International Seminar on Hydrology of Extremes 1-3 December, 1988
Roorkee (India)

VERSATILE FLOOD FREQUENCY METHODS WITH ENHANCED RELIABILITY

K. P. Singh
Principal Scientist
Illinois State Water Survey, Champaign, Illinois, U.S.A.

SYNOPSIS

Many observed annual flood series exhibit reverse curvatures when plotted on lognormal probability paper. None of the existing frequency distributions fits such a series. The physical rationale for the mixed-distribution concept is developed for fitting such shapes. Outliers and inliers greatly distort the fitted distribution parameters. Statistics for their detection at various significance levels are developed from extensive Monte Carlo experiments. A versatile flood frequency methodology with objective detection and modification of any outliers/inliers is developed and computerized. The use of mixed distribution and modification of outliers/inliers are shown to significantly reduce the uncertainty in high flood estimates.

1.0 INTRODUCTION

1.1 Many observed annual flood series exhibit reverse curvatures when plotted on lognormal probability paper. None of the most commonly used distributions in flood frequency analyses (Pearson and log-Pearson, normal and lognormal, and Gumbel and log-Gumbel) fits an observed flood series with reverse curvature. The occurrence of these curvatures may be attributed to seasonal variation in flood-producing storm types, dominance of within-the-channel or floodplain flow, and variability in antecedent basin soil moisture and cover conditions.

1.2 Distribution parameters can be significantly biased if the observed flood series has outliers and inliers. Their presence at both the high and low ends of the flood spectrum needs to be detected and objectively modified so that design flood estimates with the least bias can be computed.

1.3 Confidence limits are derived to evaluate the uncertainties inherent in computed or fitted flood frequency curves. These limits provide a measure of uncertainty of the discharge at a selected exceedance probability. The confidence band width for high flood estimates can be reduced or their reliability can be improved by objective and systematic detection and modification of any outliers at various levels of significance.

2.0 MIXED DISTRIBUTION

2.1 The magnitude of a flood peak depends largely on storm and basin characteristics. Storm characteristics of interest are the type of storm (e.g., hurricanes, thunderstorms, and frontal or air mass storms) and intensity and duration of storm. High intensity and short duration storms usually cause much higher peak flow than low intensity and long duration storms with the same total precipitation. Basin characteristics mainly include relative dominance of within-the-channel or floodplain flow, the antecedent soil moisture condition, and the vegetal cover.

2.2 The effect of channel versus floodplain flow is explained as follows. Bankfull discharge in a river corresponds to about a 2-year or median flood. With an increase in flow, the water spreads over the floodplain. For low depths of inundation, the mean velocity of the composite flow section is much lower than at the bankfull discharge. The mean velocity slowly increases with increase in depth of flooding and may surpass the bankfull flow velocity. This addition of a new storage element can lead to flattening and subsequent rise of the flood probability curve at a different slope than for the floods within the channel. Conversion of storm rainfall to runoff is largely affected by the antecedent soil moisture condition and vegetal cover.

2.3 The magnitude of annual flood peaks depends on a large number of factors which vary from season to season and within a season. The interaction between the distributions of these factors may produce a flood series resembling a conventional distribution shape or one exhibiting marked reverse curvature to be dealt with by the mixed-distribution concept. Mixed distributions have been considered in terms of rainfall and snowmelt floods [12] and floods caused by hurricanes and other storms [4].

2.4 The developed mixed distribution model considers the observed annual floods (or their logarithms) to belong to two populations with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and relative weights a and $(1-a)$:

$$p(x) = a\, p_1(x) + (1-a)p_2(x)$$

$$p_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[- \frac{(x'-\mu_1)^2}{2\sigma_1^2} \right] dx'$$

$$p_2(x) = \frac{1}{\sigma_2\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[- \frac{(x'-\mu_2)^2}{2\sigma_2^2} \right] dx'$$

in which p is the probability of being equal to or less than x, and x = log Q where Q is the annual flood. The five parameters of the mixed distribution model: μ_1, μ_2, σ_1^2, σ_2^2, and a, are a function of the mean, standard deviation, and skew of the observed annual flood series [5]. The model considers an observed annual flood series as essentially composed of two component distributions. Use of three or more component distributions increases the complexity of the problem and makes it intractable. The parameters can be estimated using a nonlinear programming algorithm [7].

3.0 OUTLIER DETECTION AND MODIFICATION

3.1 Barnett and Lewis [1] define an outlier in a set of data as an observation or a subset of observations that appears to be inconsistent with the remainder of that set of data. When the values of the highest observed floods of an annual flood series are much higher or lower than expected, these values are designated as outliers and inliers, respectively. When the values of the lowest observed floods are much higher or lower than expected, these are designated as inliers and outliers, respectively. Analyses of storms causing outliers at the high end and of droughts causing outliers at the low end can provide a physical rationale for the presence of outliers.

3.2 A procedure has been developed with an objective methodology for successive detection and modification of any outliers and inliers at both ends of the flood spectrum at 0.01, 0.05, 0.1, 0.2, 0.3 and 0.4 levels of significance, from experiments on millions of normally distributed numbers. The developed test statistic is termed a departure:

$$\text{Departure, } \Delta_i = Z_i - Z_{si}$$

in which Z_i is the theoretical standard normal deviate and Z_{si} is the sample standardized deviate corresponding to the plotting position, p_i:

$$p_i = (m_i - \alpha) / (n + 1 - 2\alpha)$$

in which p is the probability of nonexceedance, m is the rank order for the flood with series ranked from low to high, and $\alpha = 0.38$ [2]. The values of the departures for the 5 highest and 5 lowest floods in an annual flood series at various probability levels are shown in Figure 1a. The 5 highest floods are ranked from high to low and the 5 lowest floods are ranked from low to high (designated by 1, 2, 3, 4, and 5).

3.3 The observed annual flood series needs to be transformed to resemble a series distributed as $N(\mu,\sigma^2)$. This is achieved with the power transformation [3]:

$$y_i = (Q_i^{\lambda} - 1)/\lambda \quad \text{if } \lambda \neq 0, \text{ and}$$

$$y_i = \log Q_i \quad \text{if } \lambda = 0$$

in which Q is the annual flood, y is the transformed flood, and λ is the transformation parameter. The parameter λ can be estimated by using the maximum log-likelihood method [6]. The skew of the transformed series y is very close to zero but the kurtosis may be different from that for a normal distribution. The adjustment values for various values of kurtosis are given for symmetric distributions by Singh and Nakashima [7].

3.4 The detection and modification procedure begins from window 1 (Figure 1b) or significance or probability level 0.01. Any outliers/inliers detected are modified at that level. The resulting detransformed series is analyzed to test the departures and the procedure is repeated, if necessary, to ensure that there are no outliers/inliers at the 0.01 significance level. The procedure is followed sequentially from one level to the next and desired distribution statistics and floods are computed before moving to the next level. If no outliers/inliers are detected at a level, no modifications are done for that level. Windows 1, 2, 3, 4, 5, and 6 correspond to significance level (SL) of 0.01, 0.05, 0.10, 0.20, 0.30, and 0.40 or their complements, respectively.

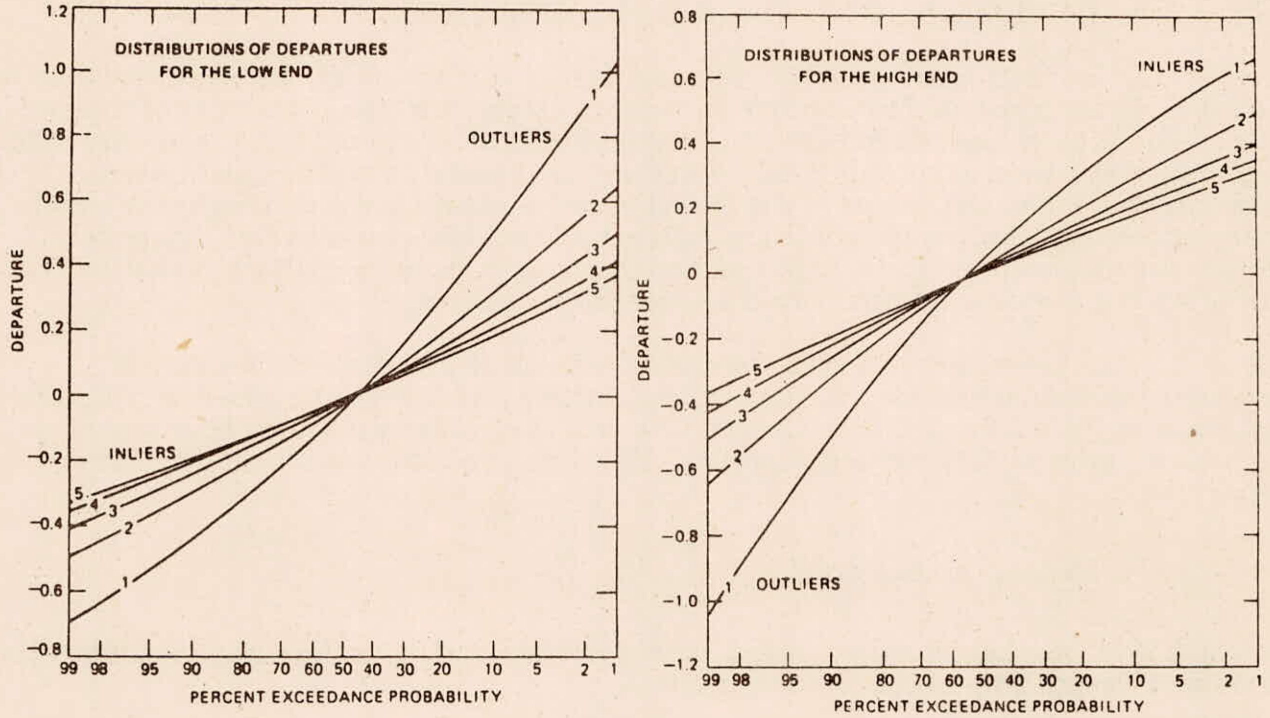

Figure la. Distribution of departures for the low and high end

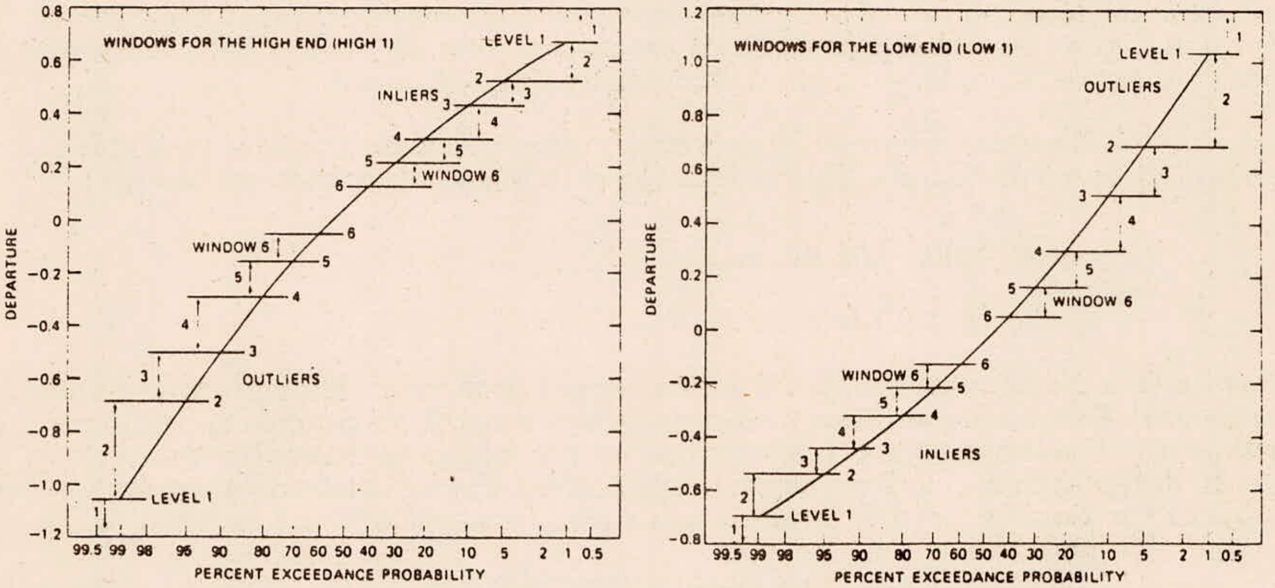

Figure lb. Levels and windows for the outliers and inliers

4.0 VERSATILE FLOOD FREQUENCY METHODOLOGY

4.1 A computer package has been developed [7,8] for objectively detecting and modifying any outliers/inliers sequentially from levels 1 to 6 and for computing various frequency floods with three methods: power transformation, log-Pearson type III, and mixed distribution method. The U.S. Water Resources Council [11] recommends log-Pearson type III for general use.

4.2 In the power transformation method, the T-yr flood is computed from:

$$Q_T = (\lambda y_T + 1)^{\frac{1}{\lambda}} \qquad \text{in which } y_T = \overline{y} + Z_T s_y$$

in which $\overline{y}$ and s_y are the mean and standard deviation of the power transformed y series and Z_T is the standard deviate corresponding to the nonexceedance probability $(1 - 1/T)$. The Z_T is obtained for both kurtosis = 3 (i.e., normal distribution) and sample kurtosis (i.e., symmetric but not normal distribution).

4.3 With log-Pearson type III, the T-yr flood is obtained from

$$Q_T = (\overline{x} + k_T s)^{10}$$

in which $\overline{x}$ and s are the mean and standard deviation of the log-transformed floods, and k_T is the frequency factor which depends on skew and recurrence interval, T.

4.4 With the mixed distribution, the x for a desired value of p is obtained by finding p for a trial value of x from the equation:

$$p\,(x) = a\ p_1(x) + (1\text{-}a)\ p_2(x)$$

and converging to the desired value of x through a fast reiterative process. The program package prints the various frequency floods, modifications of any detected outliers/inliers at various levels of significance, and distribution statistics for the three methods. These results help in deciding the best-fit distribution and in selecting the storms and droughts causing floods which are perceived as high and low outliers.

5.0 EXAMPLES

5.1 Four annual flood series from rivers in Japan, Poland, Czechoslovakia, and the U.S.A. (10) were selected to evaluate the results obtained with the mixed distribution method. The relevant information: location, years of record n, and drainage area DA in sq km, are given below:

No.	*River and gaging station location*	n, *years*	*DA, sq km*
1	Fuji River, Shimizubata, Japan	51	5110
2	Dunajec River, Nowy Sacs, Poland	50	4340
3	Zdechovka River, Zdechov, Czechoslovakia	45	4.08
4	Beetree Creek, Swannanoa, U.S.A.	45	14

5.2 Three highest (H1, H2, and H3) and three lowest (L1, L2, and L3) floods, without modification SL=0 and with modification SL=0.3, fitted power transformation or PT, log-Pearson type III or LP3, and mixed distribution or MD statistics, and 100-yr and 1000-yr floods with these methods are given in Table 1.

5.3 Fuji River, Shimizubata, Japan. All 51 observed annual floods were caused by rainfall and occurred during June to October, except one in April. Three highest floods occurred in August-September and the lowest flood in April. No high floods were detected as outliers/inliers upto 0.3 significance level. L2 and L3 were detected as inliers and modified. The fitted MD curve together with the two component distributions as well as the observed flood series and modification of inliers at the low end are shown in Figure 2.

5.4 Dunajec River, Nowy Sacs, Poland. Out of the 50 annual floods, 34 resulted from rainfall and 16 from snowmelt. The rainfall floods occurred in May through October and snowmelt floods in February to April. The highest 4 floods occurred in June and July. The second and third highest floods were modified from 3300 to 3064 and 2680 to 2600 m^3/sec. The fitted MD curve together with the two component distributions, the observed annual floods, and modifications are shown in Figure 2. For this river, the weights are practically proportional to the number of snowmelt and rainfall floods.

5.5 Zdechovka River, Zdechov, Czechoslovakia. Out of 45 annual floods, 28 resulted from rainfall and 17 from snowmelt. The rainfall floods cccurred in May through November but 19 of them occurred in June and July. The snowmelt floods occurred in December through April. The nine highest floods were all rainfall floods. The second, third, and fourth highest floods were detected as outliers at SL=0.3 and slightly modified with the exception of the second highest flood. The fitted MD curve together with the two component distributions, the observed annual floods, and outlier modifications are shown in Figure 2.

5.6 Beetree Creek, Swannanoa, U.S.A. All 45 annual floods were caused by rainfall. They occurred in all months, with maximum numbers of 9, 7, and 5 in March, December, and September, respectively. The highest flood of 38.8 m^3/s was detected as an outlier at 0.2 significance level and modified to 36.1 at 0.3 level. The three lowest floods were detected as outliers at 0.2 level and modified to values shown in Table 1. The fitted MD curve together with the two component distributions, the observed annual floods, and modifications are shown in Figure 2.

6.0 FLOOD ESTIMATE RELIABILITY ENHANCEMENT

6.1 Monte Carlo experiments were conducted to evaluate any reduction in uncertainty in flood estimate that can be achieved by objective detection and modification of any outliers/inliers. Five hundred samples (N=500) of X were generated, distributed as n(μ,σ^2), for different sample sizes n, μ, and σ. Six example sets [9] are discussed hereafter.

Set No.	*1*	*2*	*3*	*4*	*5*	*6*
N	500	500	500	500	500	500
n	25	50	100	50	50	50
μ	3.0	3.0	3.0	3.0	3.0	4.0
σ	0.3	0.3	0.3	0.2	0.4	0.3

Table 1. Highest and Lowest Floods and Fitted Distribution Statistics

Basin		1		2		3		4	
SL		0	0.3	0	0.3	0	0.3	0	0.3
Item									
H1		7300	7300	3300	3300	14.7	14.7	38.8	36.1
H2		5600	5600	3300	3064	13.0	11.0	23.6	23.2
H3		5000	5000	2680	2600	8.2	7.7	17.1	17.1
L3		400	376	186	186	0.73	0.73	2.9	3.2
L2		400	312	138	138	0.60	0.60	2.3	3.0
L1		210	210	120	120	0.60	0.57	1.9	2.6
LP3	μ	3.139	3.136	2.855	2.854	0.269	0.266	0.830	0.835
	σ	0.356	0.361	0.345	0.342	0.330	0.325	0.247	0.233
	g	-0.059	-0.096	-0.203	-0.224	0.900	0.831	0.658	0.912
PT	m	8.207	8.858	9.460	9.893	1.957*	2.050*	2.927*	1.825*
	s	1.049	1.226	1.555	1.671	0.063*	0.073*	0.084*	0.016
	kt	2.397	2.413	2.742	2.708	2.636*	2.654*	4.200*	3.315*
	λ	0.034	0.054	0.103	0.115	-0.463*	-0.435*	-0.289*	-0.530
MD	a	0.488	0.563	0.192	0.303	0.689	0.665	0.539	0.721
	μ_1	2.869	2.907	2.420	2.517	0.121	0.117	0.761	0.757
	μ_2	3.396	3.431	2.958	3.001	0.596	0.559	0.911	1.038
	σ_1	0.245	0.270	0.229	0.258	0.191	0.192	0.122	0.151
	σ_2	0.236	0.223	0.281	0.261	0.338	0.334	0.321	0.282
Q_{100}	LP3	8964	8917	4021	3925	17.60	16.24	33.24	33.71
	PT1	8840	8732	3956	3855	26.00	22.35	31.74	36.26
	PT2	7522	7496	3754	3637	18.02	17.60	39.50	40.08
	MD	7623	7530	3875	3732	16.66	15.42	36.30	35.19
Q_{1000}	LP3	16238	15937	6614	6371	52.23	45.48	67.53	73.09
	PT1	15828	15366	6479	6232	1537.89	332.57	66.22	131.96
	PT2	10995	10914	5756	5464	144.94	91.96	145.79	231.45
	MD	11903	11557	6422	6011	32.92	30.05	67.18	62.68

Notes: H and L are highest and lowest floods in m^3/s
SL = significance level
PT1 with kurtosis kt equal to 3 as normal distribution
PT2 with sample kurtosis
* statistics apply to floods multiplied by 100

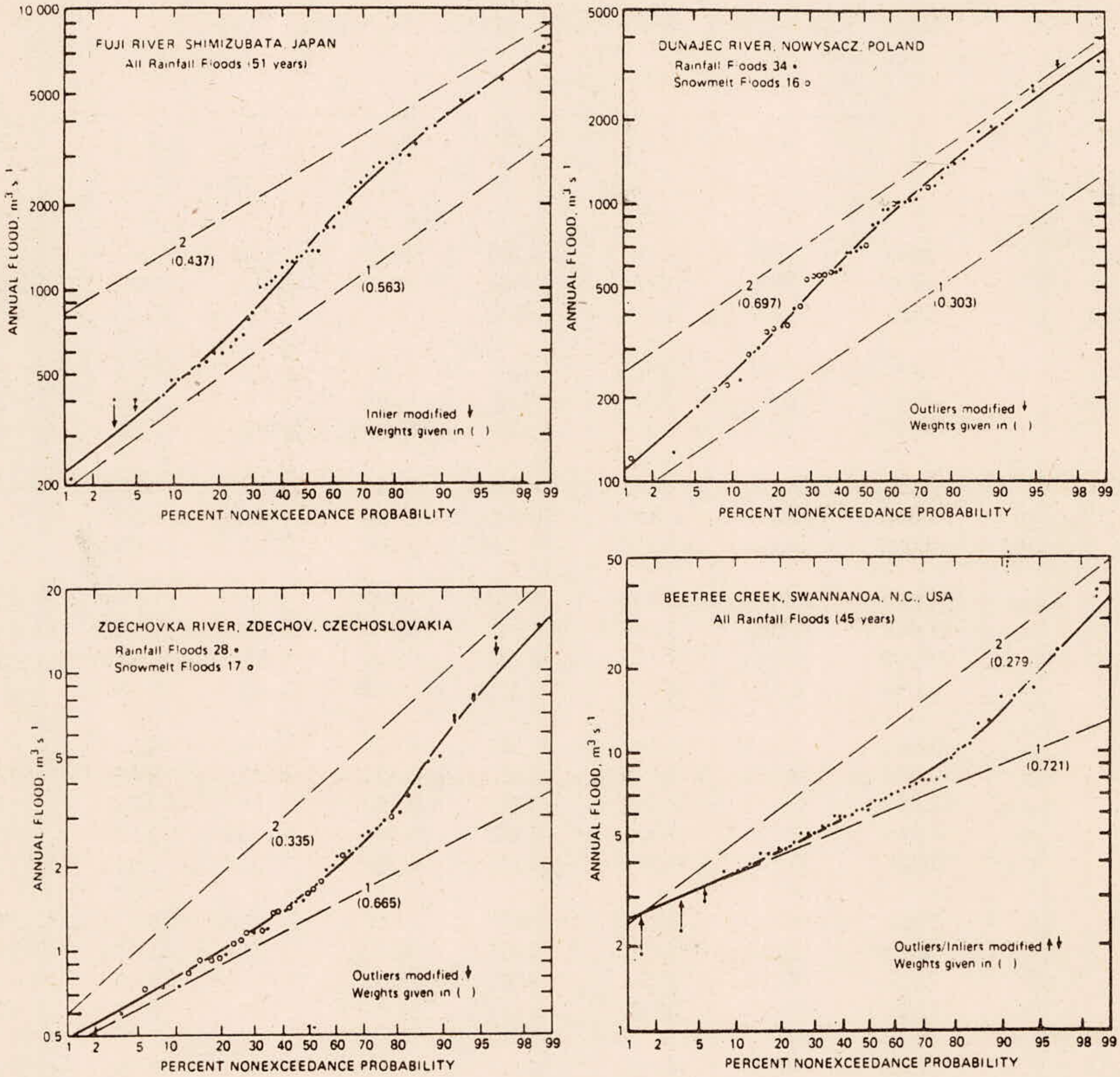

Figure 2. Observed and modified floods and the fitted mixed distribution

Values of X_T (or log Q_T) for T=2, 10, 25, 50, 100, 500, and 1000 years were obtained with each of the three methods for each sample in a set for level 0 (without any detection and modification of outliers/inliers) as well as levels 1 through 6. Each sample of 500 X_T values was analyzed to develop estimates of mean, standard deviation, skew, and kurtosis, as well as the minimum and maximum values of X_T.

6.2 The variation in standard deviation of X_T from the 6 sets was investigated in terms of sample size n, population mean μ and square root of variance σ. The relevant results for SL=0 and 0.3 are given in Table 2 for T=25, 100, 500, and 1000 years. With power transformation, 4 to 15 values of X_T for T=500 and 1000 years exceeded 8.0 whereas none exceeded 6.0 with MD or LP3. Very high values can occur with the power transformation method for high values of T combined with low values of n. Some combinations of Y_T and λ values can cause abnormally high values of Q_T and X_T.

6.3 The reduction in standard deviation of X_T from SL = 0 to SL = 0.3 greatly increases with increase in T from 100 to 1000 years with the mixed distribution method. There is very little or no improvement with the log-Pearson type III. The use of mixed distribution method with objective detection and modification of any outliers/inliers will significantly reduce the uncertainty in high flood estimates.

6.4 A detailed analysis of the results with the mixed distribution showed that 1) standard deviation of X_T for $T \leq n/2$ is practically unaffected by any detection and modification of outliers/inliers; 2) standard deviation of X_T decreases with increase in SL for a given T for T>n and increases with increase in T for a given SL; 3) for t>n and SL = 0.3, the detection and modification of outliers/inliers generally yields values of skew and kurtosis close to that for normal distribution; 4) with increasing level of modification, the maximum value decreases and minimum value increases, though by a lesser amount, indicating the effect of modification of high and low outliers; and 5) mean value of X_T decreases somewhat with increase in SL and T. It is obvious that the band width typified by the maximum and minimum values of X_T at SL = 0.3 decreases at an increasing rate from those at SL = 0 or 0.01 with increase in T for T>n. Thus, the uncertainty in high flood estimates can be greatly reduced by detection and modification of outliers/inliers and use of mixed distribution. However, the test statistics for developing confidence bands need to be obtained from extensive Monte Carlo experiments.

10

Q_T or (X_T) for Set No. 5

	SL = 0.01			SL = 0.30		
T,years	*Mean*	*Min*	*Max*	*Mean*	*Min*	*Max*
50	6,546	2,472	22,803	6,295	2,624	15,776
100	8,375	2,799	26,242	7,907	2,999	22,284
500	13,863	3,606	79,799	12,474	3,926	44,157
1000	16,827	3,972	160,694	14,859	4,355	57,016

7.0 CONCLUSIONS

1. Mixed populations and hence mixed distributions of annual floods can occur because of a host of factors such as different types of flood-producing storms, dominance of within-the-channel or floodplain flow, antecedent basin soil moisture and vegetal cover conditions, and rainfall and snowmelt floods.

Table 2. Standard Deviation of X_T for Six Data Sets and SL = 0 and 0.3

T	*SL*	*Set*	*1*	*2*	*3*	*4*	*5*	*6*
Mixed Distribution								
25	0		0.109	0.079	0.057	0.053	0.106	0.079
	0.3		0.110	0.079	0.058	0.053	0.106	0.079
100	0		0.147	0.112	0.081	0.075	0.148	0.111
	0.3		0.134	0.099	0.073	0.067	0.132	0.099
500	0		0.182	0.151	0.112	0.101	0.200	0.151
	0.3		0.163	0.119	0.089	0.080	0.160	0.119
1000	0		0.197	0.167	0.124	0.113	0.222	0.167
	0.3		0.173	0.128	0.096	0.086	0.171	0.127
Log-Pearson Type III								
25	0		0.101	0.076	0.054	0.051	0.101	0.074
	0.3		0.100	0.075	0.054	0.050	0.100	0.074
100	0		0.144	0.108	0.078	0.072	0.144	0.107
	0.3		0.141	0.105	0.079	0.070	0.141	0.104
500	0		0.200	0.150	0.110	0.100	0.200	0.150
	0.3		0.194	0.146	0.111	0.097	0.194	0.145
1000	0		0.226	0.169	0.124	0.113	0.226	0.170
	0.3		0.218	0.164	0.125	0.109	0.218	0.163
Power Transformation, kurtosis = 3								
25	0		0.110	0.077	0.054	0.051	0.103	0.074
	0.3		0.112	0.077	0.056	0.051	0.103	0.075
100	0		0.184	0.117	0.081	0.078	0.156	0.111
	0.3		0.188	0.116	0.083	0.078	0.155	0.111
500	0		0.296	0.176	0.119	0.117	0.234	0.170
	0.3		0.334	0.205	0.124	0.136	0.274	0.168
1000	0		0.371	0.214	0.139	0.142	0.285	0.202
	0.3		0.350	0.210	0.144	0.140	0.280	0.199
Power Transformation, sample kurtosis								
25	0		0.109	0.077	0.054	0.051	0.102	0.075
	0.3		0.110	0.076	0.055	0.050	0.101	0.074
100	0		0.174	0.120	0.085	0.080	0.160	0.117
	0.3		0.180	0.109	0.081	0.073	0.145	0.107
500	0		0.352	0.191	0.132	0.127	0.254	0.187
	0.3		0.284	0.163	0.118	0.109	0.218	0.161
1000	0		0.384	0.230	0.154	0.154	0.307	0.225
	0.3		0.326	0.193	0.136	0.129	0.257	0.190

2. A method has been developed and computerized to consider a mixed distribution as consisting of two normal or lognormal distributions.

3. An objective method for detection and modification of any outliers and inliers at various significance levels in a flood series has been developed and computerized. Observed annual flood series is power-transformed for application of the method.

4. A versatile flood frequency methodology has been developed and computerized for objective detection and modification of any outliers/inliers at various significance levels and for computation for various floods with power transformation, log-Pearson type III, and mixed distribution methods.

5. The results of application to flood series from various countries show the versatility and superiority of the mixed distribution method in objectively detecting and modifying any outliers/inliers and yielding reliable flood estimates.

6. The range of uncertainty for very high flood estimates is greatly reduced, or the reliability of flood estimate is greatly enhanced, with the mixed distribution and objective detection and modification of any outliers and inliers. Statistics for the confidence limits with outlier/inlier detection and modification and mixed distribution method can be developed from extensive Monte Carlo experiments.

8.0 REFERENCES

1. Barnett, V. and T. Lewis (1978), "Outliers in statistical data," John Wiley and Sons, New York, 365 pages.

2. Blom, G. (1978), "Statistical estimates and transformed beta variables," John Wiley and Sons, New York, 176 pages.

3. Box, G.E.P. and D.R. Cox (1964), "An analysis of transformation," Journal of Royal Statistical Society, Series B, Vol. 26, pp. 211-252.

4. Canterford, R.P. and C.L. Pierrehumbert (1977), "Frequency distributions for heavy rainfalls in tropical Australia," in Hydrology Symposium, Institution of Civil Engineers, Brisbane, Australia.

5. Cohen, A.C. (1967), "Estimation in mixtures of normal distributions," Technometrics, Vol. 9, No. 1, pp. 15-28.

6. Singh, K.P. (1980), "Discussion on flood frequency by power transformation," American Society of Civil Engineers Hydraulics Journal, Vol. 106, No. HY3, pp. 462-465.

7. Singh, K.P. and M. Nakashima (1981), "A new methodology for flood frequency analysis with objective detection and modification of outliers," Illinois State Water Survey Contract Report 272, Champaign, Illinois, 145 pages.

8. Singh, K.P. (1983), "Software design for a versatile flood frequency analysis," Advances in Engineering Software, Vol. 5, No. 2, pp. 107-112.

9. Singh, K.P. (1986), "Flood estimate reliability enhancement by detection and modification of outliers," in Stochastic and Risk Analysis in Hydraulic Engineering, Water Resources Publications, Colorado, pp. 253-265.

10. UNESCO (1976), "World catalogue of very large floods," The Unesco Press, Paris, 424 pages.

11. U.S. Water Resources Council (1981), "Guidelines for determining flood frequency," Bulletin 17B of the Hydrology Committee, Washington, D.C.

12. Waylen, P. and M.K. Woo (1982), "Prediction of annual floods generated by mixed processes," Water Resources Research, Vol. 18, No. 4, pp. 1283-1286.