ASM Sci. J., 11(2), 86-102

Enhancement of Boxplot Characters for Model Diagnostic of Block Maximum Extremal Events

Babangida Ibrahim Babura[1,2*], Mohd Bakri Adam[2], Anwar Fitrianto[3] and Abdul Rahim, A. S.[4]

[1]*Department of Mathematics, Federal University Dutse, Nigeria*

[2]*Institute for Mathematical Research, University Putra Malaysia*

[3]*Department of Statistics, Bogor Agricultural University, Indonesia*

[4]*Department of Economics, University Putra Malaysia*

A boxplot is an exploratory data analysis (EDA) tool for a compact visual display of a distributional summary of a univariate data set. It is designed to capture all typical observations and displays the location, spread, skewness and the tail of the data. The precision of some of this functionality is considered to be more reliable for symmetric data type and thus less appropriate for skewed data such as the extreme data. Many observations from extreme data were mistakenly marked as outliers by the Tukey's standard boxplot. A new boxplot implementation is presented which adopts a fence definition using the extent of skewness and enhances the plot with additional features such as a quantile region for the parameters of generalized extreme value (GEV) distribution in fitting an extreme data set. The advantage of the new superimposed region was illustrated in term of batch comparison of extreme samples and an EDA tool to determine search region or direction as contained in the optimisation routines of a maximum likelihood parameter estimation of GEV model. A simulated and real-life data were used to justify the advantages of the boxplot enhancement.

Keywords: Boxplot, enhancement, generalized extreme value

I. INTRODUCTION

Boxplot is considered one of the most popular exploratory data analysis (EDA) visual tool that receives a considerable amount of interest since its introduction as a schematic plot by Tukey in 1977 (Tukey, 1977). The philosophy behind boxplot construction is purposely made to utilise its simplicity in displaying important features of the univariate data set. These features constitute mainly a capture of typical observations, study symmetry or tail behaviour, identify outliers, compare parallel batches of data sets and analysis of some distributional assumptions about data. It can also be used to supplement more complex displays about univariate information. The five values of significance used in constructing boxplot are; the upper fence, lower fence, the upper hinge (upper quartile), lower hinge (lower quartile), and the median (Tukey, 1977). In a univariate data set up, the Tukey's standard boxplot was designed to capture data within the fence mark-up region given by

$$[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$$

of regular observations in X. Where X is the data set, Q_1, Q_3 are first and third quartiles of X respectively, and IQR is the interquartile range of X. Many studies are made to adjust the Tukey's fence mark-up according to data characteristics, for example, Frigge *et al.* (1989),

*corresponding author: bibabura@gmail.com

Kimber (1990), Carling (2000), Schwertman *et al.* (2004), Schwertman and de Silva (2007), Hubert and Vandervieren (2008), Bruffaerts *et al.* (2014) and Babura *et al.* (2017). The work of Babura *et al.* (2017) specifically adjusts the fence to accommodate extreme data characteristics.

The literature on boxplot characters is not limited to fence adjustment as some boxplot variants have been developed to incorporate additional display requirement to enhance visualization and analysis of some special type of dataset. For example, the variable width and notch boxplot by McGill *et al.* (1978) which embedded the notch in the boxplot to represent confidence interval around the median, the circular boxplot for circular data by Abuzaid *et al.* (2012) whereby circular display characters is reflected in the circular boxplot, K-boxplot for mixture data by Qarmalah *et al.* (2016) in which multimodality character was incorporated in the k-boxplot. Although the work of Babura *et al.* (2017) is for extreme data but limited to fence adjustment which is a diagnostic character in extreme event modelling with the ability to determine outlying observations from the dataset. So, an additional boxplot characters which reflect relative information about fitting parameters of the extreme modelling tools is considered as necessary to be develop. The framework in this paper is limited to block maximum extreme events with generalized extreme value distribution (GEV) as the modelling tool.

The paper extends the work of Babura *et al.* (2017) in visualising extreme data so that the modified boxplot can be part of the classical EDA tools such as histogram, qq plot and density plot, that are more popular in extreme model diagnostic. After introducing some important concepts and describing the methodology involved in the next sections, a simulation experiment is formulated whichresulting to enhancment of the diagnostic properties of boxplot. The classical features of boxplot were maintained with the adoption of fence definition as proposed in Babura *et al.* (2017) for a proper capture of the regular extreme observations. The fitting parameters regions using quantile estimate were proposed and embedded into the new boxplot. The enhancement enables the proposed boxplot to have an additional diagnostic feature for fitting an extreme data sample to GEV distribution model. A simulated and real-life data were used to show the advantages of this development over those found in the literature.

II. MATERIALS AND METHODS

In this section, important statistical measures and concept are described along with the implementation of the methodological framework of the research work.

A. Sample quantiles

Let $\{X_{(1)}, \ldots, X_{(n)}\}$ denote the order statistics of a sample $\{X_1, \ldots, X_n\}$ of independent identically distributed (i.i.d.) random variables from a distribution F and suppose $Q_i(p)$ denote the ith sample quantile. Then, the quantile of a distribution F is given by

$$Q_{(p)} = F^{-1}(p) = \inf\{x : F(x) \geq p\} \qquad (1)$$

There are a number of equivalent way of defining quantile estimates, in which two defi-

nitions from among the ones described by Hyndman and Fan (1996) based on order statistics are considered. These definitions have a general form which is a representation according to weighted averages of consecutive order statistics and is given by

$$\hat{Q}_{i(p)} = (1-\gamma)X_{(j)} + \gamma X_{(i+1)} \tag{2}$$

where $\frac{j-m}{n} \leq p < \frac{j-n+1}{n}$ for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$. The value of γ is a function of $j = \lfloor pn+n \rfloor$ and $g = pn+m-j$with $\lfloor . \rfloor$ denoting the greatest integer function.

B. Block maximum events and the generalized extreme value (GEV) distribution

Extreme data are records of events that are more extreme than any that have already been observed within a particular uniform block of period. The current development in global warming which signifies a considerable interest in environmental research and financial crisis a consequence from so much volatility in the financial sector, are some of the events that give rises to a universal interest in modeling and forecasting of extreme events. Jenkinson, (1955) introduced the GEV distribution for modelling the distribution of extremal events in meteorological data with unknown limiting form of extreme value distribution. The GEV distribution was described to represent the three families of extreme value distributions: Gumbel, Fréchet and Weibull type distributions. GEV distribution focus on the statistical behavior of block maximum events $M_n = max\{X_1, ..., X_n\}$ where $X_1, ..., X_n$, is a sequence of independent random variables having a common distribution function F. The following theorem describes the limiting distribution of M_n.

Theorem 1 (Coles, 2001) *If there exists sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$, *as* $n \to \infty$, *such that* $Pr\{(M_n - b_n)/a_n \leq x\} \to G(x)$ *where* G *is a non-degenerate distribution function, then* G *is a member of the GEV family*:

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{3}$$

defined on $\{x : 1+\xi(\frac{x-\mu}{\sigma}) > 0\}$ *where* $\sigma > 0$ *and* $\mu, \xi \in \mathbb{R}$,

So,$G(x)$ is said to be Freichet if $\xi > 0$, Weibull if $\xi < 0$ and Gumbel if $\xi = 0$ with re-expression of the limiting distribution as

$$G(x) = \exp\left\{-\exp\left[-\frac{x-\mu}{\sigma}\right]\right\}, -\infty < x < \infty \tag{4}$$

The quantile function which is the relationship of the GEV model with its three parameters is given by;

$$Q_{(p)} = \begin{cases} \mu - \frac{\sigma}{\xi}\left(1 - z_p^{-\xi}\right), & for\ \xi \neq 0 \\ \mu - \sigma \log z_p, & for\ \xi = 0 \end{cases} \tag{5}$$

where $z_p = \log(1-p)$.

C. Maximum likelihood estimate method for the parameters of GEV distribution

The maximum likelihood estimate (MLE) method for estimate of GEV distribution parameters is proposed by Prescott and Walden

(1980) and is regarded as the most popular and efficient among parameter estimation methods. The MLE method involves maximising the likelihood function of a distribution given by

$$L(\theta) = \prod_{i=1}^{n} g\left(x\right) \tag{6}$$

where g is a known density function with parameter vector θ. In the case of GEV distribution θ is assumed to be a rational function, such that the parameter $\theta = (\mu, \sigma, \xi)$ maximize G directly or by maximising the logarithm of likelihood functions $\log L(\theta)$ or simply $\ell(\theta)$. Now, for a sample $\{x_i\}_1^n$ of independent identically distributed block maximum observation that follows a GEV distribution the log-likelihood function for the GEV parameters when $\xi \neq 0$ is given by

$$\ell(\theta) = \ell(\mu, \sigma, \xi) = -n\log\sigma - \left(1 + \frac{1}{\xi}\right)\sum\nolimits_{i=1}^{n} \log\left[1 + \xi\left(\frac{x_i - u}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right] \tag{7}$$

$$\text{such the } 0 < 1 + \xi\left(\frac{x_i - u}{\sigma}\right) \text{ for } i = 1, 2, 3, ...n \tag{8}$$

Coles (2001) revealed that at parameter combination for which Condition (8) did not hold, corresponding to a set-up which makes at-least single observation in the data falls beyond an end-point of the distribution, the likelihood is 0 and corresponding log-likelihood equals ∞.

However, the case when $\xi = 0$ requires the Gumbel limit of the GEV distribution, which leads to the following log-likelihood function

$$\ell(\theta) = \ell(\mu, \sigma) = -n\log\sigma - \sum\nolimits_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right) - \sum\nolimits_{i=1}\exp\left\{-\left(\frac{x_i - \mu}{\sigma}\right)\right\} \tag{9}$$

Maximisation of both Equations (7) and (9) with respect to the parameter vector $\theta = (\mu, \sigma, \xi)$ leads to the MLE of the parameters of GEV distribution. The two systems in Equations (7) and (9) have no analytical solution, but for any given dataset the maximisation is attained using standard numerical optimisation algorithms. The MLE described above is implemented for comparison in the performance with real data subsection of the study using quasi-Newton method optimisation algorithm proposed by Broyden, (1970).

D. Median-unbiased and distribution-free quantiles

The distribution-free quantile is regarded as the default quantile definition in R package. In particular, Hyndman and Fan (1996) suggested their quantile definition number 8 as the best because of its advantage over other proposed definitions that possess distribution free characters. The median-unbiased property is considered the advantage of definition 8. Hyndman and Fan (1996) determined the median position $MF(X_{(k)}) \approx \frac{k-\frac{1}{3}}{n+\frac{1}{3}}$ and defined the sample quan-

tile by setting $p_k \approx \frac{k-\frac{1}{3}}{n+\frac{1}{3}}$. Consequently, p_k becomes median-unbiased of order $^o(n^{-1/2})$ (Reiss, 1989). This quantile is optimal over all other median unbiased quantile estimators and possess translation equivariant property among others (Reiss, 1989).

Another important advantage of this quantile method in the GEV modelling framework is that the estimate has far less computational requirement especially when compared to estimate in Equation (5). In the distribution-free quantile estimate, prior knowledge or estimate of the modelling parameters is unnecessary. Therefore, unless otherwise stated, the median-unbiased and distribution-free quantile estimate was choosen for all the formulated simulations in the research work that requires a quantile estimate of a random sample from the GEV distribution. This choice allows the implementation of the proposed methods over GEV samples without determining the unknown parameters from the sample in practice.

E. Three boxplot methods for fence mark-up

The boxplot construction requires an ordered sample $X = (x_{(1)}, x_{(2)}..., x_{(n)}\}$ of size n along with an estimate of the three sample quartiles Q_1, Q_2 and Q_3 from X. Usually, the boxplot hinges are versions of the first and third quartile, i.e., close to quantiles of sample X at $p = \{0.25, 0.75\}$ such that the first and third quartiles of sample X are respectively given by $Q_1 = x_{(k)}$, and $Q_3 = x_{(n-k+1)}$ where $k = \frac{1}{2}\left\lfloor \frac{n+3}{2} \right\rfloor$ with $\lfloor . \rfloor$ as the greatest integer function. The second quartile Q_2 is referred to be the median of sample X. However, any boxplot implementation requires an estimate of a robust centre spread usually with interquartile range given by $R = Q_3 - Q_1$. A boxplot method is usually determined according to a fence estimate. In this paper, three fence estimation methods was considered the classical boxplot method (Tukey, 1977), the adjusted boxplot method (Hubert and Vandervieren, 2008), and the modified boxplot method (Babura *et al.*, 2017). The classical, adjusted and modified boxplot methods are constructed in a similar way except in the fence definition.

Generally, the lower fence estimate F_L and upper fence estimate F_U can be expressed as:

$$F_L = Q_1 - h_1 IQR, \tag{10}$$

$$F_U = Q_3 + h_2 IQR, \tag{11}$$

where $h_1 = h_2 = 1.5$ for the classical boxplot method by Tukey, (1977), $\{h_1 = 1.5e^{-4m}, h_2 = 1.5e^{3m}; M \geq 0\}$ or $\{h_1 = 1.5e^{-3m}, h_2 = 1.5e^{4m}; M < 0\}$ for the adjusted boxplot method by Hubert and Vandervieren, (2008), and $\{h_1 = 1.5e^{-4\delta}, h_2 = 1.5e^{6\delta}; \delta \geq 0\}$ or $\{h_1 = 1.5e^{6\delta}, h_2 = 1.5e^{-4\delta}; \delta < 0\}$ for the modified boxplot method by Babura *et al.*, (2017). M and δ are estimates of medcouple and Bowley skewness measures respectively. Furthermore, the interval $F = [\min\{x \in X; x \geq F_L\}, \max\{x \in X; x \leq F_U\}]$ is referred as the fence cut off region.

Then the construction of a boxplot constitutes a rectangular box, which captures the middle batch of the ordered sample observations which span from the first quartile Q_1 to the third

quartile Q_3. A line is drawn to divide the box into two indicates the position of the median value. Additional lines extend outward from the two ends of the box to two adjacent fence values. The fence values are marked to capture all regular sample observations that are not flagged as outliers by the fence rule of a particular boxplot method. Finally, any data point outside the interval $[F_L, F_U]$ data points are plotted individually above or below the fence cut-off and referred to as potential outliers.

McGill *et al.* (1978) proposed an additional feature to the boxplot construction called a notch, in which the notch area of a boxplot represents confidence interval around the median typically computed as $Q_2 \pm 1.58IQRn$ for a Gaussian sample.

F. Determination of the GEV distribution parameters regions

Let F be a GEV distribution and $X \sim F(\mu, \sigma, \xi)$ where μ, σ, and ξ are the location, scale and shape parameters of F respectively. The quantile bands for the location and scale parameters of the GEV distribution are obtainable based on the following simulation processes:

- generate a sample X from F with fixed location ($\mu = 0$), scale ($\sigma = 1$) and varying the shape parameter ($\xi_i = -0.8 + 0.1(i - 1), 1 \leq i \leq 109$
- based on resampling of $X \sim F(\mu, \sigma, \xi_i)$ values of the following populations were generated;

$$A_{(\xi_i)} = \{p_i, Q_{p_i} \approx \mu\} \tag{12}$$

$$B_{(\xi_i)} = \{q_i; Q_1 - Q_{q_i} \approx \sigma\} \tag{13}$$

$$C_{(\xi_i)} = \{\delta(X_i); X_i \sim F(0, 1, \xi_i\} \tag{14}$$

where the quantile positions $pj, qj \in (0.25, 0.75), 1 \leq j \leq n$; where $n = 100$ as the sample size of X.

- the resampling process for each fixed choice of parameters was repeated 5000 times to obtain the collections that form the populations in $A_{(\xi_i)}, B_{(\xi_i)}, C_{(\xi_i)}$.

III. RESULTS AND DISCUSSION

The implementation of the simulation experiment returns a quantile band for the location and scale parameters of GEV distribution and skewness estimate of the shape parameter. These bands are reflected in the newly proposed boxplot as a location and scale parameters region for a sample from GEV distribution. The GEV distribution has three parameters namely location, scale, and shape parameters.

A. The location parameter region

Figure 1 illustrates the outcome of the simulation experiment as described earlier. Each boxplot displays the overall band of the return population $A_{(\xi_i)}$ for their corresponding values of ξ_i. All the simulated groups exhibit the same distributional characteristics in the centre and spread. The display of similar distributional character indicates that; the quantile positions p_i's are invariant over the variation in ξ of F for

estimating a quantile band of the GEV distribution location parameter. Therefore, the $Q_{(p_i)}$ band for μ is estimated to be the 95 percentile bands of its associated $A_{(\xi_i)}$. Thus, the location parameter region of F in a boxplot is marked to be the averages of lower and upper limits of $Q_{(p_i)}$ bands, corresponding to $A_{(\xi_i)}$'s respectively.

Table 1 is an extraction of the lower and upper limits of the 95 percentiles bands of $A_{(\xi_i)}$'s from the simulation result to which the overall average 95 percentile band is $[0.28, .046]$. Thus, the result of this simulation experiment is obtained to be the boxplot quantile region of the location parameter (μ) of the GEV distributed extreme dataset is obtained to be within the interval $[Q_{0.28}, Q_{0.46}]$.

B. The scale parameter region

Figure 2 illustrates a boxplot display of simulated populations $B_{(\xi_i)}$. Each boxplot displays the overall band of the return population $B_{(\xi_i)}$ for their respective values of varying ξ_i. Most of the simulated groups exhibit the same distributional characteristics in the centre and spread with little deviation for scenarios when $\xi_i \in [-0.8, -0.3]$ which can be ignored. Conclusively, the q_i's are invariant over the variation in ξ of F for estimating a quantile band of the scale parameter σ.

If $[\alpha_{1i}, \alpha_{2i}]$ is the 95 percentile bands of $B_{(\xi_i)}$, then its associated quantiles defined on $[Q_{\alpha 1i}, Q_{\alpha 2i}]$ would yield a band $[Q_{\alpha 1i} - Q_1, Q_{\alpha 2i} - Q_1]$ which capture the actual value of the scale parameter σ of F. Consequently, the overall band for the scale parameter in a boxplot is marked at $[Q_{\alpha 1} - Q_1, Q_{\alpha 2} - Q_1]$ where $\alpha 1$ and $\alpha 2$ are the averages over $\alpha 1i$'s and $\alpha 2i$'s of $B_{(\xi_i)}$'s respectively.

Table 2 shows the simulation results of $[\alpha 1i, \alpha 2i]$ for $B_{(\xi_i)}$ which consequently gives the approximate average overall band of $[0.42, 0.60]$. Thus, the result of the simulation experiment which proposed the boxplot quantile region of the scale parameter (δ) of the GEV distributed extreme data is considered as the interval $[Q_{0.42} - Q_1, Q_{0.60} - Q_1]$.

C. The skewness estimate of the shape parameter

The overall simulation band for each $C_{(\xi_i)}$ was illustrated in Figure 3 as a boxplot. As observed, a corresponding increase in lower and upper limit of the band of δ for $C_{(\xi_i)}$ as the shape parameter ξ of GEV distribution increases. To establish a relation between δ and ξ, we extract the medians of $C_{(\xi_i)}$'s and associate it with its corresponding ξ_i as a respond to a resistance fit model. If δ_i is the medians of $C_{(\xi_i)}$, we then fit ξ to be given by $\xi = a + b\delta$ where a and b resistance fit parameters.

Figure 4(a) illustrates the behaviour of the GEV distribution's shape parameter ξ in response to the extent of skewness measured as Bowley coefficient δ. In order to have a good fit for the relation illustrated in Figure 4(a), to which a response when median skewness is below 0.8 is considered. To justify this choice, we observe that due to the long tail behaviour of GEV samples with shape parameter $\xi \geq 2.5$ the boxplot visualisation of such sample becomes difficult, especially to comprehend the information around the box. Boxplot visualisation of the scenario $\delta > 0.8$ or $\xi > 2.5$ will remain a limitation in practice. Figure 4(b) shows the resistance fit

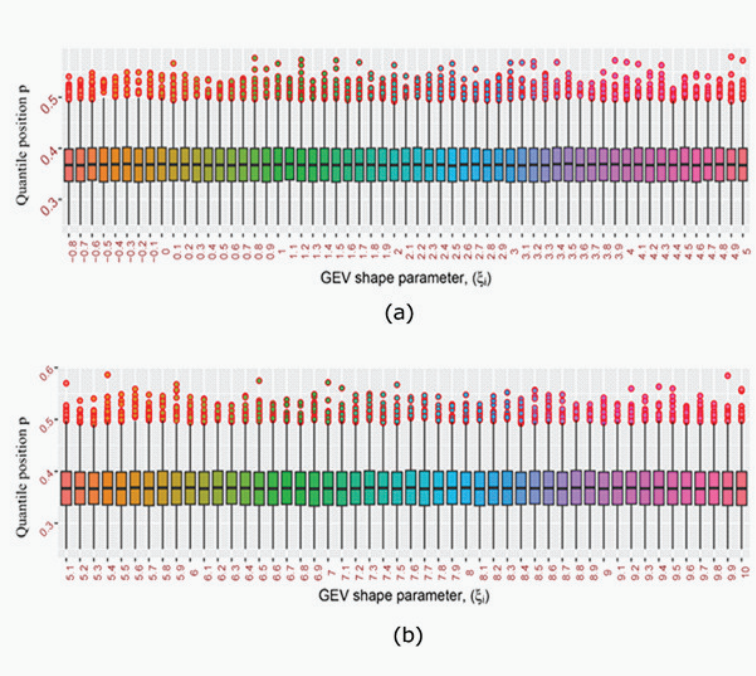

Figure 1. Boxplot display of simulation band for $A_{(\xi i)}$

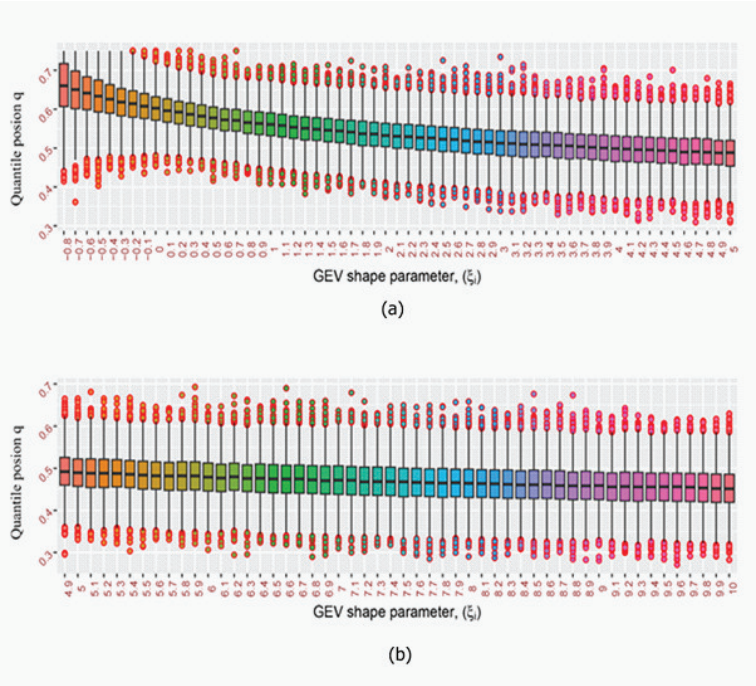

Figure 2. Boxplot display of simulation band for $B_{(\xi i)}$

based on this choice and gives an estimate of the ξ and δ given by parameters a and b to define the relation between

$$\hat{\xi} = \frac{31\delta - 4}{10}$$

Table 1. Percentile band for the quantile position p of the location parameter

ξ_i	L_p	U_p	ξ_i	L_p	U_p	ξ_i	L_p	U_p	ξ_i	L_p	U_p
0.80	0.28	0.46	2.00	0.28	0.46	4.80	0.27	0.47	7.60	0.28	0.46
0.70	0.28	0.46	2.10	0.28	0.46	4.90	0.28	0.46	7.70	0.27	0.46
0.60	0.28	0.46	2.20	0.27	0.46	5.00	0.27	0.46	7.80	0.28	0.46
0.50	0.28	0.46	2.30	0.27	0.47	5.10	0.27	0.46	7.90	0.28	0.46
0.40	0.28	0.46	2.40	0.27	0.46	5.20	0.28	0.46	8.00	0.28	0.46
0.30	0.28	0.46	2.50	0.28	0.47	5.30	0.27	0.46	8.10	0.27	0.46
0.20	0.28	0.47	2.60	0.28	0.46	5.40	0.28	0.47	8.20	0.28	0.46
0.10	0.28	0.46	2.70	0.28	0.46	5.50	0.28	0.46	8.30	0.28	0.47
0.00	0.28	0.46	2.80	0.28	0.46	5.60	0.27	0.46	8.40	0.28	0.46
0.10	0.28	0.46	2.90	0.28	0.46	5.70	0.28	0.47	8.50	0.28	0.46
0.20	0.28	0.46	3.00	0.28	0.46	5.80	0.28	0.47	8.60	0.28	0.46
0.30	0.28	0.46	3.10	0.28	0.46	5.90	0.27	0.46	8.70	0.27	0.46
0.40	0.28	0.46	3.20	0.27	0.46	6.00	0.28	0.46	8.80	0.27	0.47
0.50	0.28	0.46	3.30	0.28	0.46	6.10	0.28	0.46	8.90	0.28	0.47
0.60	0.28	0.46	3.40	0.28	0.46	6.20	0.28	0.47	9.00	0.27	0.46
0.70	0.28	0.46	3.50	0.28	0.46	6.30	0.27	0.46	9.10	0.28	0.46
0.80	0.27	0.46	3.60	0.28	0.46	6.40	0.28	0.46	9.20	0.28	0.47
0.90	0.28	0.46	3.70	0.28	0.46	6.50	0.28	0.46	9.30	0.28	0.46
1.00	0.28	0.46	3.80	0.28	0.46	6.60	0.28	0.46	9.40	0.27	0.46
1.10	0.28	0.46	3.90	0.28	0.46	6.70	0.28	0.46	9.50	0.28	0.46
1.20	0.28	0.46	4.00	0.28	0.46	6.80	0.27	0.46	9.60	0.28	0.46
1.30	0.27	0.46	4.10	0.28	0.46	6.90	0.28	0.46	9.70	0.27	0.46
1.40	0.28	0.46	4.20	0.27	0.46	7.00	0.27	0.46	9.80	0.28	0.46
1.50	0.28	0.47	4.30	0.28	0.46	7.10	0.28	0.46	9.90	0.28	0.46
1.60	0.28	0.46	4.40	0.28	0.46	7.20	0.28	0.46	10.00	0.27	0.46
1.70	0.28	0.46	4.50	0.28	0.46	7.30	0.28	0.46			
1.80	0.28	0.46	4.60	0.28	0.46	7.40	0.27	0.46			
1.90	0.28	0.46	4.70	0.27	0.46	7.50	0.28	0.46			

Lp and *Up* are respectively lower and upper limits of the 95 percentile bands of A_{ξ})

D. Implementation of the proposed boxplot with parameters regions

The location and scale parameters region described earlier in this section were superimposed on the proposed boxplot as illustrated in Figure 5.

Figure 5 Proposed modified boxplot display of GEV distribution's location and scale parameter regions along with quantile estimate of shape parameter. The location parameter region was superimposed as a rectangular box shaded in forward slide arrays of green lines that span from $Q_{0.28}$ up to $Q_{0.46}$ of the data set. Also, the visualisation of the scale parameter where similarly superimposed as the length of a thick line along the vertical edges of the box. The distance of the line spans from Q_1 up to $Q_{0.42}$ and to $Q_{0.60}$ as the lower and upper bands for the scale parameter region respectively. The region $[Q_{0.42}, Q_{0.60}]$ is marked with two adjacent rectangles along the vertical edges of the box filled with

Table 2. 95 Percentile band for quatile position q of the scale parameter

ξ_i	α_{1i}	α_{2i}	ξ_i	α_{1i}	α_{2i}	ξ_i	α_{1i}	α_{2i}	ξ_i	α_{1i}	α_{2i}
-0.80	0.51	0.75	2.00	0.44	0.63	4.80	0.40	0.59	7.60	0.37	0.56
-0.70	0.52	0.75	2.10	0.45	0.62	4.90	0.40	0.59	7.70	0.37	0.56
-0.60	0.52	0.75	2.20	0.44	0.62	5.00	0.40	0.58	7.80	0.37	0.56
-0.50	0.52	0.75	2.30	0.44	0.62	5.10	0.40	0.58	7.90	0.37	0.56
-0.40	0.52	0.73	2.40	0.44	0.61	5.20	0.39	0.58	8.00	0.37	0.56
-0.30	0.52	0.72	2.50	0.43	0.62	5.30	0.39	0.58	8.10	0.37	0.56
-0.20	0.52	0.71	2.60	0.43	0.61	5.40	0.39	0.58	8.20	0.37	0.56
-0.10	0.52	0.70	2.70	0.43	0.61	5.50	0.39	0.58	8.30	0.37	0.56
0.00	0.52	0.69	2.80	0.43	0.61	5.60	0.39	0.58	8.40	0.37	0.56
0.10	0.51	0.69	2.90	0.43	0.61	5.70	0.39	0.58	8.50	0.37	0.56
0.20	0.51	0.68	3.00	0.42	0.61	5.80	0.39	0.58	8.60	0.37	0.56
0.30	0.50	0.67	3.10	0.42	0.60	5.90	0.38	0.58	8.70	0.36	0.56
0.40	0.50	0.67	3.20	0.42	0.60	6.00	0.39	0.58	8.80	0.37	0.56
0.50	0.50	0.66	3.30	0.42	0.60	6.10	0.39	0.57	8.90	0.37	0.56
0.60	0.49	0.66	3.40	0.42	0.61	6.20	0.38	0.58	9.00	0.36	0.55
0.70	0.49	0.65	3.50	0.41	0.60	6.30	0.38	0.58	9.10	0.36	0.55
0.80	0.48	0.65	3.60	0.42	0.60	6.40	0.38	0.57	9.20	0.36	0.55
0.90	0.48	0.65	3.70	0.41	0.59	6.50	0.38	0.57	9.30	0.36	0.56
1.00	0.47	0.64	3.80	0.41	0.60	6.60	0.38	0.57	9.40	0.36	0.55
1.10	0.47	0.64	3.90	0.41	0.59	6.70	0.38	0.57	9.50	0.36	0.55
1.20	0.47	0.64	4.00	0.41	0.60	6.80	0.38	0.57	9.60	0.36	0.55
1.30	0.47	0.64	4.10	0.40	0.59	6.90	0.38	0.57	9.70	0.36	0.55
1.40	0.46	0.63	4.20	0.40	0.59	7.00	0.38	0.57	9.80	0.36	0.55
1.50	0.46	0.63	4.30	0.40	0.59	7.10	0.38	0.57	9.90	0.36	0.55
1.60	0.45	0.63	4.40	0.40	0.59	7.20	0.37	0.56	10.0	0.36	0.55
1.70	0.46	0.63	4.50	0.40	0.59	7.30	0.37	0.57			
1.80	0.45	0.63	4.60	0.40	0.59	7.40	0.37	0.56			
1.90	0.45	0.62	4.70	0.40	0.59	7.50	0.37	0.56			

α_{1i} and α_{2i} are respectively lower and upper limits of the 95 percentile bands of $B_{(\xi i)}$

backwards slides arrays of blue lines.To display the shape parameter estimate, a textual value of the estimate is imposed and situated between Q_3 and upper fence of the boxplot. The proposed improvement of boxplot display for extreme data gives a significant improvement in capturing some additional information about the fitting parameters of a GEV distribution model.

E. Performance with simulation data

Figure 6 illustrates the advantages of the new method as compared to the notch boxplot regarding the batch comparison of extreme samples. Figure 6(a), (b) and (c) show the three scenarios of extreme samples that is: Weibull type $(G(x;\mu,1,-0.5))$, Gumbel type $(G(x;\mu,1,0))$ and Freichet type $(G(x;\mu,1,0.5))$. In each case, the first two samples have the same location parameter $(\mu = 0)$ while the third sample has different location parameter $(\mu = 1)$. In all the scenarios, the notch fails to capture the actual location parameter value as against the proposed region even though they both show the population differences against the third sample.

In general performance, a better capture

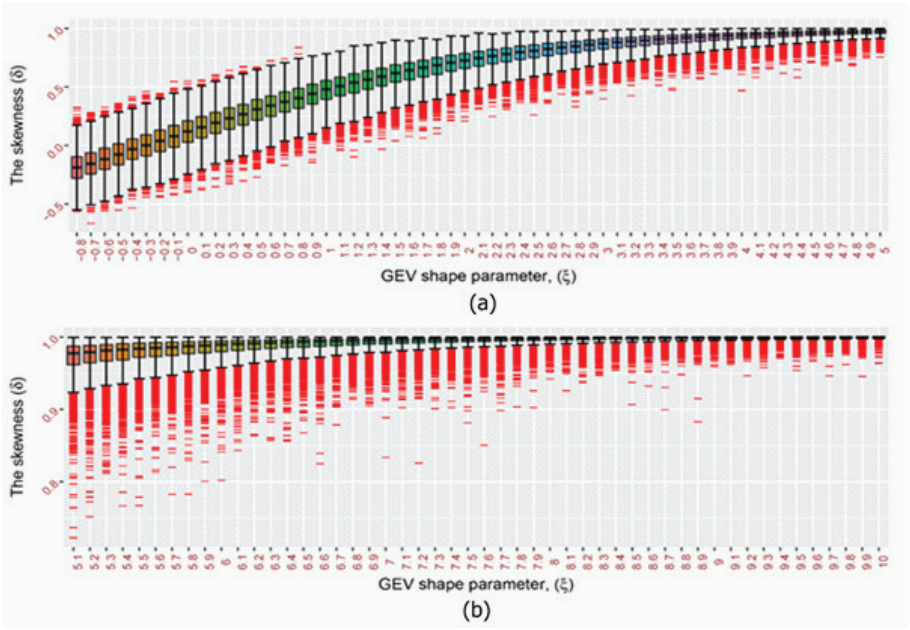

Figure 3. Boxplot display of simulation band for GEV sample skewness δ in $C_{(\xi i)}$

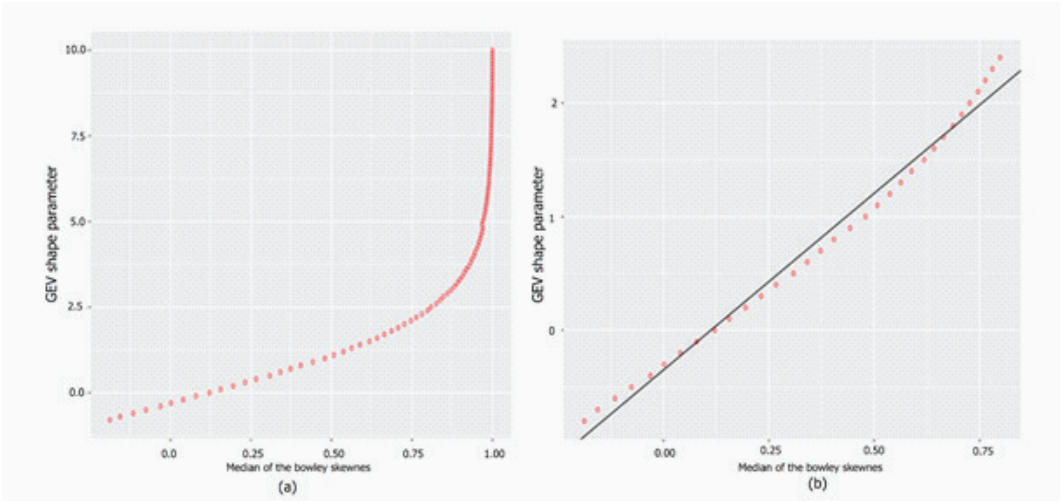

Figure 4. Median skewness of simulated GEV distribution's samples versus corresponding GEV shape parameter

(with no outlier) can be observed on the extreme data by the proposed methods as against the classical method. The illustration in Figure 6 indicates a better capture of the extreme data based on the in-cooperated fence definition and a better understanding of the actual position of the location parameter for batch comparison.

F. Performance with the real data

To explore the advantage of the proposed improvement, three batches of environmental extreme data sets were visualised. The annual maximum observed one-hour precipitation for 46 years i.e. from 1947 to 1993 source (Thomas and Nolan, 1997). The yearly maximum river flow

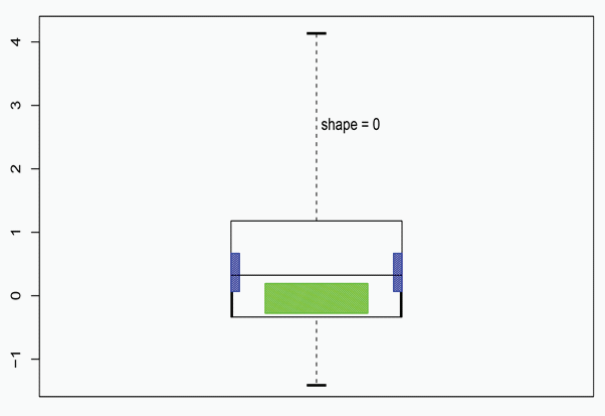

Figure 5. Proposed modified boxplot display of GEV distribution's location and scale parameter regions

discharge in cubic meters per second for 60 years source (Castillo, 1988) and the monthly maximum rainfall data values mm at Petaling Jaya record center Malaysia (source from Earth Observation Center, Institute of Climate Change, Universiti Kebangsaan Malaysia).

The three data sets were visualised with the proposed boxplot method as against the two other boxplot methods. The illustration in Figure 7 shows that the adjusted boxplot slightly improved the data capture of the classical method particularly in the upper fence region while the proposed modification not only capture the data correctly but also suggest some information about the fitting parameters in modeling the extreme data with GEV distribution. The density plot of best fitting parameters based on MLE estimate was plotted along with the density plots from the extracted parameters regions of the proposed modified boxplot.

Figure 8 is the proposed boxplot display on the left side along with density plot overlapping the dataset as a histogram on the left side. In all of the three scenarios, by chosing the upper band of the scale parameter, the quantile estimate of shape parameter and a combination with the lower and upper limits of the location parameter corresponding to a lower and upper band fit were obtained respectively to have an adequate capture of the data sets. The two fits; the lower and upper band fit have placed the best fit from MLE estimate in between them as expected. Either of the lower or upper band fit can be used to get insight into some inferential details about the data set. However, any of the band (lower/upper) can be used in identifying a search direction or as an initial point while implementing optimisation procedures required in some parameter estimation methods such as the MLE method.

The interesting thing here is the ability of the proposed parameters regions to generate an estimate that captures the data represented as a histogram and places the MLE fit between the lower band of the region to the upper band of the region which is referred to as lower band fit, and upper band fit respectively. This quantile estimate of the region is not an attempt to improve or give an alternative to the regular known parameter estimation methods. It gives an EDA idea about the actual position of the best fitting parameters. The estimates can also serve as a starting point while trying to implement optimisation procedures involved in the popular parameter estimation methods such as MLE. However, the advantage of the parameter region was demonstrated in Figure 8.

IV. CONCLUSION

The boxplot is a popular graphical EDA tool for analysis of one-dimensional data set. Among

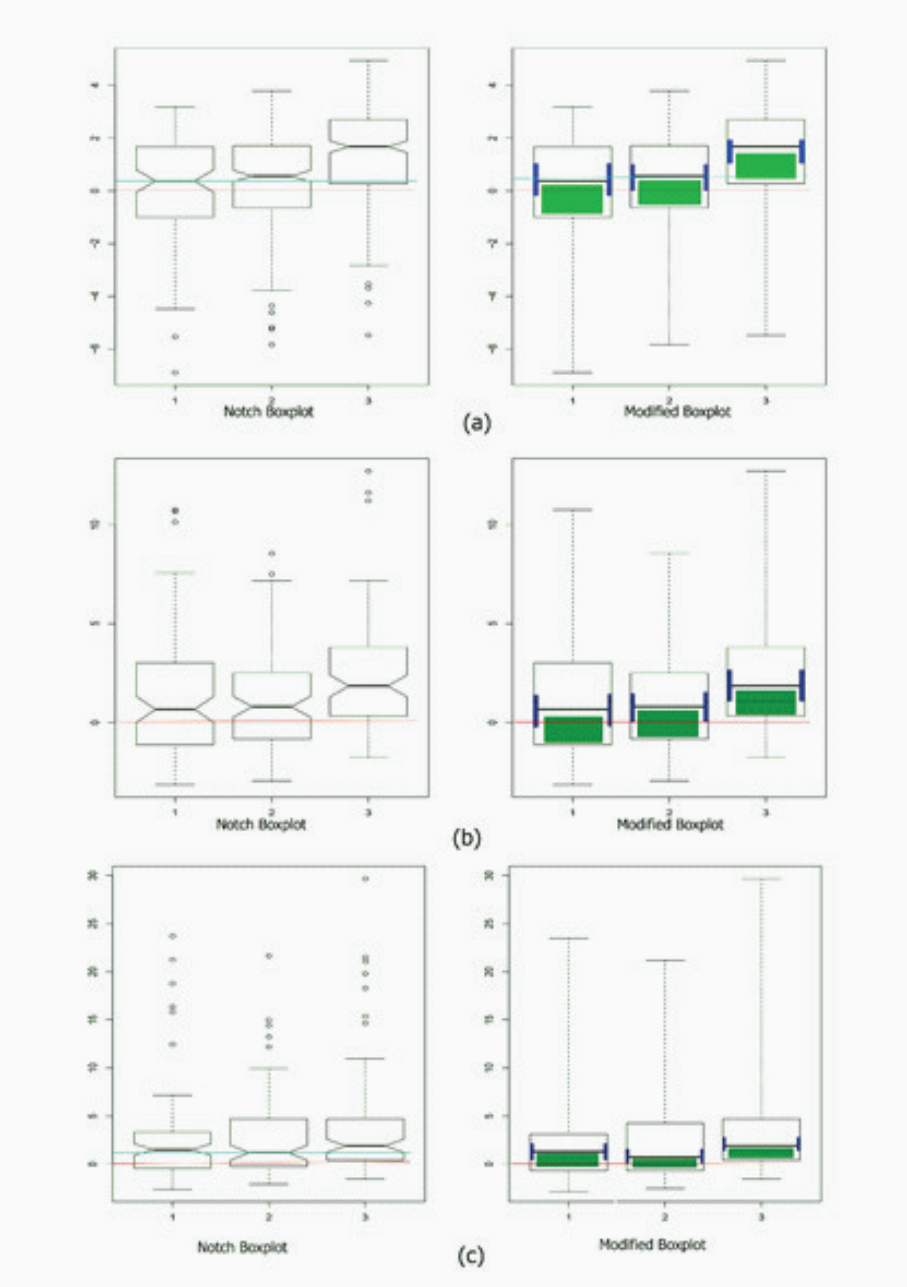

Figure 6. Comparison of newly improved modified boxplot with notch boxplot for batches of simulated GEV samples. In the figures the samples are generated from (a) A Weibull type GEV samples (b) A Gumbell type GEV Sample and (c) A Freichet type GEV samples.

its limitation is that; some of the quantile measures used does not suggest any significant properties about the data especially for data set that inherently possess skewed distribution properties like the extreme data. The functionality of boxplot was extended over an extreme sample by incorporating the fitting parameters region to the GEV distribution model. The region is a quantile estimate based on simulation experiment on samples from GEV distribution. The improve-

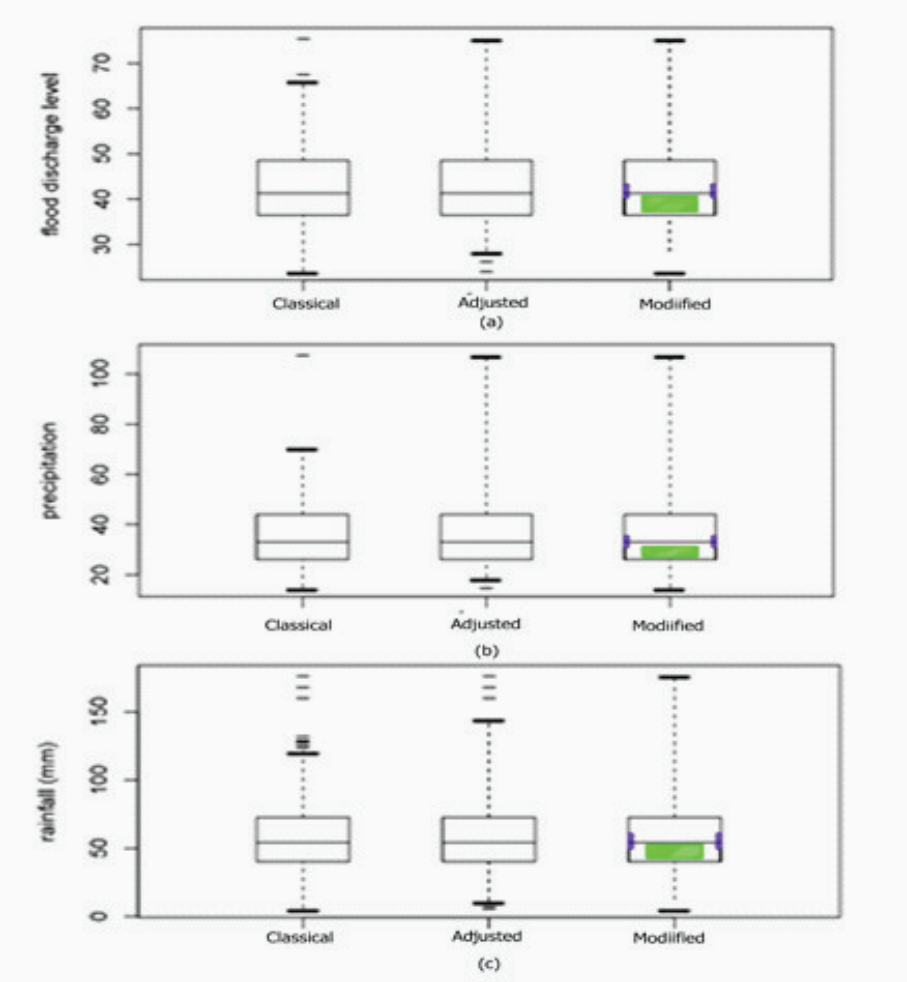

Figure 7. Comparison of three boxplot methods over some real-life data sets. Namely (a) The annual maximum observed one-hour precipitation for 46 years, (b) The yearly maximum river flow discharge in cubic meters per second for 60, (c) 31-year Monthly maximum rain

ment is not an attempt to substitute or improve the classical inferential methods used to estimate the modelling parameters. It is rather considered an EDA tool that supplements some estimation procedures which requires optimisation methods. By implication, the parameter region can be used to set an initial point and identify search direction in such optimization procedures. The advantage of the region in batch comparison of an extreme sample over the notch in the classical boxplot was shown. Also the use of the in-cooperated region to visualize population difference for batches of extreme samples was recommended. The general rules in using a boxplot should always apply despite the proposed improvement. Conclusively, the following rules should be taken into consideration while using boxplot with the proposed improvement:

- The boxplot construction requires at least 5 data points
- Use of the proposed parameters regions requires a sufficient extreme sample size of at least 30 drawn from a block-maximum

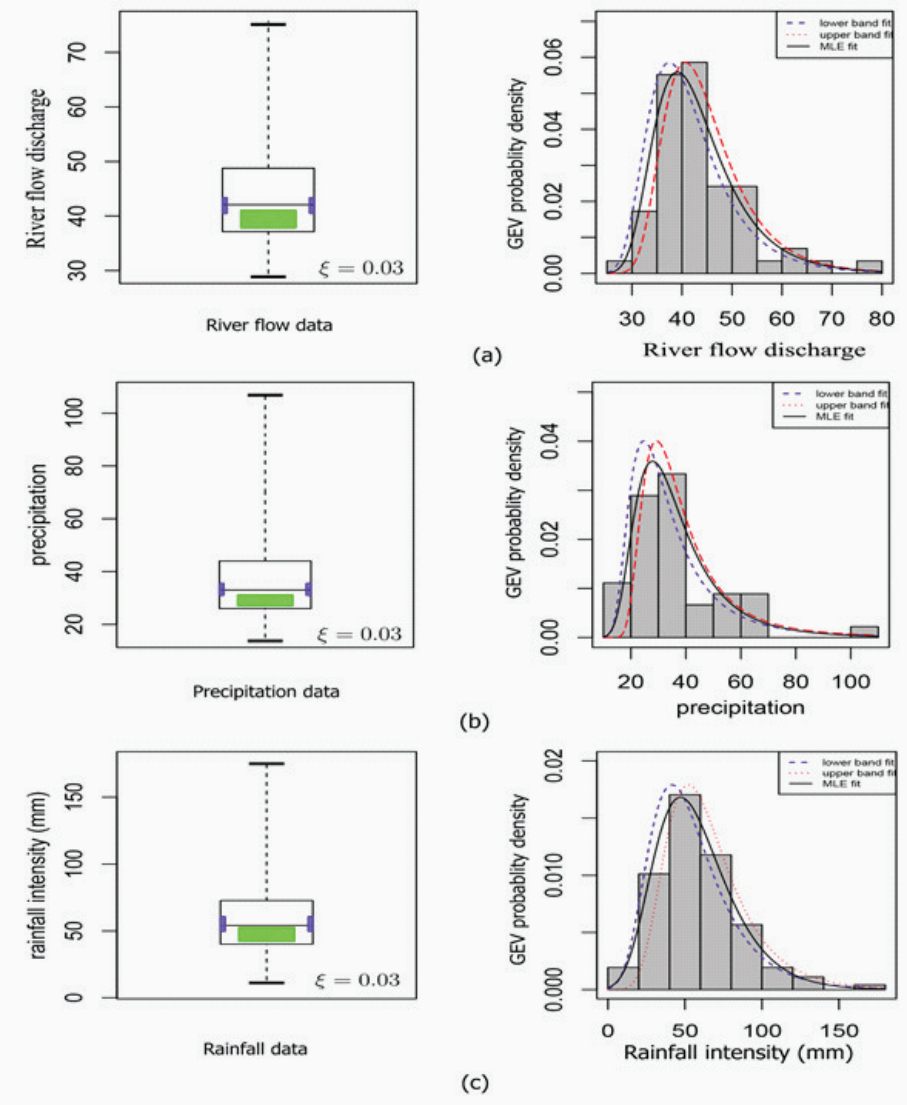

Figure 8. Visualising a block maximum extreme sample with the proposed boxplot along with its parameters region versus histogram/density fitting of the data with proposed boxplot parameter estimates for three datasets (a) River flow data (b) Precipitation data and (c) Rainfall intensity data

method

- Batch comparison of extreme samples can be visualised using the location and scale parameters region as incorporated in the proposed boxplot for a more valid conclusion

- The boxplot and the proposed enhancement in this work should remain for exploratory analysis rather than confirmatory

However, the limitation of this contribution is the non-cooperation of other extreme modeling tools such as the generalised Pareto distribution for a peak over threshold data and the r largest-order statistics process.

V. ACKNOWLEDGMENT

The authors are grateful to Universiti Putra Malaysia (Research Grant FRGS02-1-15-1741FR) and Federal University Dutse Nigeria (2013 Tetfund Fellowship) for sponsorship of the research work. The authors also acknowledge the constructive criticism by the unanimous reviewers.

VI. REFERENCES

[1] Abuzaid, A. H., Mohamed, I. B. and Hussin, A. G. (2012) Boxplot for circular variables. *Computational Statistics*, Springer, pp. 1–12.

[2] Analysis, E. D., Censored, P., From, D., Distributions, S., Source, K., Statistical, R. and Stable, S. (2015) Data Analysis for Possibly Exploratory Skewed Distributions Censored Data from. *Applied Statistics*, 39(1), pp. 21–30.

[3] Babura, B. I., Adam, M. B., Fitrianto, A. and Rahim, A. S. A. (2017) Modified boxplot for extreme data. *AIP Conference Proceedings*, 1842(1), pp. 30034. doi: 10.1063/1.4982872.

[4] Broyden, C. G. (1970) The convergence of a class of double-rank minimization algorithms 1. General considerations, *IMA Journal of Applied Mathematics (Institute of Mathematics and Its Applications)*, 6(1), pp. 76–90. doi: 10.1093/imamat/6.1.76.

[5] Bruffaerts, C., Verardi, V. and Vermandele, C. (2014) A generalized boxplot for skewed and heavy-tailed distributions. *Statistics & probability letters*, Elsevier, 95, pp. 110–117.

[6] Carling, K. (2000) Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, Elsevier, 33(3), pp. 249–258.

[7] Castillo, E. (1988) *Extreme value theory in engineering*. London: Academic Press.

[8] Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer (Lecture Notes in Control and Information Sciences). Available at: https://books.google.com.my/books?id=2nugUEaKqFEC.

[9] Frigge, M., Hoaglin, D. and Iglewicz, B. (1989) Some implementations of the boxplots. *The American Statistician*, 43(1), pp. 50–54. doi: 10.2307/2685173.

[10] Hubert, M. and Vandervieren, E. (2008) An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, Elsevier, 52(12), pp. 5186–5201.

[11] Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages. *The American Statistician*, Taylor & Francis, 50(4), pp. 361–365.

[12] Jenkinson, A. F. (1955) The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, Wiley Online Library, 81(348), pp. 158–171.

[13] McGill, R., Tukey, J. W. and Larsen, W. A. (1978) Variations of box plots. *The American Statistician*, Taylor & Francis Group, 32(1), pp. 12–16.

[14] Prescott, P. and Walden, A. T. (1980) Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, Oxford University Press, 67(3), pp. 723–724.

[15] Qarmalah, N. M., Einbeck, J. and Coolen, F. P. A. (2016) k-Boxplots for mixture data. *Statistical Papers*, Springer, pp. 1–16.

[16] Reiss, R.-D. R.-D. R. (1989) *Approximate distribution of order statistics; with applications to*

nonparametric statistics.

[17] Schwertman, N. C., Owens, M. A. and Adnan, R. (2004) A simple more general boxplot method for identifying outliers. *Computational Statistics and Data Analysis*, 47(1), pp. 165–174. doi: 10.1016/j.csda.2003.10.012.

[18] Schwertman, N. C. and de Silva, R. (2007) Identifying outliers with sequential fences. *Computational Statistics and Data Analysis*, 51(8), pp. 3800–3810. doi: 10.1016/j.csda.2006.01.019.

[19] Thomas, B. M. and Nolan, J. D. (1997) *Colorado Extreme Storm Precipitation Data Study*. Colorado Climate Center. Available at: https://ccc.atmos.colostate.edu/pdfs/Climo_97-1_Extreme_ppt.pdf.

[20] Tukey, J. W. (1977) *Exploratory data analysis*. Reading, Mass.