Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 2, pp. 55-70

CHARACTERIZATION OF THE TAIL OF RIVER FLOW DATA BY GENERALIZED PARETO DISTRIBUTION

KUMER PIAL DAS

Department of Mathematics, Lamar University, Beaumont, TX 77710, U.S.A
Email: kumer.das@lamar.edu

CHRISTOPHER ALAN SAMS

Department of Mathematics, West Orange Stark HS, Orange, TX 77631, U.S.A
Email: chsa@woccisd.net

VIJAY P SINGH

Zachry Department of Civil Engineering, Texas A & M University, TX 77843, USA
Email: vsingh@tamu.edu

SUMMARY

Extreme value theory (EVT) is a branch of statistics that seeks to access, from a given ordered sample, the probabilities of events that are more extreme than usually observed. Under very general conditions, EVT's main results characterize the distribution of the samples maxima or the distribution of values above a given threshold. Thus, two particular approaches exist, block maxima, which divides a time period into equal sections and the maximum of each section is selected. This usually leads to generalized extreme value (GEV) distribution. The second approach is peaks over threshold (POT) which selects every value that exceeds a certain threshold. Threshold excesses follow an approximate distribution within the generalized Pareto family. This study focuses on modeling the tail of a heavy tail distribution by using the threshold approach. The purpose of this study is manifold: firstly, to compare the estimates of the generalized Pareto distribution (GPD) parameters by three estimation techniques, namely, maximum likelihood estimation (MLE), method of moments (MOM), and probability weighted moments (PWM) in the light of a hydrological application; secondly, to select an appropriate model for the Mississippi River flow data by using the model validation and hypothesis testing; thirdly, to introduce the bootstrap re-sampling approach in hydrological applications; lastly, to obtain a required design value with a given return period of exceedance and probabilities of occurring extreme floods using bootstrap sampling.

Keywords and phrases: Extreme events, Peaks over threshold, Estimation techniques, Generalized Pareto distribution, Return period

AMS Classification: 62P30, 62P12, 62F99.

1 Introduction

Extreme Value Theory (EVT) focuses on the risk of low probability events that could lead to devastating losses. The goal is to make the best possible use of what little information we have about the extremity of distributions of particular interest. It is often the case that various safety measures fail us during catastrophic events, and it is during these times that we need them the most. This is largely due to the fact that most of the time, we simply ignore the risk of such events because they are associated with very low occurrence. However, through the study of EVT we can predict the frequency and cost of such events over a period of time, relative to a given location. EVT has great usefulness in many fields, especially in hydrology and in insurance. Some uses for application include: extreme floods, large wildfires, large insurance loss, equity and market risk, freak waves and a host of others. It has also been used to predict human limitations, such as fastest time possible to run 100m dash, the maximum magnitude/force of a punch without shattering one's bones, or the amount of blood a person can lose and stay alive. Although there are various applications when analyzing extremes, two practical approaches exist. Block maxima (minima) extracts annual maxima (minima), generating an annual maxima series (AMS). This type of data collection leads to generalized extreme value distribution (GEV) for fitting. However, this method could be problematic, given that the number of extreme events in a given year may be rather limited. Moreover, for a sample of moderate size, the standard GEV estimator is quite inefficient due to the possibly slow convergence toward the asymptotic theoretical distribution.

The second method, peaks over threshold (POT), extracts from a continuous record of the peak values reached for any period where values exceed a certain threshold or fall below a certain threshold. For POT data, the analysis involves fitting two distributions: one for the number of events in a basic time period and the second for the size of the exceedance. Fitting this type of data requires use of the generalized Pareto distribution (GPD). This method requires careful selection of the threshold, if it is set too high or too low, the model may prove to be invalid. POT method has popularly been used to estimate return levels of significant wave height (Mackay *et al.*, 2011; Ferreira and Guedes, 1998; Mathiesen *et al.*, 1994; Sterl and Caires, 2005), hurricane damage (Daspit and Das, 2012; Dey and Das, 2014; Dey and Das, 2016), annual maximum flood of the River Nidd at Hunsingore, England (Hosking and Wallis, 1987), earthquake magnitude (Edwards and Das, 2016) and aviation accidents (Das and Dey, 2016).

Floods have been identified as the most common hydrological disasters. In practice, flood design is estimated in two ways: (1) rainfall-runoff modeling, and (2) direct statistical analysis of streamflow extremes. In many countries the latter is the preferred method. This study focuses on estimating extreme water flow probabilities because it is important to understand the stochastic behavior of extreme water flow events for efficient planning and design of coastal defense structures. Both for economic and environmental reasons studying river flow records covering a considerable length of time is important. Over the past five decades, several approaches for estimating probabilities of extreme still water levels have been developed (Arns, 2013). However, currently no universally accepted method of analyzing extreme water level is available. In contrast, different authors and software provide different methods and often propose several alternative means for the related procedure. We have compared the effectiveness of three widely used estimation techniques to model

the study. We have also used the domain of attraction concept to verify the suitability of the model used in this study. The problem of threshold selection is crucial in hydrological applications. We have used multiple approaches to select a proper threshold. In particular, the bootstrap approach discussed in this study is novel and effective. Finally, we have used nonparametric bootstrap sampling to quantify the uncertainty of the return level estimates.

The rest of the paper is organized as follows. Section 2 provides background information on GPD. Section 3 describes and compares three estimation techniques. Section 4 presents an application of GPD in Mississippi River data.We estimate and interpret extreme return levels and the uncertainty of such estimates in Section 5. Section 6 describes findings of the study.

2 Generalized Pareto Distribution

2.1 Definition and notations

The Pareto distribution was named after the Italian civil engineer Vilfredo Pareto. Originally, this distribution was used to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of wealth of any society is owned by a smaller percentage of the people in that society. This idea is sometimes expressed as the "80-20 rule" which says that 20% of the population controls 80% of the wealth. However, over time the Pareto distribution has evolved into a family of distributions known as Generalized Pareto distribution with various applications including but not limited to wealth. There are three parameters associated with the GPD: scale parameter (denoted by α), shape parameter (denoted by k), and location parameter(denoted by μ). A random variable X has a GPD if it has a cumulative distribution function of the form

$$F(x) = \begin{cases} 1 - \left(1 - \frac{k(x-\mu)}{\alpha}\right)^{\frac{1}{k}} & \text{if } k \neq 0, \\ 1 - \exp\left(-\frac{x-\mu}{\alpha}\right) & \text{if } k = 0. \end{cases}$$

The probability density function (pdf) of a GPD is defined as:

$$f(x) = \begin{cases} (1/\alpha)\left(1 - \frac{k(x-\mu)}{\alpha}\right)^{\frac{1}{k}-1} & \text{if } k \neq 0, \\ (1/\alpha)\,\exp(-\frac{x-\mu}{\alpha}) & \text{if } k = 0, \end{cases}$$

where the pdf is positive for $x \geq \mu$, when $k \leq 0$, or for $\mu \leq x \leq \alpha/k$, when $k > 0$.

A special case of GPD is given when the shape parameter is considered as zero, the result is equivalent to the exponential distribution. The graphics in Figure 2.1 display GPD shapes with fixed location $\mu = 0$, fixed scale $\alpha = 1$, and various values for the shape parameter k. In Plot A we see a triangular fit, in plot C we have a uniform fit, E and F display heavy tailed shape. There is no useful application for D when $k > 1$ (Singh and Guo, 1995). In this study, we consider only a two-parameter (scale and shape) GPD.

The POT modeling is based on the GPD family of distributions being appropriate for describing statistical properties of excesses (Pickands, 1975). For the random variable X the foundation of the

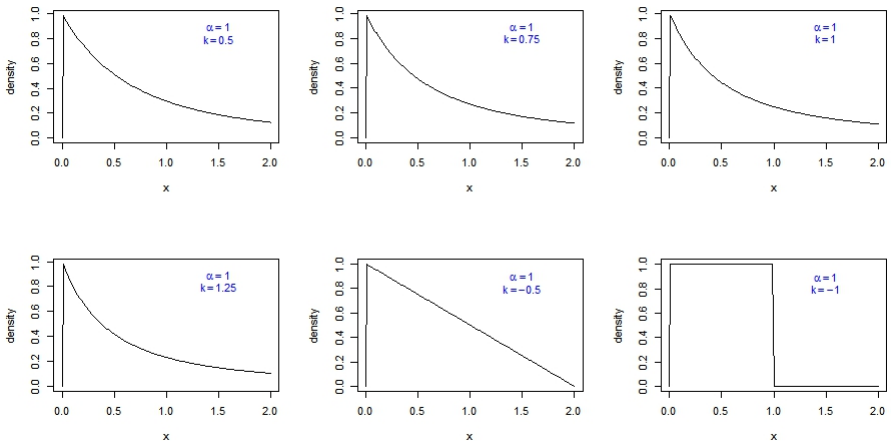

Figure 1: Densities of the generalized Pareto distribution

modeling is based on the excess distribution over a threshold u which is defined as

$$F_u(x) = P\big(X - u \leq x \,|\, X > u\big). \tag{2.1}$$

In hydrology, the use of POT modeling is critical since it can be used to model the level of water in a river or sea to avoid flooding. For instance, the level u could represent the height of a river bank, dam or levee.

2.2 Use of GPD in modeling extreme events

It is understood that many situations (e.g. natural disasters, extreme floods, heat waves, cold waves, and wild fires) are fat-tailed, which means the normal distribution is not adequate to describe the probability of severe events. In order to model these extreme events, we need a distribution which is fat tailed. This is obtained by selecting a cut-off point- a threshold- which specifically defines events in the tail. Thus, the use of GPD has surfaced. As previously mentioned, we need to carefully select the threshold; if it is too high or low the model may prove invaluable. To assist proper selection of a threshold we use the mean excess plot (Klugman, 2012). The mean excess (ME) function is a tool popularly used to aid this choice of u and also to determine the adequacy of the GPD model in practice. The ME function of a random variable X is defined as (Ghosh and Resnick, 2010)

$$M(u) = E\big[X - u \,|\, X > u\big], \tag{2.2}$$

provided $E(X_+) < \infty$, and is also known as the mean residual life function, especially in survival analysis (Ghosh and Resnick, 2010). Note that (X_+) is the function that assigns the value of X whenever $X > 0$ and otherwise assigns the value of zero.

3 Estimation Techniques

3.1 Maximum likelihood estimation (MLE)

The likelihood function of independent observations $X_1, X_2, \ldots, X_n$ from a two-parameter GPD is

$$L(x_i; k, \alpha) = \prod_{i=1}^{n} f(x_i; k, \alpha), \tag{3.1}$$

where $f = dF/dx$. The MLEs are the values of k and α, which maximize the equation (3.1). Very often, it is easier to maximize the logarithm of the likelihood function. The log-likelihood function for $k \neq 0$ states that the function can be made arbitrarily large by taking $k > 1$ and α/k close to the maximum order statistic $x_{n:n}$. Maximum likelihood estimates of α and k can be obtained by maximizing the likelihood function above. However, there are some samples for which no maximum likelihood solution exists. In order for the MLE to perform its best for the GPD, there are certain criteria that must be present. One, the sample size n, must be large because the maximum likelihood does not display its asymptotic efficiency even in samples as large as 500 (Castillo and Hadi, 1997). Two, the values of k, the shape parameter must stay within the bounds of $(-1/2)$ and $(1/2)$. When these criteria are met, then MLE would be preferred due to its effective efficiency with large samples.

3.2 Method of moments

Method of Moments (MOM) estimators of the GPD were derived by Hosking and Wallis (1987). Note that $E(1 - kX/\alpha)^r = 1/(1 + kr)$ if $1 + rk > 0$. The r^{th} moment of X exists if $k > -1/r$. Provided that they exist, the moment estimators for α and k are solutions to the following equations (Hosking and Wallis, 1987) :

$$\bar{x} = \frac{\alpha}{1+k} \tag{3.2}$$

$$s^2 = \frac{\alpha^2}{(1+k)^2(1+2k)} \tag{3.3}$$

where $\bar{x}$, and s^2 are the sample mean, and sample variance, respectively. In other words, the moment estimators for k and α are solutions of the two equations (3.2) and (3.3) above. MOM estimators for k and α denoted by $\hat{k}$ and $\hat{\alpha}$ respectively are as follows:

$$\hat{k} = \frac{1}{2}\left(\frac{\bar{x}^2}{s^2} - 1\right) \text{ and } \hat{\alpha} = \frac{\bar{x}}{2}\left(\frac{\bar{x}^2}{s^2} + 1\right).$$

Castillo and Hadi (1997), and Hosking and Wallis (1987) recommended MOM for $0 < k < 0.4$. Since the parameters are easy to compute, MOM estimates can also be used as the initial estimates in other estimation procedures which require numerical technique (Jockovic, 2012; Singh and Guo, 1995). When $k \leq (-1/2)$, the variance of the GPD does not exist.

3.3 Probability-weighted moments (PWM)

The probability-weighted moments (PWM) were introduced by Greenwood *et al.* (1979) and represent an alternative to the ordinary moments. As for the moments estimator, parameters can be expressed as a function of PWMs. The estimator is particularly advantageous for small data-sets because the probability weighted moments have a smaller uncertainty than the ordinary moments. The best performance is reached for $k \approx 0.2$ (Deidda and Puliga, 2009); for positive shape values, performances are very close to the MLE ones, while for $k < 0$ (Deidda and Puliga, 2009), PWM performances become a little worse than those of MLE. Hosking and Wallis (1987) used two definitions of PWM: unbiased (PWMU) and biased (PWMB), but the difference can be detected only for small samples. We only display the results involving PWMU in this paper. Hosking and Wallis (1987) defined the PWM estimates of the GPD parameters as :

$$\hat{k}_{PWM} = \frac{\bar{x}}{(\bar{x} - 2t)} - 2 \text{ and } \hat{\alpha}_{PWM} = \frac{2\bar{x}t}{(\bar{x} - 2t)}$$

where $t = (1/n)\sum_{i=1}^{n}(1 - p_{i:n})x_{i:n}$ with $p_{i:n} = (i - 0.35)/n$ and $x_{i:n}$ is the ith order statistics of a sample size of n.

4 Application of GPD to Mississippi River Data

In this section we will use GPD to model a real data set. Mississippi river flow data obtained from the Tarbert Landing Discharge has been used in this study. The Tarbert Landing Discharge Range is located on the Mississippi River at river mile 306.3 about 4 miles upstream of Read River Landing. The daily computed values are based and related to corresponding stage values read at the Red River Landing gage. Daily flow data from 1961 to 2011 was collected from the US Army Corps Engineers database. A total of 18,627 data points were used in this study. Note that the river discharge (flow) units are CFS (cubic feet per sec).

POT approach has been used to analyze the data. The basic descriptive statistics is shown in (Table 1). The entire data analysis is done by statistical programming language R with commonly used R packages for extreme value analysis such as *evmix*, *POT*, *extRemes*. In order to determine if data are extreme, we check for the presence of outliers by using the *fourth spread*, f_s which is also known as inerquartile range. The *fourth spread* f_s is a measure of spread that is resistant to outliers (Devore, 2010). We order the n observations from smallest to largest and separate the smaller half from the larger half using the median $\tilde{x}$. The median of the smaller half, the first quartile (Q_1) is the *lower fourth*. The median of the larger half, the third quartile (Q_3) is the *upper fourth*. Then the fourth spread f_s, is given by Devore (2010)

$$f_s = \text{upper fourth} - \text{lower fourth}.$$

Any observation farther than $1.5f_s$ from the closest fourth is an outlier. An outlier is extreme if it is more than $3f_s$ from the nearest fourth, and it is mild otherwise (Devore, 2010). Subtracting $1.5f_s$ from the lower 4^{th} gives a negative number, and none of the observations are negative, so there are

Table 1: Five-number summary (*Unit : Millions*)

Minimum	First Quartile (Q_1)	Median (Q_2)	Third Quartile (Q_3)	Maximum
0.11	0.27	0.43	0.70	1.62

no outliers on the lower end of the data. However, from Table 1 we do observe mild outliers on the upper end, in particular, all values above 1.34 are outliers since upper $4^{th} + 1.5f_s = 1.34$.

4.1 Test for independence

It is important that POT events are independent of one another. In particular, hydrological variables need to be tested for independence because of climate change and anthropogenic influence. We have used the nonparametric Von Neumann ratio test for independence in this study. This test is suggested for time series of at least 30 points. The objective is to test the null hypothesis that the series consists of independent elements. The basis for the test statistic of the Von Neumann test is given by

$$R = \frac{T\sum_{t=1}^{T}\left[X(t+1) - X(t)\right]^2}{(T-1)\sum_{t=1}^{T}\left[X(t) - \bar{X}\right]^2}$$

with $\bar{X}$ being the mean of the series $X(t)$. This R follows a normal distribution with mean $2T/(T-1)$ and variance $4(T-2)/(T-1)^2$. Thus, the test statistic for the Von Neumann test is defined as

$$C = \frac{R - 2T/(T-1)}{\sqrt{4(T-2)/(T-1)^2}},$$

which has a standard normal distribution. We investigate the independence of all values above the threshold 1.25 using the Von Neumann test discussed above and found that the hypothesis that the series consists of independent observations can not be rejected (the observed test statistic value is $C = -0.60$). However, we also found that the observations above threshold 1.17 are not independent of each other with an observed test statistic value, $C = -13.96$.

4.2 Model fitting

In order to build a model for our data, we need to select a proper threshold. Using the ME plot in Figure 4.2, we can see that the plot is almost linear everywhere, however there is a rough spot around 1.25 suggesting this may be a good estimate for our threshold. Using the value of 1.25, threshold quantities were selected that precede and exceed this value in order to observe any trend and find the threshold of best fit (see Table 2). Also, notice Figure 4.2, the curved shape of the QQ plot increasing from left to right indicates the data distribution is skewed to the right, further verifying the presence of outliers. Now that we have estimates for our model, the question arises: which parameters provide a good fit? To assess the quality of the model, we do the following:

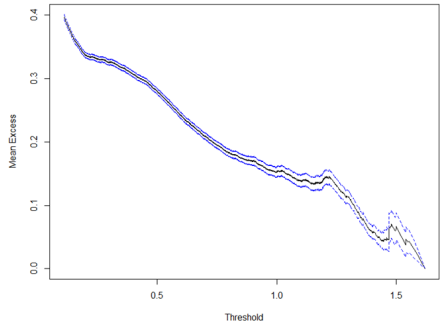

Figure 2: Mean excess plot of Mississippi river flow

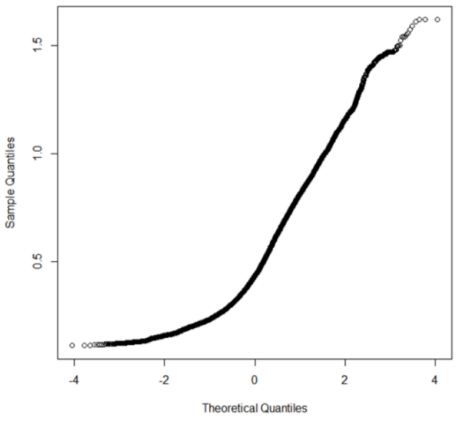

Figure 3: Normal Q-Q plot of Mississippi river flow

Table 2: Comparing Estimation Techniques (standard errors are in parenthesis), where u is the threshold (in millions) and n_T is the number of data points above u

		Scale, α			Shape, k		
u	n_T	MLE	MOM	PWMU	MLE	MOM	PWMU
1.17	392	0.1887	0.1705	0.1613	-0.3770	-0.2580	-0.1898
		(0.0117)	(0.0120)	(0.0124)	(0.0531)	(0.1187)	(0.0628)
1.21	275	0.2125	0.2404	0.2426	-0.4964	-0.6590	-0.6740
		(0.0144)	(0.0217)	(0.0228)	(0.0426)	(0.9560)	(0.1049)
1.25	227	0.1917	0.2345	0.2413	-0.4934	-0.7792	-0.8306
		(0.0143)	(0.0238)	(0.0252)	(0.0474)	(0.1187)	(0.1293)
1.29	189	0.1624	0.2009	0.2106	-0.4591	-0.7610	-0.8460
		(0.0135)	(0.0223)	(0.0241)	(0.0526)	(0.1277)	(0.1433)
1.33	155	0.1318	0.1677	0.1833	-0.4075	- 0.7523	-0.9154
		(0.0124)	(0.0205)	(0.0233)	(0.0590)	(0.1399)	(0.1661)
1.37	126	0.0928	0.1020	0.1136	-0.2804	-0.4035	-0.5627
		(0.0103)	(0.0129)	(0.0157)	(0.0718)	(0.1079)	(0.1428)
1.41	83	0.0678	0.0684	0.0771	-0.1675	-0.1787	-0.3278
		(0.0102)	(0.0104)	(0.0129)	(0.1044)	(0.1087)	(0.1487)
1.45	45	0.0405	0.4362	0.0383	0.1486	0.0755	0.1886
		(0.0108)	(0.0099)	(0.0089)	(0.2238)	(0.1692)	(0.1788)

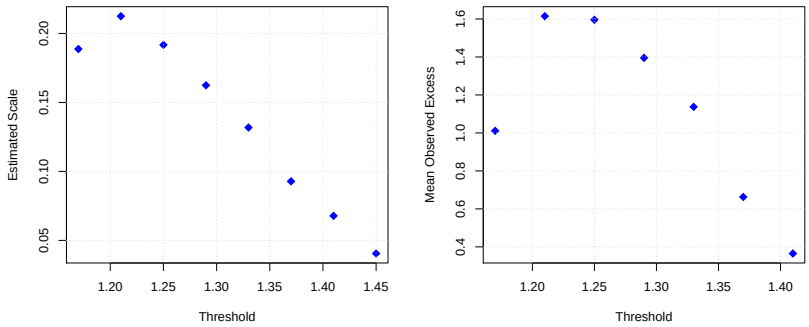

Figure 4: Model validation by MLE scale plot and MLE excess plot

1. The estimates of α are plotted versus the threshold values u. The theoretical value of α as a function of u is given by Castillo *et al.* (2005)

$$\alpha(u) = \alpha_0 - ku, \tag{4.1}$$

where k is the shape parameter, α_0 is the value of α associated with threshold, $u = 0$. Thus, we expect a linear trend in the plot.

2. Provided that $k > -1$, $u > 0$, and $\alpha - ku > 0$, we have the ME function (Castillo *et al.*, 2005)

$$E(X - u \,|\, X > u) = \frac{\alpha - ku}{1 + k}. \tag{4.2}$$

In fact, the linearity of the mean excess function characterizes the GPD class. If the ME plot is close to linear for high values of the threshold then there is no evidence against the use of a GPD model (Ghosh and Resnick, 2010). Accordingly, if the GPD is appropriate, the scatter plot of the mean observed excess over u versus u should resemble a straight line with a slope of $-k/(1 + k)$ and an intercept of $\alpha/(1 + k)$. If the points in the scatter should show a strong linear relationship, then the GPD assumption should seem reasonable Castillo (*et al.*, 2005). Figure 2 shows a clear linear trend for α and $E[X - u|X \geq u]$ versus the threshold value u when $u > 1.25$, indicating that the assumption of a GPD model for the data is reasonable for $u > 1.25$. Note that, $\alpha - ku$ is not positive for $u = 1.45$ and this has not been included in the MLE Mean Excess Plot in Figure 2. A clean linear trend for either estimated α or $E(X - u|x \geq u)$ versus threshold value has not been observed for MOM and PWM parameters, therefore MLE provides the best estimators in this case.

4.3 Hypothesis test for the domain of attraction

There are three domains of attraction in EVT: Gumbel, Fréchet, and Weibull domains. All distributions in the Gumbel domain have the exponential as the limiting distribution of their tail. This

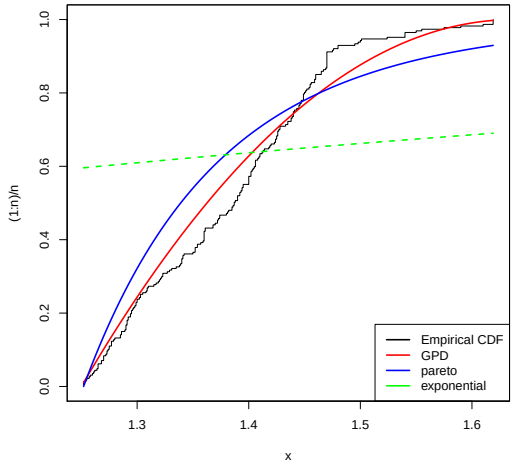

Figure 5: Model validation by density comparison

domain contains the majority of well-known distributions such as the normal, the exponential, and the gamma. The Fréchet domain contains distributions with an infinite yet heavier tail than the exponential (Beisel *et al.*, 2007). On the other hand, the Weibull domain contains distributions with lighter tails than exponential, which possess a finite upper bound (e.g., the uniform distribution.) It should be noted that the Weibull domain is not the same as the Weibull distribution, in fact, the Weibul distribution belongs to the Gumbel domain. We want to test for the domain of attraction which is exactly determined by the shape parameter k. Thus, testing $H_0 : k = 0$ versus $H_1 : k < 0$ is equivalent to testing a Gumbel versus a Fréchet domain of attraction. Similarly, testing $H_0 : k = 0$ versus $H_1 : k > 0$ is equivalent to testing a Gumbel versus a Weibull domain of attraction (Castillo *et al.*, 2005). Using a 95% confidence interval for all estimated values of k in Table 3 indicates k is likely to be negative, thus we reject H_0 in favor of H_1 supporting the distribution of the data follow a Fréchet domain of attraction (Castillo *et al.*, 2005). GPD certainly has a heavier tail than the exponential and thus the use of GPD to model the data set can be justified. Figure 4.3 also confirms the statement.

5 Estimation of Extreme Return Levels

The concepts of return period and return level are commonly used to convey information about the likelihood of extreme events such as floods, earthquakes, hurricanes etc. It is usually more convenient to interpret extreme value models in terms of return levels on an annual scale, rather than individual parameter values. The $m-$year return level is the level expected to be exceeded once every m years. If there are n_y observations per year, then the $m-$year return level, provided that m

Table 3: Maximum likelihood estimates and corresponding 95% confidence intervals (CI) for scale and shape parameters

	Scale α		Shape k	
Threshold (in millions)	MLE	95% CI	MLE	95% CI
1.17	0.1887	(0.1658, 0.2116)	-0.3770	(-0.4811, -0.2729)
1.21	0.2125	(0.1843, 0.2407)	-0.4964	(-0.5799, -0.4129)
1.25	0.1917	(0.1637, 0.2197)	-0.4934	(-0.5863, -0.4005)
1.29	0.1624	(0.1359, 0.1889)	-0.4591	(-0.5622, -0.3560)
1.33	0.1318	(0.1075, 0.1561)	-0.4075	(-0.5231, -0.2919)
1.37	0.0928	(0.0726, 0.1130)	-0.2804	(-0.3112, -0.2496)
1.41	0.0678	(0.0478, 0.0878)	-0.1675	(-0.1928, -0.1422)

is sufficiently large to ensure that $x_m > u$, is defined by

$$x_m = \begin{cases} u + (\alpha/k)\left[(mn_y\zeta_u)^k - 1\right] & \text{for } k \neq 0, \\ u + \alpha \log(mn_y\zeta_u) & \text{for } k = 0, \end{cases}$$

where $\zeta_u = Pr(X > u)$, u is the threshold value, α and k are GPD scale and shape parameter respectively. Estimation of return levels requires the substitution of parameter values by their estimates. For α and k this corresponds to substitution by the corresponding estimates with lowest standard error (Table 2). With an exception of threshold value 1.45, MLE provides the estimates with lowest standard error and that is why we use MLE estimates in our calculations. An estimate of ζ_u, the probability of an individual observation exceeding the threshold u, is also needed. This has a natural estimator of $\hat{\zeta_u} = r/n$, the sample proportion of points exceeding u. Since the number of exceedances of u follows the binomial distribution $Bin(n, \zeta_u)$, $\hat{\zeta}_u$ is also the maximum likelihood estimate of ζ_u (Coles and Simiu, 2003). Considering 1.25 as the threshold, the $m-$year return level ($\hat{x}_m$) for different values of m are obtained (Table 4) using the return level formula in Equation (5.1). For example, 50 year return level is 1.6118 million CFS, which means that 1.6118 million CFS or more water flow is likely to occur at the Tarbert Landing Discharge in the Mississippi River once every 50 years.

Table 4: Return level summary (in million CFS)

Year	5	10	20	30	50	100
Return	1.5552	1.5793	1.5965	1.6041	1.6118	1.6195

5.1 Uncertainty of the return level estimates

An accurate estimate of the uncertainty associated with parameter estimate is important to avoid misleading inference. This uncertainty is usually measured by any standard error, coefficient of variation and confidence level of the parameter of interest. The estimated confidence level is claimed to include the true parameter value with a specified probability. In the previous section we obtained the estimates of an $m-$year return level for specific values of m. The aim of this section is to calculate the uncertainty of return level estimates. We use the technique of bootstrapping to calculate this uncertainty. In many statistical problems we seek information about the value of a population parameter θ by drawing a random sample $\mathbf{Y}$ from that population and constructing an estimate $\hat{\theta}(\mathbf{Y})$ of the value of θ from that sample. The bootstrap principle is to obtain information about the relationship between θ and random variable $\hat{\theta}(\mathbf{Y})$ by looking at the relationship between $\hat{\theta}(\mathbf{y}_{obs})$ and $\hat{\theta}(\mathbf{Y}^*)$, where $\mathbf{Y}^*$ is a resample characterized by the sample $\mathbf{y}_{obs}$. $\mathbf{Y}^*$ can either be constructed by sampling with replacement from the data vector $\mathbf{y}_{obs}$, the so-called non-parametric bootstrap, or by sampling from the distribution function parameterized by $\hat{\theta}(\mathbf{y}_{obs})$, the so-called parametric bootstrap (Carpenter and Bithell, 2000). Here we discuss only non-parametric bootstrap approach.

Non-parametric resampling makes no assumptions concerning the distribution of, or model for, the data. Our data is assumed to be a vector $\mathbf{y}_{obs}$ of n independent observations, and we are interested in a confidence interval for $\hat{\theta}(\mathbf{y}_{obs})$. The general algorithm for a non-parametric bootstrap works as follows: a sample of size n is drawn, with replacement, from $\mathbf{y}_{obs}$ to obtain a bootstrap data set, denoted by $\mathbf{Y}^*$. From this data set the statistic of interest, $\hat{\theta}^* = \hat{\theta}(\mathbf{Y}^*)$ is calculated. This procedure can be repeated many times, say R times, to obtain estimates $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_R$. The series $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_R$ can be regarded as a sample from a distribution that approximates the sampling distribution of $\hat{\theta}(\mathbf{y}_{obs})$. Accordingly, standard errors, coefficient of variation and confidence intervals can be obtained from the simulated $\hat{\theta}^*_i$ series (Coles and Simiu, 2003).

For our analysis the original data set consists of 227 observations above the threshold $u = 1.25$ million. A random sample of size 227 is selected with replacement from the original data set to obtain a bootstrap data set. From this bootstrap data set the maximum likelihood estimates $\hat{\alpha}$ and $\hat{k}$ of the parameters are calculated and using these estimates the return level ($\hat{x}_m$) is determined for each $m = 5, 10, 20, 30, 50,$ and 100. This procedure is repeated 1,000 times. Therefore, for each return period (m) we have 1,000 values of return level. From this set of 1,000 values the mean return level, standard error, coefficient of variation, and 95% confidence intervals are calculated. For each of the specified return levels, the information is summarized in Table 5.2. Results in Table 5.2 shows that the standard error, the coefficients of variation and 95% confidence intervals for return level remains stable throughout the return period which indicates a stable model for the data.

Though we have illustrated our methodology in detail only for the threshold 1.25, we also applied the technique to the data for each of the following thresholds, u =1.17, 1.21, 1.25, 1.29, 1.33, 1.37, 1.41, and 1.45. A summary of this analysis (Table 5.3) shows the number of data that exceed the threshold, the 20, 30, 50, and 100-years return level estimates, together with 95% confidence intervals. It is seen that for thresholds between 1.17 and 1.41, the 30-year, 50-year and 100-year return level estimates do not differ greatly. For 20-year return level, the estimates are almost indifferent for all threshold values. This implies the model gives similar results for a threshold value between 1.17

and 1.41.

Table 5: Mean, standard error (SE), coefficient of variation (CV), and 95% confidence intervals (CI) for return level estimates from bootstrap resampling (threshold=1.25).

Period (in years)	Mean	SE	CV	95% CI
5	1.554	0.0098	0.629	(1.534, 1.571)
10	1.578	0.0095	0.600	(1.557, 1.593)
20	1.595	0.0093	0.582	(1.573, 1.609)
30	1.603	0.0092	0.576	(1.581, 1.617)
50	1.611	0.0093	0.575	(1.588, 1.624)
100	1.618	0.0094	0.580	(1.597, 1.633)

Table 6: Return level estimates for different years with 95% bootstrap confidence intervals from the GPD model for several threshold values (number of observations in tail)

Threshold	$20-$years	$30-$years	$50-$years	$100-$years
1.17(392)	1.594[1.569, 1.619]	1.605[1.579, 1.632]	1.617[1.588, 1.646]	1.629[1.598, 1.665]
1.21(275)	1.597[1.573, 1.609]	1.604[1.579, 1.616]	1.612[1.588, 1.624]	1.619[1.595, 1.633]
1.25(227)	1.597[1.573, 1.609]	1.604[1.581, 1.617]	1.612[1.588, 1.624]	1.619[1.597, 1.633]
1.29(189)	1.595[1.569, 1.609]	1.603[1.576, 1.617]	1.612[1.582, 1.625]	1.621[1.589, 1.635]
1.33(155)	1.593[1.569, 1.609]	1.602[1.575, 1.619]	1.613[1.583, 1.629]	1.622[1.592, 1.639]
1.37(126)	1.590[1.563, 1.612]	1.602[1.572, 1.625]	1.615[1.582, 1.639]	1.629[1.594, 1.658]
1.41(83)	1.589[1.557, 1.615]	1.605[1.568, 1.632]	1.622[1.581, 1.653]	1.643[1.596, 1.682]
1.45(45)	1.596[1.552, 1.627]	1.622[1.568, 1.665]	1.656[1.587, 1.726]	1.709[1.613, 1.835]

6 Concluding Remarks

Estimation of parameters for multiparameter families is challenging because it requires optimization of complicated functions or solving non-linear equations. That is why finding an appropriate and efficient estimation method for GPD parameters is always of interest. We have studied popularly used estimation techniques such as MLE, MOM, and PWM in the light of hydrological variable. Though no estimation technique is ideal for every situation, we can see MLE gives us the best fit for this data set. We have used river flow data to model flood design which has rarely been used before.The fitted model appears to have captured the key empirical features of the data set. Although

the threshold selection is complicated in hydrological examples, the bootstrap approach provides a reasonable way to choose such a threshold in this study. Our nonparametric bootstrap approach provides accurate estimate of the uncertainty associated with parameter estimation. Stable return level throughout the return period indicates a stable model for the data.

References

[1] Arns, A., Wahl, T. , Haigh, I., Jensen, J. and Pattiaratchi, C. (2013). Estimating extreme water level probabilities: A comparison of the direct methods and recommendations for best practice. *Coastal Engineering*, 81, 51-66.

[2] Beisel, C., Rokyta, D. , Wichman, H., and Joyce, P. (2007). Testing the extreme value domain of attraction for distributions of beneficial fitness effects. *Genetics*, 176, 2441-2449.

[3] Carpenter, J., Bithell, J., (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, *Statistics in Medicine*, 19, 1141-1164.

[4] Castillo, E., and Hadi, A. S. (1997). Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, 92(440), 1609-1620.

[5] Castillo, E., Hadi, A.S., Balakrishnan, N., and Sarabia, J.M. (2005). *Extreme Value and Related Models with Applications in Engineering and the Sciences.* Wiley.

[6] Coles, S. and Simiu, E. (2003). Estimating uncertainty in the extreme value analysis of data generated by a hurricane simulation model. *Journal of Engineering Mechanics*, 129(11), 1288–1294.

[7] Das, K. and Dey, A. (2016). Quantifying the risk of extreme aviation accidents. *Physica A: Statistical Mechanics and Applications*, 463, 345-355.

[8] Daspit, A. and Das, K. (2012). The generalized Pareto distribution and threshold analysis of normalized Hurricane damage in the United States Gulf Coast. *Joint Statistical Meetings (JSM) Proceedings, Statistical Computing Section*, Alexandria, VA: American Statistical Association, 2395-2403.

[9] Deidda, R. and Puliga, M.(2009). Performances of some parameter estimators of the generalized Pareto distribution over rounded-off samples.*Physics and Chemistry of the Earth*, 34, 626-634.

[10] Dey, A. K., and Das, K. (2014). Modeling extreme hurricane damage in the United States. *Joint Statistical Meetings (JSM) Proceedings, Section on Risk Analysis*, Alexandria, VA: American Statistical Association, 4356-4365.

[11] Dey, A. K., and Das, K., (2016). Modeling extreme Hurricane damage using the generalized Pareto distribution. *American Journal of Mathematical and Management Sciences*, 35(1), 55-66.

[12] Devore, J. L. (2010). *Probability and Statistics for Engineering and the Sciences, Eighth edition*. Brooks/Cole.

[13] Edwards, A. and Das, K. (2016). Using the statistical approach to model natural disasters. *American Journal of Undergraduate Research*, 13(2), 87-104.

[14] Mackay, E.B.L., Challenor, P.G. and Bahaj, A.S. (2011). A comparison of estimators for the generalized Pareto distribution. *Ocean Engineering*, 38, 1338-1346.

[15] Ferreira, J.A., and Guedes, S.C. (1998). An application of the peaks over threshold method to predict extremes of significant wave height. *Journal Offshore Mechanics Arctic Engineering*, 120, 165-176.

[16] Ghosh, S. and Resnick, S. (2010). A Discussion on mean excess plots. *Stochastic Processes and their Applications*, 120(8), 1492-1517.

[17] Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form.*Water Resour. Res.*, 15(5), 1049-1054.

[18] Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3), 339-349.

[19] Jockovic, J. (2012). Quantile estimation for the generalized Pareto distribution with application to finance. *Yugoslav Journal of Operations Research*, 22(2), 297-311.

[20] Klugman, S., Panjer, H., and Willmot, G. (2012). *Loss Models: From Data to Decisions, 4th edition.* Wiley.

[21] Mathiesen, M., Goda, Y., Hawkes, P., Mansard, E., Martin, M.J., Peltier, E., Thompson, E., and van Vledder, G. (1994). Recommended practice for extreme wave analysis. *Journal of Hydraulic Research*, 32(6), 803-814.

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

[23] Singh, V. P., and Guo, H. (1995). Parameter estimation for 3-parameter generalized Pareto distribution by the principle of maximum entropy (POME). *Hydrological Sciences Journal*, 40(2), 165-81.

[24] Sterl, A. and Caires., S. (2005). Climatology, variability and extrema of ocean waves: the Web-based KNMI/ERA-40 wave atlas. *International Journal of Climatology*, 25(7), 963-977.