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

MODERATING OUTLIERS IN RAINFALL FREQUENCY ANALYSIS

D. S. Upadhyay Surinder Kaur

India Meteorological Department.

SYNOPSIS

The high outliers in an extreme rainfall series are not be rejected as our interest lies in estimating the probable high intensity for design purposes. Maintaining the outliers in extreme series poses problem of finding a suitable distribution to describe the series. EV-II, the only befitting distribution in such cases tends to yield very high intensities for higher return periods, which is generally undesirable for practical use. As such the moderation of outlying observations seems to be an acceptable way out.

In this paper, the authors have theoretically examined the effect of i) rejecting; ii) moderating outliers on the return period values of rainfall and their standard errors. It has also been analytically deduced as to what can be the extent of minimum moderation which makes an outlier to fit in a suitable choosen extreme value distribution.

The results of derivation have been illustrated by practical exercises on daily extreme rainfall series containing outliers of different magnitudes.

1. INTRODUCTION

The outliers is an observation or a subset of observations which deviate markedly on the trend of remaining data set. The outliers (high or low) have a tendency to follow a steeper distribution function, that is why a single distribution is not able to describe the variability of a series containing outliers. In a frequency analysis outliers may be retained, modified or rejected depending on specific requirement of the situation. For the analysis of extreme rainfall, the rejection of the high outlier, which is an observed fact is not desirable. We should however be very clear about the effect of rejection or modification of the outlier on the estimation of parameters using data set provided. This is an important point. High outliers tend to over estimate X_T (T-year return period value) for larger T whereas low outlier under estimate X_T for smaller T. If we are analysing extreme rainfall for large hydrologic structures we are more concerned with high outliers in the series while for urban drainage studies of low outliers are more significant.

Barring errors in observations and processing the main cause of outlier is the large variability inherent in the nature of rainfall. Whether an observation is an outlier or not, also depends on the choice of distribution. In case of extreme rainfall series, some distributions like EV II or wakeby, generally give a good fit. But their choice is not recommended for operational studies owing to very fast rate of increase in X_T with T.

Thus, we should have a suitable test of outlier with reference to the distribution choosen for analysis. Many authors, such as Grubbs & Beck (1972), Dixon (1951) etc. have developed test statistics for outliers in series coming from normal population. Darlin (1952) gave a method in respect of uniform or chi-square distributions. Laurent (1963) and Basu (1965) suggested tests for outliers for the samples described by two parameter exponential distributions.

So far no rigrous procedure has been evolved for statistical testing of outliers in an extreme series. Some investigators used the criterian of 3 median to identify high outliers in such cases. However, if x is an EVI (x,u,α) variate, the transformations

$$y = e^{-\frac{(x-u)}{\alpha}}$$

may yield a series Y following exponential distribution. It may be noted that high outlier of X is transformed to a low outlier in series Y. The tests outlined by Basu (1965) as applicable to low outlier may be used.

While performing frequency analysis of short duration (3 to 24 hrs.) extreme rainfall data, are quite often come across the outliers of very high magnitude. In such cases, one can think of modifying them to bring closer to an EV distribution function rather than leave them to affect the parameters unreasonably. The theoretical basis is completely lost in applying a distribution functions to an outlier which falls markedly out of 2 or 3 standard error limits. It is with this aim we are proposing in this paper a practical method to modify an outlier to suit an operational frequency analysis approach. The effect of such modifications have also been examined.

2. THE EFFECT OF REJECTING OUTLIER:

Consider an extreme series $x_1, x_2, \ldots\ldots\ldots x_n$ where x_n is an outlier. If we apply EV1 distribution to the series using method of moment for estimation its parameters, the T-year return period value (X_T) is given by

$$X_T = \bar{x} + c.s \quad ......(1)$$

where, $c = -0.45 + 0.78\ Y_T$

& $\quad Y_T = -\ln\ln\left(\frac{T}{T-1}\right)$

$\bar{x}$ & s are the mean and the S.D. of the series respectively.

The standard error of X_T is given by

$$SE(X_T) = s\sqrt{\frac{0.71 + 0.12y_T + 0.65y_T^2}{n}} \qquad ....(2)$$

If the outlier x_n is removed from the series, $\bar{x}$ and s will undergo a change to $\bar{x}'$ and s' respectively, whose expressions may be given as:

$$\bar{x}' = \frac{n\bar{x} - x_n}{n-1} \qquad ....(3)$$

and, $$s' = \frac{\sqrt{n}}{n-1}\left[\sqrt{(n-1)s^2 - d^2}\ \right] \qquad ....(4)$$

where, $d = x_n - x$

Thus the revised value (X'_T) of X_T is given by

$$X'_T = X' + C.S'$$

It may be noted that C being a function of T alone will not undergo any change. Thus, the difference

$$Z_T = x_T - x'_T = (\bar{x} - \bar{x}') + c(s - s') ....(5)$$

or $$(n-1)Z_T = d + c[(n-1)s - \sqrt{n}.\sqrt{(n-1)s^2 - d^2}] ....(6)$$

$$= d + c[(n-1)s - \sqrt{n(n-1)}.s(1 - \frac{d^2}{(n-1)s^2})^{\frac{1}{2}}]$$

$$= d + c[(n-1)s - \sqrt{n(n-1)}.s(1 - \frac{d^2}{2(n-1)s^2})]$$

$$= d + \frac{C}{2s\ n-1}\left[2s^2(n-1)^{3/2} - \sqrt{n}\ \{2(n-1)s^2 - d^2\}\right]$$

Assuming $n \simeq (n-1)$, we get

$$(n-1).Z_T = d + c.\frac{d^2}{2s} \quad (+\ ve) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

$$\therefore \; x'_T < x_T$$

Differentiating Z_T with respect to T, we get

$$(n-1)\,\frac{dZ_T}{dT} = \frac{d^2}{2s}\cdot\frac{dc}{dT}$$

where, $\frac{dc}{dT} = \frac{0.78}{T(T-1)\ln(T/T-1)}$

Since $dZ_T/dT > 0$, Z_T is an increasing function of T. Further more

$$\frac{d^2Z_T}{dT^2} = \frac{0.78\,d^2}{2s}\,[\ln(\frac{T}{T-1})]^{-2}\,\frac{1}{T^2(T-1)^2}\,[1+(2T-1)\ln(1-\frac{1}{T})]$$

The term in last bracket can be approximated to $(-1 + \frac{1}{T})$, which, apparently is a -ve quantity.

$$\therefore \; \frac{d^2z_T}{dT^2} < 0$$

From these deductions, it may be stated that z_T increases with T at the decreasing rate.

Considering the effect of rejecting outlier on SE (X_T), it can be seen that

$$\frac{s'}{\sqrt{n-1}} < \frac{s}{\sqrt{n}}$$

From (2), $SE(X'_T) < SE\ (X_T)$.

A numerical illustration:

68 year annual 1-day extreme rainfall series recorded at Dharampur (Gujarat) contains an outlying observation of 987 mm (1941) against the second highest of 544 mm (1946). Means and standard deviations of original series (x,s) and the series without outlier (x',s') are given as :

$\bar{x}$ = 228 mm, $\bar{x}'$ = 217 mm

s = 130 mm, s' = 92 mm

The values of X_T, X'_T and their standard errors are provided in Table-1.

The appreciable fall in the values of standard error for all T'^s may be attributed to the reduction in standard deviation (s') by rejecting outlier. The significant diminishing of rainfall estimate (X'_T) is also noticed for higher T'^{s}.

TABLE - 1

Rainfall estimates alongwith their standard errors of annual maximum series of Dharampur in m.m.

T	Original Series		With rejected outlier		With moderated outlier	
	X_T	$SE(X_T)$	X'_T	$SE(X'_T)$	X''_T	$SE(X''_T)$
2	206	15	200	10	205	13
5	322	25	287	17	309	21
10	398	32	344	24	377	29
100	637	62	522	44	592	54
1000	872	92	698	65	804	79
5000	1036	112	820	80	951	97
10000	1107	121	874	86	1015	105

3. MODERATION OF OUTLIERS

The question of moderation rises because of the fact that X_T and X'_T both are unreasonable estimates of T-year rainfall; X_T due to poor fit of the distribution and X'_T due to omission of heaviest observed rainfall.

Some researchers suggest to reduce the weightage of x_n by combining it linearly with x_{n-1} and x_{n-2} such as

$$\hat{x}_n = \frac{x_{n-1} + x_n}{2}$$

or, $$\hat{x}_n = \frac{1}{3} \sum_{i=n-2}^{n} x_i$$

or, $$\hat{x}_n = a.x_{n-1} + b.x_n$$

This type of modification, however, has no logical basis.

In this paper, it is proposed that the modified value for outlier may be estimated as:

$$x_n = x + ks \quad ....(8)$$

The value of K is so choosen so that the series $x_1, x_2, \ldots\ldots x_{n-1}, x_n$ satisfies the following

i) $\bar{x} + Ks = X_T + i.SE(X_T)$ ....(9)

Where i = 0 and 1, depending on the magnitude of outliers. In respect of the outliers of low magnitude i = 2 may also yield desirable result. Rewriting eqn. (9) in terms of reduced variate (Y) and frequency factor (C), we get

$$x + Ks = x + Cs + \frac{is}{\sqrt{n}}\sqrt{0.71 + 0.12y + 0.67y^2}$$

or, $$K = C + i\sqrt{\frac{0.71 + 0.12y + 0.67y^2}{n}} \quad \ldots\ldots\ldots\ldots\ldots(10)$$

where C = $-0.45 + 0.78y$

ii) $F(X) = e^{-e^{-(\frac{x-u}{\alpha})}}$

If we consider the actual return period of observed rainfall from the concept of plotting position

$$T_m = \frac{n + 1}{m} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(11)$$

Where T_m is the return period of m^{th} observation of a series arranged in ascending order. The variation in T_m as given by equation (11) follows a harmonic progression. But if we examine the variations in T given by

$$T = \frac{1}{1 - F(x)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(12)$$

We see that the rate of increase in T is comparatively higher for larger values of X. As such there is sufficient reason to assume that T follows a geometric progression. Suppose the computed values of T using equation (12) corresponding to the four highest observed values in series are T_{n-3}, T_{n-2}, T_{n-1} and T_n.

Let $r_1 = \frac{T_{n-2}}{T_{n-3}}$, $\quad r_2 = \frac{T_{n-1}}{T_{n-2}}$, $\quad r_3 = \frac{T_n}{T_{n-1}}$

The geometric mean of these ratios is given by

$$r = (r_1, r_2, r_3)^{1/3}$$

or $\quad r = (T_n/T_{n-3})^{1/3}$ ....(13)

r may be regarded as the common ratio of G.P. mentioned above.

$$\hat{T}_n = r.T_{n-1} \quad ....(14)$$

Correspondingly Y can be evaluated by

$$F(y) = \frac{\hat{T}_{n-1}}{\hat{T}_n}$$

$$\text{or, } y = -l_n l_n \left(\frac{T_n}{T_{n-1}}\right) \quad ....(15)$$

K can be estimated using (10) and (15).

If the outlier x_n is replaced by $\hat{x}_n = \bar{x} + k.s$ in the series then x and s will be changed to x" and s" respectively, whose expressions may be given as

$$x'' = \frac{n\bar{x} - d + ks}{n} \qquad \text{where } d = x_n - \bar{x}$$

and $\quad n^2 s''^2 = n^2 s^2 (1 + k^2) - nd^2 - (d - ks)^2$

$$= n^2 s^2 - (d - ks)[(n + 1)ks + (n - 1)d]$$

Assuming $(n - 1) \approx n \approx (n + 1)$, we get

$$s'' = \sqrt{s^2 (1 + \frac{k^2}{n}) - \frac{d^2}{n}}$$

$$= s(1 + \frac{k^2}{n})^{\frac{1}{2}} [1 - \frac{d^2}{s^2 (n + k^2)}]^{\frac{1}{2}}$$

$$= s(1 + \frac{k^2}{2n})[1 - \frac{d^2}{2s^2 (n + k^2)}]$$

Thus, the revised values (x''_T) of x_T is given by

$$x''_T = \bar{x}'' + Cs'' \quad ....(16)$$

and, $\quad Z'_T = x_T - x''_T$

or, $\quad Z'_T = \frac{d - ks}{n} + Cs\ [1 - (1 + \frac{k^2}{2n})\ \{i - \frac{d^2}{2s^2 (n + k^2)}\}]$.....(17)

ILLUSTRATIONS

(i) One day extreme rainfall series of Dharampur (Gujarat) has been considered for the analysis. This series contains an ourlier (987 mm) which is more than 4 times the median value. In this series x_n = 987 mm, x_{n-1} = 544 mm, x_{n-2} = 499 mm, x_{n-3} = 402 mm. Median = 203 mm and N = 68 years.

Applying EVI distribution to this series, the rainfall estimates (X_T) alongwith their standard errors [$SE(X_T)$] have been computed and provided in Table-1. To see the goodness of fit of the observed sample to EVI distribution, plotting positions have been evaluated using Gringorton (1963) formula given as :

$$T_o = \frac{N + 0.12}{m - 0.44}$$

Where T_o is the return period corresponding to the observed value having m^{th} rank in a descending order series.

X_T and its confidence limits ($\pm$ 3 SE) have been plotted on a probability paper (Fig.1). It can be seen from Fig.1 that EVl does not provide a satisfactory fit to the observed highest value (987 mm) as it lies outside the confidence limits.

Applying EVI distribution to the series, the return periods corresponding to x_n, x_{n-1}, x_{n-2} and x_{n-3} are T_n = 4563, T_{n-1}=46.9, T_{n-2} = 29.8 and T_{n-3} = 10.6. Using these values in equation (13),

we get r = 7.4

From equation (14) & (15)

$\hat{T}_n$ = 7.4 x 46.9 = 349 and Y = 5.85

As the outlier has a large magnitude, we may take i = o.

From equation (10), we get

K = 4.1

Thereofore, the modified estimate of x_n is $\hat{x}_n$ = 762 mm.

With this modification, the EVI distribution is applied to the data and the results are provided in Table-1 and Fig.2. It can be seen from Fig.2 that the EVI distribution fits satisfactorily to the modified series as the moderated outlier (762mm) lies within the confidence limits. It is also apparent from Table-1 that the values of X''_T obtained with moderate outlier is more reasonable and acceptable.

ii) One day extreme rainfall series of Rajpipla (Gujarat) has been considered. The series contains an outlier (517 mm) of medium magnitude. In this series x_n = 517 mm, x_{n-1}=301 mm, x_{n-2} = 282 mm, x_{n-3} = 279 mm. Median = 121 mm and N = 68 years.

Applying EVI distribution to this series, the rainfall estimates (X_T) alongwith their standard errors [SE(X_T)] have been computed and provided in Table 2 and Fig.3. It can be seen from Fig.3 that EVI distribution does not provide a satisfactory fit to the observed series.

The return periods corresponding to x_n, x_{n-1}, x_{n-2} & x_{n-3} are

$T_n = 1197,\ T_{n-1} = 31,\ T_{n-2} = 23,\ T_{n-3} = 21$

$r = 3.8$

$T_n = r.T_{n-1} = 119$

and $Y_{119} = 4.77$

Using $i = 1,\ K = 3.8$

Hence the moderated value of outlier is given as $\hat{x}_n$ = 419 mm.

With this modification, the EVI distribution is applied to the data and the results are provided in Table-2 and Fig.4. It can be seen from Fig.4 that the EVI distribution fits satisfactorily to the modified series.

TABLE - 2

Rainfall estimates alongwith their standard errors of annual maximum series of Rajpipla in m.m.

T	Original series		Series with moderated outlier	
	X_T	SE(X_T)	$X"_T$	SE($X"_T$)
2	121	9	120	8
5	188	14	183	13
10	233	19	225	18
100	372	36	355	33
1000	509	53	482	48
5000	604	65	571	59
10000	646	71	610	64

CONCLUSIONS

This study presents a methodology for moderating outliers in rainfall frequency analysis so that the modified extreme series gives a realistic fits to the Gumbel's EV distribution. The logic is based on the pattern of variation between T and X_T. It is an observed fact that the computed return period values for the last 3 or 4 values of X in arranged series behave more like a geometrical progression than a harmonic progression as indicated by the plotting positions. Some of the salient conclusions of this study are provided below :

1) The difference between T-year return period values computed from the original series and the series after rejecting the outlier increases with T at decreasing rate.

2) The moderated value of outlier is given by $\hat{x}_n = \bar{x} + k.s$ where K can be expressed by equation (10) in section 3.

3) The frequency analysis may be carried out with modified series by recomputing the parameters of EVI distribution. The numerical illustrations carried out here show that the moderated outliers fall within 3 SE confidence bands.

4) The 1-day rainfall series of Dharampur has an outlier of very high magnitude, the highest rainfall is 987 mm, the second highest is 544 mm. The results of frequency analysis in 3 cases 1) original series; 2) the series with rejected outliers and 3) the series with moderated outliers are given in Table-1.

The advantage of moderating the series using the procedure described in this paper is apparent. When we compare these results :

a) original series does not fit to Gumbel's distribution.

b) the series with rejected outlier fits the distribution but the values of X_T are very low and have no impact of highest observed value.

c)The modified series fits the distribution without reducing X'_T.S undesirable.

ACKNOWLEDGEMENT

The authors are grateful to Dr. R.P. Sarker, Director General of Meteorology for the encouragement and his keen interest in this study. The authors are also thankful to Shri S.C. Sharma, Director and Shri A.K. Mehra, Assistant Meteorologist for valuable discussions.

REFERENCES

1. Basu, A.P., 1965, "On some tests of hypotheses relating to the exponential distribution when some outliers are present" J.Amer, Statist, ASJ., 60, 548-59.

2. Darlin, D.A., 1952 b,"On a test for homogeneity and extreme values", Ann.Math. Statist., 23, 450-456, correction 24, 135

3. Dixon, W.J., 1951, "Ratios involving extreme values",Ann. Math. Statist. 22, 68-78.

4. Grubbs, F.E. & Beck, G. 1972, "Extension of sample sizes and percentage points for significance tests of outlying observations", Technometrics 14, 847-854.

5. Laurent, A.G., 1963, "Conditional distribution of order statistics and distribution of the reduced ith order statistic of the exponential model". Ann. Math. Statist. 34, 652-657.

FIG. 4.

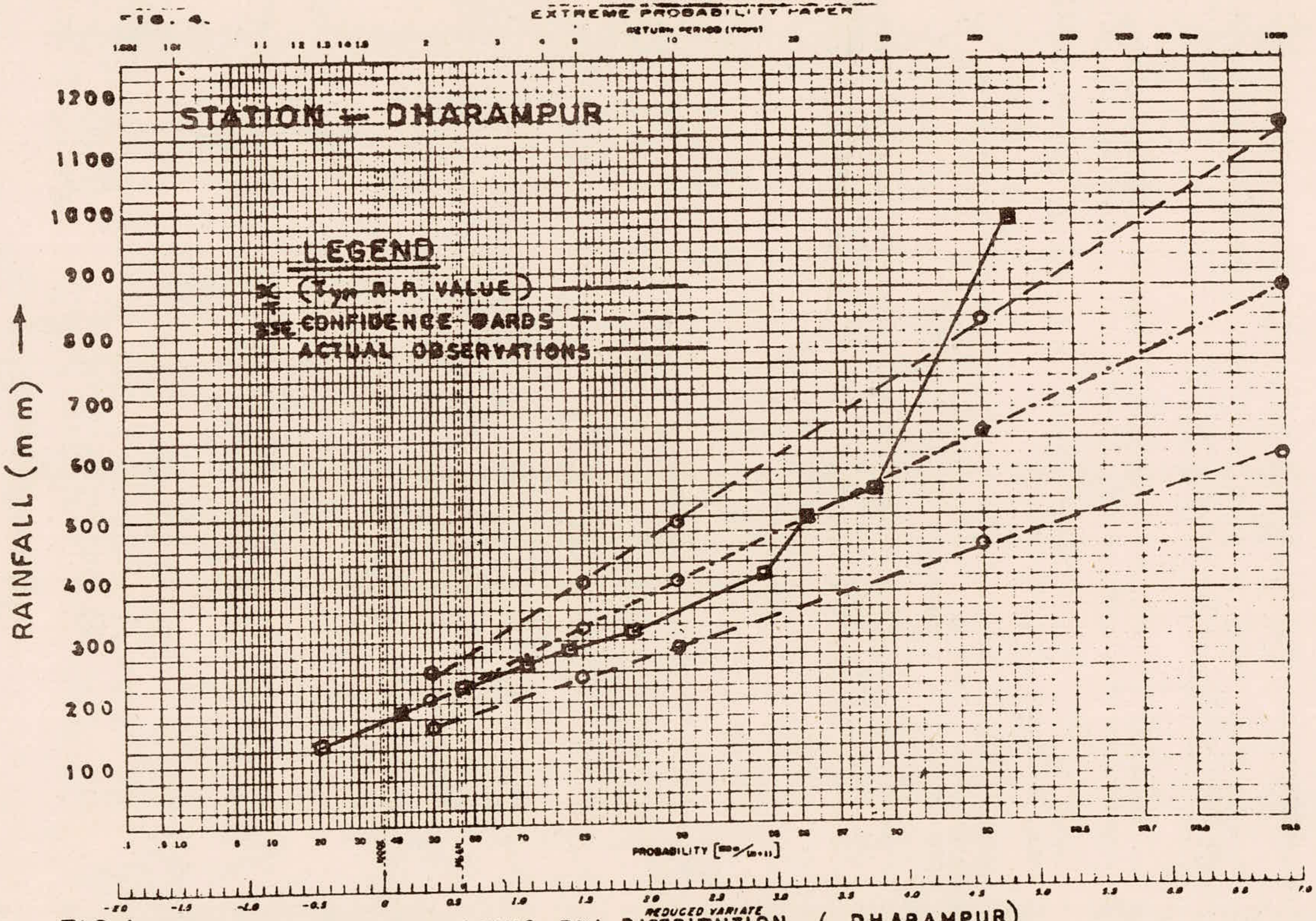

FIG:1 FREQUENCY ANALYSIS USING EVI DISTRIBUTION (DHARAMPUR)

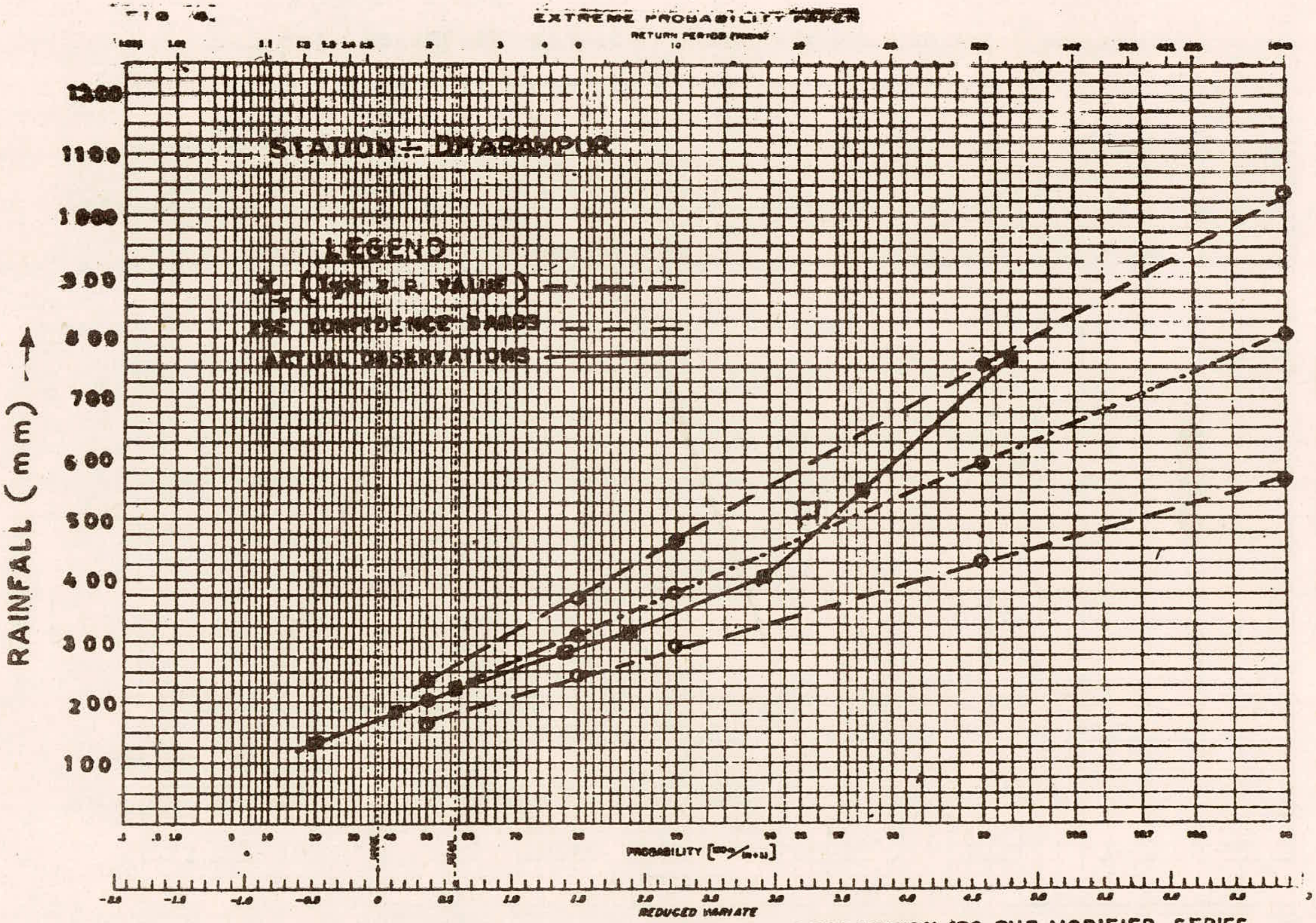

FIG :- 2 FREQUENCY ANALYSES USING EVI DISTRIBUTION TO THE MODIFIED SERIES (DHARAMPUR)

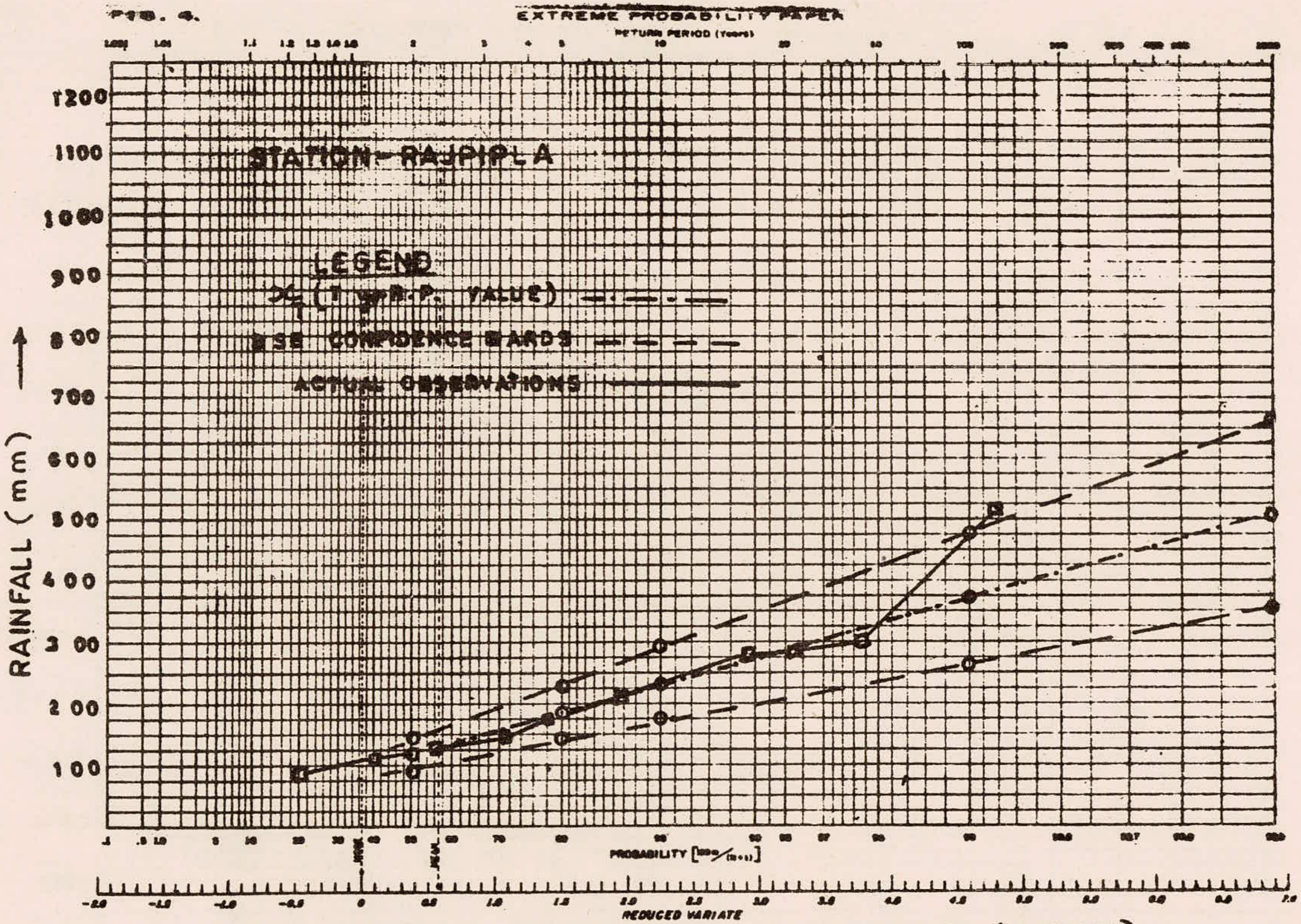

FIG.-3 FREQUENCY ANALYSIS USING EVI DISTRIBUTION (RAJPIPLA)

FIG. 4.

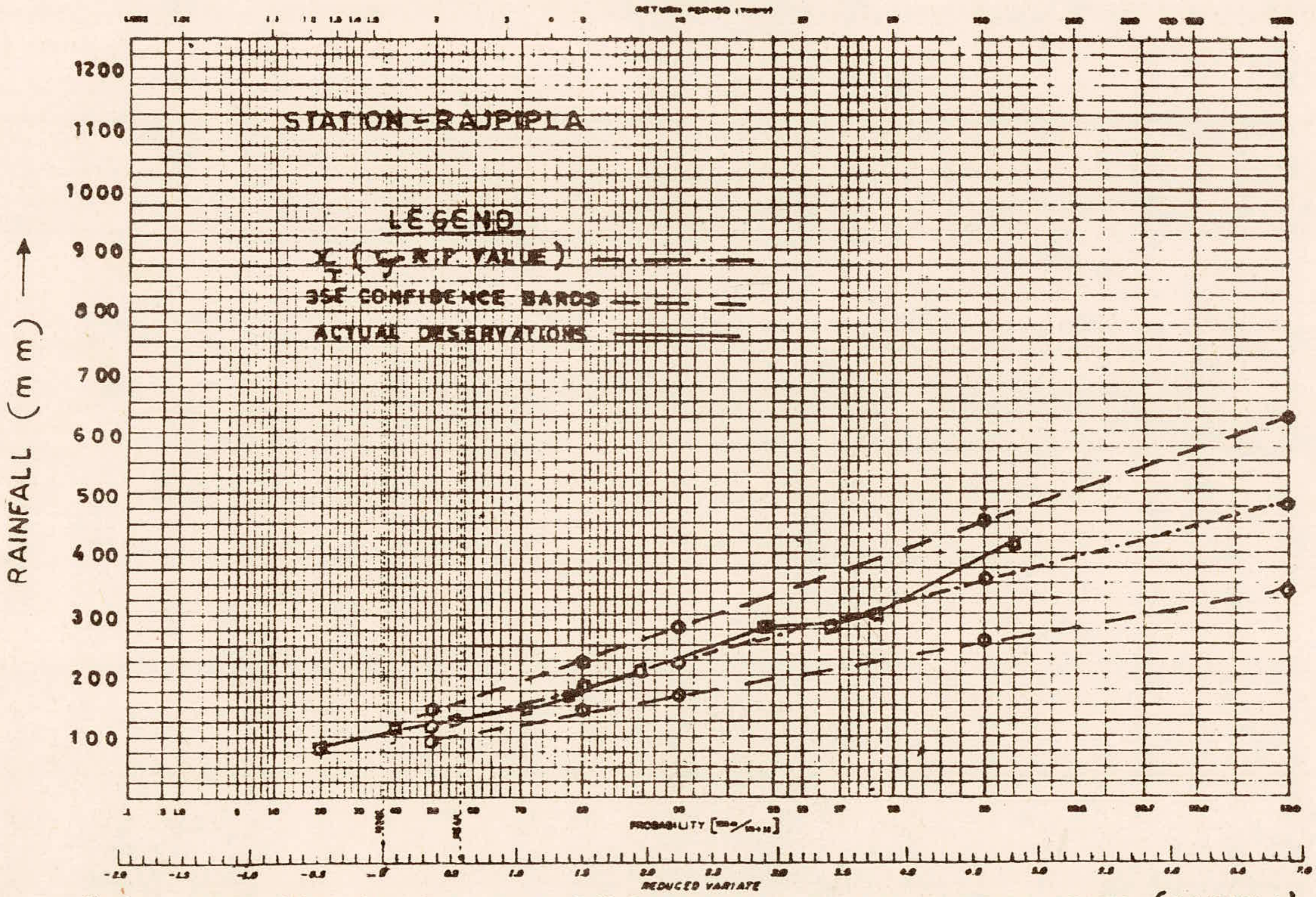

FIG-4 FREQUENCY ANALYSIS USING EVI DISTRIBUTION TO THE MODIFIED SERIES (RAJPIPLA)