-e^{-y}) \end{aligned} \tag{6.6}$$

6.3.1 *Moment-generating function*

The moments of a function such as the Gumbel distribution can be obtained through its moment-generating function (MGF). For a random variate Y with moments of all orders (as explained in section 3.3) and a probability density function $f(y)$, the MGF is defined as

$$\begin{aligned} M_Y(t) &= E(e^{tY}) \\ &= \int_{-\infty}^{+\infty} e^{ty} f(y)\,dy \end{aligned} \tag{6.7}$$

where E denotes expectation (that is, expected value) and t is a dummy variable[7]. From the series expansion of e^{tY},

$$M_Y(t) = E\{1 + tY + (tY)^2/2! + (tY)^3/3! + \ldots\} \tag{6.8}$$

6.3.2 *Statistical properties*

By substituting $z = e^{-y}$, which makes $dz/dy = -e^{-y}$, it follows from equations 6.6 and 6.7 that for the Gumbel distribution

$$M_Y(t) = \int_0^{\infty} z^{-t} e^{-z}\,dz$$

Replacing $-t$ in the right-hand side by $(1-t)-1$ it is seen, from the standard form of the gamma function given in section 3.2, that

$$M_Y(t) = \Gamma(1-t) \tag{6.9}$$

[7] Note that in the more advanced books the characteristic function $E(e^{tYt})$, which is the expectation of a complex function, is used in place of the MGF.

From equations 6.8 and 6.9, therefore, the rth moment of the Y population is given by

$$\mu_r'(Y) = \frac{d^r}{dt^r}\Gamma(1-t)\bigg|_{t=0} \tag{6.10}$$

for the Gumbel distribution. It follows, by taking the derivative of $M_Y(t)$ and by putting $t = 0$, that

$$\begin{aligned}\mu_1'(Y) &= \frac{d}{dt}\Gamma(1-t)\bigg|_{t=0} \\ &= -\Gamma(1-t)\psi(1-t)\bigg|_{t=0} \\ &= -\psi(1)\end{aligned} \tag{6.11}$$

where $\psi(t) = d[\ln\{\Gamma(t)\}]/dt$ is the psi function in subsection 3.4.1; this is also called the digamma function. From tables, $\psi(1) = -0.5772$, which is known as Euler's constant[8]. Hence,

$$E(y) = \mu_1'(Y) = 0.5772$$

where, as noted, y is a reduced variate. Also,

$$\mu_2'(Y) = \Gamma(1-t)\left[\frac{d\psi(1-t)}{dt} + \{\psi(1-t)\}^2\right]\bigg|_{t=0} \tag{6.12}$$

where $d\psi(t)/dt = \psi'(t)$ is the trigamma function and is tabulated in some books[9]. Now $\psi'(1) = \pi^2/6$. Hence, $\mu_2'(Y) = \pi^2/6 + (0.5772)^2$ and var $(Y) = \mu_2(Y) = \mu_2'(Y) - \{\mu_1'(Y)\}^2 = \pi^2/6$, which approximates to 1.6449. Similarly, by using the multigamma functions, $\psi''(1)$ and $\psi^{(3)}(1)$, it can be shown that the coefficient of skewness is 1.1396 and the coefficient of kurtosis is 5.4000.

Because the skewness and higher coefficients are the same for both Y and X populations, the only changes that arise in fitting are those of location and dispersion as given by equation 6.4. The maxima and minima of the density function $f(y)$ are found from the derivative of $f(y)$ in equation 6.6 and these are at $y = \infty$, $-\infty$ and 0, the first two of which are the minimum points and the third is a maximum point (showing that the mode is above the origin for the variate Y). The median value, obtained by setting $F(y)$ in equation 6.5 to 0.5, is 0.3665. Also, the maximum ordinate which occurs at $y = 0$ is $1/e = 0.3679$, which is a constant for both the X and Y populations. These properties and the shape of the Gumbel density function are shown in figure 6.1, in which comparison is made with the normal density function.

[8] See, for example, Abramowitz and Stegun (1964, pp. 267–71).
[9] See, for example, Tribus (1969, p. 112).

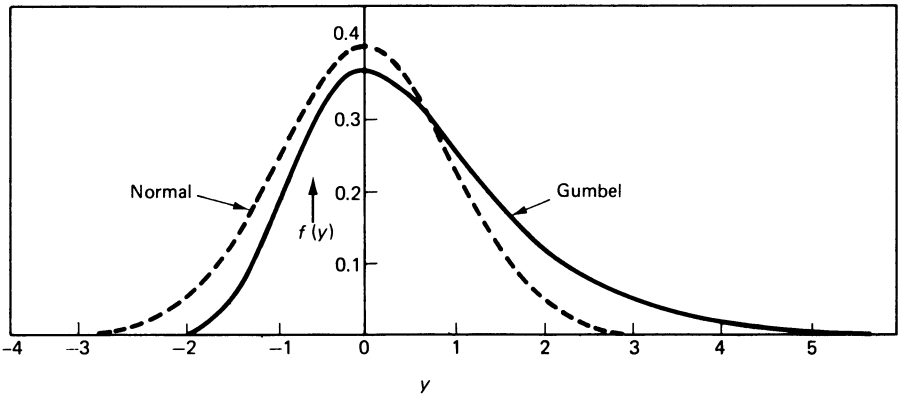


Figure 6.1 Gumbel and normal density functions compared

6.3.3 *Definition of return period*

Let the reduced random variates Y_i denote the maximum floods in years i, $i = 1, 2, 3, \ldots$, respectively. Then, if the Y_i values are serially independent, the probability that the time interval $\tilde{T}$ between exceedances of a flood magnitude y (in reduced units) equals n is given by

$$\begin{aligned}\Pr(\tilde{T} = n) &= \Pr(Y_1 < y)\Pr(Y_2 < y) \ldots \Pr(Y_{n-1} < y)\Pr(Y_n > y) \\ &= \{\Pr(Y < y)\}^{n-1}\Pr(Y > y)\end{aligned}$$

For the second equality we assume that the Y_i values are identically distributed. This geometric distribution corresponds to equation 5.6. Here, the variable n can take any value from 1 to ∞, and the expected value $E(\tilde{T})$ is found from the properties of the geometric distribution as follows.

$$\begin{aligned}E(\tilde{T}) &= \sum_{n=1}^{\infty} n\Pr(\tilde{T} = n) \\ &= \sum_{n=1}^{\infty} n\{1 - \Pr(Y > y)\}^{n-1}\Pr(Y > y) \\ &= 1/\Pr(Y > y)\end{aligned}$$

The return period T is commonly written instead of $E(\tilde{T})$ above[10].

[10] Lloyd (1970) has shown that the variance of T is $\{1 - \Pr(Y > y)\}/\{\Pr(Y > y)\}^2$. If the flood events are serially correlated, the variance is greater; the return period is then given by $T = 1/\Pr(Y_n > y \mid Y_{n-1} < y)$.

6.3.4 *Relationship between Gumbel variate and return period*

From equation 6.5 and subsection 6.3.3

$$1-\exp(-e^{-y}) = 1/T \tag{6.13}$$

Hence,

$$y = -\ln\{\ln(T)-\ln(T-1)\} \tag{6.14}$$

Now

$$\begin{aligned} y &= -\ln\{-\ln(1-1/T)\} \\ &= -\ln\{-(-1/T)+\tfrac{1}{2}(-1/T)^2-\tfrac{1}{3}(-1/T)^3+\ldots\} \end{aligned}$$

from Maclaurin's theorem. If only three terms are used in the series expansion,

$$y \approx -\ln(1/T+1/2T^2+1/3T^3) = \ln\{6T^3/(6T^2+3T+2)\}$$

Hence, the approximations

$$y \approx \ln(T-1/2) \qquad \text{or} \qquad y \approx \ln(T) \tag{6.15}$$

may be used in place of equation 6.14 for $T > 10$ years if errors up to 0.5 % and 2.5 % respectively can be tolerated; the second approximation is sufficient for all practical purposes when $T > 25$ years.

Equation 6.14 gives a non-linear relationship between the value y of the reduced variate and the return period T. A few pairs of values (y, T) are as follows: 0, 1.58 (most probable flood); 0.3665, 2 (median flood); 0.5772, 2.33 (mean flood); 1.2459, 4.00; 3.9019, 50.00; 4.6002, 100.00; 6.2136, 500.00. Engineers and hydrologists have been plotting experimental data on special types of paper since Hazen (1914), a civil engineer, originated the graphical linearisation of the normal distribution. The method is used as a verification of the suitability of one or more assumed distributions for a given sample of data.

6.3.5 *Probability paper*

Gumbel probability paper (suggested originally by Powell (1943)) may be drawn as follows. Initially, values of the return period T, such as 1.01, 1.1, 1.2, 1.3, 1.5, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50, 60, 100, 200 and 250, are selected. After computing the corresponding values of the reduced variate y by using equation 6.14, vertical lines spaced at distances directly proportional to the differences between the y values are drawn, and the return periods T are shown correspondingly against these lines (figure 6.2). In this way the y values are on a linear scale, but the return periods T are on a double-exponential scale, as given by equation 6.13. On the other hand, if a graph of x against y is drawn on arithmetic paper, a straight-line plot

$$x = y/\alpha + u \tag{6.16}$$

which follows from equation 6.4, will give the flood magnitudes x for various y

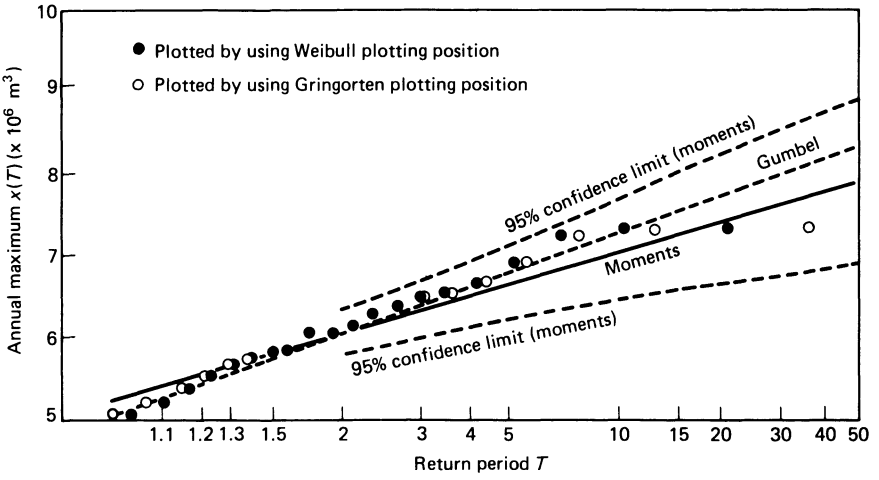


Figure 6.2 Gumbel distribution fitted to annual maxima from daily inflows to Caban Coch Reservoir for the period 1909 to 1928

values. However, it is more meaningful to relate x to the return periods T rather than to the y values. Therefore, the scales are chosen accordingly, and from equations 6.14 and 6.16 the relationship of x against T is given by

$$x(T) = u - \ln\{\ln(T) - \ln(T-1)\}/\alpha \tag{6.17}$$

This shows that a straight line could be fitted if the observed values are plotted on Gumbel paper. It provides a quick verification of fit without using the goodness-of-fit tests described in chapter 3; a good fit would justify the acceptance of the Gumbel distribution. When plotting the data, however, the true return periods associated with each of the items of data are not known. Therefore, the accepted practice is to use what is termed a plotting position[11].

6.3.6 *Plotting positions*

Let m denote the rank of N items of annual maxima which are ordered so that the first value ($m = 1$) is the largest and the smallest ($m = N$) is placed last[12]. One possible method, first applied to flood flows in California, is to take m/N as the probability of exceedance. Accordingly, $1/T(= 1 - F(x)) = m/N$, but the smallest has a probability of exceedance equal to 1, which is not found on probability paper. An alternative plotting position is to make $T = N/(m-1)$, but the drawback is that the largest flood cannot be plotted because it has a

[11] Gumbel (1958a, pp. 32–6) gives conditions for the choice of a plotting position; for example, it should be possible to plot all the observations. Also, the plotting position ought to depend on the assumed distribution.

[12] For plotting purposes, it is convenient to reverse the conventional method of ordering as defined in section 6.2.

return period of infinity, or in other words a probability of zero. Hazen (1930), a pioneer in flood studies, suggested the alternative $T = N/(m - 1/2)$ in order to plot all the data. This was objected to by Gumbel (1958b) on the grounds that the highest flood has a return period of $2N$. Instead, he recommended the use of the Weibull plotting position $T = (N + 1)/m$ for the Gumbel distribution. Subsequently, Gringorten (1963) showed that $T = (N + 0.012)/(m - 0.44)$ is a better approximation as an unbiased plotting position for this distribution. To explain the meaning of bias in this context, consider a sample of N annual maxima, from which ℓ subsamples each of length N/ℓ (with $N/\ell > 30$, say) are formed and in which the values in each subsample are ranked in order. Then, an unbiased plotting position is such that, if average values of the same rank from different samples are plotted and if ℓ is indefinitely large, these values will lie on a line which represents the distribution of the population[13]. Further comments on plotting positions from the viewpoint of the practising engineer will be given in section 6.7.

Four of the commonly used plotting positions are given in table 6.1. Also shown are the return intervals for the largest, second largest and smallest flood from a sample of 100 items.

Table 6.1 Plotting positions

Plotting position	*Usage*	*For N = 100*		
		T for m = 1 (100-year flood)	*T for m = 2 (50-year flood)*	*T for m = 100 (1.01-year flood)*
Weibull, $T = (N+1)/m$	Used by Gumbel	101	50.5	1.01
Gringorten, $T = (N + 0.12)/(m - 0.44)$	Extreme value distributions	179	64	1.01
Hazen, $T = N/(m - \frac{1}{2})$	Gamma distribution	200	66.7	1.01
Blom, $T = (N + \frac{1}{4})/(m - \frac{3}{8})$	Normal (and lognormal) distributions	160.4	61.7	1.01

6.3.7 *Method-of-moments fitting procedure*

From the relationship $Y = \alpha(X - u)$ in equation 6.4,

$$E(y) = \alpha\{E(X) - u\} \tag{6.18}$$

[13] For discussions on plotting positions, see Benson (1962b), Stipp and Young (1971) and the Natural Environmental Research Council (1975, vol. 1, chapter 2); also, Langbein (1960) gives simple derivations for some plotting positions; the treatment by Kimball (1960) and Barnett (1975) is more sophisticated.

and

$$\text{var}(Y) = \alpha^2 \text{var}(X) \tag{6.19}$$

where var denotes variance. Therefore if we substitute 0.5772 for $E(y)$ and $\pi^2/6$ for var(y) (from the results previously obtained in subsection 6.3.2) and sample estimators $\bar{x}$ and s^2, for $E(X)$ and var(X) respectively, the moment estimators of the parameters α and u are

$$\tilde{\alpha} = 1.282/s \tag{6.20}$$

and

$$\tilde{u} = \bar{x} - 0.45s \tag{6.21}$$

Hence, from equations 6.17, 6.20 and 6.21,

$$\tilde{x}(T) = \bar{x} - s[0.4500 + 0.7797 \ln\{\ln(T) - \ln(T-1)\}] \tag{6.22}$$

Using this formula a theoretical straight line could be drawn on the Gumbel graph paper of $x(T)$ against T to represent a sample of annual maxima $x_1, x_2, x_3, \ldots, x_N$ with estimated mean $\bar{x}$ and standard deviation s respectively. Two points would obviously suffice to define this line, and the Gringorten plotting position can be used to represent the data. This is, of course, on the assumption, which is not necessarily true, that the data are distributed in this way.

Example 6.1 Ranked annual maxima from mean daily inflows to Caban Coch Reservoir during the period 1909 to 1928 are as follows.

Annual maxima ($\times 10^6$ m^3)									
7.31	7.30	7.22	6.90	6.64	6.53	6.48	6.38	6.30	6.12
6.07	6.06	5.82	5.81	5.75	5.65	5.51	5.37	5.20	5.08

Plot the data on Gumbel paper, and estimate the mean and standard deviation from the sample. Using these statistics, fit a straight line to the data. Estimate the 50-year flood.

Let x_i, $i = 1, 2, \ldots, N$, denote the maxima where $N = 20$ and let $\sum$ denote $\sum_{i=1}^{N}$. The estimated mean $\bar{x} = \sum x_i/N = 6.175 \times 10^6$ m^3 and $s = \{(\sum x_i^2/N - \bar{x}^2)N/(N-1)\}^{1/2} = 0.6746 \times 10^6$ m^3 is the estimated standard deviation[14].

Equations 6.20 and 6.21 provide estimates $\tilde{\alpha} = 1.9011$ and $\tilde{u} = 5.8714$ of the parameters, and the Gringorten plotting position $T = (N + 0.12)/(m - 0.44)$

[14] This formula is used by Gumbel (1941) to compute what he terms 'the observed standard deviation'. As a matter of interest, if the x_i values are normally distributed, then in order to obtain a strictly unbiased estimate of s, the quantity $K = 2\{\Gamma(N/2)\}^2/[\Gamma\{(N-1)/2\}]^2$, in which Γ denotes the complete gamma function, should replace the divisor $N-1$ in the formula; see, for example, Holtzman (1950).

gives the return periods as follows: 35.9, 12.9, 7.86, 5.65, 4.41, 3.62, 3.07, 2.66, 2.35, 2.10, 1.91, 1.74, 1.60, 1.48, 1.38, 1.29, 1.21, 1.15, 1.08 and 1.03. The ranked values are plotted, and a straight line is then fitted by using the method of moments. This is done by calculating two values from equation 6.22, such as $\hat{x}(2) = 6.06$ and $\hat{x}(50) = 7.92$. These two points define the (full) straight line of $x(T)$ against T.

Figure 6.2 shows that, if the two largest values are disregarded, the Gumbel distribution as fitted by the method of moments provides a good fit to the data. It is quite possible that this sample is biased downwards and the two highest values represent return periods less than 20 years. Unfortunately, there is no satisfactory method of verifying this, but we could, in the first instance, fit the Gumbel distribution by different methods and see whether there is a significant change in the results.

6.3.8 *Gumbel's fitting method*

In Gumbel's fitting method, the estimation of the parameters u and α are based on the Weibull plotting position $T = (N+1)/m$, where N is the sample size and m is the rank commencing with the largest value. If we use Gumbel's notation, the procedure is to evaluate the mean $\bar{y}_N$ and the standard deviation $\sigma_N = \{\Sigma(y - \bar{y}_N)^2/N\}^{1/2}$ of the N values of the reduced variate Y, after substituting each of the values $m = 1, 2, 3, \ldots, N$ in $T = (N+1)/m$ and then calculating the corresponding y values from equation 6.14. It will, of course, be more convenient here to refer to tables of $\bar{y}_N$ and σ_N such as those provided by Gumbel (1954, 1958a). Alternatively, refer to table 6.2, which is more accurate.

Table 6.2 Expected means $\bar{y}_N$ and standard deviations σ_N of Gumbel reduced variates

N	$\bar{y}$	σ_N	N	$\bar{y}_N$	σ_N	N	$\bar{y}_N$	σ_N
16	0.5154	1.0306	33	0.5388	1.1225	50	0.5485	1.1607
17	0.5177	1.0397	34	0.5396	1.1256	51	0.5489	1.1623
18	0.5198	1.0481	35	0.5403	1.1285	52	0.5493	1.1638
19	0.5217	1.0557	36	0.5411	1.1313	53	0.5497	1.1653
20	0.5236	1.0628	37	0.5417	1.1339	54	0.5501	1.1668
21	0.5252	1.0694	38	0.5424	1.1365	55	0.5504	1.1682
22	0.5268	1.0755	39	0.5430	1.1390	56	0.5508	1.1695
23	0.5282	1.0812	40	0.5436	1.1413	57	0.5511	1.1709
24	0.5296	1.0865	41	0.5442	1.1436	58	0.5515	1.1722
25	0.5309	1.0914	42	0.5448	1.1458	59	0.5518	1.1734
26	0.5321	1.0961	43	0.5453	1.1479	60	0.5521	1.1747
27	0.5332	1.1005	44	0.5458	1.1499	70	0.5548	1.1854
28	0.5343	1.1047	45	0.5463	1.1518	80	0.5569	1.1938
29	0.5353	1.1086	46	0.5468	1.1537	90	0.0586	1.2007
30	0.5362	1.1124	47	0.5472	1.5555	100	0.5600	1.2065
31	0.5371	1.1159	48	0.5477	1.1573	∞	0.5772	1.2825
32	0.5380	1.1193	49	0.5481	1.1590			

Therefore, substituting $\bar{y}_N$ for $E(Y)$ and σ_N^2 for $\text{var}(Y)$ and the sample estimators $\bar{x}$ and s^2 for $E(x)$ and $\text{var}(x)$ respectively in equation 6.18 and 6.19, we obtain the following.

$$\tilde{\alpha}' = \sigma_N/s \tag{6.23}$$

and

$$\tilde{u}' = \bar{x} - \bar{y}_N s/\sigma_N \tag{6.24}$$

Hence, from equations 6.17, 6.23 and 6.24,

$$\tilde{x}'(T) = \bar{x} - (s/\sigma_N)[\bar{y}_N + \ln\{\ln(T) - \ln(T-1)\}] \tag{6.25}$$

Example 6.2 Plot the extreme value data from example 6.1 using the Weibull plotting position. Hence, fit a straight line using Gumbel's fitting procedure, and estimate the 50-year flood.

The ordered return periods $T = (N+1)/m$ are as follows: 21, 10.5, 7, 5.25, 4.2, 3.5, 3, 2.625, 2.333, 2.1, 1.909, 1.75, 1.615, 1.5, 1.4, 1.313, 1.235, 1.167, 1.105, 1.05. As a matter of interest, the corresponding y values are calculated as follows: 3.02, 2.302, 1.87, 1.554, 1.302, 1.089, 0.903, 0.735, 0.581, 0.436, 0.298, 0.166, 0.0355, −0.0940, −0.2254, −0.361, −0.506, −0.666, −0.855, −1.113, from which $\bar{y}_N = 0.5236$ and $\sigma_N = 1.0628$; these tally with the values given in table 6.2. (Substituting the computed values from example 6.1 of $\bar{x}$ and s in equations 6.23 and 6.24, we find $\tilde{\alpha}' = 1.5754$ and $\tilde{u}' = 5.8426$.) From equation 6.25, $\tilde{x}'(2) = 6.08$ and $\tilde{x}'(50) = 8.32$. These two points define the broken line in figure 6.2. Notice that the differences between the plotted points are mainly in the high and low values.

6.3.9 *Frequency factors for Gumbel distribution*

Following Chow (1951, 1964), we can write equations 6.22 and 6.25 in the form

$$x(T) = \mu + K(T)\sigma \tag{6.26}$$

to represent the population of annual maxima. That is to say, an annual maximum with return period T is the sum of the mean and a constant $K(T)$ times the standard deviation of the maxima. The function $K(T)$, for a particular T, depends on the form of the density function of the maxima. It is clear from equation 6.22 that, for the Gumbel distribution and the method-of-moments procedure,

$$K(T) = -[0.4500 + 0.7797 \ln\{\ln(T) - \ln(T-1)\}] \tag{6.27}$$

Some values of $K(T)$ are given in table 6.3.

Table 6.3 Values of $K(T)$ using the method of moments

T	10000	1000	500	200	100	50	20	10	2.33	1.5
$K(T)$	6.73	4.94	4.39	3.68	3.14	2.59	1.87	1.30	0	−0.38

For Gumbel's method of fitting, it follows from equation 6.25 that

$$K(T) = -[\bar{y}_N + \ln\{\ln(T) - \ln(T-1)\}]/\sigma_N \tag{6.28}$$

For a given sample of size N, $\bar{y}_N$ and σ_N are known, and this formula can be used to calculate $K(T)$ through Gumbel's method for any value of T.

6.3.10 *Confidence limits*

It can be shown that the variance of the T-year flood estimated, from a sample of size N with an estimated variance s^2, by Gumbel's fitting method is

$$\begin{aligned}\text{var}\{\hat{x}'(T)\} = (s^2/N)[1 + 1.14W(T)\{(N-1)/N\}^{1/2} \\ + W(T)^2(1.1 - 0.6/N)]\end{aligned} \tag{6.29}$$

where $W(T) = \{\bar{y}_N - y(T)\}/\sigma_N$ in which $y(T)$ is the value of y obtained from equation 6.14 and $\bar{y}_N$ and σ_N are the mean and standard deviation respectively of the N y values[15].

Then, if we assume that the sampling distribution of the T-year flood is normal, the $100(1-\alpha)\%$ confidence limits of $\hat{x}(T)$ are

$$\hat{x}'(T) \pm z_{\alpha/2}[\text{var}\{x'(T)\}]^{1/2}$$

where $z_{\alpha/2}$ is the value which a standard normal deviate exceeds with probability $\alpha/2$. Strictly speaking, the normal distribution is applicable only when N is large, say, of the order of 200 or greater and if $\hat{x}'(T)$ behaves as an arithmetic mean.

For the method-of-moments fitting procedure, the variance of the estimated T-year flood is given by

$$\text{var}\{\hat{x}(T)\} = (s^2/N)[1 + 1.14K(T) + K(T)^2\{0.6 + 0.5N/(N-1)\}] \tag{6.30}$$

for which equation 6.27 gives the $K(T)$ function

Example 6.3 Using the data in example 6.1 and equation 6.30, calculate the 95% confidence limits of the population value of $\hat{x}(T)$, for $T = 2, 5, 10, 20, 30, 50$.

$$\begin{aligned} s^2/N &= 0.6746^2/20 \\ &= 0.02276 \\ \{0.6 + 0.5N/(N-1)\} &= 0.6 + 10/19 \\ &= 1.1263 \end{aligned}$$

The calculations are given in table 6.4.

[15] Lowery and Nash (1970) and Kaczmarek (1957) give derivations; see also the World Meteorological Organisation (1974, p. 5.26).

Table 6.4 95% confidence limits by moments method

T	$K(T)$	$\text{var}\{\hat{x}(T)\}$	$\hat{x}(T)\} = \bar{x} + K(T)s$	$1.96[\text{var}\{\hat{x}(T)\}]^{1/2}$	*Upper confidence limit*	*Lower confidence limit*
50	2.592	0.2622	7.92	1.00	8.92	6.92
30	2.189	0.2024	7.65	0.88	8.53	6.77
20	1.866	0.1604	7.44	0.79	8.23	6.65
10	1.305	0.1002	7.06	0.62	7.68	6.44
5	0.7195	0.0547	6.66	0.46	7.12	6.20
2	−0.1642	0.0192	6.06	0.27	6.33	5.79

6.3.11 *Maximum likelihood method of estimation*

The ML method of estimation is described in section 3.4. From equations 6.4, 6.5 and 6.6 the probability distribution and density functions of the Gumbel distribution are given by

$$F(x) = \exp\{-e^{-(x-u)\alpha}\}$$

and

$$f(x) = \alpha e^{-(x-u)\alpha} \exp\{-e^{-(x-u)\alpha}\}$$

respectively. For a sample x_i, $i = 1, 2, \ldots, N$, the log likelihood function, $L^*(x_1, x_2, \ldots, x_N \mid u, \alpha)$, which is conditional to the values u and α of the parameters, is given by

$$L^* = -\sum(x_i - u)\alpha - \sum e^{-(x_i-u)\alpha} + N \ln(\alpha) \tag{6.31}$$

where $\sum$ denotes $\sum_{i=1}^{N}$. The partial derivatives of equation 6.31 are

$$\partial L^*/\partial\alpha = -\sum(x_i - u) + \sum(x_i - u)e^{-(x_i-u)\alpha} + N/\alpha \tag{6.32}$$

and

$$\partial L^*/\partial u = N\alpha - \sum \alpha e^{-(x_i-u)\alpha} \tag{6.33}$$

The ML estimators $\hat{u}$ and $\hat{\alpha}$ of the parameters are obtained by setting $\partial L^*/\partial\alpha = 0$, $\partial L^*/\partial u = 0$. For the ML conditions therefore, from equation 6.33

$$\exp(\hat{u}\hat{\alpha}) = N \Big/ \sum \exp(-\hat{\alpha}x_i) \tag{6.34}$$

and, from equation 6.32 after substituting from equations 6.33 and 6.34 and simplifying,

$$1/\hat{\alpha} = \bar{x} - \sum \{x_i \exp(-\hat{\alpha}x_i)\} \Big/ \sum \exp(-\hat{\alpha}x_i) \tag{6.35}$$

where $\bar{x} = \sum x_i/N$. Also, from equation 6.34,

$$\hat{u} = -(1/\hat{\alpha})\ln\{1/N)\sum \exp(-\hat{\alpha}x_i)\} \tag{6.36}$$

A simple method for solving equations 6.35 and 6.36 is to estimate an initial value of α by the method of moments and then to substitute it in the right-hand side of equation 6.35; the reciprocal of equation 6.35 will give the next trial value. Therefore, the third value of α is made equal to the weighted average of the first and second, and equation 6.35 is used again to obtain a fourth value; here, the most recent value deserves a greater weight. The routine is repeated till equation 6.35 holds, and then, if $\hat{\alpha}$ is substituted in equation 6.36, $\hat{u}$ is found[16]. All these

[16] Gumbel (1958b, pp. 231–4) explains the methods of B. F. Kimball. Elsewhere, a procedure to find numerical solutions for $\hat{\alpha}$ and $\hat{u}$ is given by Jenkinson (1969, pp. 205–9), and this is followed by the Natural Environmental Research Council (1975, vol. 1, pp. 85–9). Also, Panchang (1969) shows how to obtain a solution iteratively.

trial-and-error methods could be easily implemented through a digital computer. However, a pocket calculator was used for the short sequence in the next example.

Example 6.4 Fit the Gumbel distribution by the ML method to the data given in example 6.1. Estimate the parameters and the 50-year flood.

The first trial value of parameter α is made equal to 1.9011, which is obtained from example 6.1. After substituting this in the right-hand side of equation 6.35, the second estimate of 1.6201 is obtained from the reciprocal. Another trial value of α, say, $(2 \times 1.6201 + 1.9011)/3 = 1.7138$ used in the right-hand side of equation 6.35 gives the next estimate of 1. 7351. Finally, $\hat{\alpha} = 1.726$ is thought to be sufficiently accurate (considering the data) and, from equation 6.36, $\hat{u} = 5.861$. Also, $\hat{x}(50) = 8.12$ from equation 6.17. This estimate of the 50-year flood is higher than the value obtained by the method of moments but is less than that from Gumbel's method.

In order to place confidence limits on the $\hat{x}(T)$ values found by the ML method, it can be assumed that these are asymptotically normally distributed. The standard error function which is required is somewhat complicated; the procedure is comparable with that used by Moran (1957).

Comments on the use of the ML method such as its dependence on an assumed distribution are given in section 3.4. Of course, any other method which assumes a particular probability model will give unsatisfactory results if the probability model is itself incorrect. This point is taken up again in section 6.7. There has been criticism by Gumbel (1967) that the ML method gives undue weight to the smaller values; although this may not be a fair criticism, it should not be forgotten that engineers looking for practical means of extrapolation tend to give more attention to the larger values in the data[17].

6.3.12 *Limitations in Gumbel method*

It should be noted that the limiting form of the extreme value distribution is reached extremely slowly. On theoretical considerations, the value of n in equation 6.1 should be extremely large, perhaps greater than 10^9, for the asymptotic form (equation 6.3) to hold. On the contrary, n is taken as 365 when applied to discrete daily series[18].

In applications it is found that there is a serial dependence and periodicity in the data, from which the extremes are drawn. This is generally true of daily and shorter-interval hydrological or meteorological data. However, Watson (1954) has shown that the limiting distribution will also hold when the process is of a certain moving-average type. On the other hand, Gumbel (1967) has warned about errors of estimation arising from cycles, pseudocycles and trend-like movements. Furthermore, the theory is not strictly valid if the extreme values are not identically distributed. This happens when there are different causative

[17] There are also other fitting procedures such as Downton's method, as used, for example, by Huxham and McGilchrist (1969).

[18] See, for example, Gumbel (1958a, p. 4).

factors for floods such as frontal rains, thunderstorms and melting of snow.

There is also another operational point. Because the left extremities of density functions of the X and Y populations are at $-\infty$, some negative values can be generated by the Gumbel probability model. For the reduced random variate Y, it follows from equation 6.4 (as could be visualised from figure 6.1) that the probability of a negative value is $1/e = 0.3679$. If we consider a practical case, it is found that, by substituting in equation 6.27, $K(1.01) = -1.6424$, which is applicable for a very low value that is exceeded by the annual maxima in 99 years out of 100. Hence, if $\sigma > 0.6089\mu$ in equation 6.26, a negative flood will occur on average once in 100 years. However, for the estimated value of the standard deviation σ in example 6.1, this period is much longer than 100 years.

Finally, as for other probability distribution functions, sampling errors in the estimates of parameters may be quite large, the implication being that extrapolations may be subject to large errors. Benson (1960) showed that for a hypothetical sample of 1000 items of flood data from a known Gumbel population, parameters estimated from independently chosen subsamples (such as those from a set of 40 subsamples of 25-year periods) are vastly different. It was seen that straight-line plots representing these subsamples have widely varying intercepts and gradients, the variabilities increasing, as expected, in inverse proportion to the subsample lengths. It is found, for instance, that to estimate values of 50- and 100-year floods which are within 25 % of the correct value, 95 % of the time, minimum sample lengths of 39 and 48 years respectively are required. Again, for the estimation of a 50-year flood with a maximum error of 10 %, record lengths of 90 or 110 are required for chances of success equal to 80 % or 95 % of the time respectively.

6.4 General extreme value distribution

As already noted, the Gumbel or type I extreme value distribution is a particular type of the asymptotic or limiting distribution applicable to extreme values. The two-parameter Gumbel distribution is advantageous in the theoretical treatment of flood events, but because of the limitations in application it would be appropriate to consider also the practicability of the other two extreme value asymptotic distributions. The types II and III extreme value distributions are three parametric and their (asymptotic) forms can be obtained if we initially write $G(a_n x)$ in equation 6.1 in place of $G(x)$ and equate a_n to $n^{-\ell}$ and n^{ℓ}, where ℓ is a positive constant[19]. Readers may bypass this section on a first reading; however, the distributions are used for regional analysis in section 6.10.

6.4.1 *Type II extreme value distribution*

If we follow the notation in equations 6.2 and 6.3, the asymptotic type II extreme

[19] See Jenkinson (1955) for applications in hydrometeorology and Gumbel (1958a, b) for the theory.

value distribution is given by

$$\lim_{N\to\infty}[\Pr\{W_{(n)} \leqslant x\}] = \exp[-\{(u-\varepsilon)/(x-\varepsilon)\}^{\alpha}] \tag{6.37}$$

where Pr denotes probability of non-exceedance and $W_{(n)}$ is a random variable representing the maximum flood in any year, of which x is a particular value; $x \geqslant \varepsilon$ and $u \geqslant \varepsilon$. This is also known as the Fréchet distribution.

From equations 6.3 and 6.37

$$x = \varepsilon + (u-\varepsilon)\exp(y/\alpha) \tag{6.38}$$

where y is the type I extreme value (Gumbel) reduced variate. Because of the positive exponential form (for $\alpha > 0$), x increases faster than for the Gumbel distribution, when y is increased. Therefore the distribution can be represented by a curve which is *concave upwards* on Gumbel probability paper. Now, if x is displaced by ε, its natural logarithm will bear a linear relationship with y. If the assumption $\varepsilon = 0$ holds, a straight line of x against y can be drawn on a special type of Gumbel probability paper that has a vertical logarithmic scale to represent this distribution in its two-parameter form; alternatively we may plot $\log(x)$ against y on ordinary Gumbel probability paper.

6.4.2 *Type III extreme value distribution*

If we use similar notation as in the type II distribution, the probability of non-exceedance in this case is given by

$$\lim_{N\to\infty}[\Pr\{W_{(n)} \leqslant x\}] = \exp[-\{(\omega-x)/(\omega-u)\}^{\alpha}] \tag{6.39}$$

where $x \leqslant \omega$; $u \leqslant \omega$. The relationship between x and y is of the form

$$x = \omega - (\omega-u)\exp(-y/\alpha) \tag{6.40}$$

This is a negative exponential type, and, therefore, it can be represented by a curve which is *concave downwards* on Gumbel probability paper. Note that this is of the same type as the Weibull function given by equation 3.60.

6.4.3 *General formula for extreme value distribution*

Corresponding to equations 6.39 and 6.40, Jenkinson (1969) suggested a single equation of the type

$$x = u + (1/\alpha)\{1-\exp(-ky)\}/k \tag{6.41}$$

to represent the relationships between x and y of the three types of asymptotic extreme value distributions in which u, α and k are parameters of location, scale and shape respectively. This is called the general extreme value (GEV) distribution. By substituting the series expansion of $\exp(-ky)$ in equation 6.41 and by then dividing by k, it is seen that the special case $k = 0$ leads to the linear relationship for x against y which characterises the type I extreme value distribution as given by equation 6.4. The type II extreme value distribution is

applicable when $k < 0$, and, if $k > 0$, the type III distribution is signified; these two are represented by equations 6.38 and 6.40 respectively.

In order to evaluate the T-year flood by this method, equation 6.41 is written

$$x(T) = u + z(T)/\alpha \tag{6.42}$$

where the value $z(T)$ of the standardised variate Z is given by

$$z(T) = \{1 - \exp(-ky)\}/k \tag{6.43}$$

in which y and T are related as shown by equation 6.14.

The method of sextiles applied by Jenkinson (1969) gives approximate estimates of k, u and α as follows. The infinite population of Z values, the probability distribution of which can be given as a function of k and y through equation 6.43, is considered to be arranged in increasing order and to be divided into six groups of equal size. Denote the mean of the variates in the ith sextile group by $\mu_{Z,i}$, where $i = 1$ represents the large sextile, and the mean and standard deviation of these six mean values by μ_Z and σ_Z respectively. Then, let the corresponding X values also be divided into six groups in the same way. This is done in a sample of data by ordering the items and by dividing them into six equal or nearly equal groups. Let $\mu_{X,i}$ denote the population mean of the variates in the ith sextile group; also, let the mean and standard deviation of these six values be

$$\mu = \sum_{i=1}^{6} \mu_{X,i}/6$$

and

$$\sigma = \left\{ \sum_{i=1}^{6} (\mu_{X,i} - \mu)^2/6 \right\}^{1/2}$$

respectively. From equation 6.42, by taking expectations and by equating variances

$$\mu = u + \mu_Z/\alpha \tag{6.44}$$

and

$$\sigma = \sigma_Z/\alpha \tag{6.45}$$

The relationships between the shape parameters k (which is common to both X and Z populations), μ_Z, σ_Z and a shape ratio $r - (\mu_{Z,5} - \mu_{Z,6})/(\mu_{Z,1} - \mu_{Z,2})$ are given in table 6.5. (From the above definition, using equations 6.43, 6.7 and 6.9, $\mu_Z = \{1 - \Gamma(1+k)\}/k$. However, to calculate σ_Z and r, we need to use the inverse gamma function, explained in subsection 3.5.1.)

The shape ratio r, which is the same for the Z and X populations from equation 6.42, is estimated from the sample-based statistics $\hat{\mu}_{x,i}$ by $\hat{r} = (\hat{\mu}_{x,5} - \hat{\mu}_{x,6})/(\hat{\mu}_{x,1} - \hat{\mu}_{x,2})$. By interpolation, the corresponding estimates $\hat{k}$,

Table 6.5 Shape parameter k, mean μ_Z standard deviation σ_Z, and shape ratio r of dimensionless Z population of GEV variates

k	μ_Z	σ_Z	r
−0.5	1.54	2.85	0.08
−0.4	1.22	2.24	0.11
−0.3	0.99	1.83	0.16
−0.2	0.82	1.55	0.23
−0.1	0.69	1.35	0.32
0	0.58	1.20	0.44
0.1	0.49	1.09	0.59
0.2	0.41	1.01	0.79
0.3	0.34	0.95	1.05
0.4	0.28	0.92	1.39
0.5	0.23	0.89	1.83
0.6	0.18	0.88	2.39
0.7	0.13	0.87	3.13

$\hat{\mu}_z$ and $\hat{\sigma}_z$ are found from table 6.5, after which equations 6.44 and 6.45 are used to calculate $\hat{u}$ and $\hat{\alpha}$.

Example 6.5 Annual maxima from daily naturalised flows of the Derwent at Yorkshire Bridge for the period 1936 to 1971 are ranked in descending order in table 6.6. The sextile means $\hat{\mu}_{x,1}, \hat{\mu}_{x,2}, \hat{\mu}_{x,3}, \hat{\mu}_{x,4}, \hat{\mu}_{x,5}$ and $\hat{\mu}_{x,6}$ are also given in table 6.6. Estimate the parameters of a GEV distribution to be fitted to the data, and hence calculate $\hat{x}(60)$, $\hat{x}(40)$, $\hat{x}(20)$, $\hat{x}(10)$, $\hat{x}(5)$, $\hat{x}(3)$, $\hat{x}(2)$, $\hat{x}(1.5)$ and $\hat{x}(1.1)$. Plot these values and the given annual maxima on Gumbel probability paper.

Table 6.6

Ranked annual maxima ($\times 10^6$ m^3)					
8.68	4.27	3.49	3.09	2.58	2.30
6.28	4.17	3.47	3.05	2.47	2.28
5.59	3.89	3.44	2.86	2.46	2.15
5.42	3.76	3.40	2.83	2.44	2.13
4.54	3.59	3.20	2.63	2.40	2.12
4.50	3.58	3.12	2.59	2.38	2.02
$\hat{\mu}_{x,1}$	$\hat{\mu}_{x,2}$	$\hat{\mu}_{x,3}$	$\hat{\mu}_{x,4}$	$\hat{\mu}_{x,5}$	$\hat{\mu}_{x,6}$
5.84	3.88	3.35	2.84	2.46	2.17

$$\hat{r} = (\hat{\mu}_{x,5} - \hat{\mu}_{x,6})/(\hat{\mu}_{x,1} - \hat{\mu}_{x,2})$$
$$= 0.29/1.96$$
$$= 0.15$$

From table 6.5 for $\hat{r} = 0.15$, $\hat{k} = -0.32$. The negative sign of $\hat{k}$ shows that the type II extreme value distribution is applicable; also, $\hat{\mu}_z = 1.04$ and $\hat{\sigma}_z = 1.91$. The estimated mean and standard deviation of the sextile means are respectively $\hat{\mu} = 3.42$ and $\hat{\sigma} = 1.22$. From equation 6.45, $\hat{\alpha} = \hat{\sigma}_z/\hat{\sigma} = 1.91/1.22 = 1/0.64$, and, from equation 6.44, $\hat{u} = 3.42 - 0.64 \times 1.04 = 2.75$. The $\hat{x}(T)$ values are obtained as shown in table 6.7 from equations 6.14 and 6.41.

Table 6.7 Computations of $\hat{x}(T)$ for type II extreme value distribution

T	y	$\exp(-\hat{k}y)$	$\hat{z}(T)$	$x(T)$
60	4.09	3.70	8.43	8.14
40	3.68	3.24	7.01	7.24
30	3.38	2.95	6.10	6.66
20	2.97	2.59	4.96	5.92
10	2.25	2.05	3.30	4.86
5	1.50	1.62	1.93	3.98
3	0.90	1.33	1.05	3.42
2	0.37	1.12	0.39	3.00
1.5	−0.09	0.97	−0.09	2.69
1.1	−0.87	0.76	−0.76	2.26

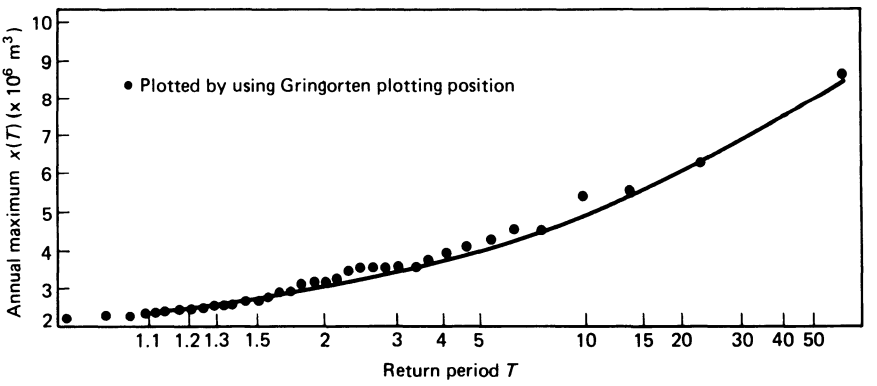


Figure 6.3 GEV distribution fitted by Jenkinson's sextile method to annual maxima from daily flows in Derwent at Yorkshire Bridge for the period 1936 to 1971

The annual maxima from table 6.6 are plotted in figure 6.3 by using the Gringorten plotting position, and the values in table 6.7 give the smooth curve.

The ML equations for the GEV distribution are, of course, more complicated than those for the Gumbel distribution[20].

[20] See Jenkinson (1969, pp. 199–205) and the Natural Environmental Research Council (1975, vol. 1, pp. 96–7).

6.5 Lognormal distribution

The general theory of the lognormal distribution, which is introduced in chapter 3, and its method of application to extreme values are given herein[21].

6.5.1 *Theoretical considerations*

Chow (1954) considered that the occurrence of a flood flow denoted, say, by the random variable X is the result of the joint multiplicative action of a vast number of meteorological and geographical effects, $X_1, X_2, X_3, \ldots, X_r$. That is, $X = X_1 X_2 X_3 \ldots X_r$. If r is infinitely large, $\log X$ is the sum of an infinite number of independent variates, and, accordingly, it is normally distributed by the central limit theorem (see section 3.1). The less satisfactory aspects of this approach are that some of the effects are interdependent (such as mean rainfall and elevation or storm intensity and catchment size, shape and orientation), and there could be great difficulty in identifying them. Because of dependency, the process should strictly be modelled by a multivariate distribution. Furthermore, in practice it is likely that the interactions of the contributory effects are of various types, such as additive, multiplicative, exponential and so on. So, the lognormal distribution can only provide an approximation to real world situations just as when other theoretical distributions are applied to flood flows.

In general, let $Y = \ln(X - \xi)$ be normally distributed with the parameters μ_Y as mean and σ_Y as standard deviation. This means that the random variable X of which an observed value given by

$$x = \exp(y) + \xi \tag{6.46}$$

is assumed to have a three-parameter lognormal distribution. It should be noted that the lognormal distribution is equally applicable when 10 (or any other number) is the base of the logarithms, which will only cause a change in scale[22]. The probability density function for the Y population is

$$f(y) = \sigma_Y^{-1}(2\pi)^{-1/2}\exp\{-(y-\mu_Y)^2/2\sigma_Y^2\} \tag{6.47}$$

The parameters μ_Y and σ_Y are obtained by the method of moments (section 3.3) as follows.

$$E\{\exp(Y)\} = \int_{-\infty}^{\infty} \exp(y) f(y)\mathrm{d}y$$
$$= \sigma_Y^{-1}(2\pi)^{-1/2}\exp(\mu_Y + \sigma_Y^2/2)$$

[21] Hazen (1914) and Horton (1914) originally applied the distribution to flood flows, and, subsequently, Chow (1954) derived the underlying theory; Kalinske (1946), Matalas (1967) and Sangal and Biswas (1970) have also made notable contributions. For an extensive treatment, see Aitchison and Brown (1957).

[22] The value y corresponds to z in equation 3.58; μ_Y and σ_Y correspond to γ and δ respectively and parameter λ is redundant as noted in section 3.6.2.

$$\times \int_{-\infty}^{\infty} \exp[-\{y-\mu_Y-\sigma_Y^2\}^2/2\sigma_Y^2]\mathrm{d}y$$

the integral part of which is equal to the unit area enclosed by a normal probability density curve, defined functionally by the parameters $\mu_Y+\sigma_Y^2$ and σ_Y.

$$E\{\exp(Y)\} = \exp(\mu_Y+\sigma_Y^2/2) \tag{6.48}$$

From equations 6.46 (by taking expectations) and 6.48

$$\mu = \exp(\mu_Y+\sigma_Y^2/2)+\xi \tag{6.49}$$

where $E(X) = \mu$ is the mean of the X population. Because $E(2Y) = 2\mu_Y$ and $\mathrm{var}(2Y) = 4\sigma_Y^2$, from equations 6.46 and 6.48

$$E\{(X-\xi)^2\} = \exp\{2\mu_Y+2\sigma_Y^2\} \tag{6.50}$$

The variance σ^2 of the X population is equal to $E(X^2)-\mu^2$. Hence, it follows from equations 6.49 and 6.50 that

$$\sigma^2 = \{\exp(2\mu_Y+\sigma_Y^2)\}\{\exp(\sigma_Y^2)-1\} \tag{6.51}$$

Also, from $E(X-\mu)^3$ and by using equations 6.46, 6.49 and 6.51, it can be shown that the skewness coefficient $\gamma_1 = \sigma^{-3}E\{(X-\mu)^3\}$ of the X population and σ_Y are related as follows.

$$\gamma_1 = \{\exp(3\sigma_Y^2)-3\exp(\sigma_Y^2)+2\}/\{\exp(\sigma_Y^2)-1\}^{3/2} \tag{6.52}$$

This formula is used to estimate σ_Y when fitting a three-parameter lognormal distribution[23].

Equation 6.49 can be easily adapted for the two-parameter lognormal distribution in which $\xi = 0$. Also, in this case if $V = \sigma/\mu$, the coefficient of variation of the X population, it can be shown that (Aitchinson and Brown, 1957)

$$\gamma_1 = V^3+3V \tag{6.53}$$

One advantage that the lognormal distribution has over the Gumbel distribution is that it is more flexible for curve fitting because the skewness is not fixed.

For the three-parameter case in which $y = \ln(x-\xi)$,

$$x(T) = \exp(\mu_Y+z'_T\sigma_Y)+\xi \tag{6.54}$$

[23] The formula is used in chapter 4; see also Matalas (1967). Sangal and Biswas (1970) suggested an alternative fitting procedure using the median ζ of the X population; if $\alpha = \xi/\mu$, $\beta = \zeta/\mu$ and $V = \sigma/\mu$,

$$2\alpha^3(1-\beta)+\alpha^2(V^2+\beta^2-5+4\beta)+2\alpha(2-\beta V^2-\beta-\beta^2)+\beta^2V^2-1+\beta^2 = 0$$

This is solved iteratively to find α and hence ξ; then we proceed as in the two-parameter case. However, Burges *et al.* (1975) have found from Monte Carlo experiments that the estimator using the median has a larger variance and bias than that based on skewness γ_1 as given by equation 6.52 except perhaps when $\gamma_1 < 0.51$.

The ML method of estimation is far more complicated; Giesbrecht and Kempthorne (1976) discuss the approach and cite earlier work.

6.5.2 *Probability paper*

Normal probability paper was devised by Hazen (1914) for determining probabilities of reservoir yield. Subsequently, Whipple (1916) used logarithmic probability paper to test, for example, the distributions of microscopic organisms in water. The normal density function for a variate X is given by equation 3.3, and equation 3.4 represents the standard normal density function $f(z)$, where $z = (x - \mu)/\sigma$.

If x is normally distributed, a graph of x against z will give on arithmetic graph paper a straight line with intercept μ and gradient σ. Each x value is associated with a z value which has a fixed probability of non-exceedance. It is more useful, on the other hand, to plot values of the integral

$$\Phi(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \{\exp(-t^2/2)\}\, dt \tag{6.55}$$

on the horizontal scale to represent corresponding z values[24]. This is given as a percentage on some types of normal probability paper in which the top and bottom scales are in units of $\{1 - \Phi(z)\}100$ and $\Phi(z)100$ respectively. As noted in section 6.1, the return period $T = 1/\{1 - \Phi(z)\}$.

Lognormal probability paper is produced in the same way except that the vertical scale is logarithmic so that the data need not be transformed to logarithms. If the lognormal law holds, it is expected that a long sequence of annual maxima will give a straight-line plot with a gradient of σ_Y, and an intercept of μ_Y on the vertical representing $\Phi(z) = 0.5$, if the vertical scale is transformed to logarithms and the horizontal scale is converted from $\Phi(z)$ to z.

Example 6.6 Ranked annual maximum daily flows in the Severn at Bewdley for the period 1940 to 1968 are given below. In order to see the fit of a two-parameter lognormal distribution, plot the values on lognormal probability paper. Fit a straight line by the method of moments, and estimate $x(10)$.

Ranked annual maximum daily flows ($m^3 s^{-1}$)

793	768	747	747	711	683	660	648	624	585
585	546	529	528	500	469	465	465	465	455
451	445	422	419	381	347	316	311	300	

The annual maxima are plotted in figure 6.4 by using the Weibull plotting position $T = (N + 1)/m$; the Blom plotting position has been recommended by

[24] See the commonly available tables of the normal distribution or the second and third columns of table 6.9.

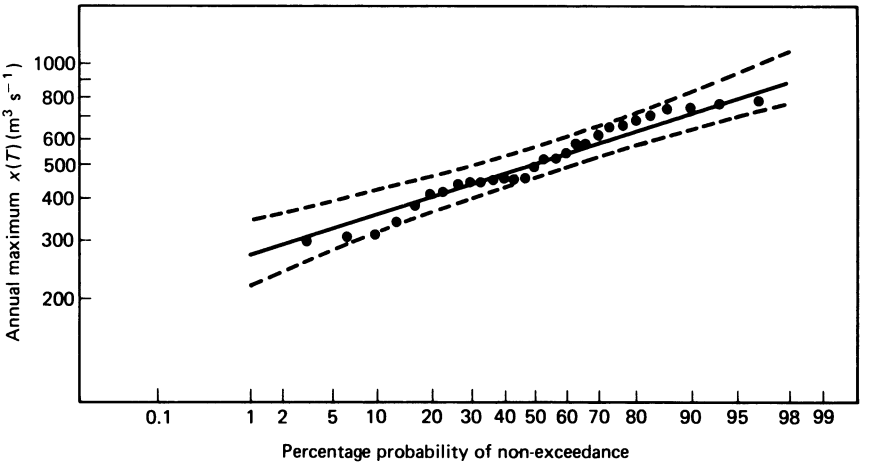


Figure 6.4 Two-parameter lognormal distribution fitted to annual maxima from daily mean flows in Severn at Bewdley for the period 1940 to 1968: broken lines denote 95% confidence limits

Gringorten (1963) for the normal (or lognormal) distribution.

The estimated mean and the standard deviation of the sample are $\bar{x} = 529.8$ and $s = 144.5$ respectively. From equations 6.49 (in which $\xi = 0$) and 6.51, the estimated mean $\bar{y}$ and the standard deviation s_y of the Y population are $\bar{y} = \ln(\bar{x}) - s_y^2/2 = 6.237$ and $s_y = [\ln\{(s/\bar{x})^2 + 1\}]^{1/2} = 0.2679$ respectively. From tables of the normal distribution, $\Phi(2.054) = 0.98$ and $\Phi(2.326) = 0.99$. Therefore, $\hat{x}(50)$, which corresponds to $\Phi(z) = 0.98$, is equal to $\exp(6.237 + 2.054 \times 0.2679) = 886$, and $x(1.01)$, for which $\Phi(z) = 0.01$, equals $\exp(6.237 - 2.326 \times 0.2679) = 274$. These two values correspond to probabilities of non-exceedance equal to 98% and 1% respectively and define the straight line in figure 6.4. The 10-year flood $\hat{x}(10)$ corresponds, of course, to a probability $(1 - 1/10) \times 100 = 90\%$ of non-exceedance. Its magnitude is 720 $m s^{-1}$ from the straight-line plot, or it is theoretically equal to $\exp(6.237 + 1.282 \times 0.2679) = 720.6$.

6.5.3 *Frequency factors*

From equation 6.26

$$K(T) = \{x(T)/\mu - 1\}/V$$

where $V = \sigma/\mu$. For the two-parameter case in which $y = \ln(x)$, $x(T) = \exp(\mu_y + z'_T\sigma_y)$, where z'_T is the value which a standard normal deviate exceeds with probability $1/T$, as given by

$$1/T = (2\pi)^{-1/2} \int_{z'_T}^{\infty} \exp(-t^2/2)\mathrm{d}t$$

and the variable Y is normal with a mean μ_Y and a standard deviation σ_Y. Substituting from equations 6.49 (in which $\xi = 0$) and 6.51, we obtain

$$K(T) = V^{-1}(\{1/(V^2+1)^{1/2}\}\exp[z'_T\{\ln(V^2+1)\}^{1/2}] - 1) \qquad (6.57)$$

This gives the frequency factor for the two-parameter lognormal distribution which is related as shown to the corresponding standard normal deviate z'_T and the coefficient of variation V of the X population. A similar, but more complicated, expression for the frequency factor of the three-parameter lognormal distribution could be given by using z'_T and the three parameters μ_Y, σ_Y and ξ.

Example 6.7 For the data given in example 6.6, estimate the $K(T)$ factor for $T = 20$ and hence $\hat{x}(20)$. Also, estimate the return period of the mean annual flood and the magnitude of the median annual flood.

From equation 6.56 and tables of the standard normal distribution, $\phi(1.6449) = 1 - 1/20$, that is, $z'_{20} = 1.6449$. The sample estimate of the coefficient of variation $V = s/\bar{x} = 144.5/529.8 = 0.273$. Hence, from equation 6.57, $K(20) = 1.829$ and, from equation 6.26, $\hat{x}(20) = 530 + 1.829 \times 144.5 = 794$. This tallies closely with the value from figure 6.4 which has a 95 % probability of non-exceedance. For the mean annual flood, $K(T) = 0$ in equation 6.57. Therefore,

$$\exp[z'_T\{\ln(V^2+1)\}^{1/2}] = (V^2+1)^{1/2}$$

Hence, $\tilde{z}'_T = \{\ln(\tilde{V}^2+1)\}^{1/2}/2 = 0.1339$. Because $1 - \Phi(0.1339) = 1/2.24$ from tables, the return period of the mean annual flood is given by $T = 2.24$, corresponding to which $\tilde{x}(2.24) = \bar{x} = 529.8$. This value has a probability of non-exceedance equal to $(1 - 1/2.24) \times 100 = 55.3\,\%$ on the horizontal scale of figure 6.4. For the median annual flood, $T = 2$ and $z'_2 = 0.0$ if the distribution is normal. Therefore, from equation 6.57, $\tilde{K}(T) = \{1/(\tilde{V}^2+1)^{1/2} - 1\}/V = -0.1292$, and, from equation 6.26, $\tilde{x}(2) = 529.8 - 0.1292 \times 144.5 = 511$, which tallies with the intercept on the 50 % probability line in figure 6.4.

6.5.4 *Confidence limits*

For the two-parameter case the logarithm of the estimate of the T-year flood $x(T)$ is related to $\bar{y}$ and s_y, the estimated values of the parameters μ_Y and σ_Y, as follows from equation 6.54.

$$\ln\{\tilde{x}(T)\} = \bar{y} + z'_T s_y$$

where z'_T which is defined by equation 6.56 has a probability of exceedance $1/T$. Now $\text{cov}(\bar{y}, s_y) = 0$, because $\bar{y}$ and s_y are independent[25]. Also, $\text{var}(\bar{y}) = \sigma_Y^2/N$ and $\text{var}(\sigma_Y) = \sigma_Y^2/2N$. These quantities are estimated by s_y^2/N and $s_y^2/2N$ respectively. Hence,

$$\text{var}[\ln\{\tilde{x}(T)\}] = s_y^2/N + z'^2_T s_y^2/2N$$

[25] Shuster (1973) shows that the statistics y and s_y^2 are independent.

where N is the sample size. The $100(1-\alpha)\%$ confidence limits of $x(T)$, the population value of the T-year flood, is given by

$$\exp\{\ln\{\tilde{x}(T)\} \pm z_{\alpha/2}(\text{var}[\ln\{\tilde{x}(T)\}])^{1/2}\} \tag{6.58}$$

where $z_{\alpha/2}$ is the value which is exceeded with probability $\alpha/2$ by a standard normal deviate[26].

Example 6.8 Using the data given in example 6.6, compute and plot the 95% confidence limits for $T = 1.01$, 1.05, 1.25, 2, 5, 20 and 100.

Now $s_y = 0.2679$ and $N = 29$; therefore, $s_y^2/N = 0.002\,475$ and $s_y^2/2N = 0.001\,237$. The confidence limits are shown in table 6.8 and figure 6.4.

Table 6.8 95% confidence limits with lognormal distribution fitted to annual maximum flows of Severn at Bewdley

T	z'_T	$(\text{var}[\ln\{\tilde{x}(T)\}])^{1/2}$	$\ln\{\tilde{x}(T)\}$	*Upper confidence limit*	*Lower confidence limit*
100	2.3263	0.0958	6.8599	1150	790
20	1.6449	0.0763	6.6773	922	684
5	0.8416	0.0579	6.4621	717	572
2	0	0.0497	6.2367	563	464
1.25	−0.8416	0.0579	6.0112	457	364
1.05	−1.6449	0.0763	5.7960	382	283
1.01	−2.3263	0.0958	5.6135	331	227

Example 6.9 Fit a three-parameter lognormal distribution to the data given in example 6.6. Here, estimate $x(50)$ and $x(1.01)$.

For the $N(=29)$ items of data from the X population, the coefficient of skewness γ_1 is estimated as follows,

$$g_1 = \{N^2\sum x^3 - 3N\sum x\sum x^2 + 2(\sum x)^3\}/N(N-1)(N-2)s^3 \tag{6.59}$$

where $s = [\{\sum x^2 - (\sum x)^2/N\}/(N-1)]^{1/2}$ and $\sum$ denotes summation of N values. Hence $g_1 = 0.2515$, and, as obtained in example 6.5, $\bar{x} = 529.8$ and $s = 144.5$ which are the sample estimates of μ and σ. From equation 6.52, s_y, the estimate of σ_y, equals 0.0835 and from equations 6.51 and 6.49, $\bar{y} = 7.451$ and $\tilde{\xi} = -1212$, which are the estimates of μ_y and ξ respectively. Hence, $\tilde{x}(50) = \exp(7.451 + 2.054 \times 0.0835) - 1212 = 845$, and $\tilde{x}(1.01) = \exp(7.451 - 2.326 \times 0.0835) - 1198 = 220$.

Because g_1 is small and ξ is negative, it does not seem worthwhile to fit this distribution here. Under more favourable conditions the theoretical straight line passing through the two points such as $\tilde{x}(5)$ and $\tilde{x}(1.01)$ may be compared with the line representing the two-parameter distribution for visual goodness of fit with the plotted points.

[26] Bias in small samples may be corrected by using the tables of Student's t distribution instead.

6.5.5 *Bias in skewness and Hazen's correction*

It should be noted that the estimator given by equation 6.59 is unbiased only if the population is normal which is not true in practice. Because g_1 is known to have a definite downward bias when calculated from a lognormal population, Hazen (1930) suggested the use of an empirical correction factor of $1 + 8.5/N$ for the coefficient of skewness[27]. Thus the revised estimator becomes

$$g_1' = \{N\sum x^3 - 3N(\sum x)(\sum x^2) + 2(\sum x)^3\}(1 + 8.5/N)/N(N-1)(N-2)s^3 \tag{6.60}$$

Example 6.10 For the data given in example 6.6, fit a three-parameter lognormal distribution using Hazen's correction for skewness. Hence, estimate $x(50)$ and $x(1.01)$.

From example 6.9, $\bar{x} = 529.8$, $s = 144.5$ and $g_1' = (1 + 8.5/29) \times 0.2511 = 0.3252$. From equations 6.52, 6.51 and 6.49 the following estimates of σ_Y, μ_Y and ξ are obtained: $s_y = 0.1077$, $\bar{y} = 7.193$ and $\tilde{\xi} = -808$. Hence, from equation 6.54, $\tilde{x}(50) = \exp(7.193 + 2.054 \times 0.1077) - 808 = 851$ and $\tilde{x}(1.01) = \exp(7.193 + 2.326 \times 0.1067) - 808 = 227$

6.5.6 *Regional skewness*

On account of sampling errors in estimates of skewness, it has been suggested that an average value should be taken over a hydrologically homogeneous region. The drawback is that such a region could be difficult to define, and in practice boundaries are marked somewhat arbitrarily. Results show that estimates of skewness for stations within a region have high variability and poor correlation with physiographic and meteorologic factors. Furthermore, such estimates are biased when outliers ('surprisingly high values') are present. More about these aspects will be found in sections 6.10 and 6.11.

State-averaged values of skewness for logarithmically transformed flood data from the United States range from 0.6 in the eastern states to -0.5 in Illinois[28]. The country has also been partitioned into 14 regions for this purpose. From flood records at 1351 selected stations, means of the skewness of untransformed data range from 3.0 in the south to 0.9 in the southwest and northeast[29]. In the

[27] Wallis *et al.* (1974) have found from Monte Carlo studies that Hazen's correction gives an unbiased estimate over a small range such as $0.5 < \gamma_1 < 2$ for the lognormal distribution. They also noted that the average bias factor in the estimated skewness is a function of the skewness and the distribution, and, subsequently, Bobée and Robitaille (1975) proposed formulae for adjustment. Regardless of bias corrections a single estimate of the coefficient of skewness is subject to high sample fluctuations, but the absolute magnitude of the statistic does not exceed $(N-2)/(N-1)^{1/2}$, as shown by Kirby (1974).

[28] See Hardison (1974).

[29] See Matalas *et al.* (1975).

United Kingdom, corresponding regional averages of skewness of flood data vary from 4.36 in the southeast to 1.04 in the northeast[30].

6.6 Pearson type III function applied to extreme values

The Pearson type III function is explained and methods of estimating the parameters are given in sections 3.2 to 3.4. The function which was applied originally to flood flows by Foster (1924) has no rigorous analytical basis, but its usefulness for curve-fitting purposes has been demonstrated[31].

6.6.1 *Frequency factors*

As shown in subsection 3.2.2, the Pearson type III function is given by

$$f(x) = (x-\xi)^{\gamma-1}\exp\{-(x-\xi)/\lambda\}/\lambda^{\gamma}\Gamma(\gamma), \qquad \xi \leqslant x \leqslant \infty \tag{6.61}$$

where γ, λ and ξ are the three parameters. By using the transformation

$$z = (x-\xi)/\lambda \tag{6.62}$$

the standard gamma function

$$f(z) = z^{\gamma-1}\exp(-z)\Gamma(\gamma) \tag{6.63}$$

is obtained. The integral of this, with finite upper limit $u(T)$ and $0 \leqslant F(u) \leqslant 1$,

$$F(u) = \int_0^{u(T)} z^{\gamma-1}\exp(-z)\mathrm{d}z\Gamma(\gamma) \tag{6.64}$$

is extensively tabulated by Wilk *et al.* (1962). The T-year flood

$$x(T) = \xi + u(T)\lambda \tag{6.65}$$

is obtained after replacing z in equation 6.62 by $u(T)$, the standard gamma variate, which is also given in table 3.3 for some values of γ and $F(u) = 1 - 1/T$. Then, substituting from equations 3.40 and 3.41 in equation 6.65, we obtain the following estimator by the method of moments.

$$\hat{x}(T) = \bar{x} + s\{u(T)g_1/2 - 2/g\} \tag{6.66}$$

where $\bar{x}$, s and g_1 are the estimators of the mean, the standard deviation and the coefficient of skewness respectively of the X population. Equation 6.66 corresponds to the general form of equation 6.26 and the frequency factors $K(T) = u(T)\gamma_1/2 - 2/\gamma_1$ are given in table 6.9 for some values of the probability $F(u)$ of non-exceedance and the coefficient of skewness γ_1. As noted from equation 3.39, $\gamma = 4/\gamma_1^2$, and this links table 6.9 to table 3.3. For more comprehensive tables, reference should be made to Harter (1969).

[30] See the Natural Environmental Research Council (1975).

[31] As shown by the Natural Environmental Research Council (1975) and by others such as Majumdar and Sawhney (1965).

Table 6.9 Frequency factors for Pearson type III function

Return period T	*Probability of non-exceedance F(u)*	*Frequency factors for coefficient of skewness* $\gamma_F =$																
		0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0001	0.0001	−3.719	−3.299	−2.899	−2.525	−2.184	−1.884	−1.628	−1.418	−1.247	−1.111	−1.000	−0.800	−0.667	−0.571	−0.500	−0.444	−0.400
1.0010	0.0010	−3.090	−2.808	−2.533	−2.268	−2.017	−1.786	−1.577	−1.394	−1.238	−1.107	−0.999	−0.800	−0.667	−0.571	−0.500	−0.444	−0.400
1.0101	0.0100	−2.326	−2.178	−2.029	−1.880	−1.733	−1.588	−1.449	−1.318	−1.197	−1.087	−0.990	−0.799	−0.667	−0.571	−0.500	−0.444	−0.400
1.0204	0.0200	−2.054	−1.945	−1.834	−1.720	−1.606	−1.492	−1.379	−1.270	−1.166	−1.069	−0.980	−0.798	−0.666	−0.571	−0.500	−0.444	−0.400
1.0256	0.0250	−1.960	−1.864	−1.764	−1.663	−1.559	−1.455	−1.352	−1.250	−1.152	−1.060	−0.975	−0.797	−0.666	−0.571	−0.500	−0.444	−0.400
1.0526	0.0500	−1.645	−1.586	−1.524	−1.458	−1.389	−1.317	−1.243	−1.168	−1.093	−1.020	−0.949	−0.790	−0.665	−0.571	−0.500	−0.444	−0.400
1.1111	0.1000	−1.282	−1.258	−1.231	−1.200	−1.166	−1.128	−1.086	−1.041	−0.994	−0.945	−0.894	−0.771	−0.660	−0.570	−0.500	−0.444	−0.400
1 2500	0.2000	−0.842	−0.850	−0.855	−0.857	−0.856	−0.852	−0.844	−0.832	−0.817	−0.799	−0.777	−0.711	−0.636	−0.562	−0.498	−0.444	−0.400
2	0.5000	0.000	−0.033	−0.067	−0.099	−0.132	−0.164	−0.195	−0.225	−0.254	−0.281	−0.307	−0.360	−0.396	−0.413	−0.413	−0.400	−0.379
5	0.8000	0.842	0.830	0.816	0.800	0.780	0.758	0.733	0.705	0.675	0.643	0.609	0.518	0.420	0.322	0.226	0.137	0.058
10	0.9000	1.282	1.301	1.317	1.329	1.336	1.340	1.340	1.337	1.329	1.318	1.303	1.250	1.180	1.096	1.001	0.900	0.795
20	0.9500	1.645	1.700	1.750	1.797	1.839	1.877	1.910	1.938	1.962	1.981	1.996	2.012	2.003	1.971	1.920	1.853	1.773
	0.9750	1.960	2.053	2.142	2.227	2.308	2.384	2.455	2.521	2.582	2.638	2.689	2.793	2.867	2.913	2.933	2.931	2.909
50	0.9800	2.054	2.159	2.261	2.359	2.453	2.542	2.626	2.706	2.780	2.848	2.912	3.048	3.152	3.226	3.274	3.298	3.300
100	0.9900	2.326	2.472	2.615	2.755	2.891	3.023	3.149	3.271	3.388	3.499	3.605	3.845	4.051	4.225	4.368	4.483	4.573
1000	0.9990	3.090	3.377	3.666	3.956	4.244	4.531	4.815	5.095	5.371	5.642	5.908	6.548	7.152	7.720	8.253	8.752	9.220
10000	0.9999	3.719	4.153	4.597	5.047	5.501	5.957	6.412	6.867	7.318	7.766	8.210	9.299	10.354	11.373	12.357	13.305	14.220

Example 6.11 Ranked annual maximum flows of the Derwent at Longbridge Weir for the period 1936 to 1962 are given below. Plot the data using the Hazen plotting position, given in table 6.1, on normal probability paper, and fit a curve to represent the Pearson type III function. Estimate $x(10)$.

Ranked annual maximum flows ($m^3 s^{-1}$)									
269	258	228	180	167	144	143	143	142	126
124	117	115	110	109	108	106	102	102	99
98	95	87	85	81	77	68			

The estimated mean and standard deviation are $\bar{x} = 129.00$ and $s = 51.84$, and the skewness coefficient g_1, estimated by equation 6.60, is equal to 2.056. The coefficient of skewness is aproximated to 2.0, and the frequency factors are found from table 6.9; calculations are given in table 6.10.

Table 6.10 Pearson type III function fitted to flood flows in Derwent at Longbridge

Return period T	*Probability of non-exceedance F(u)*	*Frequency factor K(T)*	$\hat{x}(T)$
1.001	0.001	−0.999	77.2
1.01	0.01	−0.990	77.7
1.02	0.02	−0.980	78.2
1.05	0.05	−0.949	79.8
1.11	0.10	−0.894	82.7
1.25	0.20	−0.777	88.7
2.00	0.50	−0.307	113
5	0.80	0.609	161
10	0.90	1.303	197
20	0.95	1.996	232
50	0.98	2.912	280
100	0.99	3.605	316

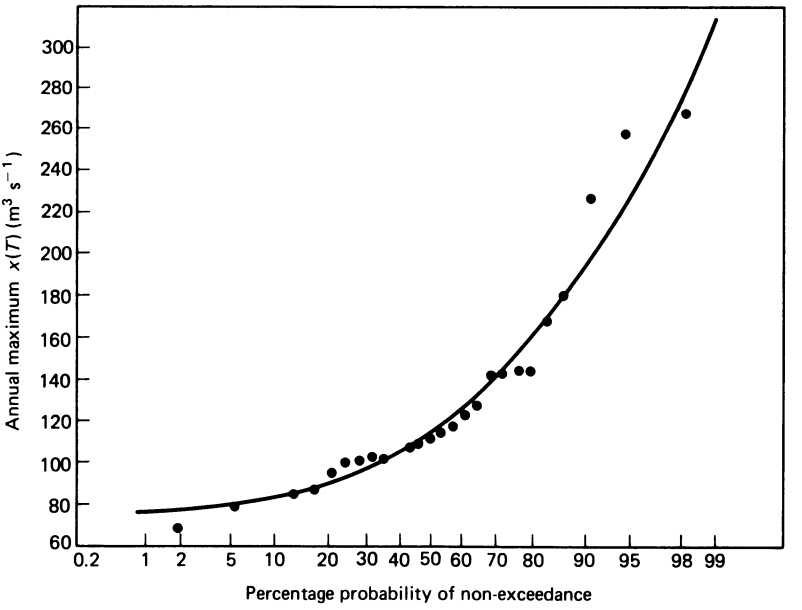
The plotted points and the theoretical curve are shown in figure 6.5. $\hat{x}(10) = 196\ m^3 s^{-3}$.

6.6.2 *Two-parameter gamma function*

The gamma probability density function

$$f(x) = x^{\gamma-1} \exp(-x/\lambda)/\lambda^{\gamma}\,\Gamma(\gamma), \qquad 0 \leqslant x < \infty \tag{6.67}$$

which has been applied to flood flows, for instance, by Moran (1957) is a simpler version of the Pearson type III function given above with the location parameter



Figures 6.5 Pearson type III distribution fitted to annual maxima for daily flows in Derwent at Longbridge Weir for the period 1936 to 1962

$\xi = 0$. The parameters γ and λ may be estimated by the method of moments from the mean $\bar{x}$ and the standard deviation s if we use equation 3.35 and the first equation of equation 3.32. Hence, $\tilde{\lambda} = s^2/\bar{x}$ and $\tilde{\gamma} = \bar{x}^2/s^2$.

Obviously, this will not fit a given sequence of data better than the type III function will. However, if goodness-of-fit tests show that neither is rejected, the two-parameter function may be used.

Example 6.12 For the data given in example 6.11, estimate $x(20)$ by using the gamma function given by equation 6.67.

As shown above, $\tilde{\lambda} = 51.84^2/129.00 = 20.83$ and $\tilde{\gamma} = (129/51.84)^2 = 6.192$. For $T = 20$, the probability of non-exceedance is given by $F(u) = 1 - 1/20 = 0.95$ with reference to equation 6.64. From table 3.3, the standard gamma variate $u(T)$ is 10.77 which is obtained by interpolation for $\gamma = 6.19$. Hence, $\tilde{x}(20) = u(T)\lambda = 10.77 \times 20.83 = 224$.

6.6.3 *Log Pearson type III function*

When the Pearson type III function is applied to the logarithms (to any base) of the flood flows, the distribution function is termed the log Pearson type III function. If $x = e^y$, then from equation 6.66

$$\tilde{x}(T) = \exp[\bar{y} + s_y\{u(T)g_y/2 - 2/g_y\}] \tag{6.68}$$

where $\bar{y}$, s_y and g_y denote the estimators of the mean, the standard devia-

tion and the skewness of the Y population for which, by comparing equations 6.26 and 6.68, $\tilde{K}(T) = u(T)g_y/2 - 2/g_y$. When referring to table 6.9 for this frequency factor, if skewness is negative which is quite possible for logarithmically transformed flood flows, the following procedure should be adopted. Replace each pair of co-ordinates $\{\gamma_Y, F(u)\}$ by $\{|\gamma_Y|, 1 - F(u)\}$, and then change the sign; for example, if $\gamma_Y = -1.0$, $K(10)$ which corresponds to $F(u) = 0.9$ is obtained from the row for $F(u) = 0.1$ and is equal to 1.128 after changing the sign. The log Pearson type III distribution was recommended for general use by the American Water Resources Council[32].

Example 6.13 Plot the data given in example 6.11 on lognormal probability paper, using the Blom plotting position from table 6.1, and fit the log Pearson type III function. Estimate $x(10)$ by this method.

$\bar{y} = 4.796$, $s_y = 0.3495$ and $g'_y = 1.0696$, which is approximated to 1.0 and table 6.9 is referred to. The calculations are given in table 6.11.

Table 6.11 Log Pearson type III function fitting to flood flows in Derwent at Longbridge

Return period T	*Probability of non-exceedance* $F(u)$	*Frequency function* $K(T)$	$\tilde{x}(T)$
1.001	0.001	−1.786	65
1.01	0.01	−1.588	69
1.02	0.02	−1.492	72
1.05	0.05	−1.317	76
1.11	0.10	−1.128	82
1.25	0.20	−0.852	90
2.00	0.50	−0.164	114
5	0.80	0.758	158
10	0.90	1.340	193
20	0.95	1.877	233
50	0.98	2.542	294

The plotted points and the theoretical curve are shown in figure 6.6 from which $\tilde{x}(10) = 193\ \mathrm{m^3\ s^{-1}}$.

Note that the log Pearson type III function has a lower limit $\exp(\xi_Y)$ when skewness is positive[33]. This is estimated by $\exp(\bar{y} - 2s_y/g'_y)$ through the method of moments. If $y = 10^x$, the limit becomes $\exp\{(\bar{y} - 2s_y/g'_y)\ln(10)\}$. For the given data, the lower limit is 57. On the contrary, if skewness is negative, the log Pearson type III variates are bound by an equal upper limit. This necessitates careful consideration in application.

[32] See Benson (1968). This was originally suggested by L. R. Beard. Confidence limits for $x(T)$ when the Pearson type III or log Pearson type III distributions are applicable involves a procedure suggested by Moran (1957) which is partly numerical. This is also given by Santos (1970) and by Condie (1977).

[33] See Gilroy (1972).

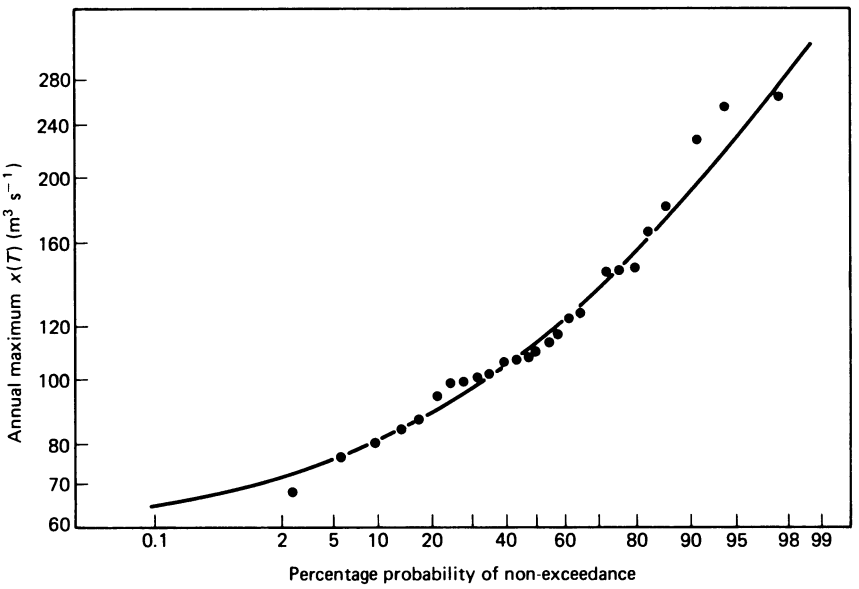


Figure 6.6 Log Pearson type III distribution fitted to annual maxima for daily flows in Derwent at Longbridge Weir for the period 1936 to 1962

6.7 Discussion on frequency methods of flood estimation

The distributions given in the preceding sections have been used extensively in the estimation of flood-flow probabilities. Although each of them has had support on theoretical or empirical grounds, it seems reasonable to think that no ordinary probability function can fully represent the complicated flood-producing factors which change in time and space owing to natural causes or man's actions. Therefore, some degree of subjectivity is unavoidable, if we consider the present state of the art. Graphical techniques do indeed provide a convenient method of choosing between different distributions; however, long extrapolations may be unreliable even if good fits are obtained.

Studies involving comparisons between probability distributions have been made recently. Firstly, six distributions were applied to records of length 40 to 97 years from 10 selected stations by the work group on flow frequency methods appointed by the Hydrological Committee of the United States Water Resources Council[34]. These are the Gumbel, log Gumbel (that is, a two-parameter type II extreme value distribution), two-parameter gamma, log Pearson type III, lognormal and lognormal modified by the Hazen method. Their recommendation was that the log Pearson type III distribution, of which

[34] See Benson (1968).

the lognormal distribution is a special case, should be used with the proviso that, if the data show evidence of a significant difference, the best-fitting distribution should be adopted.

In order to assess the relative suitabilities of the distributions, an empirical goodness-of-fit test was used by the United States work group. They thought that more information could be obtained from such a procedure than through one of the statistical goodness-of-fit tests described in chapter 3 in which a single abnormal value could cause rejection It is interesting to recall that Gumbel (1943) too had expressed some dissatisfaction over the chi-squared test when applied to flood flows. In the particular method adopted by the United States team, a sequence of data was ranked, and the items were plotted at points, say, $\{x_D(T), T\}$, where D signifies the data, on log Gumbel probability paper by using the Weibull plotting position. Then, for each station and each of the return periods $T = 2$, 5, 10, 25 and 50, the absolute differences between interpolated values on the broken straight lines joining the plotted points and the theoretical values $\hat{x}(T)$ for each function were calculated. These differences were then reduced to dimensionless units by dividing by the interpolated values $x_D(T)$. The criterion on which the log Pearson type III function was chosen is the average of these differences.

One of the criticisms levelled against the American report is that it does not show how to deal with samples containing outliers (surprisingly high values)[35]. The question of whether to include such discordant values with the rest of the sample data has been recurring over the part 100 years or more in studies on astronomy and other natural phenomena; many publications on the subject are found in the journal *Technometrics*. However, an engineer who has to face up to such a situation could benefit perhaps more from personal judgement than by using a complicated statistical fuction as formulated, for example, by Grubbs (1950). In this, associated rainfall data and the physical reasons for any extraordinary event ought to be examined. The subjective nature of decision making, implied by the word surprising, is stressed by Collett and Lewis (1976) who also point out the relevance of presentation, scale and pattern of the data in perceiving outliers; however, if an objective statistical criterion is used, the word discordant should be used in place of the word surprising. Anscombe (1960) compares a rejection rule to a domestic fire insurance policy, the choice of which depends on the answers to such questions as the following.

(1) What is the premium?
(2) How much protection does the policy give in the event of fire?
(3) How much danger really is there of a fire?

The answer to the last question will be as obvious to the prudent hydrologist as to the householder who is aware that many homes are destroyed by fire.

In the extensive report by the Natural Environmental Research Council

[35] If low flows are being examined, an outlier is, on the other hand, a surprisingly low value.

(1975), the empirical distributions of 35 annual maximum flow sequences from the United Kingdom and Ireland, selected on a reliability basis were fitted with each of 7 theoretical functions. These are the Gumbel, GEV, gamma, log gamma, Pearson type III, log Pearson type III and lognormal distributions. The record lengths ranged from 31 years to 88 years in the United Kingdom catchments; the maximum and minimum lengths of the Irish records are 44 years and 23 years respectively. Chi-squared and Kolmogorov–Smirnov goodness-of-fit tests and three indices based on probability plots, in which the standardised measure is similar to that adopted by the American group except that the divisor is the mean of the annual maxima from the record instead of the $x_D(T)$ values which are dependent on the plotting position. Not surprisingly, the three-parameter distributions such as the log Pearson type III and the GEV functions were found to fit the data better than the two-parameter functions. The Natural Environmental Research Council (1975) chose the type II extreme value function for extrapolation on a regional and national basis. It was also found, from several preliminary statistical tests carried out, that there is persistence in 2 and trend in 6 of the chosen 28 United Kingdom records, although these are not allowed for in the formulation of the theoretical functions. Such departures from ideal conditions have been encountered in applications elsewhere, and the imposed limitations should be borne in mind.

The main shortcomings in the annual maximum series method of flood estimation can be summarised as follows. Firstly, the true probability distribution, if it exists, is obscured owing to sampling errors, and, therefore, extrapolation should be treated with caution. When using graphical techniques a suitable choice of plotting position is desirable if it results in minimum bias. However, this criterion is based on repeated sampling from a hypothetical population, whereas in practical situations only a single sample is available. The point is that distortions at high return periods arising from an incorrect plotting position may be totally swamped by errors caused by an inappropriate probability model. It seems, therefore, that unwarranted emphasis can be given to the choice of a plotting position. Secondly, estimates of parameters are also subject to errors. The method of moments is generally affected by sampling errors in the estimates of moments. Moran (1957), amongst others, advocated the ML method of estimation. However, the importance associated with the ML method may not be justifiable when the assumptions on which the probability model are based are themselves incorrect.

One possible alternative is to use bayesian decision theory as explained in chapter 9. An example is given by Davis *et al.* (1972) for flood control on the Rillito Creek in Arizona; here, the decision variable is the height of dikes to be constructed. Results are, however, highly dependent on the cost or benefit function used and are also based on this assumed distribution[36]. Other possible solutions to the problem are described in section 6.12.

[36] The economic effect of floods, flood protection works and insurance are discussed by Brown (1972).

6.8 Binomial, Poisson and multinomial distributions

Another type of question which a practising engineer would ask concerns the probabilities of occurrence of floods of very high return periods within an economic life span of, say, a spillway dam. This is required in evaluating the risks involved. The answer could be given without specifying the probability distribution of the flood events and the values of the parameters, but, of course, the flood magnitude associated with the given return period is dependent on this distribution; it means that errors in estimating the true return period of a high flood magnitude will affect the value of risk.

The binomial distribution has been used for this purpose[37]. The Poisson and multinomial distributions are also applicable; the first is used as an approximation to the binomial and a joint probability, when two or more exceedance levels are considered, is calculated from the second.

6.8.1 *Binomial distribution*

The values in a serially independent annual maximum series could be thought of as either an exceedance (success) or a non-exceedance (failure) of a fixed value with probabilities of occurrence equal to p and $1-p$ respectively. This two-sided Bernoulli random variable, which was formally described by James Bernoulli in the days when probability theory was mainly applied to games of chance, leads to the binomial probability distribution

$$B(M = m \mid N,\ p) = \binom{N}{m} p^m (1-p)^{N-m} \tag{6.69}$$

of M successes in N independent identically distributed Bernoulli trials. The theory is derived in the following example.

Example 6.14 Calculate the probability of having two 10-year annual maximum flood events, in a 5-year period, assuming that the events are serially independent.

Let $x(10)$ denote the flood magnitude which has a return period of 10 years. The joint probability of having two 10-year flood events, each of which is equal to or greater than $x(10)$, with a probability of occurrence $p(=0.1)$, and three flood events each of which is less than $x(10)$ in magnitude, with probability of occurrence $1-p$ is given by $p^2(1-p)^3$ for an independent sequence. The five flood events can be arranged in 5! different ways, but the two 10-year flood events are classed together because there is no need to identify them individually; the other three are similarly included together in a separate class. Therefore, the total number of different arrangements of the two types of flood events is $5!/2!\,3! = 10$. (If P_1 denotes p and P_2 denotes $1-p$, the 10 different sequences could be denoted by $P_1P_1P_2P_2P_2$, $P_1P_2P_1P_2P_2$, $P_1P_2P_2P_1P_2$, $P_1P_2P_2P_2P_1$, $P_2P_1P_1P_2P_2$, $P_2P_1P_2P_1P_2$, $P_2P_1P_2P_2P_1$, $P_2P_2P_1P_1P_2$,

[37] See, for example, Markowitz (1971).

$P_2P_2P_1P_2P_1$ and $P_2P_2P_2P_1P_1$.) Hence, the required probability equals $\binom{5}{2}p^2(1-p)^3 = 10 \times 0.1^2 \times 0.9^3 = 0.0729$.

6.8.2 *Poisson distribution*

Under certain conditions the binomial can be approximated by the Poisson distribution. If $p = a/N$ (for example, if $p = 0.01$ which signifies a 100-year flood and $N = 40$, then $a = 0.4$), equation 6.66 can be written in the following form.

$$B(M = m \mid N, p) = \binom{N}{m}(a/N)^m(1-a/N)^{N-m}$$

$$= N(N-1)(N-2)\ldots(N-m+1)(N^m m!)^{-1}a^m$$

$$\times (1-a/N)^N(1-a/N)^{-m}$$

If m and a are fixed,

$$\lim_{N\to\infty}\{N(N-1)(N-2)\ldots(N-m+1)N^{-m}\} = 1$$

and

$$\lim_{N\to\infty}\{(1-a/N)^{-m}\} = 1$$

From a series expansion,

$$\ln\{1/(1-t)\} = -\ln(1-t)$$

$$= t + t^2/2 + t^3/3 + t^4/4 + \ldots$$

Now, if $b = (1-a/N)^N$,

$$\ln(b) = N\ln(1-a/N)$$

$$= -a - a^2/2N - a^3/3N^2 \ldots$$

Therefore,

$$\lim_{N\to\infty}(1-a/N)^N = e^{-a}$$

Hence, for large N, the binomial is approximated by the Poisson probability distribution

$$P(M = m|a) = a^m e^{-a}/m! \tag{6.70}$$

and this is justifiable if p is small, say, not more than 0.10, and N is large.

Example 6.15 The probability of at least one 100-year flood in a 40-year period is $1 - P(M = 0 \mid a) = 1 - (40 \times 0.01)^0 e^{-40\times 0.01}/0! = 0.33$. Note that the probability of no 100-year floods is subtracted from unity.

6.8.3 *Multinomial distribution*

The probability of having, in a sequence of N annual maxima, $M_1, M_2, M_3, \ldots, M_r$ events with probabilities of occurrence equal to $p_1, p_2, p_3, \ldots, p_r$ respectively is given by the multinomial distribution,

$$M(M_1 = m_1, M_2 = m_2, \ldots, M_r = m_r \mid N, p_1, p_2, \ldots, p_r)$$
$$= p_1^{m_1} p_2^{m_2} \ldots p_r^{m_r} (1 - p_1 - p_2 \ldots - p_r)^{N - m_1 - m_2 - \ldots - m_r}$$
$$\times N!/m_1!m_2! \ldots m_r!(N - m_1 - m_2 - \ldots - m_r)! \qquad (6.71)$$

where

$$\sum_{i=1}^{r} M_i \leqslant N$$

The set of probabilities p_i, $i = 1, 2, 3, \ldots, r$, can be expressed in terms of the return intervals T_i, $i = 1, 2, 3, \ldots, r$, as follows.

$$p_1 = 1/T_1$$
$$p_2 = 1/T_2 - 1/T_1$$
$$p_3 = 1/T_3 - 1/T_2$$
$$p_r = 1/T_r - 1/T_{r-1}$$

The theory can be easily derived using the same type of arguments as in example 6.14, and it is obvious that the binomial is a special case of the multinomial distribution.

Example 6.16 Calculate the probability of having, in a 5-year sequence of annual maxima, four 5-year floods of which two are 10-year floods. $N = 5$; $m_1 = 2$; $m_2 = 2$; $p_1 = 0.1$; $p_2 = 0.2 - 0.1 = 0.1$. Hence, the required probability is $0.1^2 \times 0.1^2 \times 0.8^1 \times 5!/2!2!1! = 0.0024$.

6.8.4 *Limitations*

It is important to note that these calculations are based on the assumption that the flood events are serially independent and identically distributed. Moreover, as in the choice of the basic extreme value models, the method ignores alterations in natural and environmental factors. On account of man's actions such as urbanisation, channel improvements, construction of dams and irrigation works there will be further changes in the underlying distributions[38].

Hitherto, only annual maximum series were considered. As already mentioned in section 6.7, these have limitations which casts doubts on extrapolated values. In order to increase the information found in high flows from a short sample of data, the peaks-over-threshold method, in which estimates are based on more than one value per year, is used.

[38] See, for example, Kazmann (1972, pp. 615, 616).

6.9 Peaks-over-threshold method

The peaks-over-threshold (POT) method concerns the distribution of the number and magnitude of peak flows that exceed a threshold such as x_b in figure 6.7 which shows part of a continuous record of flow in a river. Such peak flows are said to constitute a partial duration series. The threshold level x_b may be raised or lowered so as to involve a desirable number of peaks per year; this chosen number may be in the range from 3 to 5.

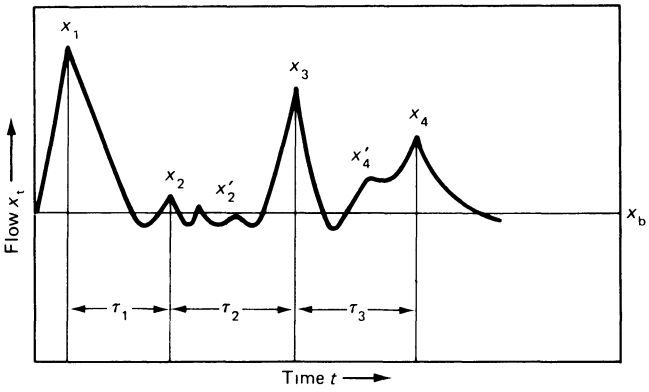


Figure 6.7 Peaks over a threshold

In order to make the analysis tractable, it is assumed that the individual peaks $x_1, x_2, x_3, x_4, \ldots$ represent independent hydrometeorological events or, in other words, that these are not serially correlated. This means that peaks such as x_2' and x_4' which do not have definite ascensions and recessions and which seem to be associated with x_2 and x_4 respectively are not considered. In practice, the selection has to be done empirically[39].

Also of interest is the distribution of the interevent (also called waiting or recurrence) times τ_i, $i = 1, 2, 3, \ldots$, between successive exceedances. The joint distribution of the τ_i values specify a stochastic process which is found by the times of peak flows exceeding x_b. The term renewal process is applicable if the τ_i values are independent and identically distributed. This cannot, of course, be strictly true because of seasonal variations. For example, the times between summer thunderstorms are different from those between cyclonic rains in winter. Moreover, if snow is contributory, the times of melting may be distributed differently.

If on average there are a peaks per year which exceed the threshold, then the number M of POT events per year is a random variable which has the Poisson probability distribution

[39] For example, the Natural Environmental Research Council (1975, vol. 1, p. 46) suggests that 'peaks should be separated in time by 3 times the time to peak and that the flow should decrease between peaks to two-thirds of the first peak'.

$$P(M = m \mid a) = a^m e^{-a}/m! \tag{6.72}$$

where m is a particular value which M takes.

Also, the probability of exceedance of the τ_i values has the exponential distribution

$$\Pr(\tau_i > \tau) = e^{-m\tau} \tag{6.73}$$

in which the parameter $1/m$ is the mean of the τ_i values, as shown below.

The magnitudes X of the peaks which exceed x_b are also assumed to be exponentially distributed so that the probability of exceedance of a particular value $x(T)$ which has a recurrence interval of T years is given by

$$\Pr\{X > x(T) \mid X > x_b\} = e^{-\lambda\{x(T)-x_b\}}$$

where λ is a constant and the vertical line inside the brackets denotes conditional to. Because the probability of exceedance is the reciprocal of the return interval, in the POT analysis the T-year flood is that which on average is exceeded once in aT events compared with once in T events in the annual maximum series. It follows that

$$e^{-\lambda\{x(T)-x_b\}} = 1/aT \tag{6.74}$$

and

$$x(T) = x_b + \{\ln(a) + \ln(T)\}/\lambda \tag{6.75}$$

For a given set of exceedances $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(N)}$, which are serially independent and ranked in ascending order so that $X_{(1)}$ is the lowest,

$$\Pr(x_{(1)} < x) = 1 - e^{-(x-x_b)/(1/N\lambda)} \tag{6.76}$$

The expectation of a variate with an exponential distribution $F(x) = 1 - e^{-\lambda x}$ is given by

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x}\,dx$$

$$= [-xe^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx,$$

integrating by parts.

Because $\lim_{x\to\infty}(xe^{-\lambda x}) = 0$, for the same reasons given after equation 3.30, $E(X) = 1/\lambda$. Correspondingly, it follows from equation 6.74 that

$$\mu = x_b + 1/\lambda \tag{6.77}$$

where μ is the mean of the $X_{(i)}$ population with the sample estimator

$$\bar{x} = \sum_{i=1}^{N} x_{(i)}/N$$

Also, from equation 6.76, the expected value of the lowest item is given by

$$E(X_{(1)}) = x_b + 1/N\lambda \tag{6.78}$$

Now $\bar{x}$ and $x_{(1)}$ are sufficient estimators of μ and x_b respectively, where according to the definition by Fisher (1922) an estimator t, say, is sufficient for a parameter θ, say, if the distribution of a sample, given t, does not depend on θ. Therefore, the following estimator for λ is obtained from equation 6.77.

$$\hat{\lambda}' = 1/(\bar{x} - x_{(1)}) \tag{6.79}$$

By taking expectations and by substituting from equations 6.77 and 6.78

$$E(\hat{\lambda}') = \lambda N/(N-1)$$

Then the following unbiased estimator for λ is obtained by substituting the last result in equation 6.79 for $E(\hat{\lambda}')$.

$$\hat{\lambda}' = \{(N-1)/N\}/(\bar{x} - x_{(1)}) \tag{6.80}$$

Again, for x_b, the following unbiased estimator is obtained from equation 6.78.

$$\hat{x}_b = x_{(1)} - 1/N\hat{\lambda} \tag{6.81}$$

The POT method is useful for estimating the magnitudes and frequency of events which have low return periods T. Its main importance is in the design of cofferdams or culverts. However, for $T > 10$, Langbein (1949) has shown that $x(T)$ calculated from the POT method differs very little from that calculated from an annual maximum series.

Example 6.17 Peak daily flows in the River Derwent at Yorkshire Bridge which exceed 19 $m^3 s^{-1}$ are tabulated below for the period 1933 to 1937.

Daily flows ($m^3 s^{-1}$)			
1933		1936	
Feb. 1	21.16	Mar. 8	20.46
Mar. 3	30.71	Mar. 9	20.51
Mar. 4	28.39	Sep. 7	28.21
Nov. 15	19.06	Nov. 9	20.12
1934		Nov. 12	20.23
		Nov. 15	22.17
None		Nov. 17	19.92
		Dec. 14	26.73
1935		1937	
Feb. 15	36.62	Jan. 6	29.55
Feb. 16	41.04	Feb. 14	20.65
Oct. 9	28.54	Mar. 17	22.88
Oct. 27	35.09	Mar. 18	33.15
Oct. 28	30.65	Mar. 19	24.87
Oct. 29	23.11	Dec. 2	32.35
Oct. 30	23.39	Dec. 22	29.17
Oct. 31	24.21		
Nov. 4	22.96		
Nov. 17	27.87		
Nov. 20	21.37		
Nov. 21	21.99		

Using a suitable threshold value, estimate the following.

(1) The probability of at least two exceedances of the threshold in 1 year.
(2) The probability of having a 3-year period without any exceedances.
(3) The magnitude of a 5-year flood.

(This sample is too short for practical purposes and is only used here merely to explain the procedure.) The following POT values are taken as serially independent values with an average of three per year; 21.16, 30.71, 41.04, 28.54, 35.09, 22.96, 27.87, 21.99, 28.21, 22.17, 26.73, 29.55, 33.15, 32.35, 29.17. The lowest value $x_{(1)} = 21.16$ and the mean $\bar{x} = 28.71$.

(1) Using equation 6.72, $a = 3$, $P(M = 0 \mid 3) = 3^0 e^{-3}/0! = 0.0498$ and $P(M = 1 \mid 3) = 3^1 e^{-3}/1! = 0.1494$. The probability of at least two exceedances per year is $1 - P(M = 0 \mid 3) - P(M = 1 \mid 3) = 0.8008$ which tallies with the visual evidence of four cases out of five.

(2) By using equation 6.73, $P(\tau_i > 3) = \exp(-3 \times 3) = 0.0001$, which is an exceedingly low probability of having a 3-year period without an exceedance.

(3) From equations 6.80, 6.81 and 6.75, $\hat{\lambda} = (14/15)/(28.71 - 21.16) = 0.1236$, $\hat{x}_b = x_{(1)} - 1/N\hat{\lambda} = 20.62$ and $\hat{x}(5) = 20.62 + 8.09 \ln(3 \times 5) = 42.53$.

6.10 Regional flood frequency analysis

The limitations in single-site data are, in summary, that a sequence may be too short to represent the population of flood events adequately, even without considering possible non-stationarities. In addition, the critical values in the records may be subject to serious errors of measurement. On account of such deficiencies, hydrologists have resorted to regionalisation, that is, to combining the information in several records from a homogeneous zone or region. This would hopefully lead to a more realistic estimation of floods of given return periods.

Initially, there is the problem of defining the boundary of such a region. One way to demarcate a region is so that the hydrologic or response characteristics of the catchment areas within it are comparable. These may be assessed through unit hydrographs, lag times and flow duration curves. Alternatively, physical and climatological characteristics may be the overriding criteria in the choice. Finally, and this is probably the easiest method, regions could be defined through existing geographical boundaries; also, areas of similar soils or geology and land use maps have been employed particularly in the United States[40].

There are two main objectives in regional analysis. The first is to extrapolate flood estimates to sites with scanty or no data. In regional studies a multiple regression formula of the type

$$x(T) = aB^bC^cD^d \ldots K^k$$

[40] See, for example, the numerous references in Schulz *et al.* (1973).

is generally assumed for $x(T)$ in which $B, C, \ldots, K$ are the parameters or factors and $b, c, d, \ldots, k$ are regression constants[41]. Then a linear regression equation is obtained through a logarithmic transformation. Initially, a distribution such as the log Pearson type III is fitted to the observed flood data, separately for each station. From the station curves, estimated values $\hat{x}(T)$ are obtained for a particular value of T and are regressed by using a step-forward method, with catchment and other characteristics. These variables are tested in turn for significance before they are included in the equation. The procedure, at any step, is to select the independent variable which maximises the squared partial correlation coefficients, given the $\hat{x}(T)$ and the variables selected before[42].

In the report by the Natural Environmental Research Council (1975), the mean flood, which has a return period of 2.33 years for the Gumbel distribution, is correlated to significant catchment characteristics. The parameters used are as follows: catchment area (km^2); STMFRQ, the number of stream junctions, shown on a 1:25 000 map divided by A; S1085, stream gradient which is calculated from 10 % to 85 % of the stream length from the gauge; RSMD, net 1-day rainfall (which is rainfall less a weighted mean soil moisture deficit) with a 5-year return period; LAKE, proportion of catchment draining through a lake; SOIL, an index of catchment soils in the range of 0.15 to 0.50 calculated from $0.15S_1 + 0.3S_2 + 0.4S_3 + 0.45S_4 + 0.5S_5$, where S_1, S_2, S_3, S_4 and S_5 are the fractions of the catchment area covered by five soil types in increasing order of perviousness; URBAN, the urban fraction of the catchment[43].

For example, the regression equation for the central region of the United Kingdom is

$$\hat{x}(2.33) = 0.0213(\text{AREA})^{0.94}(\text{STMFRQ})^{0.27}(\text{S1085})^{0.16} \times (\text{SOIL})^{1.23}(\text{RSMD})^{1.03}(1 + \text{LAKE})^{-0.85} \quad (6.82)$$

At the same time a dimensionless flood sequence is obtained for a region after dividing the observed values from the various catchments by the estimated mean for the particular catchment. In this way floods from different catchments can be compared directly; originally, engineers used catchment area as a divisor. A region curve is then drawn from the ordered set of data on normal and Gumbel probability paper by using appropriate plotting positions. These curves are similar to figures 6.2 or 6.5 except that the vertical axes are marked in units of $x(T)/x(2.33)$, in which $x(2.33)$ is the (Gumbel) mean annual maximum flood. It is noted from the report by the Natural Environmental Research Council

[41] For example, Benson (1962a) used the following parameters for the northeastern United States: N-year annual peak discharge; drainage area; main-channel slope; percentage of surface storage area plus 0.5 %; N-year rainfall intensity; average January degrees below freezing; orographic factor.

[42] Standard methods of regression are explained for example by Fryer (1966). Most computers have routines for this type of work.

[43] Regional studies have also been made for the United Kingdom by Nash and Shaw (1966) and by Cole (1966).

(1975) that the type II extreme value distribution seems to provide the best fit for distributions of flood events on a regional basis. Accordingly, parameters of this distribution are calculated for ten regions in the United Kingdom and one for the whole of Ireland. The limits of the regional values of parameters in equation 6.41 are $0 \leqslant k \leqslant -0.325$, $0.77 \leqslant u \leqslant 0.87$ and $0.18 \leqslant \alpha \leqslant 0.28$ with mean values of -0.2, 0.8 and 0.24 respectively. By using the second approximation of equation 6.15, the national (United Kingdom) equation suggested by the Natural Environmental Research Council (1975) is

$$\hat{x}(T)/\hat{x}(2.33) = -0.4 + 1.20T^{0.2} \tag{6.83}$$

It should be noted that the presence of an outlier, say, x_{max}, which is an extremely high flow such as a 1000-year flood within a short sample, will give an upward bias to the mean flow as given by equation 6.82 or by a similar regression equation. Now, the median flood flow x_{med} is known to be a more stable statistic than the mean $x(2.33)$; the Natural Environmental Research Council (1975) found that $x(2.33)/x_{med} \approx 1.07$ for United Kingdom data and recommended that, if in a particular case $x_{max} > 3x_{med}$, $\hat{x}(2.33)$ should be equated to $1.07x_{med}$.

Regionalisation is sometimes used to extend floods temporarily in order to estimate the frequency of floods of high return periods. According to Kritsky and Menkel (1969), hydraulic structures in the Soviet Union are designed to pass maximum floods which occur on average once in 1000 or 10 000 years. This has been achieved by combining flood records from the Volga, Dneiper and other river basins. However, because of spatial correlation between flood events, the return period of a critical flood event could be much less than, say, the hypothetical 1000-year period obtained by combining 20 records of length 50 years. The influence of correlation, in this so-called station–year method, is examined by Carrigan (1971).

Finally, if we return to the general regional approach, its main shortcoming is that the highest floods within a region are often caused by a single meteorological event. The same could also apply to the second, third and other critical floods. When this happens there seems to be little virtue in using regionalisation because we cannot obtain more information than in a single-station analysis. At the other extreme, if the crucial floods are caused by local convective precipitation, orographic effects or the melting of snow rather than through cyclonic systems which are often widespread, the standard error in the regression may be too high, and the method is of doubtful value for spatial extrapolation, on account of significant differences between the flood-producing characteristics of the individual catchments.

6.11 Probable maximum precipitation

The inadequacies in the frequency approach are discussed in previous sections. Even if long records are available, there is uncertainty regarding estimated values. For instance, regardless of the largest observed flow, it is inevitable that a

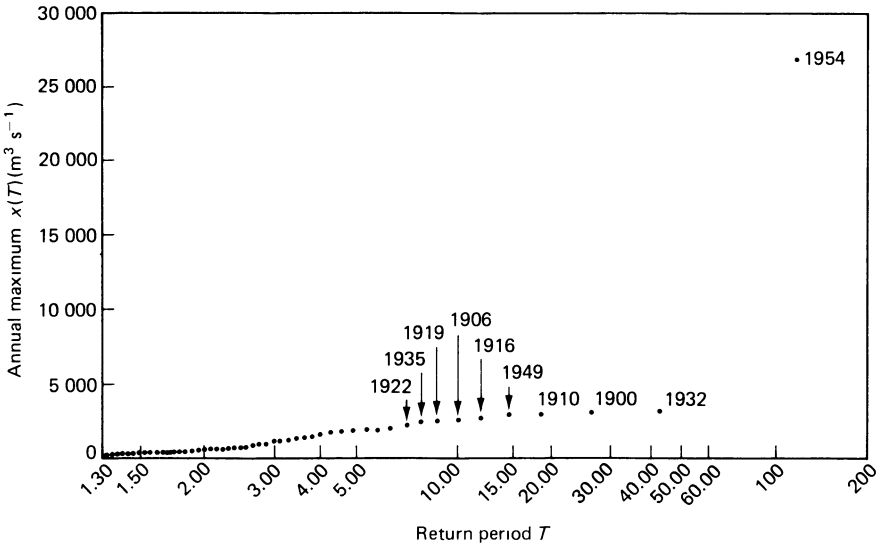


Figure 6.8 Annual maximum peak flows in Pecos near Shumla, Texas, for the period 1900 to 1968: Gringorton plotting position is used; return periods of less than 1.3 years are not shown. Years in which the ten highest flows occurred are given. Note the flood which occurred on 28 June 1954 as shown at the top. Prior to October 1954, the gauging station was 13 miles downstream at Pecos, Texas

much larger flood will occur in the future, and it is in the application to abnormally high flows that frequency methods are the least satisfactory. An example is shown in figure 6.8 which is a Gumbel plot from 67 years of annual maxima in the Pecos River near Shumla, Texas. The maximum flow in 1954 is an outlier which, if ignored and calculated on the basis of the other 66 items of data, has a return period of more than 10^{11} years. This is clearly a flood event which cannot be accounted for by conventional methods! Other examples from the United States are the floods in Virginia during August 1969 due to hurricane Camille, which were about ten times those recorded earlier, and this was followed by the catastrophic events in Rapid City, South Dakota. However, the largest flood damage, estimated at three billion dollars, was caused in June 1972 by hurricane Agnes in the eastern United States during June 1972, and the greatest flooding elsewhere during recent decades occurred in Bangladesh during November 1970 as a result of a tropical cyclone[44].

Although such freak events are possible almost everywhere, it is rational to assume from a knowledge of physics that there is an upper limit to maximum floods, however impractical its definition might seem, in the same way that other natural phenomena have their own ends or bounds. To quote Horton (1936), 'A

[44] Information on other outliers in flood data from the United States is given by Hardison (1973).

small stream cannot produce a major Mississippi flood just as an ordinary barnyard fowl cannot lay an egg a yard in diameter: it would transcend nature's capabilities under the circumstances.' We could also add other impossible cases such as a man of age 200 years, a woman 10 feet tall and a snake 1 mile long.

In this sense, the unbounded right tails of commonly postulated frequency functions for flood flows are not realistic. The question then arises what the upper limit should be, and this will be of particular interest in the design of large dams, the failure of which can have a serious effect on lives and property. In order to find a practical solution to the problem, hydrometeorologists have developed the technique of probable maximum precipitation (PMP).

It is easy to imagine that, when observed storms are transposed from neighbouring catchments to an area above a particular observation site, extreme flood flows which exceed the magnitudes of observed events could occur at this site. Storm maximisation obtained by considering dew points, wind velocities, condensation of cloud particles and other criteria related to storm efficiency increases this effect. However, all the meteorological factors associated with maximum floods cannot be accounted for, because of the limitations in the knowledge of atmospheric processes and also because of the lack of data. Because of these shortcomings, the approach is subjective and it has aroused a great deal of controversy[45]. Nevertheless, in the United States, design floods are based on the PMP method if the dam heights are greater than 60 feet, and a United States committee has considered the safety of large dams on this basis[46]. Then, the unit hydrograph method is used to obtain the probable maximum flood from the PMP. The method is also followed in Australia. The following is an outline of the basic principles.

Because the upper limit of high floods cannot be satisfactorily defined, no structure which is designed to cope with these extraordinary events could be absolutely safe. On the other hand, the design of, say, a spillway dam that can pass the flood caused by the highest possible precipitation is conceivable, if the flood is obtained by maximising all the factors simultaneously, but the cost of such a structure would be prohibitive. Besides, there is uncertainty regarding these 'maximum' factors. For engineering expediency, therefore, PMP has been defined as the magnitude of rainfall over a catchment area that would result in a flood flow of which there is 'virtually no risk of being exceeded'[47]. There have also been other definitions, and a discussion on these is given by Alexander (1965). In this context it is important to note here that in some areas the melting of snow is an important contributory factor.

As for the term storm transposition used in hydrometeorology, with its area of extent and physical boundaries, a region of meteorological homogeneity is best regarded as one in which every catchment within it can have precipitation events with similar inflow wind movement and storm mechanisms but with

[45] See, for example, Gumbel (1958b), Yevjevich (1968) and Benson (1973).
[46] This is reported, for instance, by Gray (1974).
[47] See Myers (1969).

variations in the total moisture charge and frequency of occurrence. There is also another important point to bear in mind. This concerns the types of storms which we are justified in transposing. Whereas thunderstorms lend themselves easily to transposition with hardly any reservation regarding distance, hurricanes are effective only in certain coastal areas[48]. Also, there are limits to transposition in mountainous zones, and storms observed in one catchment cannot be considered to occur in another if the difference in elevation is excessive (say, more that about 1500 feet). By the same token, the shape and orientation of rainfall patterns associated with frontal rains should not be altered. Therefore, it follows that a hydrometeorological analysis of this type requires careful judgement.

When faced with inadequate data, PMP analysts estimate the moisture of a storm from surface dew points. This approximation is reasonable for heavy storms when the column of air is saturated and the vertical temperature gradient is equivalent to the saturated pseudoadiabatic lapse rate which is a decrease of about 0.5 °C per 100 m above the surface. The importance of the dew point stems from the fact that there is an increase of about 9 % in precipitation for every 1 °C increase in the dew point; dew points over oceanic surfaces are of special significance. Maps giving the variation in precipitation water with dew points and elevations (of orographic barriers to inflowing air) are given with examples by Weisner (1970) and by the Tennessee Valley Authority (1961). In elementary applications of the method recorded precipitation–depth–duration curves are increased directly in proportion to the amounts of water that can be precipitated in the two catchment areas[49].

It has also been suggested that an empirical factor $K(T)$ times the standard deviation should be used in addition to the mean of a maximum precipitation sequence as an initial approximation to the PMP in the form given by equation 6.26. For instance, Hershfield (1961) found that, in a key group of 24-hour United States stations, $K(T)$ has an upper limit of 15. Meanwhile, Alexander (1963), in order to provide a measurably probabilistic basis to the problem, related the return period T_c of PMP in a catchment area A_c to the rank r, in descending order where $r = 1, 2, 3, \ldots, N_e$, of observed maximum precipitation events in the homogeneous zone, of area A_h, as follows.

$$T_c = N_e A_h / r A_c$$

[48] See, for example, Lane (1948, chapter 1).

[49] For practical application in the Tennessee valley area, see the Tennessee Valley Authority (1961). A manual for the estimation of PMP is given by the World Meteorological Organisation (1973). Calculations involving other criteria are also given by Weisner (1970) and elsewhere by Miller (1973). As regards national maps of maximum precipitation and other aspects of analysis and design, reference may be made, for example, to Chow (1964, sections 9, 21, 25), Linsley *et al.* (1949), Berry *et al.* (1945) and the Natural Environmental Research Council (1975).

6.12 Other methods and comments

A different approach to flood estimation is possible through the generation of large samples of data by means of the daily flow time series models explained in chapter 4. Kottegoda (1972, 1973) has examined the possibilities through a linear autoregression model; the work of Green (1973) and Quimpo (1967) are relevant. Of more recent origin is the shot noise model of Weiss (1977) and the model of Treiber and Plate (1975) based on a deterministic system function (see chapter 4). One of the main purposes in this approach is to estimate parameters of a probability model from very large samples of data. However, there are problems regarding the correct formulation of daily models and the estimates of their parameters, and on average the uncertainties in this method may balance those in conventional frequency methods of flood estimation. Nevertheless, the output should be useful for simulation of complex systems.

Monte Carlo methods could also indicate improved decision-oriented methods to counteract uncertainty in flood estimation, although practicalities are yet not clear. For instance, extensive computer studies were made by Slack *et al.* (1975) on the choice of distribution between normal, Gumbel, lognormal or Weibull distributions for high-flow data generated on the basis of these distributions. Their criterion is the minimum expected design loss with square root, linear and quadratic loss functions and variable scaling factors; the sample space was defined through skewness (in the third and fourth distributions), sample size and return period. If $\hat{x}(T)$ and $x(T)$ denote the estimated and true values respectively of the design flood (in the authors' notation) an underdesign loss occurs if $\hat{x}(T) < x(T)$ and vice versa. On the basis of expected opportunity losses, the normal does not seem to be disadvantageous, regardless of whether we identify the underlying distribution of floods or not. However, with limited information on skewness and detailed information on the relative scale of overdesign to underdesign losses, a substantial reduction in opportunity losses occurs. In a subsequent work, it was found that the assumed distribution which minimises the expected design loss is quite stable with respect to N, the sample size[50].

Because longer records of rainfall are usually available, attempts have been made to obtain improved estimates of frequencies of high flows from rainfall events. However, antecedent conditions, for instance, are highly variable, and because gauged rainfall data may not be representative of catchment rainfall there is high scatter in plots of rainfall against river flow. On the other hand, there is a central tendency for the return intervals in the two sides to be theoretically equal in the long run, but in a practical situation this property is not of much use[51].

If we return now to the POT approach examined in section 6.9, perhaps its main drawback is that the data are not identically distributed. As an improvement, Todorovic and Rouselle (1971) formulated a seasonal model (see

[50] See Wallis *et al.* (1976).

[51] See Larson and Reich (1973) and the discussion of their paper by Whittaker (1973).

also Todorovic (1978)). For this type of analysis, harmonic fitted cumulative sums of the mean number of exceedances of the threshold value in 17 periods of 20 days and 1 of 25 days within an annual cycle are initially computed. The probability distribution of the largest POT value in, say, the summer season is given by

$$\Pr(X_{su} < x) = \exp(-M_{sp}{}^{-x/\mu_{sp}} - M_{su}{}^{-x/\mu_{su}})$$

in which M_{sp} is the difference between the mean number of exceedances at the end and start of the spring season and μ_{sp} is the expectation of the POT value during the spring season; M_{su} and μ_{su} have similar connotations with respect to the summer season. The original work of Todorovic and his coworkers is commendable; nevertheless, the main problem of estimating the magnitudes of flood peaks for specified high return periods still remain, on account of the fact that the distributions of the annual maximum and partial duration series merge rather quickly.

On the subject of annual maxima flows, Singh and Sinclair (1972) proposed an empirical five-parameter distribution comprising two normal distributions in order to model the reverse curvatures frequently seen in probability plots. In spite of better fits to sample data which is anticipated, there could be serious doubts about the true form of the population distribution as estimated by this method. The idea of mixed distributions is intuitively correct, but empirical curve-fitting methods cannot provide permanent solutions. Indeed, the future of objective treatment of high flows must lie on a rigorous mathematical and physical approach without restrictive assumptions.

6.13 Final remarks and summary

As mentioned before, great uncertainty is associated with the estimation of the probabilities of rare floods. This seems to be inevitable because, firstly, there is insufficient information at present to define empirically the right tails of density functions of high flows. Secondly, because of the underlying complexities that are unaccounted for, theoretical models are inadequate for dealing with the important problems. Therefore, a definite set of rules cannot be given in the foreseeable future for flood estimation, and any decisions taken will be subject to personal bias. More confidence could, of course, be placed in the estimation of average or more likely events.

In the hydraulic design of a structure, such as a culvert for which the criterion is a high flow with a return period of about 5 to 10 years, the POT method should normally provide satisfactory answers when the available sequence of data is sufficiently long, perhaps more than 30 years. Large floods which affect the design of structures such as dams could be estimated through a probability function chosen from a selected few that fit the data. This may suffice for practical purposes when samples are sufficiently long and return intervals are commensurate with sample lengths. An indication of the likely errors which arise

even in such cases is given in subsection 6.2.1; these errors would escalate when incorrect probability models are chosen or on account of non-stationarities. If the estimation involves an extrapolation far beyond the data sample, then the regional method is suggested which is also the best way by which floods at ungauged sites could be estimated. It is important here to bear in mind the limitations which this entails, such as bias and standard errors due to lack of representative data and incorrectly defined regional boundaries respectively.

This means that estimates of floods of high return intervals are generally subject to serious errors. As regards very high floods that are a threat to life and property, the most feasible method of tackling this problem at present is by the PMP technique. Although the concept is subjective and the method tends to become arbitrary in practice, it helps to provide an engineering solution which takes into account the relevant information and uncertainties.

References

Abramowtiz, M., and Stegun, L. (eds) (1964). Handbook of mathematical functions. *Natl Bur. Stand., Appl. Math. Ser. Publ.*, No. 55; *Handbook of Mathematical Functions*, Dover Publications, New York

Aitchison, J., and Brown, J. A. C. (1957). *The Lognormal Distribution*, Cambridge University Press, London

Alexander, G. N. (1963). Using the probability of storm transposition for estimating the frequency of rare floods. *J. Hydrol.*, **1**, 46–57

——(1965). Discussion of 'hydrology of spillway design: large structures—adequate data'. *J. Hydraul Div., Proc. Am. Soc. Civ. Eng.*, **91** (HY1), 211–19

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, **2**, 123–47

Barnett, V. (1975). Probability plotting methods and order statistics. *Appl. Statist.*, **24**, 95–108

Benson, M. A. (1960). Characteristics of frequency curves based on a theoretical 1000-year record. *U.S., Geol. Surv., Water-Supply Paper*, No. 1543-A, 51–94

——(1962a). Factors influencing the occurrence of floods in a humid region of diverse terrain. *U.S., Geol. Surv., Water-Supply Paper*, No. 1580-B

——(1962b). Plotting positions and economics of engineering planning. *J. Hydraul. Div., Proc. Am. Soc. Civ. Eng.*, **88** (HY6), 57–71. (1963). Discussion closure. *J. Hydraul. Div., Proc. Am. Soc. Civ. Eng.*, **89** (HY6), 251–2

—— (1968). Uniform flood-frequency estimating methods for federal agencies. *Water Resour. Res.*, **4**, 891–908. (1969). Comments. *Water Resour. Res.*, **5**, 910–11. (1970). Comments. *Water Resour. Res.*, **6**, 998–9

—— (1973). Thoughts on the design of design floods. *Floods and Droughts, Proceedings of the 2nd International Hydrology Symposium, 11–13 September* 1972, Water Resource Publications, Fort Collins, Colorado, pp. 27–33

Berry, F. A., Bollay, E., and Beers, N. R. (eds) (1945). *Handbook of*

Meteorology, McGraw-Hill, New York

Bobée, B., and Robitaille, R. (1975). Correction of bias in the estimation of the coefficient of skewness. *Water Resour. Res.*, **11**, 851–4

Brown, J. P. (1972). *The Economic Effect of Floods*, Springer, Berlin

Burges, S. J., Lettenmaier, D. P., and Bates C. L. (1975). Properties of the three-parameter log normal probability distribution. *Water Resour. Res.*, **11**, 229–35

Bury, K. V. (1975). *Statistical Models in Applied Science*, Wiley, New York

Carrigan, P. H., Jr. (1971). A flood frequency relation based on regional record maxima. *U.S., Geol. Surv., Prof. Paper*, No. 434-F

Chow, V. T. (1951). A general formula for hydrologic frequency analysis. *Trans. Am. Geophys. Un.*, **32**, 231–7. (1952). Discussion. *Trans. Am. Geophys. Un.*, **33**, 277–82

—— (1954). The log-probability law and its engineering applications. *Proc. Am. Soc. Civ. Eng.*, **80**, paper 536, 1–25

——(1964). *Handbook of Applied Hydrology*, McGraw-Hill, New York

Cole, G. (1966). An application of the regional analysis of flood flows. *Proceedings of the Symposium on River Flood Hydrology, March* 1965, Institution of Civil Engineers, London, session B, paper 3

Collett, D., and Lewis, T. (1976). The subjective nature of outlier rejection procedures. *Appl. Statist.*, **25**, 228–37

Condie, R. (1977). The log Pearson type 3 distribution: the *T*-year event and its asymptotic standard error by maximum likelihood theory. *Water Resour. Res.*, **13**, 987–91

Court, A. (1952). Some new statistical techniques in geophysics. *Adv. Geophys.*, **1**, 45–85

Davis, D. R., Kisiel, C. C., and Duckstein, L. (1972). Bayesian decision theory applied to design in hydrology. *Water Resour. Res.*, **8**, 33–41

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. A*, **222**, 309–68

Fisher, R. A., and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Philos. Soc.*, **24**, 180–90

Foster, H. A. (1924). Theoretical frequency curves and their applications to engineering problems. *Trans. Am. Soc. Civ. Eng.*, **87**, 142–203

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathematique*, **6**, 92–117

Fryer, H. C. (1966). *Concepts and Methods of Experimental Statistics*, Allyn and Bacon, London

Fuller, W. E. (1914). Flood flows. *Trans. Am. Soc. Civ. Eng.*, **77**, 564–617

Giesbrecht, F., and Kempthorne, O. (1976). Maximum likelihood estimation in the three-parametric lognormal distribution. *J. R. Statist. Soc. B*, **38**, 257–63

Gilroy, E. J. (1972). The upper bound of a log-Pearson type 3 random variable

with negatively skewed logarithms. *U.S., Geol. Surv., Prof. Paper*, **800**, B273–5

Gray, D. A. (1974). Safety of dams—bureau of reclamation. *J. Hydraul. Div., Am. Soc. Civ. Eng.*, **100** (HY2), 267–77

Green, N. M. D. (1973). A synthetic model for daily streamflow. *J. Hydrol.*, **20**, 351–64

Gringorten, I. I. (1963). A plotting rule for extreme probability paper. *J. Geophys. Res.*, **68**, 813–4

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Statist.*, **21**, 27–58

Gumbel, E. J. (1941). The return period of flood flows. *Ann. Math. Statist.*, **12**, 163–90

——(1943). On the reliability of the classical chi-square test. *Ann. Math. Statist.*, **14**, 253–63

——(1954). Statistical theory of extreme valves and some practical applications. *Natl Bur. Stand., Appl. Math. Ser., Publ.*, No. 33

——(1958a). *Statistics of Extremes*, Columbia University Press, New York

——(1958b). Theory of floods and droughts. *J. Inst. Water Eng.*, **12**, 157–84 (1959). Communications. *J. Inst. Water Eng.*, **13**, 71–102

——(1967). Extreme value analysis of hydrologic data. *Statistical Methods in Hydrology, Proceedings of the 5th Hydrology Symposium, McGill University*, 1966, Queen's Printer, Ottawa, pp. 147–81

Hardison, C. H. (1973). Probability distribution of extreme floods, highways and the catastrophic floods of 1972. *Proceedings of the 52nd Annual General Meeting of the Highway Research Board*, National Academy of Sciences, Washington, D.C., No. 479, pp. 42–5

——(1974). Generalized skew coefficients of annual floods in the United States and their application. *Water Resour. Res.*, **10**, 745–52

Harter, H. L. (1969). A new table of percentage points of the Pearson type III distribution. *Technometrics*, **11**, 177–87

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Trans. Am. Soc. Civ. Eng.*, **77**, 1539–640

——(1930). *Flood Flows*, Wiley, New York

Hershfield, D. M. (1961). Estimating the probable maximum precipitation. *J. Hydraul. Div., Am. Soc. Civ. Eng.*, **87** (HY5), 99–116

Holtzman, W. H. (1950). The unbiased estimate of the population variance and standard deviation. *Am. J. Psychol.*, **63**, 615–17

Horton, R. E. (1914). Discussion on 'Flood flows' by W. E. Fuller. *Trans. Am. Soc. Civ. Eng.*, **77**, 663–70

——(1936). Hydrologic conditions as affecting the results of the application of method of frequency analysis to flood records. *U.S., Geol. Surv., Water-Supply Paper*, No. 771, 433–50

Huxham, S. H., and McGilchrist, C. A. (1969). On the extreme value distribution for describing annual flood series. *Water Resour. Res.*, **5**, 1404–5

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or

minimum) values of meteorological elements. *Q. J. R. Meteorol. Soc.*, **81,** 158–71

——(1969). Estimation of maximum floods. *World Meteorol. Organ., Tech. Note*, No. 98, chapter 5, 183–257

Kaczmarek, Z. (1957). Efficiency of the estimation of floods with a given return period, vol. 3, International Association of Scientific Hydrology, Toronto, pp. 145–59

Kalinske, A. A. (1946). On the logarithmic-probability law. *Trans. Am. Geophys. Un.*, **27,** 709–11

Kazmann, R. G. (1972). *Modern Hydrology*, 2nd edn, Harper and Row, New York

Kendall, M. G., and Stuart, A. (1977). *The Advanced Theory of Statistics*, vol. 2, 4th edn, Griffin, London

Kimball, B. F. (1960). On the choice of plotting positions on probability paper. *J. Am. Statist. Assoc.*, **55,** 546–60

Kirby, W. (1974). Algebraic boundedness of sample statistics. *Water Resour. Res.*, **10,** 220–2

Kottegoda, N. T. (1972). Flood evaluation—can stochastic models provide an answer? *Proceedings of the International Symposium on Uncertainties in Hydrology and Water Resources Systems*, vol. 1, 11–14 *December* 1972, *University of Arizona, Tucson*, pp. 105–14

——(1973). Flood estimation by some data generation techniques. *Floods and Droughts, Proceedings of the 2nd International Hydrology Symposium*, 11–13 *September* 1972, Water Resource Publications, Fort Collins, Colorado, pp. 189–99

Kritsky, S. N. and Menkel, M. F. (1969). On principles of estimation methods of maximum discharge. *Floods and their Computation*, vol. 1, International Association of Scientific Hydrology, Belgium, No. 84, pp. 29–41

Lane, F. W. (1948). *The Elements Rage*, Country Life, London

Langbein, W. B. (1949). Annual floods and the partial-duration flood series. *Trans. Am. Geophys. Un.*, **30,** 879–81

——(1960). Plotting positions in frequency analysis. *U.S., Geol. Surv., Water-Supply Paper*, No. 1543-A, 48–51

Larson, C. L., and Reich, B. M. (1973). Relationships of observed rainfall and runoff intervals. *Floods and Droughts, Proceedings of the 2nd International Hydrology Symposium*, 11–13 *September* 1972, Water Resource Publications, Fort Collins, Colorado, pp. 34–43

Linsley, R. K., Kohler, M. A., and Paulhus, J. L. H. (1949). *Applied Hydrology*, McGraw-Hill, New York

Lloyd, E. H. (1970). Return periods in the presence of persistence. *J. Hydrol.*, **10,** 291–8

Lowery, M. D., and Nash, J. E. (1970). A comparison of methods of fitting the double exponential distribution. *J. Hydrol.*, **10,** 259–75

Majumdar, K. C., and Sawhney, R. P. (1965). Estimates of extreme values by different distribution functions. *Water Resour. Res.*, **1,** 429–34

Markowitz, E. M. (1971). The chance a flood will be exceeded in a period of years. *Water Resour. Bull.*, **7**, 40–53

Matalas, N. C. (1967). Mathematical assessment of synthetic hydrology. *Water Resour. Res.*, **3**, 937–45

Matalas, N. C., Slack, J. R., and Wallis, J. R. (1975). Regional skew in search of a parent. *Water Resour. Res.*, **11**, 815–26

Miller, J. F. (1973). Probable maximum precipitation—the concept, current, procedures and the outlook. *Floods and Droughts, Proceedings of the 2nd International Hydrological Symposium*, 11–13 *September* 1972, Water Resource Publications, Fort Collins, Colorado, pp. 50–61

Moran, P. A. P. (1957). The statistical treatment of flood flows. *Trans. Am. Geophys. Un.*, **38**, 519–23

Myers, V. A. (1969). The estimation of extreme precipitation as the basis for design flood—resume of practice in the United States. *Floods and their Computation*, vol. 1, International Association of Scientific Hydrology, Belgium, No. 84, pp. 84–104

Nash, J. E., and Shaw, B. L. (1966). Flood frequency as a function of catchment characteristics. *Proceedings of the Symposium on River Flood Hydrology, March* 1965, Institution of Civil Engineers, London, session C, paper 6

Natural Environmental Research Council (1975). *Flood Studies Report*, Natural Environment Research Council, London

Panchang, C. M. (1969). Improved precision of future high floods. *Floods and their Computation*, vol. 1, International Association of Scientific Hydrology, Belgium, No. 84, pp. 51–9

Powell, R. W. (1943). A simple method of estimating flood frequencies. *Civ. Eng.*, **13**, 105–6

Quimpo, R. G. (1967). Stochastic model of daily river flow sequences. *Colo. St. Univ., Fort Collins, Hydrol. Papers*, No. 18

Sangal, B. P., and Biswas, A. K. (1970). The 3-parameter lognormal distribution and its application in hydrology. *Water Resour. Res.*, **6**, 505–15

Santos, A., Jr. (1970). The statistical treatment of flood flows. *Water Power*, **22**, 63–7

Schulz, E. F., Koelzer, V. A., and Mahmood, K. (eds) (1973). *Floods and Droughts, Proceedings of the 2nd International Hydrology Symposium*, 11–13 *September* 1972, Water Resource Publications, Fort Collins, Colorado

Schuster, J. (1973). A simple method of teaching the independence of $\overline{X}$ and s^2. *Am. Statist.*, **27**, 29–30

Singh, K. P., and Sinclair, R. A. (1972). Two-distribution method for flood-frequency analysis. *J. Hydraul Div., Proc. Am. Soc. Civ. Eng.*, **98** (HY1), 29–44

Slack, J. R., Wallis, J. R., and Matalas, N. C. (1975). On the value of information to flood frequency analysis. *Water Resour. Res.*, **11**, 629–47

Stripp., J. R., and Young, G. K., Jr. (1971). Plotting positions for hydrologic frequencies. *J. Hydraul. Div., Proc. Am. Soc. Civ. Eng.*, **97** (HY1), 219–22

Tennessee Valley Authority (1961). *Floods and Flood Control*, Tennessee Valley

Authority, Knoxville, Tennessee

Todorovic, P. (1978). Stochastic models of floods. *Water Resour. Res.*, **14,** 345–56

Todorovic, P., and Rouselle, J. (1971). Some problems of flood analysis. *Water Resour. Res.*, **7,** 1144–50

Treiber, B., and Plate, E. J. (1975). A stochastic model for the simulation of daily flows. *Proceedings of the International Symposium and Workshop on the Application of Mathematical Models in Hydrology and Water Resources System*, International Association of Scientific Hydrology, Bratislava, preprints

Tribus, M. (1969). *Rational Descriptions, Decisions and Designs*, Pergamon, New York

Wallis, J. R., Matalas, N. C., and Slack, J. R. (1974). Just a moment! *Water Resour. Res.*, **10,** 211–19

——(1976). Effect of sequence length *n* on the choice of assumed distribution of floods. *Water Resour. Res.*, **12,** 457–71

Watson, G. S. (1954). Extreme values in samples from *M*-dependent stationary stochastic processes. *Ann. Math. Statist.*, **25,** 798–800

Weisner, C. J. (1970). *Hydrometeorology*, Chapman and Hall, London

Weiss, G. (1977). Shot noise models for synthetic generation of multisite daily streamflow data. *Water Resour. Res.*, **13,** 101–8

Whipple, G. C. (1916). The element of chance in sanitation. *J. Franklin Inst.*, **182,** 205–27

Whittaker, J. (1973). Discussion on 'Relationship of observed rainfall and runoff recurrence intervals' by C. L. Larson and B. M. Reich. *Floods and Droughts, Proceedings of the 2nd International Hydrology Symposium*, 11–13 *September* 1972, Water Resource Publications, Fort Collins, Colorado, pp. 108–9

Wilk, M. B., Gnanadesikan, R., and Hugett, M. J. (1962). Probability plots for the gamma distribution. *Technometrics*, **4,** 1–20

Wilson, E. M. (1974). *Engineering Hydrology*, 2nd edn, Macmillan, London

World Meteorological Organisation (1973). Manual for estimation of probable maximum precipitation. *Oper. Hydrol. Rep.*, No. 1, *World Meteorol. Organ., Geneva, Publ.*, No. 332.

——(1974). Guide to hydrometeorological practices. *World Meteorol. Organ., Geneva, Publ.*, No. 168

Yevjevich, V. (1968). Misconceptions in hydrology and their consequences. *Water Resour. Res.*, **4,** 225–32. (1969). Comments and reply. *Water Resour. Res.*, **5,** 535–41