Nordic Hydrology, 9, 1978, 31-42

Analysis of Hydrologic Data Using Pearson Type III Distribution

V. M. Shaligram and V. S. Lele
Central Water & Power Research Station (CWPRS),
Khadakwasla, Pune (Poona) 411024, India

Computation of reliable and precise long term estimates from the available short term hydrologic records is often a challenging task for the design engineers/hydrologists.

Peak flow magnitudes relating to 'sixteen' streams were utilised for deriving long term estimates using Maximum likelihood method (Gumbel 1941 and Panchang et al. 1962). The results revealed that most of peak flows were underestimated and their departures from the respective prototype magnitudes were of the order of 10 to 40 per cent.

These peak flow magnitudes were graduated by the Pearson type III distribution with the result that a perfect calibration (departure within 1 per cent only) was achieved for three streams. For remaining streams the departures were reduced and were within 20 per cent.

The precision of the estimates, deduced from the use of the Pearson type III distribution, was ascertained by evaluating the confidence intervals for these estimates. For three cases the confidence intervals (deduced from the latter distribution) were smaller than their counterparts deduced from the Gumbel's distribution. For the remaining streams' data, the confidence intervals were nearly double those obtained with the Gumbel's distribution. Thus to achieve conformity between prototype and model (using the Pearson type III distribution), one has to be content with even slightly less precise estimates, but which are, otherwise, realistic.

Introduction

In early designs of hydraulic structures such as a dam or a weir across a stream, the magnitudes of hydrologic data were, generally estimated on the basis of certain empirical formulae. These empirical relations often yielded estimates which were either too large or too small, thereby making the design uneconomical. Hence the problem of computing reliable and precise long term estimates of hydrologic data has ever remained a challenging task for the design engineer/hydrologist. An attempt has been made, on the basis of the analysis of peak flow data, to describe a method for proper calibration of the short term as well as long term estimates of the hydrologic data.

Lacuna in the empirical formulae, used in earlier designs, was that they were lacking in themselves the concept of frequency of return period of a hydrologic data. Gumbel (1941) had, for the first time, postulated this concept and had put it in use to the hydrologic data in the succeeding years.

According to Gumbel, the frequency f of the hydrologic data, ranked m in a total of n years is given as:

$$f = \frac{m}{n+1} \qquad (1)$$

Eq. (1) pre-supposes that all the available hydrologic records are first arranged in the descending magnitudes and each record is then assigned a rank. It is possible to express a return period T, of the hydrologic data (which is reciprocal of the frequency, f), from Eq. (1) as follows:

$$T \equiv \frac{n+1}{m} \qquad (2)$$

Eq. (2) states that the hydrologic data, ranked m, (in a total of n years), will occur once in T years.

Extreme Value Distribution

Since each record of hydrologic data emerges as at the highest 365 daily magnitudes recorded during a year, their distribution may therefore be considered as the extreme value distribution. Gumbel (1958) has developed the function of the extreme value distribution which has the following form:

$$P(x)dx = a\, e^{-a(x-u)}\, e^{-e^{-a(x-u)}}\, dx \qquad (3)$$

Eq. (3) contains the Naperian constant e (= 2.71828) and two parameters a and u which may be evaluated using the recorded hydrologic data on streams. Integration of the distribution function (3) yields:

$$P_0 = e^{-e^{-a(x_0 - u)}} \tag{4}$$

Eq. (4) thus yields the probability P_0 of any one year's hydrologic data say x, being smaller than x_0. In terms of the return period T, the Eq. (4) takes the form:

$$x = u - \frac{1}{a} \ln \left(\ln \frac{T}{T-1}\right) \tag{5}$$

From the knowledge of the parameter values a and u, it is possible to compute the estimated value of a hydrologic data for a desired return period (T), utilising Eq. (5).

Maximum Likelihood Method

A number of methods are available for deducing the parameters a and u in Eq. (5). However, when these parameters are deduced by the method of maximum likelihood, Eq. (5) yields estimates, which are precise. Panchang et al. (1962) have decribed, in detail, the derivation and the computational procedure involved in deducing the parameter values a and u. Only the important computational steps are described below:

Since the evaluation of the parameter a involves an iterative procedure, an intelligent start is made by assuming an initial value for a as:

$$a = \frac{\pi}{\sqrt{6}} \frac{1}{s} \tag{6}$$

Eq. (6) contains the standard deviation(s) of the n values of the peak flow (x's). Using this value for the parameter a, the functions $f(a)$ and $f'(a)$ are evaluated as:

$$f(a) = \sum x_i e^{-ax_i} - \left(\bar{x} - \frac{1}{a}\right) \sum e^{-ax_i} \tag{7}$$

$$f'(a) = - \sum x_i^2 e^{-ax_i} + \left(\bar{x} - \frac{1}{a}\right) \sum x_i e^{-ax_i} - \frac{1}{a^2} \sum e^{-ax_i} \tag{8}$$

In Eqs. (7) and (8) each summation has been carried over the n values of peak flow (x's), whose mean is $\bar{x}$. An increment, h, to the starting value of a is then computed:

$$h = -\frac{f(a)}{f'(a)} \tag{9}$$

Second approximation for the parameter a is then:

$$a' = a + h \tag{10}$$

By replacing the value of a by a', Eqs. (7), (8) and (9) are recomputed. The iterative procedure is continued till the value of the increment (h) falls below the desired level of accuracy. This is, generally achieved, in four to five iterative steps.

From the knowledge of the final value of a (derived in the preceeding sub-para) the parameter u is deduced uniquely as:

$$u = \frac{1}{a} [\ln n - \ln \sum e^{-ax_i}] \qquad (11)$$

Summation in Eq. (11), also extends over the n values of x's as stated earlier.

From the known values of the parameters a and u, one can avail oneself of Eq. (5) to compute the estimated value (X_T) of peak flow for the desired return period T. Precision of such estimate (X_T) is given as follows:

$$S_{X_T} \equiv \frac{1}{a\sqrt{n}} [1 + \frac{6}{\pi^2} (1-C-R)^2]^{\frac{1}{2}} \qquad (12)$$

Eq. (12) contains C as the Euler's constant (=0.5772) and R, a function of the desired return period T, (R=ln ln $T/(T$-1).

Results of Computation

Adopting the hydrologic data of peak flow, recorded on sixteen streams, the values of the parameters a and u were computed. The values of return period T were computed using Eq. (2) by substituting therein $m=1$ (for the first highest record) and $m=2$ (for the second highest record). With these values of the return periods and the pre-determined values of the parameters a and u Eq. (5) was utilised for deriving the estimated peak flow values (x's). Results of these computations are listed in Table 1. In addition, Table 1 lists the following:

- prototype magnitudes for the first highest records,
- percentage departures of the estimates from the prototype magnitudes,
- confidence intervals for the estimates,

Analysis

From Table 1, one notes the following:

- the departures are within 10 per cent for the data on four streams,
- the departures are between 11 and 20 per cent for the data on five streams,
- for the Sabarmati at Dharoi, the departure is 60 per cent, for remaining six streams, these are between 21 and 40 per cent.

Table 1 - Model and prototype conformity of peak flow magnitudes (m^3/s). Figures following the estimates refer to their respective confidence intervals. Figures in parantheses refer to percentage depatures from first highest record.

S. No.	River-site	Return period (years)	Prototype 1st highest record	Deduced from the model of	
				Extreme value	Pearson type III
1	Baitarni-Akhuapada	85	9203	8921+ 587 (-3)	9153+ 1096 (-1)
2	Krishna-Vijawada	66	30044	26488+1515 (-12)	28141+ 3101 (-6)
3	Godavari-Doulaishwaram	54	80136	55022+3709 (-31)	64700+11103 (-19)
4	Penner-Nellore	53	13564	10769+1051 (-21)	13675+ 2716 (+1)
5	Sutlej-Bhakra	45	9203	7867+ 638 (-15)	8054+ 1026 (-12)
6	Sone-Dehri	33	34235	37237+3740 (+9)	32231+ 2981 (-6)
7	Mahanadi-Naraj	33	42334	55065+4867 (+30)	43885+ 2696 (+4)
8	Yamuna-Tajewala	33	15942	10214+1113 (-36)	12656+ 2899 (-20)
9	Ravi-Madhopur	31	17472	10420+1168 (-40)	14854+ 3384 (-15)
10	Hatmati-Himatnagar	29	1702	1095+ 154 (-36)	1397+ 339 (-18)
11	Mahanadi-Sambalpur	28	27396	32228+2541 (+18)	27653+ 1377 (+1)
12	Damodar-Rhondia	25	14767	13469+1563 (-9)	12598+ 1604 (-15)
13	Mutha-Khadakwasla	22	3211	2557+ 255 (-20)	2809+ 483 (-13)
14	Ohio-Mill Creek Reading (USA)	21	164	168+ 20 (+2)	152+ 53 (-7)
15	Sabarmati-Dhario	19	8240	3266+ 561 (-60)	4371+ 1788 (-47)
16	Tapi-Kathore	15	25485	22069+3559 (-13)	22864+ 4200 (-10)

Hence, for nearly 50 per cent of streams' data analysed, the departures are more than 20 per cent suggesting thereby, that these estimates are inadequately calibrated.

Hydrologic Data

From the foregoing analysis, it is clear that a distribution should be such that it would calibrate adequately the hydrologic data. Before turning towards any appropriate distribution, the behaviour of the hydrologic data was first studied. For each of the sixteen streams under study, the usual statistical parameters, namely, the mean ($\bar{x}$), the standard deviation(s) the coefficient of variation (Cv) and the coefficient of skewness (Cs) were evaluated. These values are listed in Table 2. Range of values of coefficient of variation was between 0.217 and 1.305: similar range for the coefficient of skewness was between 0.179 and 3.513. Ratios of the coefficients of skewness (Cs) to the coefficients of variation (Cv) are also given in Table 2. For the most of the streams, the value of this ratio is more than 2. This fact, therefore, indicates that, in general, the hydrologic data, under investigation, possess considerable skewness.

Table 2 - Statistical parameters derived from the recorded peak flows(x's) in m^3/sec mean ($\bar{x}$), standard deviation(s), coefficient of variation(Cv) and coefficient of skewness(Cs)

S. No.	River-site	Period	Total years	$\bar{x}$	s	Cv	Cs	Ratio: Cs/CV
1	Baitarni-Akhuapada	1874-1957	84	3591	1834	0.511	1.212	2.37
2	Krishna--Vijayawada	1894-1958	65	14765	4576	0.310	1.384	4.47
3	Godavari-Daulaishwaram	1905-57	53	29610	11509	0.389	2.198	5.65
4	Penner-Nellore	1903-54	52	3764	3486	0.926	1.639	1.51
5	Sutlaj-Bhakra	1912-55	44	3942	1649	0.418	1.076	2.57
6	Sone-Dehri	1920-54	32	17873	7385	0.413	0.179	0.43
7	Mahanadi-Naraj	1926-58	32	29258	8594	0.294	-0.387	-1.32
8	Yamuna-Tajewada	1927-59	32	4668	3204	0.686	2.006	2.92
9	Ravi-Madhopur	1929-58	30	4989	4003	0.802	2.213	3.44
10	Hatmati-Himatnagar	1922-49	28	396	435	1.098	1.654	1.51
11	Mahanadi-Sambalpur	1926-53	27	20231	4386	0.217	-0.272	-1.25
12	Damodar-Rhondia	1933-56	24	6690	2978	0.445	0.766	1.72
13	Muta-Khadakwasla	1940-60	21	1577	602	0.382	1.608	4.21
14	Ohio-Mill Creek, Reading (USA)	1941-60	20	90	35	0.594	0.442	1.14
15	Sabarmati-Dharoi	1935-52	18	1386	1809	1.305	3.513	2.69
16	Tapi-Kathore	1940-53	14	11610	6898	0.594	0.853	1.44

Skewed Distribution

For 'adequately' calibrated estimates, therefore, one must have a distribution which would appropriately account for the skewness inherent in the hydrologic data. A basic equation, relating peak flow magnitude x, to the return period, generally takes the form:

$$x = \bar{x} + K_s \qquad (13)$$

Eq. (13) involves the usual satistical parameters, the mean ($\bar{x}$), the standard deviation(s), and the frequency factor (K) which increases with the return period, with the functional relationship dependent on the assumed distribution. For the Pearson type III distribution, the relationship depends on the skewness (Cs) of the recorded data. Hence this distribution is suitable for present analysis. Central Technical Unit, Soil Conservation Service, USA (1968) have provided the new tables of percentage points of the Pearson type III distribution which related the factor K with the skewness. The variation of the factor K with the skewness, is shown in Fig.1, for seven specific values, (namely 2,5,10,25,50,100 and 200 years), of the return periods.

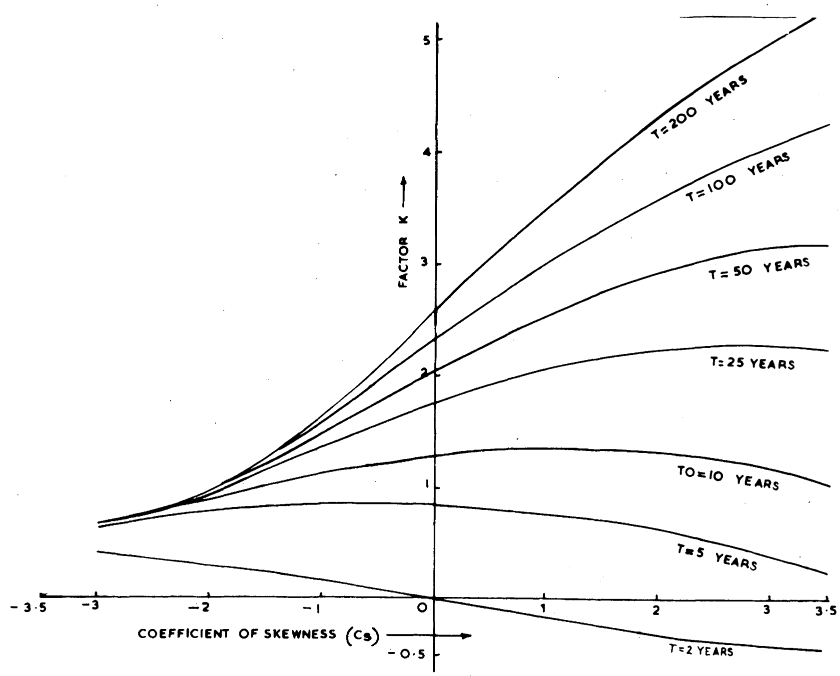

Fig. 1. Relationship between factor K and coefficient of skewness (Cs). Pearson type III distribution.

Revised Calibration

Following the procedure similar to the one described in the section on results of computation, the return period values, computed for $m=1$ and $m=2$ were used. For these return period values, the frequency factor K was suitably interpolated for the pre-determined values of the skewness (Cs) from Fig.1. Eq. (13) was then used for computing the estimated peak flow magnitudes for each of the sixteen streams. Results of these computations are listed in Table 1. Table 1 also gives the percentage departures of the estimates and the confidence intervals for the estimates deduced from the Pearson type III distribution.

Confidence intervals for the estimates based on this distribution are computed as follows:

$$S'_{X_T} = \frac{S}{\sqrt{n}} \left[1 + \frac{K^2}{2} \left(1 + \frac{3}{4} C_s^2\right) + K\, C_s \right]^{\frac{1}{2}} \qquad (14)$$

where s and Cs have the same meaning as described earlier and K is determined from the knowledge of the return period T. Bobee (1973) has shown that the confidence intervals derived from the expression (14) are only approximate and has provided a table for the correction factor K_1, which also varies with the return period T as well as with the skewness (Cs). Thus the corrected values for the confidence intervals were derived by multiplying the rough estimate S'_{X_T} with K_1.

Results of computations reported in Table 1 revealed the following:

- perfect calibration is achieved (departures being within 1 per cent) in three cases,
- the departures are between 6 to 20 per cent for the data on twelve streams,
- for the Sabarmati at Dharoi, the departure is 47 per cent.

Improving Goodness of Fit

Comparison of the estimates and their confidence intervals (as deduced from the extreme value distribution and the Pearson type III distributiion), which are listed side by side in Table 1 suggests the following:

- departures between the estimates and the prototype magnitudes are smaller when deduced from the Pearson type III distribution
- for the three streams data, the confidence intervals are smaller for estimates deduced from the Pearson type III distribution
- for remaining streams data the confidence intervals deduced from the Pearson type III distribution are nearly 1.5 to 2.5 times the values derived from the extreme value distribution.

Thus, the use of Eq. (13), which is based on the distribution of the Pearson type III has improved the goodness of fit, for the estimates within the range of records. Furthermore, these estimates are 'adequately' calibrated since their departures from the prototype magnitudes are, mostly within 20 per cent.

Model and Prototype Conformity

Foregoing analysing, has revealed that within the range of records, the use of the Pearson type III distribution yielded estimates which were adequately calibrated and were also reliable. However, these estimates are in general not as precise as those deduced from the extreme value distribution. The computations were then extended to obtain the long term estimates say for the return periods of 50, 100 and 200 years. These results, as deduced from both types of distributions, are given in Table 4. A comparison of the long term estimates with the corresponding first highest recorded magnitudes reveal the following:

- the values derived from the extreme value distribution are so much under estimated for certain streams that the prototype magnitudes (for which return periods are within 85 years) would be expected to occur once in 200 years.
- the estimates on two streams, (Sabarmati at Dharoi and Ravi at Madhopur) are so small that their prototype magnitudes could be expected to occur once in 6330 years and 991 years respectively; corresponding magnitudes for the Pearson type III distribution are worked out at 85 years and 41 years respectively.

The above analysis thus, suggests that at the risk of having wider bands of confidence intervals for the estimates of peak flows one can always have appropriately calibrated and reliable estimates, when deduced from the Pearson type III distribution. Similarly inference can be drawn from the analysis reported in Table 3 which is based on the second highest recorded peak flows.

Conclusions

For deducing the long term estimates from the available short term hydrologic records, the extreme value distribution, first postulated by Gumbel, is still in vogue. These estimates, were found to be inadequetly calibrated within the range of records. Since, the hydrologic data, in general, is skewed, the goodness of fit of the prototype and model values, was improved when the Pearson type III distribution was adopted for the calibration. Such adequately calibrated estimates

Table 3 - Model and prototype conformity of peak flow magnitudes (m^3/s) Figures following the estimates refer to their respective confidence intervals Figures in parantheses refer to percentage departures from second highest records.

S. No.	River-site	Return period (years)	Prototype 2nd highest record	Deduced from the model of Extreme value	Pearson type III
1	Baitarni-Akhuapada	42.5	8496	7714+ 488 (-9)	8179+ 488 (-4)
2	Krishna-Vijayawada	33.0	27043	24192+1303 (-11)	25526+2484 (-6)
3	Godavari-Doulaishwaram	27.0	58692	49715+3167 (-17)	56153+7090 (-6)
4	Panner-Nellore	26.5	13394	9272+ 897 (-31)	11493+1879 (-14)
5	Sutlaj-Bhakra	22.5	7815	7000+ 541 (-14)	7197+ 730 (-8)
6	Sone-Dehri	16.5	29110	32556+3124 (+12)	29523+2234 (+1)
7	Mahanadi-Naraj	16.5	40918	48973+4078 (+20)	4120+2071 (+2)
8	Yamuna-Tajewala	16.5	12743	8981+ 932 (-31)	10437+1889 (-18)
9	Ravi-Madhopur	15.5	16027	9757+ 976 (-39)	12211+2422 (-24)
10	Hatmati-Himatnagar	14.5	1291	908+ 128 (-28)	1121+ 244 (-13)
11	Mahanadi-Sambalpur	14.0	26901	29177+2122 (+9)	26353+1170 (-2)
12	Damodar-Rhondia	12.5	11964	11643+1293 (-3)	11138+1259 (-7)
13	Mutha-Khadakwasla	11.0	2977	2268+ 210 (-24)	2430+ 351 (-18)
14	Ohio-Mill Creek, Reading (USA)	10.5	160	151+ 17 (-6)	137+ 41 (-14)
15	Sabarmati-Dhoroi	9.5	2464	2651+ 457 (+8)	3104+1348 (+26)
16	Tapi-Kathore	7.5	21464	18351+2859 (-14)	19216+ 309 (-10)

Table 4 - Long-term peak flow estimates (m^3/s) deduced from distributions of the extreme value and the Pearson type III.
Figures following the estimates refer to their respective confidence intervals.

S. No.	River-site	Return period (years)	Deduced from the model of Extreme value	Pearson type III
1	Baitarni-Akhuapada	50	7931+ 510	8416+ 909
	(1874-1957)	100	8852+ 595	9380+ 1166
		200	9769+ 651	10324+ 1441
2	Krishna-Vijyawada	50	25569+1416	27118+ 2806
	(1894-1958)	100	27857+1643	29687+ 3646
		200	30136+1869	32219+ 4539
3	Godavari-Doulaishwaram	50	54438+3653	63786+10858
	(1905-57)	100	59711+4191	72240+14945
		200	64965+4729	80741+19387
4	Penner-Nellors	50	10642+1048	13502+ 2629
	(1903-54)	100	12129+1189	15651+ 3549
		200	13610+1359	17781+ 4474
5	Sutlej-Bhakra	50	7996+ 651	8187+ 1052
	(1912-55)	100	8854+ 736	9006+ 1338
		200	9710+ 850	9804+ 1064
6	Sone-Dehri	50	40017+4106	33735+ 3214
	(1920-54)	100	44628+4701	36015+ 3877
		200	49223+5324	38132+ 4572
7	Mahanadi-Naraj	50	58679+5381	45083+ 2931
	(1926-58)	100	64740+6995	48281+ 3975
		200	70658+6996	48281+ 3975
8	Yamuna-Tajewala	50	11042+1218	14004+ 2777
	(1927-59)	100	12414+1416	16229+ 4934
		200	13781+1586	18454+ 6344
9	Ravi-Madhopur	50	11401+1303	16894+ 4399
	(1929-58)	100	12816+1501	19847+ 6062
		200	14226+1699	22817+ 7869
10	Hatmati-Himatnagar	50	1240+ 170	1613+ 869
	(1922-48)	100	1423+ 198	1883+ 608
		200	1606+ 227	2150+ 769
11	Mahanadi-Sambalpur	50	34749+1677	29111+ 2379
	(1926-53)	100	37751+3342	29543+ 1990
		200	40724+3767	30404+ 2310
12	Damodar-Rhondia	50	15269+1841	13947+ 2161
	(1933-56)	100	17045+2124	15231+ 2690
		200	18833+2379	16462+ 3246
13	Mutha-Khadakwasla	50	2892+ 309	3254+ 720
	(1940-60)	100	3173+ 355	3621+ 949
		200	3456+ 401	3985+ 1194
14	Ohio-Mill Creek Reading	50	195+ 25	171+ 72
	(1941-60)	100	217+ 28	184+ 87
		200	239+ 32	196+ 104
15	Sabarmati-Dharoi	50	4104+ 708	7093+ 3921
	(1935-52)	100	4700+ 821	8983+ 6059
		200	5293+ 906	10900+ 8481
16	Tapi-Kathora	50	28328+4786	28696+ 6891
	(1940-53)	100	31887+5494	31794+ 8390
		200	35432+6230	34783+10617

sometimes possess confidence intervals which are wider than those deduced from the Gumbel distribution. Even then these estimates are to be accepted on the basis of their calibration.

Acknowledgements

The authors are deeply indebted to Shri V.S. Thakar, Research Officer, Central Water & Power Research Station, Pune, for kindly going through the manuscript and making valuable suggestions.

References

Bobee, B., (1973) Sample of T-Year Events Computed by Fitting a Pearson Type 3 Distribution, *Journal of Water Resources Research, vol. 9,* No.5.

Central Technical Unit, Soil Conservation Service, USA, (1968) New Tables of Percentage Points of the Pearson Type III Distribution.

Gumbel, E.J., (1941) The Return Period of Flood Flows *Annals of Mathematical Statistics, vol. XIII,* pp. 163-190.

Gumbel, E.J., (1958) *Statistics of Extremes,* pp. 29-37, Columbia University Press, New York.

Panchang, G.M., and Aggarwal, V.P. (1962) Peak flow Estimation by Method of Maximum Likelihood, Technical Memorandum HL02, CWPRS, POONA (PUNE).

First received: 11 October, 1977
Revised version received: January 23, 1978

Address:
Shri V.M. Shaligram,
690 Budhwar Peth,
Bajirao Road,
Pune 411 002 (India),

V.S.Lele,
697 Narayan Peth,
Laxmi Road,
Pune 411 030 (India),