International Journal of Research Studies in Science, Engineering and Technology
Volume 7, Issue 5, 2020, PP 13-18
ISSN 2349-476X

Selection of Best Fit of Extreme Value Family of Distributions for Frequency Analysis of River Flow Data

N. Vivekanandan*

Central Water and Power Research Station, Pune, Maharashtra, India

***Corresponding Author:** N. Vivekanandan, Scientist-B, Central Water and Power Research Station, Pune 411024, Maharashtra, India*

ABSTRACT

For planning, design and management of civil and hydraulic structures, estimation of Peak Flood (PF) for a particular return period is needed to be carried out. This can be achieved through Flood Frequency Analysis (FFA), which involves fitting probability distributions to the series of observed Annual Peak Flood (APF) data. This paper details a study on adoption of Extreme Value Family of Distributions (EVD) such as Extreme Value Type-1, Extreme Value Type-2, Generalized Extreme Value and Generalized Pareto for FFA. Parameters of the EVD are determined by method of moments, maximum likelihood method and L-Moments, and also used for estimation of PF. The adequacy of fitting EVD to the APF data series is evaluated by Goodness-of-Fit (viz., Chi-Square and Kolmogorov-Smirnov) and diagnostic (viz., mean absolute error and model efficiency) tests. This paper presents the procedures adopted in FFA with illustrative example and the results obtained from the study.

Keywords: *Chi-Square, Generalized Extreme Value, Kolmogorov-Smirnov, L-Moments, Mean Absolute Error, Model Efficiency, Peak Flood*

INTRODUCTION

Determination of the frequencies and magnitudes of those events are important for flood plain management and design of hydraulic structures, civil protection plans, etc. However, length of available records is not enough large to define the risk of peak flood, extreme rainfall, low-flow, drought, etc. In these cases, Flood Frequency Analysis (FFA) involves fitting probability distributions to the Annual Peak Flood (APF) data series is considered as an alternative tool to arrive at a design value [1].

A number of probability distributions belong to the normal, gamma and extreme value families of distributions will be generally adopted in FFA. The Normal family of distributions consists of Normal, 2-paramer Log-Normal (LN2), 3-parameter Log Normal (LN3) and Generalized Normal (GNO) while the Gamma family of distributions consists of Exponential, Gamma, Generalized Gamma, Pearson Type-3 (PR3) and Log Pearson Type-3 [2]. Likewise, Generalized Extreme Value (GEV), Extreme Value Type-1 (EV1), Extreme Value Type-2 (EV2) and Generalized Pareto (GPA) distributions are the members of Extreme Value family of Distributions (EVD) [3]. Generally, Method of Moments (MoM) is used in determining the parameters of the probability distributions. Sometimes, it is difficult to assess the exact information about the shape of a distribution that is conveyed by its third and higher order moments. Also, when the sample size is small, the numerical values of sample moments can be very different from those of the probability distribution from which the sample was drawn. It is also reported that the estimated parameters of distributions fitted by MoM are often less accurate than those obtained by other parameter estimation procedures viz., Maximum Likelihood Method (MLM), method of least squares and probability weighted moments [4]. To address these shortcomings, the application of alternative approach, namely L-Moments (LMO) is used for FFA [5]. Bhuyan et al. [6] applied generalized version of LMO (LH-moments) for Regional FFA (RFFA) of river Brahmaputra. They have found the RFFA based on the GEV distribution by using level one LH-moment give better results over LMO. It was reported by Malekinezhad et al. [7] that GEV (LMO) is better suited for modelling APF of three different regions in Iran. Badreldin and Feng [8] carried out the RFFA for the Luanhe basin using LMO and cluster techniques.

Haberlandt and Radtke [9] carried out FFA using APF data for three mesoscale catchments in northern Germany. Markiewicz et al. [10] adopted Generalized Exponential (GE) and inverse Gaussian distributions in frequency analysis of annual maximum flows for Polish rivers. They described that the GE occupies as front runner among all distributions commonly used for FFA of Polish data and can be included into the group of the alternative distributions. Kossi et al. [11] carried out RFFA for Volta River Basin (VRB) using LMO of five probability distributions. By using LMO diagrams and Goodness-of-Fit (GoF) test (i.e., Z-statistic), they found that the GEV and the GPA distributions are better suited to yield accurate flood quantile in VRB. Amr et al. [12] compared the performance of several parameter estimators of GPA distribution through Monte Carlo simulation. Kolbjorn [13] used APF data from four selected Norwegian catchments, and historical flood information to provide an indication of water levels for the largest floods in the last two to three hundred years. Ul Hassan et al. [14] applied the GEV, PR3, EV1, GLO (Generalized Logistic) and LN3 distributions for estimation of flood at five gauging sites of Torne River. Moreover, when different distributional models are used for FFA, a common problem that arises is how to determine which model fits best for a given set of data. This can be answered by formal statistical procedures involving Goodness-of-Fit (GoF) and diagnostic tests; and the results are quantifiable and reliable. Qualitative assessment is made from the plots of the observed and estimated PF. For quantitative assessment on discharge data within the observed range, Chi-square (χ^2) and Kolmogorov-Smirnov (KS) tests are applied. A diagnostic test of Mean Absolute Error (MAE) and Model Efficiency (MEF) is used for the selection of best fit probability distribution of EVD for estimation of PF. This paper presents a study on EVD adopted in FFA of river flow data and illustrates the applicability of GoF and diagnostic tests procedures in identifying which distribution is better suited for estimation of PF.

METHODOLOGY

The procedures involved in FFA of river flow data are: (i) prepare the observed APF data series from daily river flow data series; (ii) determination of parameters of EVD (viz., GEV, EV1, EV2 and GPA) by MoM, MLM and LMO; (iii) estimate the PF for different return periods by adopting EVD (iv) check the adequacy of fitting EVD through GoF and diagnostic tests to identify the best fit of EVD to arrive at a design value; and (v) analyse the FFA results and suggestions made thereof. Table 1 presents the Cumulative Distribution Function (CDF), quantile estimator (q_T) and estimators of the parameters of EVD [15] adopted in FFA. Theoretical descriptions of the determination of parameters of EVD by MoM, MLM and LMO are available in the text book titled 'Flood Frequency Analysis' by Rao and Hamed (2000).

Goodness-of-Fit Tests

GoF tests are essential for checking the adequacy of probability distributions to the APF data series in the estimation of PF. Out of a number GoF tests available, the widely accepted GoF tests are χ^2 and KS, which are used in the study. The theoretical descriptions of GoF tests statistic are given as below:

$$\chi^2 = \sum_{j=1}^{NC} \frac{(O_j(q) - E_j(q))^2}{E_j(q)} \quad \ldots (1)$$

where, $O_j(q)$ is the observed frequency value of q for j^{th} class, $E_j(q)$ is the expected frequency value of q for j^{th} class and NC is the number of frequency classes [16]. The rejection region of χ^2 statistic at the desired significance level (η) is given by $\chi^2_C \geq \chi^2_{1-\eta, NC-m-1}$. Here, m denotes the number of parameters of the distribution and χ^2_C is the computed value of χ^2 statistic by EVD.

$$KS = \max_{i=1}^{N} |F_e(q_i) - F_D(q_i)| \quad \ldots (2)$$

where, $F_e(q_i)=i/(N+1)$ is the empirical CDF of q_i, $F_D(q_i)$ is the computed CDF of q_i, q_i is the observed APF for i^{th} observation and N is the number of observations [17].

Test criteria: If the computed values of GoF tests statistic given by the distribution are less than that of the theoretical values at the desired significance level then the distribution is considered to be acceptable for FFA at that level.

Diagnostic Tests

Sometimes the GoF test results would not offer a conclusive inference thus posing a problem for the user in selecting a suitable probability distribution (with parameter estimation method) of EVD for their application. In such cases, a diagnostic test in adoption to GoF is applied for making inference. The selection of best fit probability distribution of EVD for estimation

of PF can be performed through MAE and MEF, which is defined as below:

$$MAE = \left((1/N) \sum_{i=1}^{N} |q_i - q_i^*| \right) \quad \ldots (2)$$

$$MEF\ (\%) = \left(1 - \frac{\sum_{i=1}^{N} (q_i - q_i^*)^2}{\sum_{i=1}^{N} (q_i - \bar{q})^2} \right) * 100 \quad \ldots (3)$$

where, q_i is the observed PF of i^{th} sample, q_i^* is the estimated PF of i^{th} sample and $\bar{q}$ is the average of observed PF [18]. A distribution with minimum MAE and better MEF is considered as better suited distribution in comparison with the other distributions of EVD adopted in FFA for estimation of PF.

Table1. *CDF, Quantile estimator and estimators of the parameters of EVD*

Dis-tri-bution	CDF	Quantile estimator (q_T)	Estimators of the parameters		
			MoM	MLM	LMO
GEV	$F(q) = e^{-\left(1-\frac{k(q-\xi)}{\alpha}\right)^{1/k}}$	$q_T = \xi + \frac{\alpha[1-(-\ln(F))^k]}{k}$	$\bar{q} = \xi + \frac{\alpha(1-\Gamma(1+k))}{k}$ $S_q = \frac{\alpha}{k}\left(\Gamma(1+2k) - \Gamma(1+k)^2\right)^{1/2}$ $c_s = (\text{sign } k)\frac{\Gamma(1+3k)+3\Gamma(1+k)(1+2k)-2\Gamma^3(1+k)}{\left\{\Gamma(1+k)-\Gamma^2(1+k)^2\right\}^{3/2}}$	Estimators are obtained by using modified Newton-Raphson algorithm (Hosking, 1990).	$z = (2/(3+\tau_3)) - (\ln(2)/\ln(3))$ $\tau_3 = (2(1-3^{-k})/(1-2^{-k})) - 3$ $k = 7.817740z + 2.930462z^2 + 13.641492z^3 + 17.206675z^4$ $\xi = \lambda_1 + (\alpha(\Gamma(1+k)-1)/k)$ $\alpha = \lambda_2 k/(1-2^{-k})\Gamma(1+k)$
EV1	$F(q) = e^{-e^{-\left(\frac{q-\xi}{\alpha}\right)}}$	$q_T = \xi + \alpha[-\ln(-\ln(F))]$	$\xi = \bar{q} - (0.5772157)\alpha$ $\alpha = \left(\frac{\sqrt{6}}{\pi}\right) S_q$	$\xi = -\alpha \ln\left[\sum_{i=1}^{N} \exp(-q_i/\alpha)\Big/ N\right]$ $\alpha = \bar{q} - \left[\sum_{i=1}^{N} q_i \exp(-q_i/\alpha)\Big/ \sum_{i=1}^{N} \exp(-q_i/\alpha)\right]$	$\xi = \lambda_1 - (0.5772157)\alpha$ $\alpha = \frac{\lambda_2}{\ln(2)}$
EV2	$F(q) = e^{-\left(\frac{q}{\alpha}\right)^{-k}}$	$q_T = \alpha e^{[-\ln(-\ln(F))]/k}$	By using the logarithmic transformation of the observed data, parameters of EV1 are initially obtained by MoM, MLM and LMO; and further used to determine the parameters of EV2 from $\alpha = \exp(\xi)$ and k=1/(scale parameter of EV1).		
GPA	$F(q) = 1 - \left(1 - \frac{k(q-\xi)}{\alpha}\right)^{1/k}$	$q_T = \xi + \frac{\alpha(1-(1-F)^k)}{k}$	$\bar{q} = \xi + (\alpha/(1+k))$ $s_q^2 = \alpha^2/(1+2k)(1+k)^2$ $C_s = 2(1-k)(1+2k)^{1/2}/(1+3k)$	$\sum_{i=1}^{N} \frac{(q_i-\xi)/\alpha}{1-(k(q_i-\xi)/\alpha)} = \frac{N}{1-k}$ $\sum_{i=1}^{N} \ln[1-(k(q_i-\xi)/\alpha)] = -Nk$ $\xi \le$ lowest value of observed q_i	$\xi = \lambda_1 - (\alpha/(1+k))$ $\tau_3 = (1-k)/(3+k)$ $k = (1-3\tau_3)/(\tau_3+1)$ $\alpha = (1+k)(2+k)\lambda_2$

In Table 1, ξ, α, k are the location, scale and shape parameters respectively; $\bar{q}$, S_q, s_q^2 and ψ are the average, standard deviation, variation and Coefficient of Skewness of the observed data; F(q) (or F) is the CDF of q (i.e., APF); ϕ^{-1} is the inverse of the standard normal distribution function, $\phi^{-1} = (P^{0.135} - (1-P)^{0.135})/0.1975$ where in P is the probability of exceedance; sign(k) is plus or minus 1 depending on the sign of k ; λ_1, λ_2 and λ_3 are the first, second and third L-moments respectively; L-Skewness (is a measure of the lack of symmetry in a distribution) and given by $\tau_3 = (\lambda_3/\lambda_2)$; q_T is the estimated PF for a return period (T). A relation between the terms F, P and T is defined by F (or F(q)) =1-P=1-1/T.

Application

In this paper, a study on FFA for Kuppa barrage site by adopting MoM, MLM and LMO of EVD was carried out. The barrage is located on river Baspa at village Kuppa near Sangla and the power house is located near village Karcham about 800 m upstream of the confluence of rivers Satluj and Baspa. Figure 1 shows the location map of the study area.

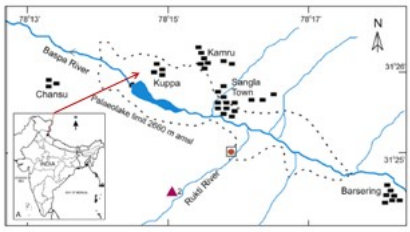

Figure1. *Location map of the study area*

The APF data series for the period 1991 to 2016 was extracted from the daily river flow data series and also used for FFA. The descriptive statistics viz., average, standard deviation, coefficient of skewness and coefficient kurtosis of the observed APF is noted to be 261.1 cumecs, 66.5 cumecs, 0.415 and 2.028 respectively.

Results and Discussions

By applying the procedures of FFA, as described above, parameters of EVD were determined by MoM, MLM and LMO with the aid of statistical software and also used for FFA.

The estimated PF at Kuppa barrage by EVD (using MoM, MLM and LMO) are presented in Table 2 while the plots are shown in Figures 2 and 3. For river flow data of Kuppa barrage, MLM is noted to be not feasible for determination of parameters of GPA distribution and hence FFA results of GPA (MLM) are not presented in Table 2. From FFA results, it is noted that the estimated PF obtained from EV2 (using MLM) is comparatively higher than the corresponding values of other distributions for return periods from 20-year and above.

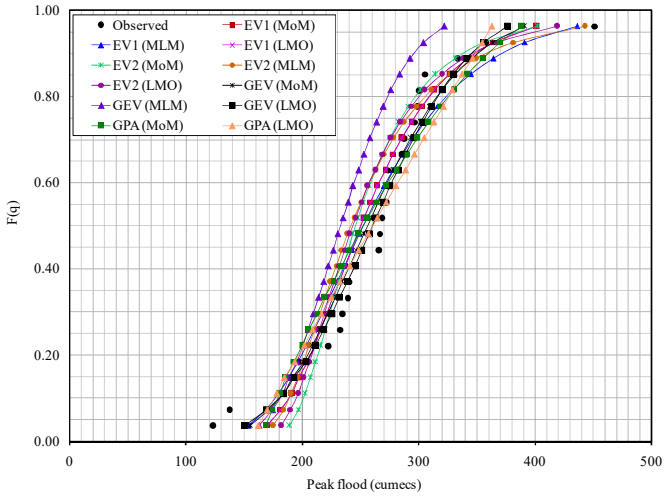

Figure2. *CDF plots of estimated peak flood by EVD distribution with observed APF*

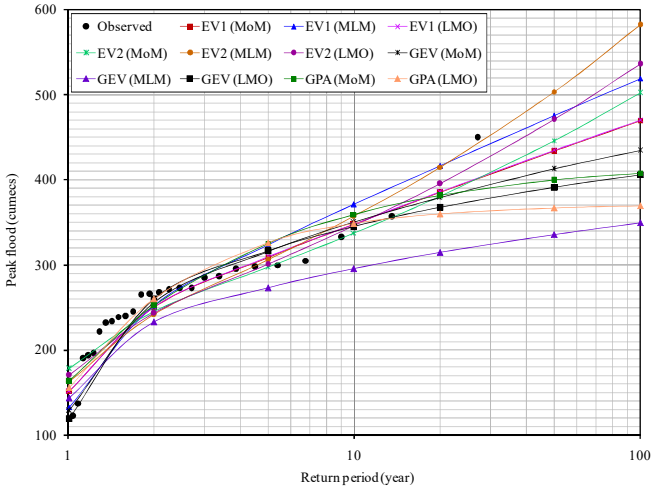

Figure3. *Plots of estimated peak flood by EVD with observed APF*

Analysis Based on GoF Tests

By using MoM, MLM and LMO estimators of EVD (viz., GEV, EV1, EV2 and GPA), GoF tests statistic values were computed and are presented in Table 3. From χ^2 test results, it is noted that the computed values are less than its theoretical values (3.84 for GEV and GPA, 5.99 for EV1 and EV2) at 5% significance level, and at this level, all four distributions are found to be acceptable for FFA. Also, from Table 3, it is noted that the computed values of KS test statistic by EVD are less than its theoretical value of 0.240 at 5% significance level, and at this level, EVD is found to be acceptable for FFA.

Analysis Based on Diagnostic Tests

The selection of suitable probability distribution (with parameter estimation method) for FFA was carried out by using MAE and MEF though GoF tests results confirmed the applicability of GEV, EV1, EV2 and GPA distributions for FFA for Kuppa barrage. The diagnostic tests values of EVD were computed and are presented in Table 4. From the diagnostic tests results, it is noted that the MAE obtained from GEV (LMO) distribution is minimum when compared with the corresponding values of EV1, EV2 and GPA (using MoM, MLM and LMO). The MEF obtained from GEV and EV1 adopted in FFA was found to be 91.3% and 91.2% respectively.

By considering the diagnostic tests results and quantitative assessment through GoF tests, it is identified that GEV (LMO) is better suited probability distribution for estimation of PF at Kuppa barrage.

Table2. *Estimated PF (cumecs) for different return periods by GEV, EV1, EV2 and GPA distributions*

Return period (year)	GEV			EV1			EV2			GPA		
	MoM	MLM	LMO	MoM	MLM	LMO	MoM	MLM	LMO	MoM	MLM	LMO
2	255.9	232.9	260.2	250.2	252.3	250.1	245.6	241.7	243.2	252.1	NF	260.4
5	315.8	273.2	316.3	309.0	323.8	309.2	297.4	305.9	300.6	325.5	NF	326.0
10	350.1	295.7	345.1	347.9	371.2	348.3	337.7	357.4	345.8	359.4	NF	348.5
20	379.4	314.6	367.9	385.2	416.6	385.8	381.3	415.1	395.7	381.6	NF	360.0
50	412.8	335.7	391.7	433.6	475.4	434.4	446.3	503.7	471.0	399.7	NF	367.1
100	434.8	349.3	406.0	469.8	519.5	470.8	502.2	582.2	536.6	408.1	NF	369.6
200	454.5	361.3	417.8	505.9	563.4	507.0	564.9	672.7	611.1	413.6	NF	370.8
500	477.4	375.0	430.5	553.5	621.3	554.9	659.6	814.0	725.5	418.1	NF	371.6
1000	492.6	383.9	438.2	589.5	665.1	591.0	741.7	940.1	826.0	420.2	NF	371.9

Table3. *Computed values of GoF tests statistics by GEV, EV1, EV2 and GPA distributions*

GoF Tests	GEV			EV1			EV2			GPA		
	MoM	MLM	LMO	MoM	MLM	LMO	MoM	MLM	LMO	MoM	MLM	LMO
χ^2	0.769	1.385	1.692	0.769	1.769	0.769	1.292	1.323	1.385	0.769	NF	1.692
KS	0.114	0.109	0.102	0.154	0.156	0.153	0.203	0.221	0.205	0.294	NF	0.185

Table4. *Diagnostic test values given by GEV, EV1, EV2 and GPA distributions*

Diagnostic tests	GEV			EV1			EV2			GPA		
	MoM	MLM	LMO	MoM	MLM	LMO	MoM	MLM	LMO	MoM	MLM	LMO
MAE (cumecs)	13.26	31.63	12.31	14.29	16.46	14.30	17.44	17.72	17.62	19.40	NF	18.16
MEF (%)	92.1	90.8	91.3	91.1	90.9	91.2	86.8	89.1	88.7	86.7	NF	85.4

CONCLUSIONS

The paper presents the study carried out for FFA of river flow data for Kuppa barrage by adopting EVD (viz., GEV, EV1, EV2 and GPA). The parameters of the distributions were determined by MoM, MLM and LMO, and also used for estimation of PF. The intercomparison of the results was carried out and the following conclusions were drawn from the study:

- For the return period of 20-year and above, it was found that the estimated PF by EV2 (MLM) is comparatively higher than the corresponding values of other distributions.
- Qualitative assessment through plots indicated that the pattern of the fitted lines of the estimated PF by EV2 are in the form of exponential curve.
- The χ^2 and KS test results confirmed the applicability of GEV, EV1, EV2 and GPA distributions (using MoM, MLM and LMO) for FFA.
- On the basis of quantitative ad qualitative assessments, the study suggested that the estimated PF by GEV (LMO) could be used as a design value for designing civil and hydraulic structures.
- For the case of economical design of hydraulic structure with little risk involvement, PF obtained from GEV (LMO) distribution may be considered.
- For the case of risk involved in the operation and management of hydraulic structures, PF obtained from EV1 (LMO) distribution may be used for design purposes.

However, by considering the data length (i.e., 26-years) of river flow data used in FFA, the study suggested that the estimated PF beyond 100-year may be cautiously used due to uncertainty in higher order return periods.

ACKNOWLEDGMENTS

The author is grateful to the Director, Central Water and Power Research Station, Pune for providing research facilities to carry out the study. The author is thankful to Central Water Commission, New Delhi, for the supply of stream flow data used in the study.

REFERENCES

[1] Guevara, E., Engineering design parameters of storms in Venezuela, Hydrology Days, pp. 80-91, 2003.

[2] Naghavi, B., Yu, F.X., and Singh, V.P., Comparative evaluation of frequency distributions for Louisiana extreme rainfall. Water Resources Bulletin, 29(2): 211-219, 1993.

[3] Rao, A.R., and Hamed, K.H., Flood frequency analysis, CRC Publications, New York, 2000.

[4] CEH, FLOODS version 1.1, Regional flood frequency analysis software manual, Water Resources Section, Centre for Ecology and Hydrology (CEH), Wallingford, U.K., 2001.

[5] Hosking, J.R.M., L-moments: Analysis and estimation of distributions using linear combinations of order statistics, Royal Statistical Society (Series B), 52 (1): 105-124, 1990.

[6] Bhuyan, A., Borah, M., and Kumar, R., Regional flood frequency analysis of North-Bank of the River Brahmaputra by using LH-Moments, Water Resources Management, 24 (9): 1779-1790, 2010.

[7] Malekinezhad, H., Nachtnebel, H.P., and Klik, A., Regionalization approach for extreme flood analysis using L-moments, Agricultural Science and Technology (Iran), 13 (Supplementary Issue): 1183–1196, 2011.

[8] Badreldin, G.H.H., and Feng, P., Regional rainfall frequency analysis for the Luanhe Basin using L-moments and cluster techniques, International Conference on Environmental Science and Development, 5-7 Jan 2012, Hong Kong, Vol. 1, pp. 126–135, 2012.

[9] Haberlandt, U., and Radtke, I., Hydrological model calibration for derived flood frequency analysis using stochastic rainfall and probability distributions of peak flows, Hydrology and Earth System Sciences, 18 (1): 353-365, 2014.

[10] Markiewicz, I., Strupczewski, W.G., Bogdanowicz, E., and Kochanek, K., Generalized exponential distribution in flood frequency analysis for Polish Rivers, PLoS ONE, 10 (12): 1-15, 2015.

[11] Kossi, K., Barnabas, A.A., Bernd, D., and Fabien, C.C.H., Regional flood frequency analysis in the Volta river basin, West Africa, Hydrology Journal, 3 (1): 1-15, 2016.

[12] Amr, G., Evan, G.R.D., Greg, G.G., and Monireh, F., Assessment of the combined effects of threshold selection and parameter estimation of generalized pareto distribution with applications to flood frequency analysis, Water, Vol. 9, Paper ID: 692, doi:10.3390/w9090692, 2017.

[13] Kolbjorn, E., Donna, W., Péter B., Lars, R., and Erik, H., Use of historical data in flood frequency analysis: a case study for four catchments in Norway, Hydrology Research, 49(2): 466-486, 2018.

[14] Ul Hassan, M., Hayat, O., and Noreen, Z., Selecting the best probability distribution for at-site flood frequency analysis; a study of Torne River, SN Applied Science 1(12): Article ID: 1629, 2019.

[15] USWRC, Guidelines for determining flood flow frequency, United States Water Resources Council, Bulletin No. 17B (Revised), Washington, DC, New York, 1982.

[16] Charles Annis, P.E., Goodness-of-Fit tests for statistical distributions, 2009, http://www.statistical engineering.com/goodness.html.

[17] Zhang, J., Powerful goodness-of-fit tests based on the likelihood ratio, Journal of Royal Statistical Society, 64(2): 281-294, 2011.

[18] Singh, R.D., Mishra, S.K., and Chowdhary, H., Regional flow duration models for 1200 ungauged Himalayan watersheds for planning micro-hydro projects, ASCE Journal of Hydrologic Engineering, 6(4): 310-316, 2001.

Citation: *N. Vivekanandan*. "Selection of Best Fit of Extreme Value Family of Distributions for Frequency Analysis of River Flow Data" International Journal of Research Studies in Science, Engineering and Technology, 7(5), 2020, pp. 13-18.*