Asian Journal of Water, Environment and Pollution, Vol. 18, No. 1 (2021), pp. 31-41.
DOI 10.3233/AJW210005

Modelling Unusual Behaviour of Rainfall Using Truncated GEV Distribution in a Mixture Framework

R.S. Jagtap* and U.V. Naik-Nimbalkar[1]

Central Water and Power Research Station, Khadakwasla, Pune – 411024, India
[1]Department of Statistics & Centre for Advanced Studies in Statistics
Savitribai Phule Pune University, Pune – 411007, India
✉ rsj.wadi@gmail.com

Received September 1, 2018; revised and accepted December 15, 2020

Abstract: A truncated generalised extreme value (GEV) distribution in a mixture framework is proposed for the analysis of abnormal rare events in heterogeneous data representing environmental phenomena. The proposed extremal mixture model produced a better understanding of the extremal rainfall behaviour in the Mula-Mutha-Bhima subbasin in India. It gave some realistic extrapolation of quantiles corresponding to a very low probability of exceedance useful in water resources planning and design of civil infrastructure. The proposed model could be useful for the class of problems characterising extreme events and heterogeneity in fields like hydrology, environment and so on.

Key words: Extreme values, heterogeneity, maximum likelihood estimate, mixture modelling, quantile, truncated GEV distribution.

Introduction

Reliable information on the magnitude and occurrence of extreme quantiles associated with environmental random variables would certainly aid planners and decision-makers in effectively coping with calamitous scenarios. Such random variables are, therefore, usually kept under a continuous monitor. Eventually one needs to assess probabilities of occurrences of adverse events that should especially be accurate on regions related to extreme events (Mendes and Lopes, 2004). Dependable information on the tail portion of the probability distribution of annual rainfall would contribute to an effective short and medium-term planning of water resources in a basin. Estimation of quantiles of rainfall corresponding to different return periods is an important aspect for deciding the design flood for reservoirs, protection of constructions along river banks, etc. The generalised extreme value (GEV) distribution has been extensively applied in hydrology, climatology, etc as it encompasses the three main classes of tail behaviour associated with the fat-tailed Fréchet type and the thin and short-tailed Weibull and Gumbel types.

Recent studies expressed the need of revising the standard set of statistical assumptions of independence, stationarity and homogeneity underlying the hydrological process. A study of climatic processes over Central India by Roxy et al. (2017) indicates a decline in the total rainfall during 1950-2015 despite the concurrent rise in the magnitude and frequency of extreme rainfall events. A need is expressed to include appropriate covariates based on say, time or climate indices to tackle the likely non-stationarity in data series; whereas mixture distributions could be used to model heterogeneous

*Corresponding Author

datasets. In this paper, we explore mixture models to analyse heterogeneous datasets in the quest for efficient estimation of high quantiles and better assessment of extremal behaviour of random quantities.

Uncertainty in understanding the behaviour of random variables may reduce if distinct probability distributions are used for modelling the mechanisms generating the 'usual' and 'unusual' behaviour of a process. We have proposed Normal distribution for the bulk part and truncated GEV distribution for an extreme portion in a mixture framework to model heterogeneous datasets. A robust maximum likelihood criterion and graphical diagnostics are used for assessing the model performance. The proposed model produced realistic extrapolation of quantiles corresponding to longer return periods. A comparative study on modelling of extremal behaviour of rainfall using the proposed mixture model and some other existing models such as Skew Normal distribution (Azzalini, 1985) and flexible skew-generalised normal (FGSN) distribution (Ma and Genton, 2004) is presented.

A brief on basic extreme value distributions and mixture models is given in section Basic Extreme Value Distributions. The section Proposed Normal-Truncated GEV Model in a Mixture Framework explains the proposed extremal model, maximum likelihood estimation (MLE), properties of an estimator and the assessment of model performance. This is followed by the sections Models Considered for Case Study and Applications, which briefs about various models considered for comparative study and the goodness of fit criteria and presents the application of the proposed model to various data series in Mula-Mutha-Bhima (MMB) subbasin in India, respectively. Finally, the sections Results and Discussion and Conclusions summarise the article.

Basic Extreme Value Distributions

The cumulative distribution function (cdf) of three extremal distributions, Gumbel, Fréchet and Weibull unified into the GEV distribution is given by:

$$F_{\mathrm{GEV}}(x)=\begin{cases}\exp\left[-\left\{1+\gamma\left(\dfrac{x-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]; \gamma\neq 0\\ \exp\left[-\exp\left\{-\left(\dfrac{x-\alpha}{\beta}\right)\right\}\right]; \gamma=0\end{cases} \quad (1)$$

defined on $x{:}x\in R$ and $1+\gamma\left(\dfrac{x-\alpha}{\beta}\right)>0$ where γ is a shape parameter, $\beta > 0$ a scale parameter and α a location parameter (Coles, 2001).

The return level x_T alias the p^{th} quantile is the magnitude of a random variable that can be equaled or exceeded at least once in a span of T years. Return level x_T for various return periods, T can be computed by solving:

$$F(x_T)=P(X\le x_T)=1-\frac{1}{T}=p, \quad (2)$$

where p is the probability of non-exceedance. We may treat the quantile as a high quantile when the probability of non-exceedance, p is large, and it may be seen as an extreme quantile when p is relatively still larger.

Return level x_T using the GEV distribution yields,

$$x_T=\begin{cases}\alpha+\dfrac{\beta}{\gamma}\left\{\left[-\ln\left(1-\dfrac{1}{T}\right)\right]^{-\gamma}-1\right\}, \gamma\neq 0\\ \alpha+\beta\ln\left[-\ln\left(1-\dfrac{1}{T}\right)\right], \qquad \gamma=0\end{cases} \quad (3)$$

Another classical model characterising the behaviour of extremal events over a high threshold, u is the generalised Pareto distribution (GPD) with cdf given by (Pickands, 1975).

$$F_{GPD}(x)=\begin{cases}1-\left\{1+\gamma\left(\dfrac{x-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}; \gamma\neq 0\\ 1-\exp\left\{-\left(\dfrac{x-\alpha}{\beta}\right)\right\}; \quad \gamma=0\end{cases} \quad (4)$$

defined on $x : x \in R, x > u$, and $1+\gamma\left(\dfrac{x-\alpha}{\beta}\right)>0$ where γ is a shape parameter, $\beta > 0$ a scale parameter and α a location parameter.

Extreme Value Analysis using Mixture Models

The motivation in analysing extremal behaviour using mixture models is due to the assumption that the extreme and non-extreme elements in a data series may have been generally governed by different physical processes. A comprehensive account of theory on mixture distributions and their applications is found in McLachlan and Peel (2000). Escalante-Sandoval (2007) employs a simple mixture of Gumbel and GEV distribution for analysing heterogeneous flood samples

in Mexico. Behrens et al. (2004) propose a flexible mixture model with normal or gamma for the bulk part and the GPD for modelling the data above the threshold. In the present work, truncated GEV distribution is proposed in a mixture framework to examine extremal behaviour in heterogeneous datasets.

Proposed Normal-Truncated GEV Model in a Mixture Framework

A mixture of normal distribution $F_N(x|\mu, \sigma^2)$ for a bulk part below the threshold, u and the truncated GEV distribution $F_{tGEV}(x|u, \alpha, \beta, \gamma)$ for the tail fraction above the threshold is proposed. The mixture cdf F in the threshold framework is defined by:

$$F(x|\mu,\sigma^2,u,\alpha,\beta,\gamma,\varnothing_u)=\begin{cases}(1-\varnothing_u)\dfrac{\Phi(x|\mu,\sigma^2)}{\Phi(u|\mu,\sigma^2)}; & x\le u\\[2ex](1-\varnothing_u)+\varnothing_u\dfrac{\exp\left[-\left\{1+\gamma\left(\frac{x-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]-\exp\left[-\left\{1+\gamma\left(\frac{u-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]}{1-\exp\left[-\left\{1+\gamma\left(\frac{u-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]}; & x>u,\gamma\neq 0\\[2ex](1-\varnothing_u)+\varnothing_u\dfrac{\exp\left[-\exp\left\{-\left(\frac{x-\alpha}{\beta}\right)\right\}\right]-\exp\left[-\exp\left\{-\left(\frac{u-\alpha}{\beta}\right)\right\}\right]}{1-\exp\left[-\exp\left\{-\left(\frac{u-\alpha}{\beta}\right)\right\}\right]}; & x>u,\gamma=0\end{cases} \quad (6)$$

where $\Phi(u|\mu,\sigma^2)=\int_{-\infty}^{u}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}dx$ is the cdf of $N(\mu, \sigma^2)$.

The maximum likelihood (ML) method is used to estimate the parameters of the proposed model. Let $X = (X_1, X_2, \ldots X_n)$ be a random sample of size n from the proposed mixture model with probability density function (pdf) as $f(x|\mu, \sigma^2, u, \alpha, \beta, \gamma, \varnothing_u)$. Out of the n realised X_i's, some n_b X_i's are $\le u$ and the remaining n_u X_i's are $> u$ such that $n = n_b + n_u$. The realised values are denoted by lowercase letters x_i. The likelihood function corresponding to the proposed extremal mixture model is:

$$L(\mu,\sigma^2,u,\alpha,\beta,\gamma,\varnothing_u|X)=\Pi_{i=1}^{n} f(x_i|\mu,\sigma^2,u,\alpha,\beta,\gamma,\varnothing_u) \quad (7)$$

$$F(x|\mu,\sigma^2,u,\alpha,\beta,\gamma,\varnothing_u)=\begin{cases}(1-\varnothing_u)\dfrac{F_N(x|\mu,\sigma^2)}{F_N(u|\mu,\sigma^2)}; & x\le u\\(1-\varnothing_u)+\varnothing_u F_{tGEV}\\(x|u,\alpha,\beta,\gamma); & x>u\end{cases} \quad (5)$$

where μ and σ^2 are the mean and variance of normal distribution; α, $\beta > 0$ and γ are the location, scale and shape parameters of truncated GEV distribution; $\varnothing_u$ is the probability of an observation being above the threshold u. Furthermore, we have treated threshold, u itself as one of the parameters in the model. The advantage of the proposed mixture is that it makes use of the threshold framework as well as the classical extreme value type GEV distribution in the tail region.

The cdf of the proposed Normal-truncated GEV distribution in a mixture framework is

$$L(\mu,\sigma^2,u,\alpha,\beta,\gamma,\varnothing_u|X)=\Pi^{n_b}_{\{X_i\le u\}}(1-\varnothing_u)\frac{f_N(x_i|\mu,\sigma^2)}{F_N(\mu|\mu,\sigma^2)} I_{(-\infty,u)}(x_i)\times\Pi^{n_u}_{\{X_i>u\}}\varnothing_u f_{tGEV}(x_i|u,\alpha,\beta,\gamma)I_{(u,\infty)}(x_i) \quad (8)$$

$$\text{For } \gamma\neq 0;\ \ln L(\mu,\sigma^2,u,\alpha,\beta,\gamma,\varnothing_u|X)=n_b\ln(1-\varnothing_u)-n_b\ln\left\{\Phi(u|\mu,\sigma^2)\right\}-n_b\ln(\sigma\sqrt{2\pi})-\frac{1}{2}\Sigma_{i=1}^{n_b}\left(\frac{x_i-\mu}{\sigma}\right)^2+n_u\ln\varnothing_u-n_u\ln\beta-n_u\ln\left(1-\exp\left[-\left\{1+\gamma\left(\frac{u-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]\right)-\left(\frac{1}{\gamma}+1\right)\Sigma_{i=1}^{n_u}\ln\left\{1+\gamma\left(\frac{x_i-\alpha}{\beta}\right)\right\}-\Sigma_{i=1}^{n_u}\left\{1+\gamma\left(\frac{x_i-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}} \quad (9)$$

$$\text{For } \gamma = 0;\ \ln L(\mu, \sigma^2, u, \alpha, \beta, \varnothing_u \mid X) = n_b \ln(1-\varnothing_u) - n_b \ln\{\Phi(u \mid \mu, \sigma^2)\} - n_b \ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n_b}\left(\frac{x_i-\mu}{\sigma}\right)^2 + n_u \ln\varnothing_u - n_u \ln\beta - n_u \ln\left(1-\exp\left[-\exp\left\{-\left(\frac{u-\alpha}{\beta}\right)\right\}\right]\right) - \sum_{i=1}^{n_u}\left(\frac{x_i-\alpha}{\beta}\right) - \sum_{i=1}^{n_u}\exp\left\{-\left(\frac{x_i-\alpha}{\beta}\right)\right\} \quad (10)$$

To obtain the MLEs of the parameters μ, σ^2, u, α, β, γ, $\varnothing_u$; the log-likelihood function *lnL* (μ, σ^2, u, α, β, γ, $\varnothing_u$ |*X*) is differentiated with respect to each parameter and equated to zero to get a system of simultaneous equations. Since no closed-form solution exists for the system of ML equations, the MLEs were obtained by the direct maximisation of the log-likelihood function by developing codes in *R* software.

Properties of Estimators of Parameters of the Proposed Model

The log-likelihood of the proposed model is a sum of the likelihoods based on the normal and the truncated GEV distribution. For the normal distribution, the regularity conditions for asymptotic properties of MLEs hold. However, since the range of the GEV distribution depends on unknown parameters, the regularity conditions are not necessarily satisfied. Nevertheless, the desirable asymptotic properties of efficiency and normality of MLEs hold for the GEV distribution for $\gamma < 0.5$ (Smith, 1985).

Estimation of Extreme Quantiles

After getting the estimates of parameters of the proposed extremal mixture model, the return level x_T alias the p^{th} quantile is computed as given below.

Return level, x_T for $p \le 1 - \varnothing_u$ is computed using the relation

$$F(x_T \mid \mu, \sigma^2, u, \varnothing_u) = \frac{(1-\varnothing_u)\Phi(x_T \mid \mu, \sigma^2)}{\Phi(u \mid \mu, \sigma^2)} = 1 - \frac{1}{T} = p;\ p \le 1-\varnothing_u \quad (11)$$

Whereas return level, x_T for $p > 1 - \varnothing_u$ is derived using the relation

$$F(x_T \mid u, \alpha, \beta, \gamma, \varnothing_u) = (1-\varnothing_u) + \varnothing_u F_{tGEV}(x \mid u, \alpha, \beta, \gamma) = 1 - \frac{1}{T} = p;\ p > 1-\varnothing_u \quad (12)$$

The cdf F of the proposed model is an explicit function; by solving, we get Eq. 13.

$$x_T = \begin{cases} \alpha + \dfrac{\beta}{\gamma}\left[\left\{-\ln\left(\dfrac{(p-1+\varnothing_u)\left(1-\exp\left[-\left\{1+\gamma\left(\dfrac{u-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]\right)}{\varnothing_u} + \exp\left[-\left\{1+\gamma\left(\dfrac{u-\alpha}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]\right)\right\}^{-\gamma} - 1\right]; & \gamma \neq 0 \\ \alpha - \beta\ln\left[-\ln\left\{\dfrac{(p-1+\varnothing_u)\left(1-\exp\left[-\exp\left\{-\left(\dfrac{u-\alpha}{\beta}\right)\right\}\right]\right)}{\varnothing_u} + \exp\left[-\exp\left\{-\left(\dfrac{u-\alpha}{\beta}\right)\right\}\right]\right\}\right]; & \gamma = 0 \end{cases} \quad (13)$$

Normal-Truncated GEV Re-parameterised Mixture

The Normal-truncated GEV re-parameterised mixture is a slight variation of the extremal mixture model described in previous section wherein the threshold itself has been taken as the location parameter of truncated GEV distribution which results in the reduction of one parameter in the model. The cdf of the normal-truncated GEV re-parameterised mixture is given below:

$$F(x\mid\mu,\sigma^2,u,\beta,\gamma,\varnothing_u)=\begin{cases}(1-\varnothing_u)\dfrac{\Phi(x\mid\mu,\sigma^2)}{\Phi(u\mid\mu,\sigma^2)}; & x\le u\\[2ex](1-\varnothing_u)+\varnothing_u\dfrac{\exp\left[-\left\{1+\gamma\left(\dfrac{x-u}{\beta}\right)\right\}^{-\frac{1}{\gamma}}\right]-\exp(-1)}{1-\exp(-1)}; & x>u,\gamma\neq 0\\[2ex](1-\varnothing_u)+\varnothing_u\dfrac{\exp\left[-\exp\left\{-\left(\dfrac{x-u}{\beta}\right)\right\}\right]-\exp(-1)}{1-\exp(-1)}; & x>u,\gamma=0\end{cases} \quad (14)$$

The return level, x_T for $p > 1 - \varnothing_u$ has been derived as

$$x_T=\begin{cases}u+\dfrac{\beta}{\gamma}\left\{\left(-\ln\left[\dfrac{(p-1+\varnothing_u)\{1-\exp(-1)\}}{\varnothing_u}+\exp(-1)\right]\right)^{-\gamma}-1\right\}; & \gamma\neq 0\\[2ex]u-\beta\ln\left(-\ln\left[\dfrac{(p-1+\varnothing_u)\{1-\exp(-1)\}}{\varnothing_u}+\exp(-1)\right]\right); & \gamma=0\end{cases} \quad (15)$$

Models Considered for Case Study

The proposed Normal-truncated GEV distribution in a mixture framework has been applied in the study of extremal behaviour of rainfall datasets in the MMB subbasin. Following models useful for the analysis of datasets with the likely presence of heterogeneities were considered for comparative study in addition to the unimodal GEV and skew normal (SN) distributions.

(a) Proposed Normal-truncated GEV distribution in mixture framework (NormtGEV6p)
(b) Proposed Normal-truncated GEV re-parameterized mixture (NormtGEV5p)
(c) Normal-GPD mixture (NormGPD)
(d) Flexible Generalized Skew Normal distribution (FGSN)

Models mentioned in (a) and (b) above are described in the heading Proposed Normal-Truncated GEV Model in a Mixture Framework whereas models (c) and (d) are summarized below.

Normal-GPD Mixture (NormGPD)

Normal-GPD mixture is an extremal mixture developed by Behrens et al. (2004) wherein the GPD model is used for modelling data above the threshold, *u*. The cdf of the Normal-GPD mixture model is given below.

$$F(x\mid\mu,\sigma^2,\alpha,\beta,\gamma)=\begin{cases}F_N(x\mid\mu,\sigma^2); & x\le u\\ F_N(u\mid\mu,\sigma^2)+[1-F_N(x\mid\mu,\sigma^2)]F_{GPD}(x\mid\alpha,\beta,\gamma); & x>u\end{cases} \quad (16)$$

Flexible Generalized Skew Normal Distribution (FGSN)

The pdf of skew-normal distribution *SN* (μ, σ^2, λ) with location μ, scale σ, and skewness parameter λ, an arbitrary real number which regulates the shape of the pdf (Azzalini, 1985) is given by

$$f(y\mid\mu,\sigma,\lambda)=\frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right);\quad y\in R \quad (17)$$

Ma and Genton (2004) proposed a flexible class of skew-symmetric distributions with pdf *FGSN* (μ, σ, λ_1, λ_2, θ) as the product of a symmetric density and a skewing function useful in capturing skewness, heavy tails and multimodality are given below.

$$f(y\mid\mu,\sigma,\lambda_1,\lambda_2,\theta)=\frac{1}{\sigma[1-\Phi(\theta)]}\phi\left(\left|\frac{y-\mu}{\sigma}\right|+\theta\right)\Phi\left[\frac{\lambda_1(y-\mu)}{\sqrt{\sigma^2+\lambda_2(y-\mu)^2}}\right];\ y\in R \quad (18)$$

where $\phi(.)$ and $\Phi(.)$ are the standard normal pdf and cdf, $\mu \in R$, $\sigma \ge 0$; θ, $\lambda_1 \in R$; $\lambda_1 \ge 0$.

Goodness of Fit Criterion

The goodness of fit is assessed graphically using empirical and modelled density plots and quantitatively using Negative Log-Likelihood (NLL) and the Akaike Information Criterion corrected for small sample sizes (AICc). The likelihood measures the support for various models by treating the data as known and model parameters as unknown. Alternately, one can seek infimum of NLL = $-lnL$. The $\text{AICc} = \text{AIC} + \dfrac{2k(k+1)}{(n-k-1)}$, where AIC = $2k - 2\ ln\ L$, where k is the number of independently adjusted parameters with a penalty for additional parameters in more complicated models

(Akaike, 1974). The most parsimonious model with infimum AICc value will be selected.

The normalised mean squared error criterion, $\mathrm{NRMSE} = 100 \dfrac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q\mathrm{est}_i - Q\mathrm{obs}_i)^2}}{sd(Q\mathrm{obs}_i)}$ is used to measure errors in quantiles estimated using different models; where $Q\mathrm{obs}_i$ is the observed quantile assigned using Weibull plotting position formula, $Q\mathrm{obs}_i$ is the estimate of quantile using the model, n is the sample size and sd is the standard deviation of observed values.

Application

The proposed extremal mixture models *viz.*, NormtGEV6p, NormtGEV5p and some other existing models viz., NormGPD, FGSN, GEV and SN enlisted in the previous section were applied for assessing the extremal behaviour of rainfall in the Mula-Mutha-Bhima (MMB) subbasin. The MMB subbasin is located in the north-western part of Krishna basin in India covering parts of Western Ghats and Deccan plateau with a geographical area 6317 km^2. Major portions of precipitation is received through monsoon-type rains which are highly uncertain in the region. The details of two types of annual rainfall data series from MMB subbasin maintained by the Water Resources Department, Maharashtra State used are described in headings Series Type I and II.

Series Type I: Annual Maximum Daily Rainfall for Catchments in MMB Subbasin

The annual maximum daily rainfall data series were derived based on the time series data of areal rainfall for some catchments in the MMB subbasin as shown in Figure 1. These annual block maxima rainfall series were subjected to frequency analysis for different return periods. Such extreme value analysis of hydro-meteorological parameters is useful in designing vital civil infrastructure to withstand meteorological extremities that are likely to occur during the life-time of the structure.

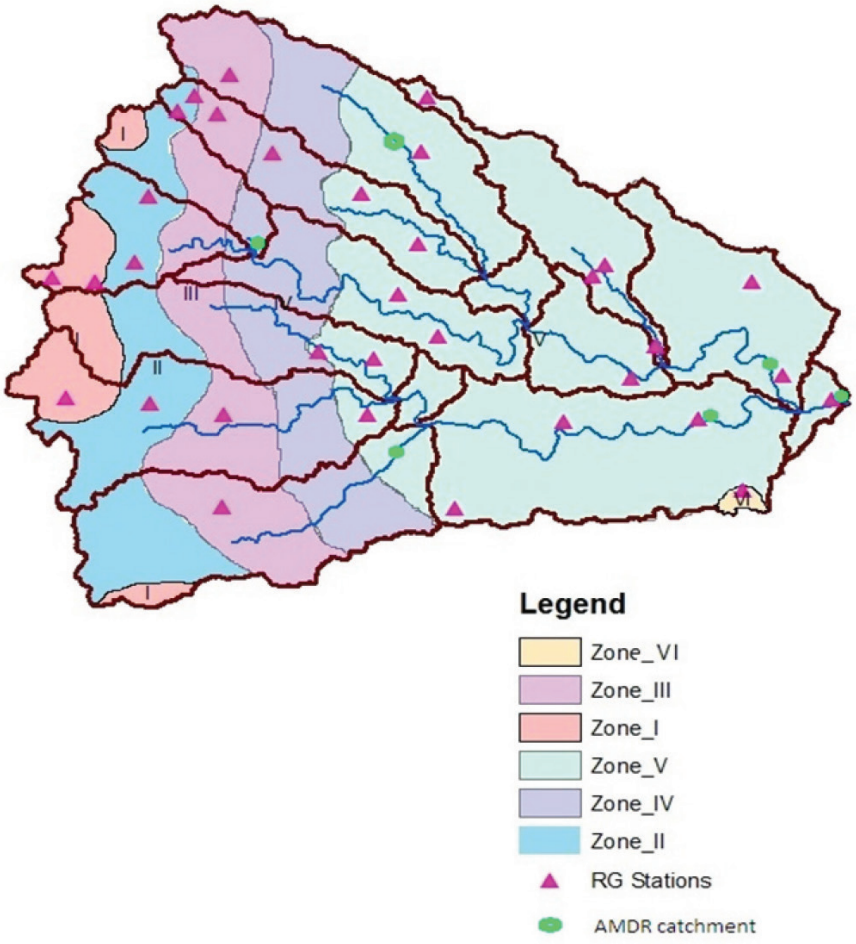


Figure 1: Isohyetal Zones, catchments and rain gauging stations in MMB subbasin.

Series Type II: Annual Rainfall for Isohyetal Zones in MMB Subbasin

Due to very high temporal and spatial variation, the MMB subbasin has been delineated into six zones *viz.*, I to VI which is homogenous within as far as possible, by drawing isohyetal lines of 2500 mm, 2000 mm, 1500 mm, 1000 mm, and 500 mm as shown in Figure 1. Annual rainfall for isohyetal zones and whole MMB subbasin has been compiled using available rainfall data for the period 1971 to 2007. Reliable information on high quantiles of annual rainfall for various return periods would aid in better assessment of behaviour of water resources and planning in a river basin.

Although the annual rainfall series does not strictly resemble the widely used block maxima series, GEV distribution is better suited to model annual rainfall series due to its skewed nature and to get better estimates of high quantiles. Reeve (1996) derived extreme monsoon rainfall for various return periods by fitting normal, exponential, Gumbel, Weibull and GEV distributions to the annual total monsoon rainfall series in India aiming such analysis could address questions germane to the planning of flood assessment and drought management activities.

The GEV distribution has been extensively used in the analysis of the block maxima series. Usually, the annual maximum daily series in hydrology is derived from daily time series data to make the series independent or nearly independent. If we consider the duration of rainfall equal to 1-year in place of 1-day, then there would be one observation per year and the block maxima would be annual rainfall in this case. The GEV distribution would thus be applicable as per the classical extreme value theorem in a stricter sense also. Further, the annual total rainfall in the study region is characterized by extreme intermittency typically driven by a very small number of intense events. The application of GEV distribution for modelling annual precipitation in Southern California is found in the study by Cannon (2010). The annual rainfall data series in the MMB subbasin were thus subjected to frequency analysis for different return periods.

Preliminary investigation using graphical tools and Wald-Wolfowitz Run test indicated the presence of heterogeneities, deviation from stationarity assumption; which could possibly be attributed to natural climate variability or anthropogenic climate change and highly uncertain nature of monsoon type climate of the region. The presence of heterogeneity in data series articulated the need for suitable probabilistic structure for modeling such datasets.

Results and Discussion

Various computations related to the fitting of models enlisted in the section named Application have been executed using codes developed in *R* software (R Core Team 2015). Some representative tables and figures depicting findings were brought out in the paper. Estimates of NLL, AICc and NRMSE criterion used for comparing the models fitted to series type II are given in Table 1. Proposed truncated GEV models in mixture framework produced a better fit for four data series (type II) in comparison with other models. The NRMSE indicated better estimates of extreme quantiles using proposed models in six out of seven data series, thus indicating their usefulness in analyzing complex, especially heterogeneous real-world datasets. Likewise, the estimates of NLL, AICc and NRMSE for series type I are given in Table 2. The proposed truncated GEV model in the mixture framework produced a comparatively better fit for five out of six data series (type I). The fair range of estimates of extreme quantiles indicates the usefulness of the proposed model in analyzing heterogeneous datasets.

The empirical density plots of annual rainfall for the representative data series (type II) given in Figure 2 suggest multimodal distribution and the presence of heterogeneity. The plots have indicated a better match between the empirical and modeled densities using proposed models. The tail portion of data distribution in the upper region has been adequately represented by the proposed models. The FGSN distribution also produces bimodal and smooth fit; however, the fit was dependent on the initial estimates given for its shape parameters in the optimization algorithm. Thus, the complex probabilistic structure could be better explained by the proposed NormtGEV model as compared with other models.

Irregularities encountered occasionally in the model fitting process in *R* computations could be attributed to small sample sizes and improper solutions to log-likelihood equations during the optimization algorithm and were overcome using practical essence and empirical knowledge on physical processes governing the datasets. The estimates of extreme quantiles for different return periods using the NormtGEV model are given in Table 3 for data series type II in the MMB subbasin.

It is more likely that the practice of allocation of water to domestic, agriculture, industrial sectors, etc is being based on annual water yield assessed from monsoon rains in the MMB subbasin region. It is understood

Table 1: Assessment of models fitted to annual rainfall in MMB subbasin (Series Type II)

Isohyetal zone →	*I*	*II*	*III*	*IV*	*V*	*VI*	*MMB*
NLL Criterion							
NormGEV6p	303.15	293.13	284.07	278.88	241.50	162.19	264.06
NormGEV5p	306.05	291.83	284.16	279.47	239.79	162.15	265.01
NormGPD	304.70	294.28	285.92	278.40	240.03	159.71	262.71
FGSN	304.69	292.99	286.86	279.54	242.11	161.28	264.68
AICc Criterion							
NormGEV6p	621.10	601.05	582.95	572.56	497.80	341.04	542.93
NormGEV5p	624.04	595.59	580.25	570.88	491.51	337.46	541.95
NormGPD	621.33	600.49	583.77	568.73	491.99	332.58	537.35
FGSN	621.32	597.92	585.66	571.02	496.16	335.72	541.30
NRMSE Criterion							
NormGEV6p	15.240	18.956	11.022	18.260	18.371	16.767	18.894
NormGEV5p	18.387	14.578	20.251	17.531	19.024	10.441	19.214
NormGPD	16.171	18.256	17.255	17.637	22.126	17.857	19.411
GEV	15.299	17.823	17.280	19.431	19.962	14.939	17.964

Table 2: Assessment of models fitted to annual maximum daily rainfall in MMB subbasin (Series Type I)

Catchment	*Budhawadi*	*Chaskaman*	*Dattawadi*	*Khamgaon*	*Rakshewadi*	*Pargaon*
NLL Criterion						
NormGEV6p	146.51	181.57	130.70	107.96	104.60	104.22
NormGEV5p	144.45	181.15	129.95	107.98	107.86	104.22
NormGPD	144.95	183.14	130.77	107.05	109.96	106.09
FGSN	147.55	185.32	131.87	110.65	107.11	105.53
AICc Criterion						
NormGEV6p	309.22	378.13	277.83	233.17	226.14	224.86
NormGEV5p	301.76	374.37	272.90	229.48	229.06	221.44
NormGPD	302.76	378.34	274.54	227.64	233.25	225.18
FGSN	307.97	382.71	276.73	234.83	227.55	224.05
NRMSE Criterion						
NormGEV6p	24.565	22.104	18.460	33.537	8.233	15.991
NormGEV5p	27.567	19.264	20.909	125.775	26.090	16.596
NormGPD	37.034	20.903	16.056	22.101	22.588	17.084
GEV	25.657	13.987	18.156	25.321	23.962	19.029

that different sectors have started facing water scarcity owing to rapidly rising future water demands on account of population growth, urbanization, industrialization, etc. Design basis values of annual rainfall *viz.,* 1559.9 mm for 5-year return period and 1762.2 mm for 10-year return period given in Table 3 could aid in effectively planning the water resources in MMB subbasin for some years ahead instead of the practice of planning based on just on a year-on-year basis.

The representative plots of extreme value estimate based on data series type I and II using NormtGEV6p, NormtGEV5p, NormGPD and GEV models for the MMB subbasin are given in Figure 3. It is noted that the difference in the return level estimates pronounces

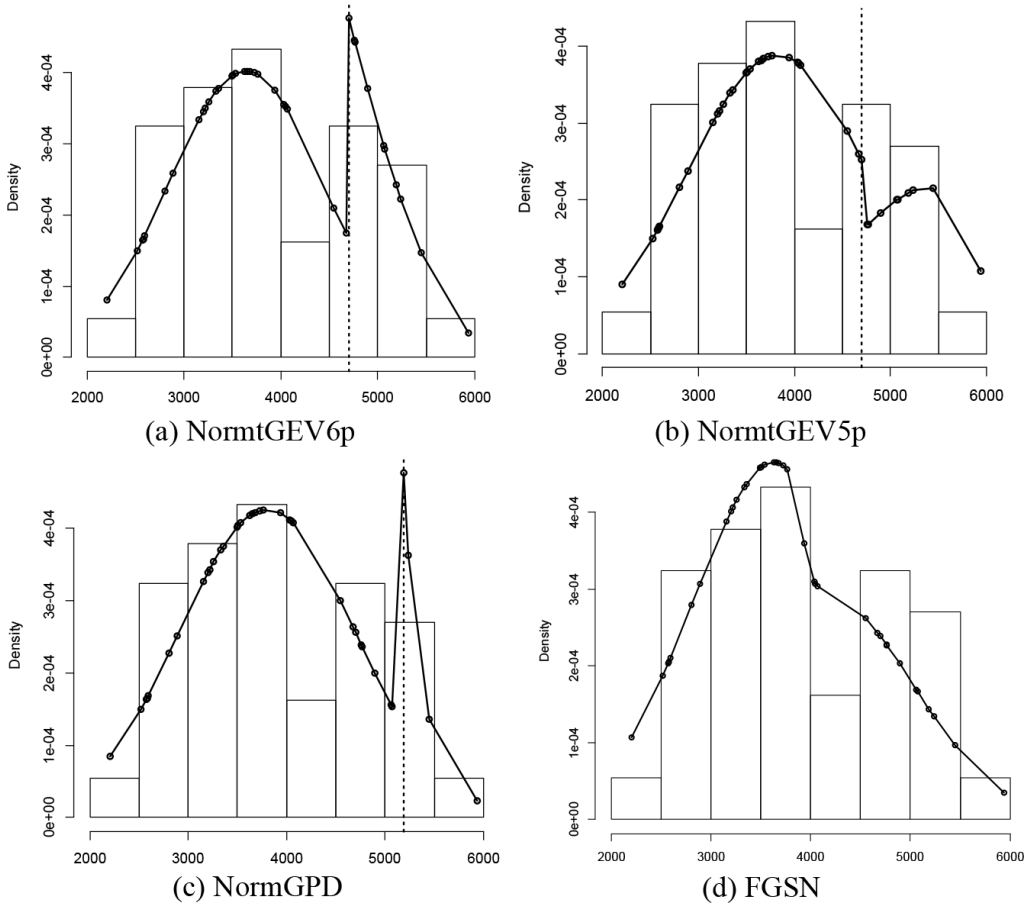


Figure 2: Empirical (Histogram) and modeled density for annual rainfall (mm) in Zone I of MMB subbasin (Series Type II).

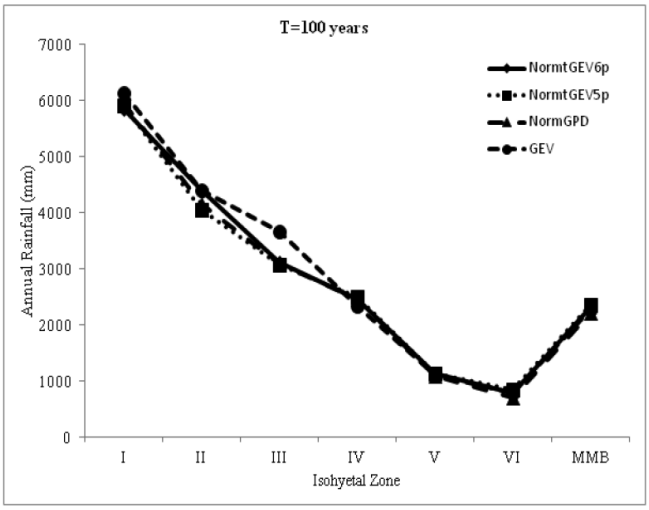
Table 3: Quantile estimates of annual rainfall (mm) in MMB subbasin for various return periods using NormtGEV6p model (Series Type II)

Return period (years)	*Isohyetal zone*						*MMB subbasin*
	I	II	III	IV	V	VI	
2	3887.9	2427.7	1735.3	1048.6	662.3	314.1	1336.8
5	4861.4	2958.5	2505.0	1387.9	789.2	451.6	1559.9
10	5175.2	3348.7	2788.2	1686.0	886.3	539.3	1762.2
20	5427.6	3695.6	2943.6	1943.6	969.6	617.7	1943.0
25	5498.4	3800.5	2978.1	2020.4	994.3	641.5	1998.1
50	5691.5	4109.3	3055.2	2244.4	1065.4	711.8	2161.6
100	5850.3	4395.4	3101.7	2449.1	1129.4	777.3	2315.5

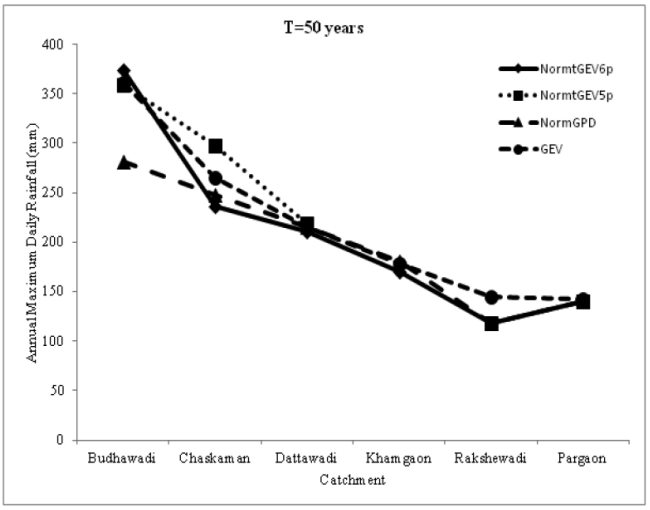
for a lower probability of exceedance. Use of a better model for estimating extremes corresponding to a very low probability of exceedance is thus very important. As the proposed model produced a comparatively better fit, their utility in obtaining realistic estimates of high order quantiles is vital for heterogeneous datasets.

Conclusions

Random variables representing natural phenomena realising data series with heterogeneity require intricate probabilistic structure for their realistic analysis. The proposed extremal mixture model for the analysis of



(a) Series type II: Annual Rainfall (mm) for Isohyetal zones.



(b) Series type I: Annual Maximum Daily Rainfall (mm) for catchments.

Figure 3: Extreme value estimates using NormtGEV6p, NormtGEV5p, NormGPD and GEV models for MMB subbasin.

unusual events is based on a truncated GEV distribution in a mixture framework. Extreme and non-extreme components of annual rainfall have been elegantly represented by the proposed model as compared with other existing models. The maximum likelihood estimates for very good and large sample properties were used for estimating parameters of the proposed model. The proposed model successfully extrapolated quantiles delivering some realistic estimates corresponding to a very low probability of exceedance. The model resulted in a better understanding of extremal behaviour, thereby helped in reducing uncertainty in the estimation of annual rainfall. In this study, quantiles from the upper extreme region of the probability distribution are analyzed; quantiles in the lower extreme region of the distribution could be analyzed on similar lines. Besides, one can simultaneously model the lower and upper tails of the distribution.

Acknowledgements

The authors wish to thank Dr (Mrs) V.V. Bhosekar, Director, CWPRS, Pune for her encouragement and continuous support.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19:** 716-723.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12:** 171-178.

Behrens, C.N., Lopes, H.F. and D. Gamerman (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modeling*, **4(3):** 227-244.

Cannon, A.J. (2010). A flexible nonlinear modelling framework for non-stationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, **24:** 673-685.

Coles, S. (2001). An Introduction to Statistical Modelling of Extreme Values. Springer: London.

Escalante-Sandoval, C. (2007). A Mixed distribution with EV1 and GEV components for analyzing heterogeneous samples. *Engineering, Research and Technology*, VIII. **3:** 123-133 pp.

Hosking, J.R.M. and J.R. Wallis (1997). Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, Cambridge, UK, 224 pp.

Ma, Y. and M.G. Genton (2004). Flexible class of skew-symmetric distribution. Scandinavian Journal of Statistics, 31, Issue 3, Sept.

McLachlan, G.J. and D. Peel (2000). Finite Mixture Models. Wiley, New York.

Mendes, B.V.M. and H.F. Lopes (2004). Data driven estimates for mixtures. *In:* Computational Statistics & Data Analysis, **47:** 583-598. Elsevier.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, **3:** 119-131.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.

Reeve, D.E. (1996). Estimation of extreme Indian monsoon rainfall. *International Journal of Climatology,* **16:** 105-112, Shorter contribution.

Roxy, M.K., Ghosh, S., Amey Pathak, A., Athulya, R., Mujumdar, M., Murtugudde,R., Terray P. and M. Rajeevan (2017). A threefold rise in widespread extreme rain events over central India. *Nature Communications*, **8:** 708, DOI: 10.1038/s41467-017-00744-9.

Smith, R.L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, **72:** 67–90.